Research article

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

Céline Gomez, Stéphane Dussert, Perla Hamon, Serge Hamon, Alexandre de Kochko and Valérie Poncet*

Address: UMR DIAPC, IRD, 911 avenue Agropolis, BP 64501, 34394 Montpellier CEDEX 5, France

Email: Céline Gomez - celine.gomez@ird.fr; Stéphane Dussert - stephane.dussert@ird.fr; Perla Hamon - perla.hamon@ird.fr; Serge Hamon - serge.hamon@ird.fr; Alexandre de Kochko - alexandre.dekochko@ird.fr; Valérie Poncet* - valerie.poncet@ird.fr

* Corresponding author

Abstract

Background: Among Coffea species, C. canephora has the widest natural distribution area in tropical African forests. It represents a good model for analyzing the geographical distribution of diversity in relation to locations proposed as part of the "refuge theory". In this study, we used both microsatellite (simple sequence repeat, SSR) and restriction fragment length polymorphism (RFLP) markers to investigate the genetic variation pattern of C. canephora in the Guineo-Congolean distribution zone.

Results: Both markers were first compared in terms of their informativeness and efficiency in a study of genetic diversity and relationships among wild C. canephora genotypes. As expected, SSR markers were found to have a higher genetic distance detection capacity than RFLP. Nevertheless, similarity matrices showed significant correlations when Mantel's test was carried out (r = 0.66, p < 0.0001). Finally, both markers were equally effective for group discrimination and phylogenetic studies, but SSR markers tended to outperform RFLP markers in discriminating the source of an individual among diversity groups and in putative hybrid detection. Five well defined genetic groups, one in the Upper Guinean forests, the four others in the Lower Guinean forests, were identified, corresponding to geographical patterning in the individuals.

Conclusion: Our data suggested that the Dahomey Gap, a biogeographical barrier, played a role in wild C. canephora differentiation. Climatic variations during the Pleistocene and/or Holocene probably caused the subgroup differentiation in the Congolese zone through the presence of a mosaic of putative refugia. Recent hybridization between C. canephora diversity groups, both for spontaneous individuals and cultivars, was further characterised according to their geographic dissemination or breeding history as a consequence of human activities.

Background

The species richness in some African zones has attracted attention on the origin of diversification in tropical forests [1-4]. In particular, the Guineo-Congolean regional center of endemism (Zone I on White's map, 1983) includes about 8,000 spp., about 80% of which are endemic. This zone also corresponds to one of the 34 biodiversity hotspots defined by Mittermeier et al. [5]: the Guinean
forests, divided into the Upper Guinean and Lower Guinean forests by the Dahomey Gap. Analyses of patterns of geographical distribution of species richness and endemism of extant organisms could partly help in understanding these species composition patterns [e.g. [3]]. Species richness in the tropics has been attributed to the gradual accumulation and/or preservation of species over a long geological period in stable equatorial climates (the “museum model”) [6-8] and/or to high speciation rates in response to late Tertiary geological events and unstable Pleistocene climates [9,10]. In fact, all major glacial advances in the Arctic resulted in great paleoenvironment and vegetation variations in the African tropics [2]. At least for the most recently evolved taxa, a certain amount of speciation and a great deal of subspeciation is said to have occurred during the late Pleistocene climatic fluctuations, especially since the last severe glaciation which cumulated around 18,000 B.P. [11]. To explain how plants and animal species survived during particularly cold and arid periods, Maley [2] proposed putative refugia, where the respective ancestors of the present species formed isolated populations, leading to allopatric, or geographical speciation [12]. Some of the refugia were located along the West African coast from Sierra Leone to Côte d’Ivoire, from southwestern Cameroon to western Gabon, in the eastern part of the Democratic Republic of the Congo (formerly Zaire) and along the Zaire river [2]. Through a study of the African Rubiaceae genera distribution, Robbrecht [13] gave additional support in favour of the refuge area concept, and demonstrated the importance of fluvial refugia in some taxa.

Molecular phylogenies and species-level diversity could be used to further unravel the forest diversification patterns. A population-genetics rather than a species-level approach has also been proposed to test the role of the relatively recent time frame of Pleistocene events [14]. Climatic changes have affected the genetic structure of many tree species in Europe [e.g. [15-17]]. For Africa, major advances are under way for vertebrates [e.g. [14,18]], but very few studies have focused on tree species of West and Central Africa [19-22], and only one has attempted to establish the relationship between the molecular genetic variation of a tree species, i.e. the shea tree (Vitellaria paradoxa), and the possible evolution of vegetation after the last glaciations in the Sudano-Sahelian region of Africa [19].

*Coffea* species (Rubiaceae) are endemic to intertropical forest zones in Africa, Madagascar, Mauritius, Comoros and Réunion [23-27]. There are over 103 accepted species, including the two most commonly cultivated species, *C. arabica* L. and *C. canephora* Pierre ex A. Froehn [28]. Within *C. arabica*, a predominantly self-pollinating (autogamous) species, there is a genetic structure with low differentiation between accessions from the east and west side of the Great Rift Valley in Ethiopia as recently revealed with microsatellites [29]. The hybrid origin of *C. arabica* with *C. canephora* and *C. eugenioides* as the likely progenitors [30] was probably recent, and its colonization of Ethiopia likely occurred after the formation of the Great Rift Valley [31]. The systematic position and geographic distribution of *C. arabica* is isolated among diploid *Coffea* species [27,28]. On the contrary, *C. canephora* is one of the two *Coffea* species, along with *C. liberica* Hiern[28,32], with the widest natural distribution area of the genus. Moreover, these species are both diploid, allogamous (self-incompatible), and belong to the same phylogenetic clade [27]. They present overlapping geographical distributions, which extend west to east from Guinea to Uganda, and north to south from Cameroon to Angola [28,33]. This feature represents an exceptional resource for understanding the evolution and adaptation of tropical trees in these regions.

Few studies have been undertaken to assess the genetic diversity of natural *C. canephora* populations. Allozymic surveys revealed marked geographical clustering for two groups: the “Guinean” group, composed of populations from Côte d’Ivoire, and the “Congolese” group, consisting of two subgroups, SG1 and SG2, with populations from the Central African Republic and Cameroon [33,34]. Based on RFLP data, five wild *C. canephora* groups were further distinguished and the diversity structure was found to be conserved even when cultivated material, known as Robusta coffee, was included in the analysis [35]. Due to the increase in the number of microsatellite markers suitable for coffee species analyses [e.g. [36-38]], PCR-based assays are becoming increasingly attractive and compatible with the requirements of evolutionary studies or conservation genetics on large sets of genotypes.

Polymerase chain reaction (PCR)-based marker systems like simple sequence repeats (SSRs) have been widely used in recent years, replacing restriction fragment length polymorphism (RFLP) in DNA fingerprinting [39]. However, few studies have been carried out to compare the efficiency of RFLP and SSR markers for characterizing genetic diversity [40]. Both techniques offer the advantage of implementing nuclear co-dominant, locus-specific markers dispersed throughout the genome. Polymorphism detected by RFLP assays reflects restriction size variations, while microsatellite variants differ in the number of short (1–6 bp) amplified tandem repeats [41]. Their use in the assessment of population genetic structure has both advantages and drawbacks. While RFLPs assay are time consuming and labor intensive, among PCR-based markers, microsatellites require sequence information for primer pair design but, once developed, they are highly transferable across species, especially within the genus...
Because of their extremely high level of polymorphism, they are probably the most efficient markers for fingerprinting, assignment tests or paternity analyses [44]. Differences in the resolution power of RFLP and SSR loci are thus expected because of differences in their mutation rates and processes. The simultaneous use of these molecular methods could help in inferring the signature of natural histories of organisms at different time scales, i.e., evolutionary history and historic migration patterns. Moreover, the recent development of Bayesian approaches in assignment tests, as implemented in STRUCTURE[45], has greatly increased the potential for understanding population structure.

In the present study, we investigated Coffea canephora genetic diversity across the West and Central African geographic range of the species. Spontaneous individuals are naturally distributed in two postulated refugia areas, i.e., the Upper Guinea and Lower Guinea/Congo regions, separated by the Dahomey Gap. Using both SSRs and RFLP loci, we evaluated the relative efficiency of these DNA-based marker systems, associated with different mutation rates, in resolving genetic diversity, population structure and gene flow among samples of C. canephora.

The final aims were:

(i) to analyse the C. canephora genetic variability structure and assess any relationship between the molecular variation of the species and the possible signature of the past evolution of vegetation in the Guineo-Congolian region of Africa

(ii) to evaluate the impact of human agricultural activities on gene flow, dispersal and migration of both wild and cultivated stocks through the detection of putative interdiversity group hybrids

(iii) to define core sets of C. canephora accessions that best capture the species diversity in terms of alleles.

Results

Efficiency of RFLP and SSR markers for polymorphism detection

In the 107 sampled Coffea canephora accessions (Figure 1, table 1, and table 2), genetic polymorphism indexes (Na, H0, and Hi) at the 16 SSR loci and 8 RFLP loci were found to be highly variable throughout all geographic samples and types (Table 3). A total of 154 alleles across the SSR loci were detected, giving an average of 9.6 alleles per locus, ranging from 3 at M804 to 20 at M368. In comparison, the RFLP loci analysed gave 59 alleles with an average of 7.4 alleles per locus. The PIC value and gene diversity were quite even for SSR and RFLP loci, i.e., 0.62 vs. 0.59 and 0.65 vs. 0.63, respectively.

Structure of C. canephora natural diversity

Dissimilarity matrices were constructed on the whole sample based on shared-allele distances and revealed that the average genetic dissimilarities for RFLP (0.604) and SSR (0.626) markers were very similar. The average genetic dissimilarities obtained for the wild genotypes were also similar for both marker types, RFLPs (0.581) and SSRs (0.595). The correlation coefficient obtained with Mantel’s test matrix correspondence test was, indeed, statistically significant (r = 0.66, p < 0.0001).

Genetic distance estimates obtained from the wild genotypes were used to derive Neighbour-joining trees for both the RFLP and SSR data. The topology of each tree (Figure 2) was unique but both trees discriminated the five a priori genetic groups (A, B, C, D, and E) from Dussert et al. [35]. Nevertheless, the SSR tree was not completely congruent with the RFLP tree. One difference was that group C, which occupied an intermediate position on the tree based on SSRs, occurred at a more distant position from groups [B-E-A] within group D based on RFLPs. Another contrast was noted with respect to the A individuals which clustered together near the E group based on SSRs and one of its individuals fell into the group E, based on RFLPs. The internal branches were generally longer for the RFLP tree. Despite these differences, both tree topologies reflected the same distinct clades, corresponding to a geographical patterning in the individuals. Plants from Côte d’Ivoire mostly clustered with plants from Guinea (group D). Individuals from northwest Congo, southeast Cameroon, and southwest Central African Republic fell within the same clade (group C), in line with their geographic proximities in Central Africa. The plants of Central African group C clustered near plants from West Africa (group D). From the eastern part of the geographical distribution, plants from northeast Congo clustered with plants from southeast Cameroon and south Central African Republic (group E), near plants from the southern border of Central African Republic (group B). The genotypes of a population from northwest Congo and a population from southwest Cameroon (group A) clustered near the B and E clades. Several individuals (marked with asterisks) appeared to be classified with groups other than their a priori groups. We performed a population structure analysis to further assess the group memberships of the plants and to detect actual migrants, hybrids or misclassified individuals.

The structure analysis using STRUCTURE with admixture showed that the five groups (K = 5) were genetically distinct based on SSR data (Figure 3). The results indicated that most of the plants had a high membership in their own cluster, with 97, 97, 69, 90, and 92% mean ancestry for the A, B, C, D, and E groups, respectively. In comparison, with the RFLP data, plants from groups B, C, and D
Table 1: Geographic and genetic origin of the wild *C. canephora* genotypes

| Location | Name | Long. | Latit. | Coll. # | A priori genetic group | Putative hybrids |
|----------|------|-------|--------|---------|------------------------|-----------------|
| Côte d'Ivoire | Ira 1 | 07.40 | -07.46 | 0009, 0028 | D | |
| | Bafingdala | 07.51 | -07.41 | 0056 | D | |
| | Ira 2 | 07.50 | -07.42 | 0105, 0159 | D | |
| | Gbapleu | 07.34 | -08.18 | 0121, 0128 | D | |
| | Gao | 06.57 | -07.39 | 0136 | D | |
| | Bossematie | 06.30 | -03.30 | 0145, 0146 | D | |
| | Logbonou | 08.04 | -05.15 | 0186, 0194 | D | DA, DE |
| | Fourougbankoro | 08.29 | -05.45 | 0213, 0233 | D | |
| | Goazra | 07.00 | -05.36 | 0236 | D | |
| | Marahoué | 06.54 | -06.12 | 0245, 0250 | D | |
| | Kouin | 07.30 | -07.18 | 0292, 0293 | D | E |
| | Géoulé | 07.36 | -07.56 | 0313, 0315 | D | E |
| | Gbapleu 1 | 0319, 0321 | | E | E |
| | Gbapleu 2 | 0328 | | D | |
| | Gbapleu 3 | 0336 | | D | |
| | Dobia | 0345, 0350 | | D | DEA, DE |
| | Sabrégué | 0354, 0358 | | D | DE |
| | Pélézi 2 | 0362, 0395 | | D | D |
| | Gonaté | 0404 | | D | |
| Guinea | Piné 19 | 08.57, 08.57 | -08.06, -08.06 | 0803, 0808 | D | |
| Cameroon | Koto | 04.22 | 09.34 | 0651 | E | |
formed separated clusters with high membership in their own cluster (87, 91 and 88% on average, respectively). However, mixed ancestry between groups A and E at even proportions (A and E contributing to 50 and 44% ancestry on average) was observed, corresponding to a grouping by geographical regions. For this mixed group, we considered that plants possessing <80% ancestry in the [A+E] cluster were putative hybrids. Despite this discrepancy, the level of admixture per group was very low and similarly estimated by both markers (Table 4). Eight genotypes were identified as admixed by SSRs, with over 20% ancestry from other groups. Five of them were also identified by RFLPs. Some of these putative hybrids, 4/7 and 2/5 detected by SSRs and RFLPs, respectively, were confirmed with the reassignment test implemented in GENECLASS2 (Table 4). The five individuals identified as putative

| Membership to genetic groups, as previously defined by Dussert et al. [35] and putative hybrids detected in the present study are indicated for each accession. |
|---|
| 1Berthaud [33]; 2Le Pierrès et al. [26]; 3Anthony et al. [24]; 4Berthaud and Guillaumet [23]; 5de Namur et al. [25] |

| Table 1: Geographic and genetic origin of the wild C. canephora genotypes (Continued) |
|---|
| 22 | Nguila | 04.43 | 11.40 | 0658 | C | DA |
| 23 | Mwamepen | 03.29 | 14.48 | 0662 | 0663 | E | A | AE |
| 24 | Bitonga | 03.19 | 15.16 | 0664 | C |
| 25 | Boumba | 02.03 | 15.10 | 0665 | 0666 | E |
| 26 | Nguilili | 02.04 | 15.36 | 0678 | 0683 | C |
| 27 | Bombi | 04.56 | 13.10 | 0685 | C |
| Central African Republic 4 |
| 28 | Loukoussou | 0501 | B |
| 29 | Doungba | 0504 | 0511 | E | B |
| 30 | Yombi | 0516 | E |
| 31 | Libengué | 0518 | B |
| 32 | Carnot | 0602 | 0604 | C | C | CE |
| 33 | Ndouguié | 9004 | C |
| Congo 5 |
| 34 | Ouesso 2 | 01.35 | 14.48 | 0716 | C |
| 35 | Ouesso | 01.37 | 14.53 | 0721 | C |
| 36 | Sembé-Souanké | 01.56 | 14.11 | 0723 | A |
| 37 | Souanké-gabon | 02.07 | 14.00 | 0725 | 0727 | E |
| 38 | Boyélé 5 | 0729 | 0730 | E |
| 39 | Impfondo 1 | 0738 | E |
| 40 | Impfondo 2 | 0739 | 0740 | E |
Table 2: Geographic and genetic origin of the cultivated *C. canephora* genotypes.

| Type of introduction | Group name | Donor or collector | Country of origin | Coll # | A priori genetic group | Final group |
|----------------------|------------|-------------------|-------------------|-------|------------------------|-------------|
| Donation from field genebanks | "Aboisso" | Aboisso¹, Gabon | Côte d'Ivoire | C135 | A | A |
| | | Côte d'Ivoire | | C307 | A | A |
| | | | | C318 | E | DE |
| | | | | C329 | E | AE |
| | "C10 Man" | Unknown | Rep. of Congo | C429 | E | E |
| | | | | C439 | E | E |
| | "INEAC" | INEAC⁴, Rep. of Congo | | C002 | E | E |
| | | | | C078 | E | E |
| | | | | C032 | E | E |
| | | | | C054 | E | E |
| | | | | C057 | E | AE |
| | | | | C062 | E | E |
| | | | | C095 | E | DE |
| | | | | C003 | A | AE |
| | | | | C006 | E | E |
| | | | | C015 | E | E |
| | "Kouilou of Madagascar" | Bingerville², Côte d'Ivoire | Gabon | CK07 | A | A |
| | | | | CK14 | A | A |
| | | | | CK29 | A | AE |
| | "Niaouli" | Bingerville², Côte d'Ivoire | Togo | C008 | A | A |
| | | | | C017 | A | A |
| Collection in plantations | "Guinea" | Guinea⁶ | | 0855 | D | DE |
| | | | | 0856 | D | DE |
| | | | | 0850 | E | DE |
| | | | | 0851 | E | DE |
| | | | | 0852 | E | E |
| | | | | 0909 | E | DE |
| | | | | 0915 | E | DE |
| | | | | 0848 | E | DE |
| | | | | 0853 | E | D- |
| | | | | 0881 | E | DE |
| | | | | 0882 | E | DE |
| | "Hybrid" | Côte d'Ivoire | | C107 | D | DE |
| | | | | C126 | E | DE |
| | | | | C181 | D | DE |
| | | | | C636 | D | DE |
| | "Côte d'Ivoire" | Côte d'Ivoire⁵,⁶ | | 0317 | D | DE |
| | | | | 0318 | D | DE |
| | | | | 0163 | A | AE |
| | | | | 0164 | A | A |
| | | | | C155 | A | D |
| | "Togo" | Togo | | 0693 | A | A |
| | | | | 0695 | A | A |
| | "Tanzania" | Tanzania | Unknown | 0270 | Unknown | E- |
| | | | | 0279 | E | CE |
| Unknown | "Robusta A1" | Unknown | | C077 | A | A |

Membership to a priori genetic groups, as previously defined by Dussert et al. [35], and final group definition as identified in the present study are indicated.

¹ Introduction at Aboisso (Côte d'Ivoire) by Beynis in 1910, of material cultivated in Gabon [57]
² Introduction at the garden at Bingerville (Côte d'Ivoire) in 1914 of material cultivated in Togo [57]
³ Introduction at CRA Bingerville (Côte d'Ivoire) in 1951, of material selected in Madagascar and originating from Gabon [57]
⁴ Introduction in Côte d'Ivoire in 1935 of material selected at INEAC in the Republic of Congo (Cordier [57])
⁵ Le Pierrès et al. [26]; ⁶Berthaud [33]
hybrids by structure analysis using either marker type originated from the a priori D group of Côte d'Ivoire, and contained >20% group E or group A ancestry (Table 4). On the RFLP tree, these individuals nested within the same clade, with an intermediate position on the between groups [A-B-E] and [C-D]. They occurred at the base of the D clade on the SSR tree. The individual AC663, classified a priori in group A, and identified as a putative AE hybrid with RFLPs, fell into the group E cluster in the RFLP tree.

The genetic differentiation of wild C. canephora groups, evaluated with $F_{ST}$ values, were found to be higher for RFLPs compared to SSRs, with 0.50 vs. 0.33, respectively (Table 5). Pairwise RFLP-based or SSR-based $F_{ST}$ values ranged from 0.29 to 0.67 and from 0.20 to 0.50, respectively, with substantially lower $F_{ST}$ values within regions than between regions. Group D (West Africa) showed the highest levels of differentiation detected with all other groups with both marker types, whereas the A-B and E-B comparisons (groups from Central Africa), gave the lowest $F_{ST}$ values with RFLP and SSR data, respectively. Mantel's tests revealed that pairwise $F_{ST}$ estimates among samples calculated from SSRs were not significantly differ-

| Locus | Map location | Total sample N = 107 | Wild sample N = 61 |
|-------|--------------|----------------------|-------------------|
|       | N<sub>a</sub> | H<sub>O</sub> | H<sub>E</sub> | PIC | N<sub>a</sub> | H<sub>O</sub> | H<sub>E</sub> | PIC |
| M257  | B            | 6                    | 0.25  | 0.55  | 0.47  | 6                    | 0.21  | 0.59  | 0.52  |
| M809  | H            | 5                    | 0.15  | 0.25  | 0.23  | 4                    | 0.11  | 0.16  | 0.15  |
| M904  | M            | 3                    | 0.06  | 0.07  | 0.07  | 1                    | 0.00  | 0.00  | 0.00  |
| M821  | F            | 13                   | 0.28  | 0.83  | 0.81  | 13                   | 0.28  | 0.83  | 0.81  |
| M314  | G            | 5                    | 0.25  | 0.49  | 0.43  | 5                    | 0.18  | 0.42  | 0.36  |
| M368  | -            | 20                   | 0.81  | 0.93  | 0.92  | 15                   | 0.76  | 0.91  | 0.91  |
| M764  | -            | 13                   | 0.65  | 0.83  | 0.81  | 11                   | 0.63  | 0.82  | 0.80  |
| M394  | C            | 9                    | 0.41  | 0.80  | 0.77  | 8                    | 0.39  | 0.75  | 0.71  |
| M779  | -            | 12                   | 0.75  | 0.85  | 0.83  | 12                   | 0.63  | 0.82  | 0.80  |
| M782  | -            | 4                    | 0.06  | 0.25  | 0.22  | 4                    | 0.03  | 0.12  | 0.11  |
| M495  | -            | 9                    | 0.35  | 0.65  | 0.60  | 9                    | 0.33  | 0.62  | 0.57  |
| M259  | N            | 6                    | 0.52  | 0.59  | 0.55  | 6                    | 0.48  | 0.60  | 0.56  |
| M367  | -            | 15                   | 0.30  | 0.88  | 0.87  | 12                   | 0.20  | 0.85  | 0.83  |
| M755  | -            | 13                   | 0.42  | 0.78  | 0.76  | 11                   | 0.23  | 0.68  | 0.65  |
| M387  | J            | 10                   | 0.33  | 0.75  | 0.71  | 8                    | 0.26  | 0.74  | 0.70  |
| M856  | I            | 11                   | 0.49  | 0.88  | 0.87  | 11                   | 0.41  | 0.85  | 0.84  |
| Total |             | 154                  | 9.63  | 0.38  | 0.65  | 0.62                | 136   | 8.5   | 0.32  | 0.61  | 0.58  |

| Locus | Map location | Total sample N = 107 | Wild sample N = 61 |
|-------|--------------|----------------------|-------------------|
|       | N<sub>a</sub> | H<sub>O</sub> | H<sub>E</sub> | PIC | N<sub>a</sub> | H<sub>O</sub> | H<sub>E</sub> | PIC |
| gA13  | 10           | 14                   | 0.51  | 0.73  | 0.70  | 11                   | 0.43  | 0.66  | 0.61  |
| gA19  | 9            | 10                   | 0.43  | 0.78  | 0.75  | 9                    | 0.38  | 0.72  | 0.68  |
| gA29  | 5            | 5                    | 0.27  | 0.47  | 0.44  | 5                    | 0.26  | 0.50  | 0.47  |
| gA10  | 3            | 8                    | 0.48  | 0.78  | 0.76  | 7                    | 0.42  | 0.77  | 0.74  |
| gA14  | 3            | 5                    | 0.30  | 0.59  | 0.53  | 5                    | 0.31  | 0.56  | 0.52  |
| gA61  | 7            | 8                    | 0.50  | 0.70  | 0.66  | 6                    | 0.48  | 0.69  | 0.65  |
| cr167 | 4            | 7                    | 0.43  | 0.60  | 0.57  | 7                    | 0.34  | 0.58  | 0.55  |
| gA72  | 7            | 2                    | 0.16  | 0.41  | 0.33  | 2                    | 0.13  | 0.34  | 0.28  |
| Total |             | 59                   | 7.37  | 0.38  | 0.63  | 0.59                | 52    | 6.50  | 0.34  | 0.60  | 0.56  |

Linkage group locations for mapped markers are given. The number of alleles ($N_a$), observed heterozygosity ($H_O$), expected heterozygosity ($H_E$, or gene diversity) and PIC (polymorphism information content) values are also given.

1. Linkage group of (C. canephora × C. heterocalyx) × C. canephora genetic map from Coulibaly et al. [60]
2. Linkage group of C. canephora genetic map from Paillard et al. [59], with estimated distances: gA10-gA14 = 21 cM, gA61-gA72 = 30 cM
ent from those calculated from RFLPs ($R = 0.72$, $N = 1000$ permutations).

**Cultivated sample origin**

When the cultivated genotypes were included in the neighbour-joining tree construction, the topology of the trees in five groups could not be identified appropriately with RFLPs and the clusters were even less separated using SSRs (data not shown). Since some of the cultivated plants might be of putative hybrid origin (Table 2, “hybrid” group), we assessed the cultivated genotype origin by population structure analysis.

Assignment tests using GENECLASS were first performed to assess genotypic similarities of the cultivated plants with respect to the five wild genetic reference groups. Using RFLPs, of the 46 cultivated plants, 10 (22%), 13 (28%) and one (2%) were assigned to groups A, E, and D, respectively, whereas the remaining individuals (48%) were all unassigned. In comparison, SSRs only assigned 8 (17%), 9 (20%) and one (2%) individuals to groups A, E, and D, respectively, whereas 28 (61%) remained unassigned. Considering the overall dataset, assigned cultivated individuals, except for three of them, were correctly classified within their *a priori* group of origin with both RFLP and SSR data. Genetic groups B and C were not represented by cultivated genotypes. None of the *a priori* group D individuals were reassigned to D, suggesting a possible mixed ancestry origin.

**Figure 1**

**Wild Coffea canephora sampling locations in West and Central Africa.** Codes are given in Table 1.
Figure 2

Unrooted trees of individual wild *Coffea canephora* genotypes using the neighbour-joining method and shared-allele distance among (A) 16 microsatellite loci and (B) 8 RFLP loci. The first letter of the individual labels, A, B, C, D and E, indicates the a priori diversity group of Dussert et al. [35]. Individuals who appear to be misclassified in the present tree are marked (*). Circled names represent plants identified as being mis-classified or putative hybrids by the subsequent genetic structure analysis.
Genetic admixture analysis was conducted with STRUCTURE using the overall dataset with both cultivated and wild plants, while excluding wild individuals from group B. The four groups (K = 4) observed, corresponding to groups A, C, D, and E, were found to be genetically distinct with either RFLPs or SSRs. The mean proportion of population membership of wild genotypes to their own clusters was 72, 84, 90, and 92%, for groups A, C, D, and E, respectively, using RFLPs; and 69, 82, 89, and 93% for groups A, C, D, and E, respectively, using SSRs. This confirmed that the previously defined genetic groups were still responsible for the observed population structure, even after the addition of cultivated individuals.

Admixture analysis with RFLPs of individual genotypes identified hybrids among individuals from a priori group A (3/14 = 21%) and a priori group E (8/24 = 33%). All plants from a priori group D were detected as hybrids between groups D and E. In comparison, the SSR analysis detected slightly more hybrids with 4/14 = 29% in a priori group A, 13/24 = 54% in a priori group E, and all plants from a priori group D.

Most of the hybrids 13/17 = 76% and 13/23 = 57%, for RFLPs and SSRs, respectively, were identified as originating from hybridization between a priori groups D (West Africa) and E (Central Africa). These hybrids shared an equivalent average fraction of ancestry from both groups (49%D-45%E with RFLPs and 43%D-48%E with SSRs), suggesting that most of them could be classified as first generation hybrids.

Figure 3
The genetic structure of wild Coffea canephora (N = 61) based on (A) 16 polymorphic SSR loci (B) 8 polymorphic RFLP loci. Result of STRUCTURE analysis using K = 5 and assignment to the five a priori diversity groups (A to D) from Dussert et al. [35]. The colour bars represent percentages of ancestry from the different groups observed in one individual.
The combined data analysis of cultivated coffee trees using both marker types and both population structure analyses allowed us to identify the putative hybrid origin of many cultivated plants (25/46 = 54%) (Table 2).

Principal coordinate plots representing the genetic similarity between wild and cultivated *C. canephora* for SSRs are presented in Figure 4 on the basis of SSR data. The variance explained by the first two axes was greater for RFLPs than for SSRs (45.2% vs. 16.7%). However, on both plots, wild individuals were genetically distinct, with no overlap of the five genetic groups. The distribution of cultivated individuals overlapped that of wild groups, with putative hybrids located at intermediate positions.

### Core set selection

We selected core sets of both wild and cultivated accessions from *C. canephora* that capture the maximum number of SSR or RFLP alleles for sample sizes 12 and 25. Core sets of 12 captured 48/59 and 103/154 of the RFLP and SSR alleles, respectively. Gene diversity in each core set was equivalent to that found in the entire sample (Table 6). Each core set contained wild and cultivated accessions from the different groups, and a substantial fraction of putative hybrids.

### Discussion

Our analysis of *C. canephora* genetic diversity through RFLP and SSR markers revealed a marked geographic structure in subgroups, which could be interpreted with a broad scope of regional and historical influences. The detection of inter-diversity group putative hybrids allowed us to evaluate the influence of humankind on the original natural distribution and the occurrence of gene flow between wild and cultivated stocks.

The choice of the appropriate marker for any specific study depends principally on the purpose of the research and the biology and genetic structure of the species. The recent development of assignment tests based on Bayesian approaches [45-47] has greatly increased the potential for understanding population structure across its diversity.

---

**Table 4: Missclassified wild individuals or putative hybrids detected by Geneclass and Structure analyses.**

| Individuals | A priori Genetic Group | Geneclass | Structure |
|-------------|------------------------|-----------|-----------|
|             |                        | RFLP     | SSR       | RFLP Ancestry (%) | CI | SSR Ancestry (%) | CI |
| AC663       | A                      | E 1.445  | E 43.5    | 0–100             |   | A 39         | 0–100 |
|             |                        | B 14.2   |           | 0–61               |   | B 12         | 0–61  |
| CC604       | C                      | -        | E 3.224   | -                  |   | C 39         | 0–100 |
| CC658       | C                      | -        | D 5.010   | -                  |   | D 21         | 0–91.7|
| DC186       | D                      | -        | D 64      | 36.1–86.5          |   | D 56.3       | 24.8–83.6|
|             |                        | E 16     |           | 0–52.8             |   | E 33.5       | 0–68.3|
|             |                        | A 16     |           | 0–54.2             |   | A 21.2       | 0–71.9 |
| DC194       | D                      | -        | D 37.5    | 0–66.1             |   | D 57.9       | 29.6–84|
|             |                        | E 29.6   |           | 0–85.8             |   | E 35.3       | 0–50.3|
|             |                        | A 31.6   |           | 0–94.2             |   | A 35.3       | 0–50.3|
| DC345       | D                      | A 0.579  | D 49.9    | 12.1–77.9          |   | D 64.3       | 42.6–83.7|
|             |                        | E 0.526  |           | 0–66.6             |   | E 21.2       | 0–51.2|
| DC350       | D                      | -        | D 54.5    | 0–87.1             |   | D 42.4       | 19.8–65.5|
|             |                        | E 20.7   |           | 0–75.5             |   | E 51.6       | 14.2–76.9|
|             |                        | A 21.7   |           | 0–91.7             |   | A 35.3       | 0–50.3|
| DC358       | D                      | -        | D 49.9    | 12.1–77.9          |   | D 64.3       | 42.6–83.7|
|             |                        | E 21     |           | 0–66.6             |   | E 21.2       | 0–51.2|
|             |                        | A 21.5   |           | 0–68.5             |   | A 35.3       | 0–50.3|

The alternative group and log10 L_home/L_max from the Geneclass analysis is given. The Structure results are given for the percentage of admixture (mean ancestry) of each of the five diversity groups and the 90% CI. Cells are kept empty for individual non-detected hybrids, i.e. over 80% membership to their own cluster.

---

**Table 5: Pairwise $F_{ST}$ estimates of genetic divergence in wild *C. canephora* groups (N = 53) obtained with SSRs (below the diagonal) and RFLPs (above the diagonal).**

| SSR\RFLP | A    | B    | C    | D    | E    |
|----------|------|------|------|------|------|
| A        | 0.29 | 0.50 | 0.67 | 0.30 |
| B        | 0.33 | 0.51 | 0.67 | 0.30 |
| C        | 0.30 | 0.54 | 0.54 | 0.41 |
| D        | 0.50 | 0.32 | 0.54 | 0.52 |
| E        | 0.31 | 0.25 | 0.34 | 0    |
Principal component analysis of wild and cultivated C. canephora accessions based on their SSR polymorphism. The first axis represents 10% of the variation and the second one represents 6.5% of the variation. The wild genetic distance values. The dissimilarity values taken over the higher mutation rates of SSRs also had an effect on genetic distance values. The dissimilarity values taken over the lower mutation rates as compared to SSRs, whose mutation rates can range from $10^{-3}$ to $10^{-4}$ depending on their genomic position [48]. Levels of polymorphism detected with RFLP and SSR markers have been compared in soybean [40], also indicating the hypervariability of SSR loci and greater expected heterozygosity.

The higher mutation rates of SSRs also had an effect on the association diversities. Indeed, the differences in mutation rates would facilitate the determination of phylogenetic relationships and classifications, cluster analysis and population structure analysis in the C. canephora gene pool was further assessed. Indeed, the differences in mutation rates would also likely affect the power of the different markers to detect population differentiation.

In our data, overall and pairwise $F_{ST}$ estimates obtained with RFLPs were higher than those obtained with SSRs. However, in spite of the differences, SSR and RFLP multilocus $F_{ST}$ estimates were not significantly different ($R = \ldots$).
when computed over the entire set of samples. In fact, the high level of polymorphism, typical of SSRs, may induce downward bias in the population differentiation estimates. The degree of differentiation assessed through SSRs, and thus the \( F_{ST} \) values, are expected to be lower than those calculated using RFLPs [49]. A marked population subdivision similar to that noted in our study was also found at RFLP loci in brown trout [50] and soybean [40]. Moreover, the consistency of our results obtained from both markers is congruent with the findings of the other study comparing SSR and RFLP variation [40].

Finally, the RFLP and SSR multilocus diversity structure analysis divided wild coffee trees into largely concurrent five groups, with main branches on the individual neighbour-joining tree. Comparable diversity structuring was obtained, e.g. in terms of distinguishing geographical origins from West Africa and Central Africa.

At a smaller geographical scale, RFLPs and SSRs showed a different resolution power in detecting the genetic structure in the wild samples. The Bayesian analysis using STRUCTURE revealed five clearly distinct groups with SSR data, while samples from the southwest Cameroon/northwest Congo region—\textit{a priori} groups A and E—were clustered with the RFLP data. A higher membership in their own cluster was obtained with SSRs (89\% vs. 72\% mean ancestry over all wild groups with SSRs and RFLPs, respectively). Moreover, the SSR data allowed the identification of slightly more admixed plants, putatively hybrids, both within the wild and cultivated pools. The GENECLASS assignment test, which has been shown to be effective even if clusters are not in HWE, gave congruent results.

In conclusion, SSR markers outperformed RFLP markers in terms of discriminatory power in cluster analyses and assignment tests. Moreover, SSRs were also more efficient in discriminating the source of an individual genotype among putative diversity groups on a local scale. This advantage of SSR markers when analysing genetic affinities at individual levels was also demonstrated in a study.

| RFLP | Total: \( N_a = 59 \) \( GD_{107} = 0.63 \) | SSR | Total: \( N_a = 154 \) \( GD_{107} = 0.65 \) |
|------|---------------------------------|------|---------------------------------|
| Core set of 12 | Statut | Core set of 25 | Core set of 12 | Statut | Core set of 25 |
| _C270 | (c) H | _C270 | BC511 | (+) B | BC511 |
| DCC181 | (c) H | DCC181 | ECC002 | (c) E | ECC002 |
| ECC057 | (c) H | ECC057 | ECC429 | (c) E | ECC429 |
| ECC439 | (c) E | ECC439 | EC279 | (c) H | EC279 |
| CC9004 | (+) C | CC9004 | ACC135 | (c) A | ACC135 |
| ECC053 | (c) H | ECC053 | BCS501 | (+) B | BCS501 |
| CC721 | (+) C | CC721 | _C270 | (c) H | _C270 |
| AC164 | (c) A | AC164 | CC664 | (+) C | CC664 |
| EC315 | (+) E | EC315 | DC9 | (+) D | DC9 |
| BC501 | (+) B | BC501 | ECC095 | (c) H | ECC095 |
| EC319 | (+) E | EC319 | EC851 | (c) H | EC851 |
| BCS518 | (+) B | BCS518 | DC146 | (+) D | DC146 |
| (c) H | EC848 | (c) C | CC716 |
| (c) H | EC866 | (c) A | ACC008 |
| (c) H | EC915 | (+) H | CC658 |
| (c) H | CC658 | (+) H | DC186 |
| (c) H | EC881 | (+) E | EC739 |
| (c) E | EC651 | (c) H | EEC057 |
| (+) D | DC803 | (c) H | EEC882 |
| (+) C | CC716 | (+) D | DC136 |
| (c) H | EC850 | (c) A | ACK07 |
| (c) H | EC882 | (c) H | ACK29 |
| (c) E | ECC429 | (+) H | CC604 |
| (+) E | EC740 | (c) A | AC164 |
| (c) H | EC279 | (+) H | AC663 |

| Core set of 12 | Statut | Core set of 25 |
|----------------|--------|----------------|
| \( N_a = 48 \) | 36.9\% | \( GD = 0.68 \) |
| \( N_a = 58 \) | 59.6\% | \( GD = 0.72 \) |
| \( N_a = 103 \) | 19.6\% | \( GD = 0.66 \) |
| \( N_a = 137 \) | 38.3\% | \( GD = 0.69 \) |

Corresponding gene diversity (GD) values, cultivated status: wild (+) or cultivated (c), genetic group (A to E) or putative hybrid (H), No. of alleles \( (N_a) \) and percentage of total variability revealed in the whole sample (107 individuals) are given.
C. canephora genetic diversity and history

Organisation of C. canephora natural diversity

The genetic structure analysis carried out on the African samples of wild C. canephora using RFLPs and SSRs revealed marked separation between the West and Central African samples corresponding, in their composition, to the Guinean and Congolese groups of Berthaud [33]. This marked separation might be related both to the large geographical distances and to historical events. Indeed, the last glaciations and the subsequent migration from various Pleistocene refugia had caused large-scale changes in vegetation patterns, most notably around the Dahomey Gap and Cross River [2]. The western forests from Guinea and Côte d’Ivoire are separated from the Central forests of Cameroon, Central African Republic and Congo by the current 300 km wide Dahomey Gap, which is known to be an important biogeographic barrier. This area is thought to have become an extremely arid and much wider area during the last glacial maxima, around 18,000 BP, separating the forest refuges of southwest Ghana and west Cameroon [2]. Divergence through this geographical isolation might have led to genetic differentiation of the C. canephora populations. When analysing C. liberica, another Coffea species with the same geographical distribution, N’Diaye et al. [32] reported that the two varieties C. liberica var. liberica Bull. ex Hierne and C. liberica var. dewevrei (De Wild. & T. Durand) Lebrun, had high genetic differentiation and were characterised by marked reproductive barriers between the two varieties, with a pollen viability of their F1 hybrids similar to that of interspecific hybrids. However, although highly differentiated, the West (group D) and Central African (A, B, C, and E) groups of C. canephora present fully interfertile individuals (see below). This suggested that the two main C. canephora diversity groups do not present a state in the speciation process as advanced as that of the two C. liberica varieties. Phylogenetic studies have indeed revealed the role of the Dahomey Gap on lineage origins in the Upper Guinea regions [27,51]. But a similar distinct evolutionary split between populations in Upper Guinea and those of Lower Guinea across the Dahomey Gap has also been observed in shea tree species [19] and in Fire-crested Alethe birds [52].

The refugia scenario concerning the African Guineo-Congolian rainforest is supported by the fact that a similar speciation pattern has been observed in various genera [reviewed in [4]], particularly in African Rubiaceae genera [13]. Isolation of populations in refugia was also suggested to influence divergent adaptations and diversification at the species level [e.g. [14,18]]. Within the Central African zone, the distribution of the four C. canephora groups (A, B, C, and E) showed a complex pattern. This regional pattern of differentiation could be interpreted as evidence of cycles of fragmentation and subsequent expansion of forest habitats. The Congolese region consisted of a mosaic of several Pleistocene refugia [2] that might explain the pattern of genetic diversity in the Central African zone. Although it is difficult to precisely localise the refugia, the reconstructed hypothesized areas [2] were found to be related to species diversification. For example, Anthony et al. [14] suggested a role of these Pleistocene refugia in structuring gorilla genetic diversity. For C. canephora, the refugia origin of sub-group E is most likely located in the Congo-Zaïre basin (Figure 5). Meanwhile, the C. canephora group C might derive from the expansion of the Biafran forest refugia, inside curve of the Gulf of Guinea, from Cross River to Sanaga River (west Cameroon sensu lato). The late Holocene phase of dramatic climatic disturbance could also have been favourable for C. canephora diversification in Central Africa. Culminating about 2,500 years ago, it led to a catastrophic reduction in central African rainforests, in the region of south Cameroun, south Central African Republic, Gabon and Congo, and still exerts a major influence on the present forest vegetation distribution [53]. This event involved a brutal extension of savannas, favourable for the expansion of pioneering species such as oil palm [54]. The return of wetter conditions favourable for forest reinvansion began around 2,000 years B.P. from residual forests. This climatic disturbance caused fragmentation of the ancient Okoume (Aucoumea klaineana) distribution area into two subgroups, as suggested by the two tree varieties observed at the molecular level [21]. This process could also have produced the spatial genetic structuring of C. canephora in the Congolese zone. Both the intensity and length of these recent range expansion episodes could explain why coffee trees did not accumulate enough differentiation to lead to reproductive barriers and speciation.

Impact of human cultivation on inter-group gene flow and hybridization

The genetic structure analyses conducted in the present study seem to be efficient for reliable detection of admixtures and individual identification. Indeed, deviations from a priori diversity group classifications, are in line with issues concerning the introduction of non-indigenous wild or cultivated C. canephora plants or with the putative hybrid origin of cultivars.

Assignment tests confirmed that the landraces have remained genetically very close to the original wild populations. Indeed, since the 19th century, many local cultivars were cultivated spontaneous forms of C. canephora directly collected in adjacent forest populations. This was the case in the Democratic Republic of the Congo, in Côte
d’Ivoire (for “Kouilou” genotypes), Togo and Benin (for “Niaouli” genotypes), and Central African Republic (for “Nana” genotypes). Even improved material from INEAC (Table 2) has undergone few breeding cycles. Moreover, contact between local spontaneous forms and adjacent plantations could have promoted intermixing of individuals both in the wild and cultivated germplasm. In fact, according to the domestication syndrome definition (Hammer 1984), Robusta coffee trees should be considered as "cultivated" or "semi-domesticated" rather than "domesticated", since no heritable distinctive traits could be discerned when compared to their wild relatives.

The putative hybrid origin of some cultivars is suggested by the present data. In particular, most of the a priori group D (from Côte d’Ivoire and Guinea) and some of a priori group E (from southeast Cameroon and south Central African Republic) cultivated forms turned out to be putative F1 hybrids between these two genetic groups. Humans have contributed to this phenomenon since cultivated material originating from Central Africa, in particular from the Congo region, has been introduced in West African countries, such as in Côte d’Ivoire since 1930 (Portères, 1937). The low level of genetic introgression we observed in the putative hybrids suggests a recent contact between the local and exogenous material. Hybrids between the Guinean and Congolese group material had been previously identified by Berthaud [33] in cultivated variants on the basis of their vigour and yield. The use of intergroup hybrids and the efficiency of reciprocal recurrent selection was subsequently demonstrated [55]. This contrasts markedly with the cultivation history of C. arabica, whose varieties originated from a narrow genetic base [56]. Polymorphism was further reduced during selection cycles and by the predominant autogamy of C. arabica.

In this study, the putative intergroup hybrids, corresponding to various introgression levels detected and supposedly of wild origin, probably resulted from either culture escapes or from cross-pollinisation with neighbouring plantations.

Consequently, the core sets we propose contain both wild and cultivated accessions from the different groups, and a non-negligible fraction of putative hybrids. They provide good reference sets for further identification of spontaneous or cultivated material.

**Conclusion**

The high degree of concordance between the SSR and RFLP data for diversity group differentiation indicates that our set of markers provided adequate genome coverage for accurate germplasm characterization. Even though these loci have likely evolved at different rates, they revealed comparable diversity structure with five well-differentiated groups corresponding to geographical patterning in the individuals. The molecular variation was
correlated with the natural distribution in two postulated refugia areas, *i.e.* the Upper Guinea and Lower Guinea/Congo regions, separated by the Dahomey Gap.

A more intense and larger scale sampling would be required for more detailed geographical mapping of the diversity and more genetically precise refugia identification, especially in the Central African mosaic zone.

SSR genotyping provided highly informative data for multi-locus discrimination of individuals and putative hybrid detection. These data could be related to the recent history of coffee agricultural activities. Inter-group hybridizations were detected at a non-negligible frequency, especially between *a priori* group D (*Guinean* in West Africa) and E (*Congolese* in Central Africa). The previous observation of heterosis of Guinean-Congolese cultivars suggests that the high diversity included in these diversity groups represents an excellent genetic reservoir that could be tapped for adapted stock breeding schemes. This study also provides an excellent basis for determining the appropriate scale of wild population conservation and management.

**Methods**

**Study species and sampled genotypes**

*Coffea canephora* is a widespread species throughout western-central Africa [33]. More than 700 wild genotypes were collected by ORSTOM (now IRD, Institut de Recherche pour le Développement, France) in collaboration with CIRAD, FAO (Food and Agriculture Organization), IPGRI (International Plant Genetic Resources Institute), and MNHN (Muséum National d'Histoire Naturelle), between 1975 and 1987, in five African countries: Côte d'Ivoire and Guinea in West Africa; and Cameroon, Congo and Central African Republic in Central Africa [23-26,33]. These genotypes are conserved in the only reference collection for wild forms of *C. canephora*, *i.e.* the Divo collection, Côte d'Ivoire. In parallel, CIRAD assembled a collection of cultivated material, also conserved at the Divo experimental station. This collection contains more than 600 accessions of diverse origins: local varieties and populations, forms taken from village plantations, and selected material [26,33,35,57].

In this study, a total of 107 *Coffea canephora* genotypes (61 wild and 46 cultivated) were selected from the initial set analysed in Dussert et al. [35]. The wild genotypes were sampled in order to have a representation of each of the 40 forest populations studied throughout the range of the species (Figure 1, Table 1). For the cultivated material, a random proportional sampling was done for each of the 10 principal origins identified in collection (Table 2).

Total genomic DNA was extracted according to the method described by Ky et al. [58].

**RFLP analysis**

Two restriction enzymes were used to digest the genomic DNA: EcoRI and HindIII. Out of the 26 initially tested probes, eight were retained for their polymorphic and mono-locus characteristics. Selected RFLP probes corresponded to previously mapped loci distributed on six linkage groups of the *C. canephora* linkage map (Table 3) [59].

**SSR analysis**

Sixteen SSR loci (Table 3) were amplified as previously described [38,42] using a touchdown PCR profile optimised for each set of primers: touchdown 60°C to 55°C or touchdown 55°C to 50°C. PCR products were detected on an IR2 Automated DNA Sequencer (LI-COR, model 4200L-2, Lincoln, NE, USA) using an M13 primer coupled to the infrared tag IRD700 or IRD800 after migration on 25 cm 6.5% KBplus (LI-COR, CAT# 827-05607) polyacrylamide gels. The gel images were processed by SAGA GT™ software (LICOR Biotech) to estimate the size of amplicons according to a 50–350 bp size standard (LI-COR, CAT# 829-05343, 829-05344).

The sixteen SSR loci were evenly distributed throughout the *Coffea* genome and nine of them mapped on nine different linkage groups of the intespecific map ([*C. heterocalyx* × *C. canephora*] × *C. canephora*) (Table 3) [60].

**Data analysis**

**Genetic diversity**

For each SSR and RFLP locus, we assessed genetic polymorphism within total or wild samples by calculating the observed number of alleles (*Na*), observed and expected heterozygosity (*Ho* and *He*), and the polymorphism information content (PIC) using the POWERMARKER v3.25 software [61]. In these analyses, individual wild plants of possible hybrid origin, as determined by genetic structure analysis (see below), were excluded from the wild pool.

**Genetic affinities among individual genotypes, Cluster analyses**

Neighbour-joining trees were constructed using the shared-allele distances for both the RFLP and SSR data using POWERMARKER v3.25 software [61]. Bootstrapping was conducted with 2000 replicates and the trees were implemented in the PHYLIP package [62] to obtain a consensus tree, visualized in TREEVIEW (taxonomy.zoology.gla.ac.uk/rod/treeview.html). The Mantel matrix correspondence test was used to compare individual genetic distances generated by each marker type using POWERMARKER v3.25.

In order to display patterns in the individual genetic distances, a principal coordinate analysis (PCO) was performed on both the RFLP and SSR dataset. For each individual, we calculated the frequencies of each allele (0, 0.5, and 1) at each locus, and used this data to perform a
principal component analysis (PCA), "French PCA" in STATISTICA V6.1 [http://www.statsoft.com]. This analysis was computed for the whole sample, with cultivated and putative hybrids, as determined by genetic structure analysis (see below), included as additional passive elements.

**Genetic structure**

For statistical investigation of genetic structure of the *C. canephora* wild sample and detection of intermediate types (hybrids between diversity groups), two different Bayesian analyses were performed.

The genetic structure was first investigated for both RFLP and SSR datasets with the Bayesian approach in STRUCTURE v2.1 [45,47]. Parameters were set at K = 5 for the number of groups, 30,000 for the burn-in time and 1,000,000 for the number of runs, with five repetitions. The number of clusters (K = 5) was confirmed as the value that maximized the increase in the posterior probability of the Ln P(D) data according to the formula [Ln P(D|K)-Ln P(D|K-1)] as suggested by Garnier et al. [63]. We assessed the proportion of qi membership of each genotype to the five genetic groups, i.e. the proportion of its genome drawn from each group. We assigned each individual genome to one group when the average proportion of membership was qi > 0.80, i.e. over 80% ancestry to their own cluster. In the case of admixed individuals, we jointly assigned them to two or more groups if the proportion of membership to each one was 0.20 < qi < 0.80. For each individual, we calculated a 90% CI of the qi parameter.

We also used the "leave one out" procedure and the method proposed by Cornuet et al. [64]. The Bayesian-based maximum likelihood test implemented in GENELCLASS v2.0 [46] has been shown to be effective in genotyping assignment, even when populations deviate from Hardy-Weinberg equilibrium [65,66]. Each individual to be reassigned was removed from its source group and the frequency estimates of each locus were modified accordingly (Monte Carlo simulations of 1,000 independent individuals for each candidate group). Differences in log-likelihood values were computed to assign individuals to one group with a risk of 0.01. The reference groups used were the *a priori* A, B, C, D, and E groups. This software was also used to assign cultivated plants to the wild genetic groups.

**Genetic differentiation**

For the analysis of genetic differentiation within the wild sample, individual plants of possible hybrid origin were excluded. The partition of the genetic variation between wild genetic groups for both RFLP and SSR data was estimated with the *F*~ST~ of Weir and Cockerham implemented in GENETIX [67]. Significance levels of pairwise *F*~ST~ values were calculated using permutation tests (N = 1000). The correlation between the two *F*~ST~ matrices generated by each marker type was investigated by Mantel's test of matrix correspondence in GENETIX.

**Core set**

To assist in the use or conservation of wild and cultivated *Coffea canephora* germplasm, we defined core sets of accessions that capture the maximum RFLP or SSR diversity using the principal component score strategy (PCSS) [68]. Based on Khi-2 distances, a factorial analysis is applied to transform initial data into factor scores. Iterative selection of individuals maximising subset variability is based on their relative contribution to the generalised sum of squares (GSS), expressed in percentage.

**Authors' contributions**

CG carried out the PCR amplification experiments and the genotyping, participated in the analyses and drafted the manuscript. SD obtained all the specimens, carried out the RFLP experiments and helped in analyzing the diversity data, and also drew up some of the figures. PH, SH and AK helped in analyzing the data. AdK and SH coordinated the project. VP served as the principal investigator of the project, participated in its design and coordination, helped in the data analysis, assisted in the drafting of the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**

CG received French government funding for her PhD thesis. The authors thank Jérome Minier for technical support on the SSR experiments and Jean Maley for passionate discussions. We are grateful to Madalena Simas Branco (CIBIO/UP) for her critical review of the manuscript.

**References**

1. White F: The guineo-congolian region and its relationships to the other phytocenosis. *Bull Jard Bot Natl Belgu* 1979, 49(1):1-55.
2. Maley J: The African rain forest: main characteristics of changes in vegetation and climate from the upper Cretaceous to the Quaternary. *Proc R Soc Edinburgh* 1996, 104B:31-73.
3. Linder HP: Plant diversity and endemism in sub-Saharan tropical Africa. *Journal of Biogeography* 2001, 28(2):169-182.
4. Plana V: Mechanisms and tempo of evolution in the African Guineo-Congolian rainforest. *Philos Trans R Soc Lond B Biol Sci* 2004, 359(1450):1585-1594.
5. Mittermeier RA, Rodrigues P, Hoffman M, Pil-Grim J, Brooks T, Mittermeier CG, Lamoreux J, Fonseca GAB: Hotspots revisited: Earth's biologically richest and most threatened terrestrial ecoregions. *Mexico City, Mexico* 2004.
6. Stebbins GL: *Flowering Plants: Evolution Above the Species Level*. Cambridge, MA 1974.
7. Fjeldså J, Lovett JC: Geographical patterns of old and young species in African forest biota: the significance of specific montane areas as evolutionary centres. *Biodiversity and Conservation* 1997, 6(3):325-346.
8. Evans B, Kelley DJ, Tinsley RC, Melnick DJ, Cannatella DC: A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. *Mol Phylogenet Evol* 2004, 33(1):197-213.
9. Haffer J: Speciation in Amazonian forest birds. *Science* 1969, 165(3889):131-137.
10. Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM: Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science* 2001, 293(5538):2242-2245.
11. Mayr E, Ohara RJ: The Biogeographic Evidence Supporting the Pleistocene Forest Refuge Hypothesis. Evolution 1986, 40(1):55-67.

12. Salomon M: Evolutionary biogeography and speciation: essay on a synthesis. Journal of Biogeography 2001, 28:12-27.

13. Robberecht E: Generic Distribution Patterns in Subsaharan African Rubiaceae (Angiospermae). Journal of Biogeography 1998, 25:31-314.

14. Anthony NM, Johnson-Bawe M, Jeffery K, Clifford SL, Abernethy KA, Tutin CE, Lahm SA, White LJT, Utley JF, Wickers EJ, et al.: Postglacial migration of Populus nigra L.: lessons learnt from chloroplast DNA. For-estry and Ecology Management 2005, 206(1-3):71-90.

15. Hewitt GM: Genetic consequences of climatic oscillations in the Quaternary. Philos Trans R Soc Lond B Biol Sci 2004, 359(1442):183-195.

16. cottrell JE, Krystufek V, Tabbener HE, Milner AD, Connolly T, Sign L, Fitch S, Bligne A, Teygue F, Achard P, et al.: Postglacial migration of Populus nigra L.: lessons learnt from chloroplast DNA. For-est and Ecology Management 2005, 206(1-3):71-90.

17. Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R, combes MC, Berthaud J, Noirot M, Hamon S: Coffee, Coffea canephora. In Diversité génétique des plantes tropicales cultivées Edited by: Hamon P, Seguin M, Perrier X, Glaszmann JC-Montpellier, France: CIRAD, 1999:75-79.

18. Lorenzen ED, De Neergaard R, Arctander P, Siegismund HR: The role of hotspots but not melting pots of genetic diversity. Evolutionary biogeography and speciation: essay on a synthesis. Journal of Biogeography 1998, 25:31-314.

19. Lashermes P, Combes MC, Robert J, Trouslot P, D’Hont A, Anthony F, Cheddadi R, combes MC, Berthaud J, Noirot M, Hamon S: Coffee, Coffea canephora. In Diversité génétique des plantes tropicales cultivées Edited by: Hamon P, Seguin M, Perrier X, Glaszmann JC-Montpellier, France: CIRAD, 1999:75-79.

20. Lowe AJ, Gillies AC, Wilson J, Dawson IK: Development of genomic microsatellite markers in Coffea canephora and their transferability to other coffee species. Genome 2007, 50(12):1156-1161.

21. Estoup A: The comparison of RFLP, RAPD, AFLP and SSR (micro-satellite) markers for germplasm analysis. Molecular Breeding 1996, 2(3):225-238.

22. Estoup A, Tailliez C, Cornuet JM, Solignac M: Inference of population structure from random amplified polymorphic DNA analysis. Mol Ecol 2000, 9(7):831-841.

23. Mulloko-Ntoumoure N, Petit RJ, White L, Abernethy K: Chloroplast DNA variation in a rainforest tree (Aucoumea klaineana, Burseraceae) in Gabon. Mol Ecol 2000, 9(3):339-353.

24. Fontaine C, Lowen PN, Sanou H, Maley J, Bouyer J: Genetic diversity of the shea tree (Vitellaria paradoxa C.F. Gaertn), detected by RAPD and chloroplast microsatellite markers. Heredita 2004, 93(6):639-648.

25. Lowe AJ, Gillies AC, Wilson J, Dawson IK: Conservation genetics of bush mango from central/west Africa: implications from random amplified polymorphic DNA analysis. Mol Ecol 2000, 9(7):831-841.

26. Mulloko-Ntoumoure N, Petit RJ, White L, Abernethy K: Chloroplast DNA variation in a rainforest tree (Aucoumea klaineana, Burseraceae) in Gabon. Mol Ecol 2000, 9(3):339-353.

27. Brouat C, McKey D, Douzery EJ: Differentiation in a geographical mosaic of plants coevolving with ants: phylogeny of the Acanthocystis complex (Fabaceae: Caesalpinioideae) using amplified fragment length polymorphism markers. Mol Ecol 2004, 13(21):1157-1171.

28. Berthaud J, Guillaumet JL: Les cafésiers sauvages en Centrafricain- Q: Résultats d’une mission de prospection (Janvier-février 1975). Café Cacao Thé 1978, 22:171-186.

29. Anthony F, Couturon E, Namur de C: Les cafésiers sauvages du Cameroun. Résultats d’une mission de prospection effectuée par l’ORSTOM en 1983. 11e Colloque ASIC, Lomé, Togo 1985:495-501.

30. Namur de C, Couturon E, Eta P, Anthony F: Résultats d’une mis- sion de prospection des cafésiers sauvages du Congo. 12ème Colloque ASIC, Montreux, France 1988:397-404.

31. Le Pierrès D, Charmetan P, Yapo A, Leroy T, Couturon E, Bonemms S, Tehe H: Les cafésiers sauvages de Côte-d’Ivoire et de Guinée: bilan des missions de prospection effectuées de 1984 à 1987. 13ème Colloque ASIC, Poapa, Colombie 1989:420-428.

32. Maurit O, Davis AP, Chester M, Mvungi E, Jaufeerally-Fakim Y, Fay MF: Towards a Phylogeny for Coffea (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. Ann Bot (Lond) 2007, 100(7):1565-1583.

33. Davis AP, Govaerts R, Brison DM, Soffelen P: An annotated tax- onomic conspectus of the genus Coffea (Rubiaceae). Botanical Journal of the Linnean Society 2006, 152(4):465-512.

34. Silvestrini M, Junqueira MG, Favarin AC, Guerreiro O, Maluf MP, Sil- veira MB, Colombo GA: Genetic diversity and structure of the Brazilian and Yemenian Coffea arabica L. accessions using microsatellites markers. Genet Resour Crop Evolution 2007, 54(6):1367-1379.

35. Lashermes P, Combes MC, Robert J, Trouslot P, D’Hont A, Anthony F, Cheddadi R, combes MC, Berthaud J, Noirot M, Hamon S: Coffee, Coffea canephora. In Diversité génétique des plantes tropicales cultivées Edited by: Hamon P, Seguin M, Perrier X, Glaszmann JC-Montpellier, France: CIRAD, 1999:75-79.

36. Estoup A, Angers B: Microsatellites and minisatellites for molecular ecology: theoretical and experimental considerations. In Advances in Molecular Ecology Edited by: Carvalho G. Amsterdam: NATO Press. 1998.

37. Rouxet F: Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 1997, 145(4):1219-1228.

38. Estoup A, Rouxet F, Michalakis Y, Cornuet JM, Adragna M, Guit- orand R: Comparative analysis of microsatellites and allozyme markers: a case study investigating microgeographic differentiation in brown trout (Salmo trutta). Mol Ecol 1998, 7(10):339-352.

39. Nicolas V, Queroz S, Verheyen E, Verheyen W, Mbounba JF, Dillen M, Colyn M: Mitochondrial phylogeny of African wood mice, genus Hylomyscus (Rodentia, Muridae): implications for their taxonomy and biogeography. Mol Phylogenet Evol 2006, 38(3):779-793.

40. Sengal RN, Jones HI, Smith TB: Molecular evidence for host specificity of parasitic nematode microfilariae in some African rainforest birds. Mol Ecol 2005, 14(13):3977-3988.

41. Maley J: La destruction catastrophique des forêts d’Afrique centrale survenue il y a environ 2500 ans exercer encore une
influence majeure sur la répartition actuelle des formations végétales. In Plant systematics and phytogeography for the understanding of african biodiversity Proceedings of the XVIIIth AETFAT Congress, held in 2000 at the National Botanic Garden of Belgium Volume 1. Edited by: Robbrecht E, Degreve J, Friis I. Sys Geogr Pl; 2001:777-796.

54. Maley J, Chepstow-Lusty A: *Elaeis guineensis* Jacq. (oil palm) fluctuation in central Africa during the late Holocene: climate or human driving forces for this pioneering species? Vegetation History and Archaeobotany 2001, 10:117-120.

55. Leroy T, Montagnon C, Cilas C, Yapo A, Charmetant P, Eskes AB: Reciprocal recurrent selection applied to *Coffea canephora* Pierre 3. Genetic gains and results of first cycle intergroup crosses. Euphytica 1997, 95(3):347-354.

56. Anthony F, Combes MC, Astorga C, Bertrand B, Graziosi G, Lashermes P: The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. Theor Appl Genet 2002, 104(5):894-900.

57. Cordier L: Les objectifs de la sélection cafféine en Côte d’Ivoire. Café Cacao Thé 1961, 5(3):147-159.

58. Ky CL, Barre P, Lorieux M, Trouslot P, Akaffou S, Louarn J, Charrier A, Hamon S, Noirot M: Interspecific genetic linkage map, segregation distortion and genetic conversion in coffee (*Coffea* sp.). Theor Appl Genet 2000, 101(4):669-676.

59. Paillard M, Lashermes P, Petiard V: Construction of a molecular linkage map in coffee. Theor Appl Genet 1996, 93(1-2):41-47.

60. Coulibaly I, Revol B, Noirot M, Poncet V, Lorieux M, Carasco-Lacombe C, Minier J, Dufour M, Hamon P: AFLP and SSR polymorphism in a *Coffea* interspecific backcross progeny ([*C. canephora* × *C. heterocalyx*] × *C. canephora*). Theor Appl Genet 2003, 107(6):1148-1155.

61. Liu K, Muse SV: PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 2005, 21(9):2128-2129.

62. Felsenstein J: PHYLIP (Phylogeny Inference Package) version 3.6. In Distributed by the author Department of Genome Sciences, University of Washington, Seattle: 2005.

63. Garnier S, Albert P, Audiot P, Prieur B, Rasplus JY: Isolation by distance and sharp discontinuities in gene frequencies: implications for the phylogeography of an alpine insect species, *Carabus solieri*. Mol Ecol 2004, 13(7):1883-1897.

64. Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M: New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 1999, 153(4):1989-2000.

65. Rannala B, Mountain JL: Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci USA 1997, 94(17):9197-9201.

66. Paetkau D, Calvert W, Stirling I, Strobeck C: Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol 1995, 4(3):347-354.

67. Bellkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F: GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. 4.05th edition. Laboratoire Génome, Populations, Interactions, CNRS UMR Université de Montpellier II, Montpellier (France); 5171:1996-2004.

68. Hamon S, Dussert S, Deu M, Hamon P, Seguin M, Glaszmann JC, Grivet L, Chantereau J, Chevalier MH, Flori A, et al.: Effects of quantitative and qualitative Principal Component Score Strategies (PCSS) on the structure of coffee, rice, rubber tree and sorghum core collections. Genetics Selection Evolution 1998, 30:237-258.