OXFORD

Journal of AOAC INTERNATIONAL, 103(2), 2020, 325–334

doi: 10.1093/jaocint/qsz003
Advance Access Publication Date: 18 March 2020
Article

SPECIAL GUEST EDITOR SECTION

Authentication of *Coffea arabica* Varieties through DNA Fingerprinting and its Significance for the Coffee Sector

Solène Pruvot-Woehl,[1] Sarada Krishnan,[2] William Solano,[3] Tim Schilling,[4] Lucile Toniutti,[1] Benoit Bertrand,[5] and Christophe Montagnon[1,*]

[1]World Coffee Research, 5 Avenue du Grand Chêne, 34270 Saint Mathieu de Tréviers, France, [2]Denver Botanic Garden, 909 York Street, Denver, CO 80206, USA, [3]Tropical Agricultural Research and Higher Education Center (CATIE), Turrialba 30501, Costa Rica, [4]World Coffee Research, Babel Community, 70 Avenue de la République, 13000 Marseille, France, [5]Centre de coopération internationale en recherche agronomique pour le développement (CIRAD), UMR IPME, Montpellier, France

*Corresponding author's e-mail: christophe.montagnon@rd2vision.com

Abstract

Background: Locating the optimal varieties for coffee cultivation is increasingly considered a key condition for sustainable production and marketing. Variety performance varies when it comes to susceptibility to coffee leaf rust and other diseases, adaptation to climate change and high cup quality for specialty markets. But because of poor organization and the lack of a professional coffee seed sector, most existing coffee farms (and even seed lots and nurseries) do not know which varieties they are using. DNA fingerprinting of coffee planting material will contribute to professionalize the coffee seed sector.
Objective: The objective of this paper is i) to check in a large scale the robustness of the existing coffee DNA fingerprinting method based on eight Single Sequence Repeats markers (SRR) and ii) to describe how it can help in moving the needle towards a more professional seed sector.
Method: 2533 samples representing all possible genetic background of Arabica varieties were DNA fingerprinted with 8 SRR markers. The genetic diversity was analyzed and the genetic conformity to varietal references was assessed.
Results: The DNA fingerprinting method proved to be robust in authenticating varieties and trace back the history of *C. arabica* breeding and of the movement of *C. arabica* varieties. The genetic conformity of two important coffee varieties, Marseillesa and Gesha, proved to be 91% and 39% respectively.
Conclusions: DNA fingerprinting provides different actors in the coffee sector with a powerful new tool—farmers can verify the identity of their cultivated varieties, coffee roasters can be assured that marketing claims related to varieties are correct, and most of all, those looking to establish the a more professional and reliable coffee seed sector have a reliable new monitoring tool to establish and check genetic purity of seed stock and nursery plants.
Highlights: While *C. arabica* is primarily self-pollinating, even fixed line varieties appear to be drifting away from their original genetic reference due to uncontrolled cross pollination. A set of 8 SSR markers applied to the largest possible genetically diverse set of samples prove to discriminate between a wide range of varieties Figures confirm that genetic non conformity of coffee varieties can represent up to 61% of checked samples.

Received: 22 October 2019; **Accepted:** 22 October 2019

Coffee production contributes to the income of some 12.5 million, value per annum households around the world (1). The coffee industry as a whole is estimated to generate some 74 billion USD. However, for a long period of time, little attention has been given to cultivated coffee varieties. Being considered a commodity, only coffee species were clearly identified: *Coffea arabica* producing Arabica coffee and *C. canephora* producing the coffee known as Conillon when produced in Brazil and Robusta anywhere else in the world. However, during the last decade, focus on coffee varieties has gained interest. According to Montagnon et al. (2), three main reasons explain this new interest in coffee varieties: i) the rust disease (*Hemileia vastatrix*) crisis in Latin America in early 2012 (3), which shed light on both the vulnerability of farmers growing rust-susceptible, low-yielding varieties, and the vulnerability of the coffee supply in general, ii) the growing evidence for the impact of climate change on the future of coffee growing (4) and the inability of current varieties to cope with higher abiotic stress such as extreme drought or heat (2), and iii) the growth of the specialty coffee market searching for top aromatic quality niches, often related to specific varieties. The most iconic example of the value-generation possibilities of specialty-market-bound specific varieties is the Gesha variety, which in 2018 hit a new world record auction price of 803 USD/lb when the commodity price was 1.11 USD/lb (5).

Although the importance of varieties were not until recently high on the coffee industry agenda, coffee breeders have always been active in creating improved varieties for both Robusta (6, 7) and Arabica (2, 8). Nevertheless, in many countries, the proportion of improved varieties in widespread cultivation is low, with the notable exceptions of Colombia and Honduras (Arabica), and Vietnam (Robusta), where significant renovation and replanting schemes have taken place. Even in major or well-known producing countries like Brazil or Costa Rica, most coffee land is still cultivated using varieties selected in the 1950s, such as Caturra, Catuai or Mundo Novo (9). However, there is growing demand, namely for improved Arabica varieties for both mainstream high production, such as the new generation of F1 hybrids (10), or for specific aromatic quality capable of fetching high prices (2).

Because of past limited appetite for improved varieties, the coffee seed sector has remained poorly organized in most parts of the coffee world. In order for coffee producers to benefit from genetic improvement, and to meet the growing demand for specific varieties, there is great need to professionalize the coffee seed sector (2). The few academic studies focusing on the coffee seed sector addressed East Africa and concluded that the informal exchange of seeds was the main way to access planting material (11–13). Recent research has led to improvements in techniques for the efficient mass multiplication of varieties, namely new Arabica F1 hybrid varieties (10), but it was no under the remit of the efficiency of the seed sector.

World Coffee Research (WCR) identified this coffee sector gap as a major constraint for the long-term sustainability of coffee production. WCR hence decided to take action and proposed tools to professionalize the sector. One major constraint observed is the uncertainty about the genetic conformity of the planting material. After decades of informal seed exchange, sometimes over borders, it is very difficult to ascertain the true-to-type-ness of cultivated varieties. Molecular markers have long been used to describe the genetic diversity of *C. canephora* (14, 15) and *C. arabica* (16). However, no specific method to check genetic conformity of cultivated varieties is available. Several papers report on genetic diversity analysis of coffee cultivars in some countries: Brazil (17–20), Nicaragua (21) or Puerto Rico (22).

In this article, we present the first globally relevant genetic authentication method for Arabica coffee. The main objectives are: i) to check the robustness of the method, and ii) to describe how it can help in moving towards a more professional seed sector.

Materials and Methods

Coffee Samples

WCR maintains a DNA fingerprinting database composed of 2533 Arabica samples gathered since 2014. The samples can be grouped according to their source, to their genetic category or to their geographical origin. There are three main sources of samples:

(1) WCR research populations (22%) analyzed in the process of genetic diversity and varietal experiments. This includes the analysis of 100 accessions representing a core collection of Arabica germplasm (a subset of the germplasm collection of Tropical Agricultural Research and Higher Education Center (CATIE), 24 accessions collected by Schilling and Krishnan in South Sudan in 2012 (T.Schilling, World Coffee Research, personal communication) and varieties included in various experiments.
(2) Nursery verification (10%): WCR Verified^{SM} is a scheme that verifies the production practices of nurseries and seed lots to ensure healthy, genetically-pure plant production. One step of nursery verification is to check the genetic conformity of varieties from applicant seed lots and nurseries. Current varieties that have been tested include Marsellesa, CR95 and Centroamericano.
(3) Anonymized samples (68%) from individuals willing to check their own material.

The geographical origin of these samples is Central and North America (47%), Africa (30%), South America (13%) and Asia (10%).

The different genetic categories were chosen based on the history of *C. arabica* movement and breeding [*see* for instance (23, 24)]. The center of origin of *C. arabica* is the South Western forests of Ethiopia and the Boma Plateau of South Sudan. During the 15th century, coffee was produced in Yemen from some seeds taken out of Ethiopia. In the late 17th and early 18th centuries, seeds were taken out of Yemen to: i) Bourbon Island (today French Reunion Island) giving the Bourbon variety, and ii) to India, and from India to Indonesia, giving the Typica variety. Bourbon and Typica were used in The Americas and Asia to grow cultivated coffee. In the late 19th and early 20th centuries, coffee cultivation started in East and Central Africa with varieties introduced from Bourbon Island but also back from Americas (Bourbon and Typica). Interestingly, Indian germplasm (taken out of Yemen 200 years before) was also introduced. Finally, in East Africa, a few Ethiopian or Sudanese landraces escaping these countries through informal routes were also introduced. Two of the most famous such landraces were Gesha (Ethiopia) and Rume Sudan (Sudan). From these introductions some breeding work were initiated in Kenya (Scott Agricultural Laboratory) and in Tanzania (Lyamungu Research Station). Mulungu, in today's Democratic Republic of Congo, was a later breeding center that supported the Kivu region (Rwanda and Burundi). In the 1960s, fear about the vulnerability of coffee production to common diseases prompted

germplasm-collecting missions, primarily in Ethiopia, led by the FAO (and ORSTOM), and the distribution of collected materials to numerous genebanks. Spurred by the same concern, breeding programs in Latin America redoubled efforts to create varieties resistant to coffee leaf rust and coffee berry disease, primarily using the Timor Hybrid (a natural cross between *C. arabica* and *C. canephora* discovered in the 1920s). It gave rise to a number of introgressed varieties, commonly referred to as Catimor and Sarchimor (24). Finally, F1 hybrids were created in the 2000s as the latest generation of Arabica coffee varieties (25).

In summary, there were only three pathways for *C. arabica* out of Ethiopia and Sudan. The original Yemen pathway, the official FAO (26) and ORSTOM (27) surveys in the 1960s and a few "escapes" through individual initiatives (Gesha, Rume Sudan...).

Based on this history, the following categories of *C. arabica* material were considered (Table 1):

- Out of Ethiopia Yemen pathway
  - Typica/Bourbon (458 samples): All the varieties matching the Bourbon or Typica references or very closely related to one of those two.
  - East African varieties (132 samples): All the varieties that were selected in East Africa since the 1940s until present: SL series, K series...
  - Kivu region varieties (129 samples): All the varieties that were selected in the Kivu regions since the 1940s until present: BM series, Mibirizi, Mulungu...
- Ethiopian landraces (406 samples): Ethiopian accessions that were surveyed in the 1960s by the FAO (26) and ORSTOM missions (27) and a few cultivated landraces outside Ethiopia that did not follow the historical Yemen pathway.
- Sudanese landraces (24 samples): in situ survey of leaves by Schilling and Krishnan in 2012 (T. Schilling, World Coffee Research, personal communication) and the Rume Sudan variety.

Table 1. Number of samples and unique allelic phenotypes for the different categories of coffee genetic material

Categories	Number of samples	Number of unique phenotypes
Typica/Bourbon	458	25
East African varieties	132	34
Kivu varieties	129	63
Ethiopian landraces	406	269
Sudanese landraces	24	11
Introgressed varieties	1150	282
F1 hybrid varieties or experimental crosses	234	68
Total	2533	752

- Introgressed varieties (1150 samples): originating from a breeding program and deriving from the Timor hybrids (mainly Sarchimors and Catimors).
- F1 hybrids (234 samples): selected in the 2000s.

For two varieties (Marsellesa and Gesha), the percentage of individuals matching the reference was evaluated. Those two varieties were chose because: i) we had a significant number of samples in the WCR database (299 and 88 for Marsellesa and Gesha, respectively, and ii) Marsellesa is a recent variety registered under the International Union for the Protection of New Varieties of Plants by Ecom and whose reference was provided by Ecom; the story of Gesha is well referenced [*see* (28) for instance] and the reference is the T.02722 accession held by CATIE in Turrialba (Costa Rica).

DNA Extraction and SSR Marker Analysis

All the operations of DNA extraction and SSR marker analysis were performed by the ADNiD laboratory of the Qualtech company in the South of France (http://www.qualtech-groupe.com/en/).

Cell lysis was performed from few milligrams of leaf tissue with 1 mL of SDS buffer. Deproteination was done using potassium acetate and DNA was then purified with magnetic beads followed by elution in TA buffer.

The same eight SSR primer pairs (Table 2), selected after Combes et al. (29), have been used since the beginning of building up the WCR reference database.

PCR was performed in a 15 μL final volume comprising 30 ng genomic DNA and 7.5 μL of 2× PCR buffer (Type it, Qiagen), and 1.0 μM each of forward and reverse primer (10 μM). Amplifications were carried out in a thermal cycler (Eppendorf) programmed at 94°C for 5 min for initial denaturation, followed by 94°C for 30 s, the annealing temperature depending on the primer used then for 30 s, and 72°C for 1 min, for 35 cycles, followed by a final step of extension at 72°C for 5 min. The final holding temperature was 4°C. PCR samples were run on a capillary gel electrophoresis [Applied Biosystems™ 3130XL with an internal standard (homemade)].

The SSR profiles were first established through the GeneMapper™ Software 6 and then visually inspected.

Robustness of the method was checked with 10 biological repetitions of three different clones (F1 hybrids) for which a 100% repeatability was observed (data not shown).

Data Analysis

Because *C. arabica* is tetraploid, the presence/absence (1/0) was coded for each allele. Strictly speaking, this is a SSR

Table 2. List of primer sequences and PCR product size (in base pairs) used for the allelic phenotyping of *C. arabica*

SSR code	Primer sequence (forward)	Primer sequence (reverse)	Product size
Sat11	ACCCGAAAGAAAGAACCAA	CCACACAACTCTCCTCATTC	143–145
Sat225	CATGCCATCATCAATTCCAT	TTACTGCTCATCATTCCGCA	283–317
Sat235	TCGTTCTGTCATTAAATCGTCAA	GCAAATCATGAAAATAGTTGGTG	245–278
Sat24	GGCTCGAGATATCTGTTTAG	TTTAATGGGCATAGGGTCC	167–181
Sat254	ATGTTCTTCGCTTCGCTAAC	AAGTGTGGGAGTGTCTGCAT	221–237
Sat29	GACCATTACATTTCACACAC	GCATTTTGTTGCACACTGTA	137–154
Sat32	AACTCTCCATTCCCGCATTC	CTGGGTTTTCTGTGTTCTCG	119–125
Sat47	TGATGGACAGGAGTTGATGG	TGCCAATCTACCTACCCCTT	135–169

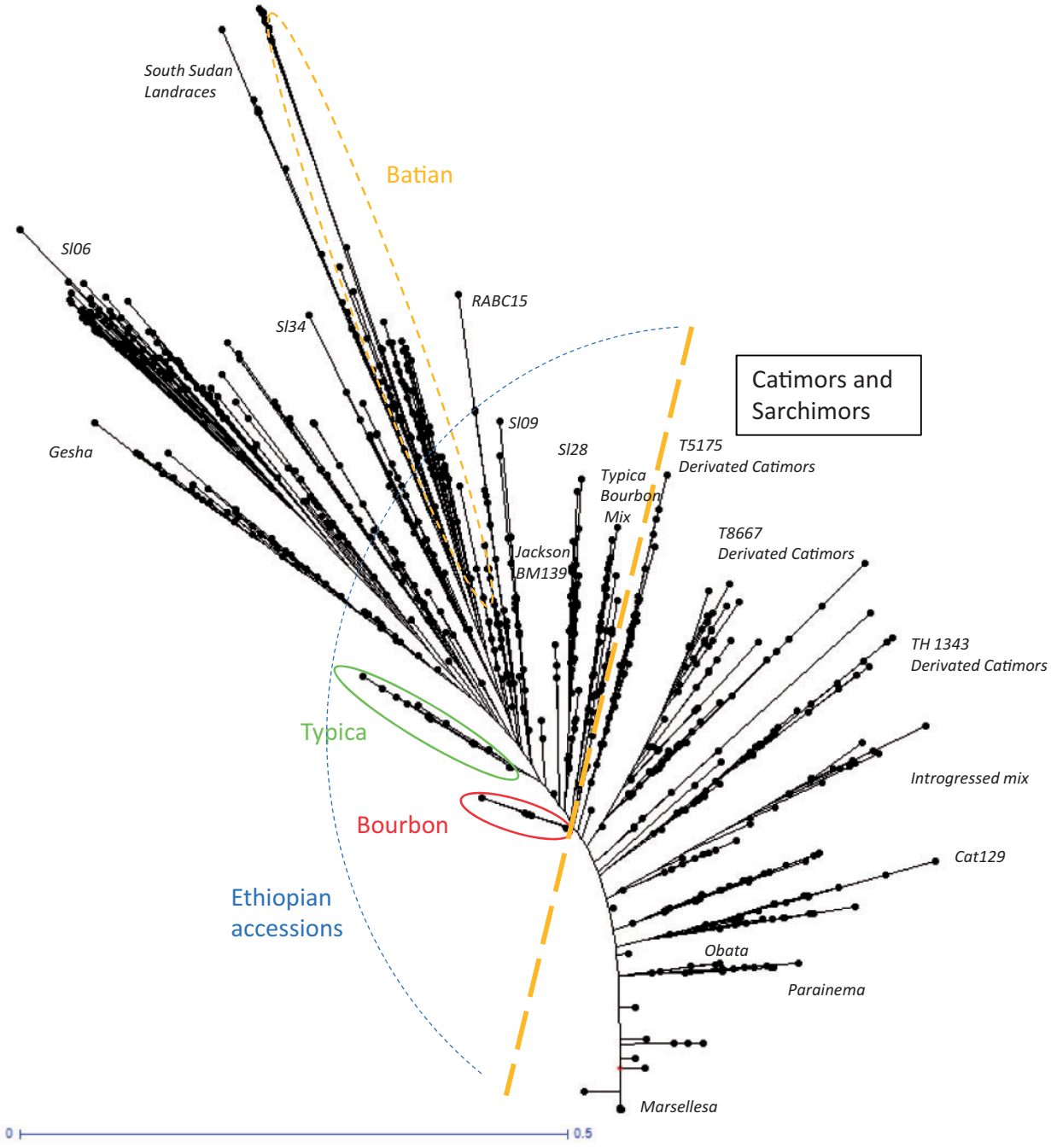

Figure 1. Neighbor-Joining tree from the single allelic data (0/1) of 2533 *C. arabica* samples from the WCR DNA fingerprinting database. The area of some genetic categories or position of some varieties are indicated. One point might represent more than one sample.

allelic phenotype rather than genotype. Indeed, the phenotype AB could be either of the following genotypes: AABB, ABAB, AAAB, ABBB. One practical consequence is that two samples having the same allelic phenotype have not necessarily the same genotype. It means that the percentage of samples matching the reference for Marsellesa and Gesha based on allelic phenotype is an upper limit of the strict genotype matching.

DARwin6 software (30) was used with single data files. A dissimilarity matrix was calculated using Dice Index. The genetic diversity tree was constructed using the Neighbor-Joining method (31).

Results

A total of 95 different alleles were identified across the eight markers. The number of alleles per marker ranged from three to 20. Out of the 95 alleles, 35 (40%) had a frequency ranging between 0.05 and 0.95. Hence, 60% of the alleles were either rare or close to monomorphic.

Out of a total of 2533 samples in the database, 752 corresponded to a unique allelic phenotype over the eight markers (Table 1). This reflected the fact that most of the Typica/Bourbon were either very close to the Typica reference or the Bourbon reference. For other categories, this was mainly due to the relatively

high representation of some specific varieties under specific scrutiny, such as Marsellesa for introgressed varieties, Gesha for Ethiopian landrace or Centroamericano for F1 hybrids.

The overall Neighbor-Joining tree is presented Figure 1. It is not a genetic diversity tree of the *C. arabica* species, but rather a genetic diversity tree of cultivated varieties. Hereafter, the different genetic categories are further explored from a variety authentication perspective.

Introgressed Varieties

All Catimors and Sarchimors have in common a Bourbon parent (Caturra or Villa Sarchi) and a Timor Hybrid. Each single variety was selected after a process of genealogical selection through selfing. The number of selfing generations is not always well documented: it is often between six and nine selfing generations.

All Catimors and Sarchimors are located in the bottom of the tree, below the orange dot line. It is hence easy to differentiate Catimors/Sarchimors varieties from the remaining categories. The precise authentication of varieties depends on the degree of fixation (selfing generation). For instance, Marsellesa, CR95 or Lempira are fixed and homogeneous lines that are easy to authenticate. Small deviations from the reference can be clearly identified as residual segregation that always occurs even in well-fixed varieties.

Other varieties, such as Cat129 in East Africa, became famous and widely used in East Africa without following a controlled and diffusion multiplication system. As a result, the Cat129 fingerprint is more a single branch of the tree than a single reference. Still, this branch is individualized from other varieties.

Central American Catimors that are derived from the T.8667 populations, itself derived from the CIFC HW26 cross involving TH 832/1, are all grouped in a large branch of the tree. Although individual varieties such as CR95 (Costa Rica) can be clearly identified, some samples seem to originate from cross pollinations between different varieties of the same lineage.

In comparison, the T.5175 populations, also descending from an initial HW26 cross, is a narrow branch.

Obata (Brazil), Parainema (Honduras) and Marsellesa (Central America) are Sarchimors derived from the same T.5296 population, itself derived from an initial cross between Villa Sarchi and TH 832/2. Those are well-fixed varieties.

Batian (Kenya) and RABC15 (Rwanda) are two introgressed varieties that are not derived from a single Bourbon variety (Caturra or Villa Sarchi) but from multiple crosses involving many traditional old East African, Indian lines and even Ethiopian or Sudanese landraces (e.g., Rume Sudan). Batian is a composite variety made from different fixed lines whereas RABC15 is a single fixed line derived from Indian Sln6 population. As a result, both varieties are located in the region of old Scott Labs varieties (e.g., SL34, SL14) and Rume Sudan, which compose most of their genetic background. As a multi-line variety, Batian forms a branch whereas RABC15 has an expected narrow genetic fingerprint.

Apart from well-identified introgressed varieties, some samples appear to originate from uncontrolled crosses between various other introgressed varieties. In such cases, it is often difficult to trace back the original genetic composition of the samples.

F1 Hybrids

All released F1 hybrid varieties are clones (10), but the recent Starmaya variety which is distributed by seeds from a male sterile parent (32). It is hence the easiest situation to authenticate a F1 variety because it has a single clear reference. Most recently released F1 hybrids in Central America (Centroamericano, Milenio, Casiopea and Mundo Maya) have clearly a unique allelic phenotype. Interestingly, Centroamericano and Milenio are full Sibs, both clones have been selected from the same cross (T.5296 × Rume Sudan) and were still discriminated with the eight SSR markers.

Typica and Bourbon

Typica and Bourbon form two close and very narrow branches. Their respective reference allelic phenotype is well identified. The tall Bourbon variety and its derived dwarf single gene mutation varieties such as Caturra or Villa Sarchi have exactly the same allelic phenotype with our set of markers. Hence, DNA fingerprinting alone cannot differentiate between those varieties. An additional visual observation of the tree can ascertain the difference between tall and dwarf.

Numerous samples correspond to trees originating from crosses between Bourbon and Typica (Typica/Bourbon mix), most likely uncontrolled crosses that occurred some time in the long history of growing Bourbon and Typica side by side.

East African and Kivu Region Varieties

Most cultivated varieties in East Africa and the Kivu region form a genetic cluster related to Typica and Bourbon. Although the allelic phenotypes of some old varieties such as SL28 or SL34 are clearly identified, cultivated material is found to have mainly originated from different old varieties. Jackson and BM139 from Rwanda are found to be "population" varieties. Trees forming these varieties are genetically diverse and correspond to a genetic cluster rather than to a single reference.

Ethiopian and Sudanese Landraces

Not surprisingly Ethiopian landraces cover the entire genetic variability of the database, except the "Catimors/Sarchimors" area. The Gesha variety is confirmed to be an Ethiopian landrace. The Gesha reference is the T.02722 accession of the CATIE germplasm collection in Costa Rica and most cultivated varieties identified as Gesha are forming an individual genetic cluster (Figures 1 and 3).

South Sudanese landraces compose a uniform genetic group, including the old Rume Sudan landrace. Because Rume Sudan is an important contributor in the genealogy of Batian, South Sudanese landraces cluster closely to Batian.

Focus on Genetic Conformity of Two Varieties: Marsellesa and Gesha

Some varieties in the WCR database are highly represented either because they are part of large nursery verification programs (such as Marsellesa) or because interest in them an important candidate for DNA fingerprinting (such as Gesha).

Out of 299 supposed Marsellesa samples in the WCR database, 91% were confirmed Marsellesa (Figure 2): 82% were an exact match with the Marsellesa reference, and 9% were closely related to Marsellesa (i.e., there is an acceptable level of residual segregation of Marsellesa). However, 8% of samples were nonrelated various Catimors and 1% were Typica or Bourbon.

Out of 88 supposed Gesha samples in the WCR database (Figure 3), 39% were an exact match with the Gesha reference (accession T.2722 from CATIE), and 24% were closely related to

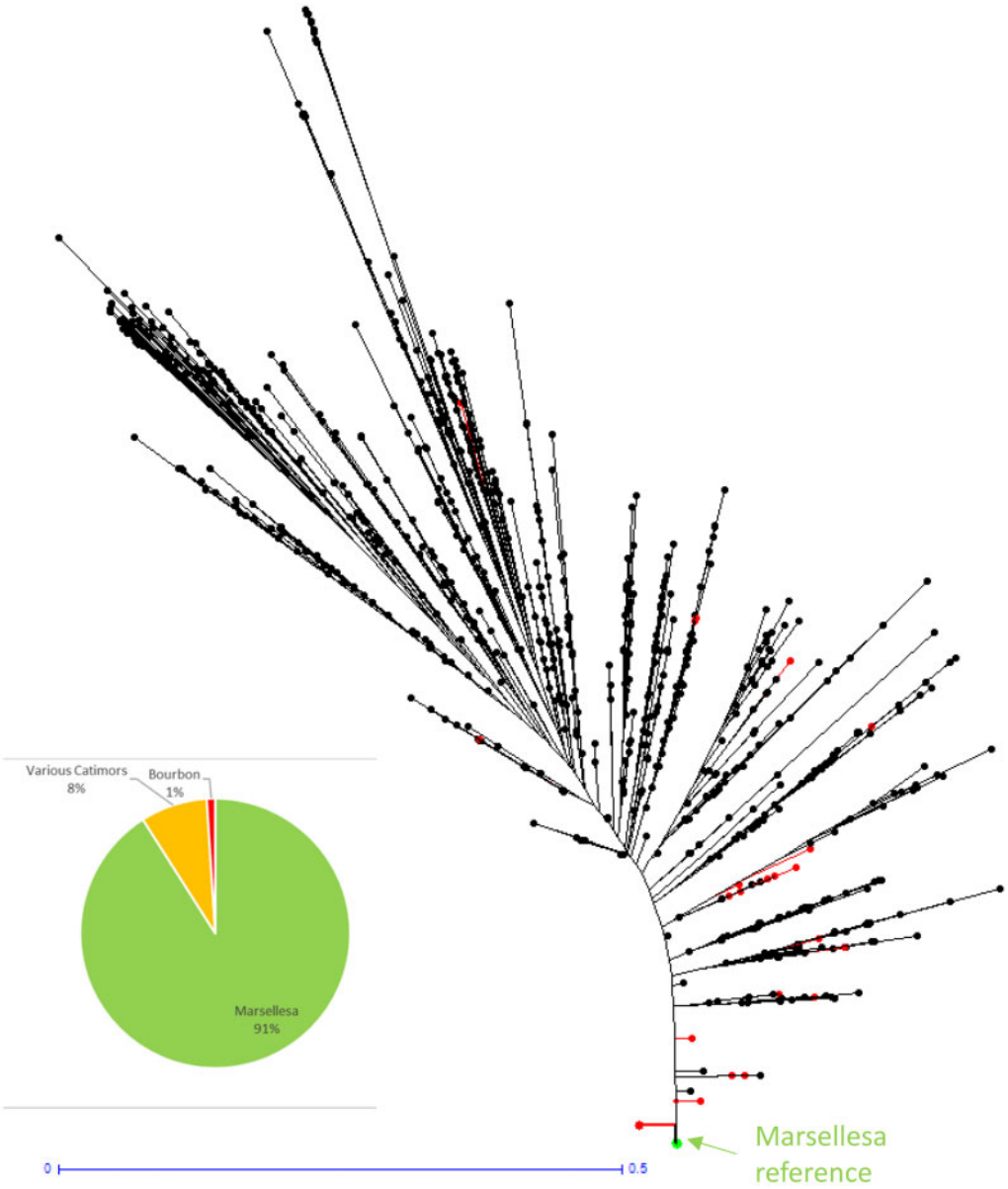

Figure 2. Representation and genetic conformity evaluation of the 299 Marsellesa samples in the Neighbor-Joining tree from the single allelic data (0/1) of 2533 *C. arabica* samples from the WCR DNA fingerprinting database. The black points represent the position of supposed Marsellesa samples. One point might represent more than one sample.

Gesha, forming a Gesha "cluster". However, 37% of samples had an unrelated genetic background; some were old East African varieties such as SL34, and some were Typica/Bourbon mix.

Discussion

Plant genotyping to DNA fingerprint plant varieties is not new (33). However, to our knowledge, our study represents the first attempt to apply DNA fingerprinting to *C. arabica* coffees varieties on such a large scale. In Vieira et al. (17), DNA fingerprinting of *C. arabica* was restricted to Brazil and samples were taken out of a germplasm collection. DNA fingerprinting of plant varieties has been performed in numerous different crops, including tropical tree crops such as cocoa (34, 35) or rubber tree (36). However, all those studies were either to identify duplicates in germplasm collections or to describe the genetic fingerprinting of varieties in germplasm collections. We are not aware of any published study looking at the DNA fingerprinting of tropical tree crops from a seed sector point of view, evaluating the genetic conformity of supposed varieties in the field or of planting material to be distributed.

The present study is the first of its kind to bring knowledge on *C. arabica* DNA fingerprinting of varieties not only from germplasm collections but from cultivated coffee trees around the world.

The first result is that the set of eight SSR markers, representing a total of 95 alleles, that has been used is powerful to discriminate between *C. arabica* varieties. In some instances, it was not possible to distinguish the allelic phenotype of well-defined different varieties. This was the case for Bourbon, Caturra and Villa Sarchi. Caturra and Villa Sarchi are dwarf

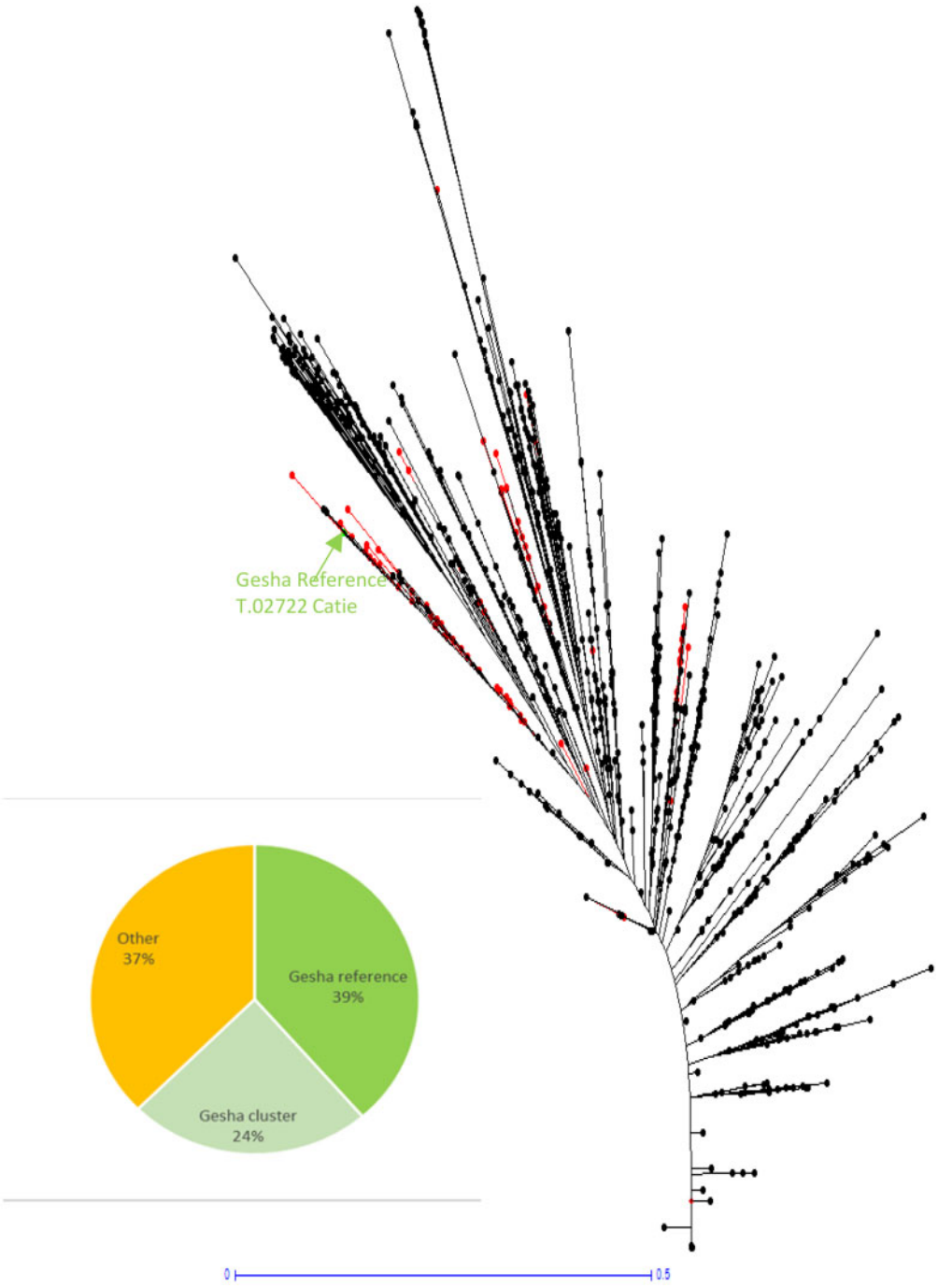

Figure 3. Representation and genetic conformity evaluation of the 88 Gesha samples in the Neighbor-Joining tree from the single allelic data (0/1) of 2533 *C. arabica* samples from the WCR DNA fingerprinting database. The black points represent the position of supposed Gesha samples. One point might represent more than one sample.

mutants of the tall Bourbon variety. It is a single gene mutation that is very unlikely to be spotted through eight SSR markers. Hence, when a Bourbon allelic phenotype is identified, it cannot be concluded which variety it is between Bourbon, Caturra or Villa Sarchi. Only a visual inspection of the tree (tall or dwarf) will help in deciding. Another example where the current markers do not discriminate is between CR95 (Costa Rica) and Lempira (Honduras). These two distinct varieties have the same allelic phenotype; they descend from the same breeding population (T.8667). Those were the only cases when it was not possible to distinguish well-described varieties. Recently, new markers are being tested to fine-tune discrimination in those rare situations.

Our study reveals that a significant share of cultivated Arabica coffee trees in the field are mixes derived from one or several generations of uncontrolled pollinations between existing varieties. This is the case in Central America for Catimors and Sarchimors. (In this situation, it is possible to locate variety references because the initial selections from the 1990s are held in various research center collections.)

In East Africa and the Kivu region, the situation is different. Recent varieties officially released by national breeding programs such as Batian or Ruiru 11 in Kenya are often conforming. The story is different with varieties such as Cat129, which gained a reputation for disease resistance and has been widely planted in Southern East Africa, but which has never been formally released or distributed by a research institute, with the controls that typically accompany such a release. Consequently, the allelic phenotype for Cat129 is well defined; however, a lot of outliers can be found.

Older East African and Kivu region varieties present an interesting case. "True" SL28 and SL34 are widespread, but so are populations that evolved from varieties like the SLs and K7, which were initially well defined, but have drifted and now contain a mix of the initial alleles. It is not necessarily a risky situation. It might even be argued that the current cultivated trees were selected along generations and their allelic richness might be an added value for resilience.

Old varieties such as BM71, BM139 or Jackson in Kivu can be described today as population varieties, which also represent a mix of alleles. Each of these varieties covers a genetic fingerprint which is not a single reference but rather a cluster. There is sometimes overlap between the clusters. In those situations, the best one can tell from the DNA fingerprint of a sample is that it is, for instance, "compatible" with BM139 or Jackson.

The Ethiopian and Sudanese landraces are easily fingerprinted, and the fingerprints are numerous and different. The Gesha or the Rume Sudan reference has a unique allelic phenotype.

Experience with the WCR genetic database points to the conclusion that a recently-selected variety in a region with a relatively organized research and nurseries network shows a satisfactory genetic conformity. The best example of this currently is the Marsellesa variety, with 91% of genetic conformity.

However, when varieties are older and/or the research and nurseries network is poorly organized, the percentage of genetic conformity can drastically decrease. The figure of 39% of genetic conformity for Gesha can be interpreted in different ways. It can be considered as low, but one might also consider that it is high in the absence of any formal channel of Gesha seed distribution. Furthermore, 24% of supposed Gesha are still close to the Gesha reference, forming a Gesha cluster. Nevertheless, the reputation and diffusion of Gesha is very recent (over the past 10–15 years). Without a formal seed sector, it is very likely that the share of genetic conformity of Gesha will shrink rather than increase.

One could wonder why there is such high genetic nonconformity when *C. arabica* is a self-pollinating species and the great majority of cultivated varieties are fixed lines. In agreement with possible reasons given for cocoa by Turnbull et al. (34), the main source of error is likely human: mislabeling, erroneous (good faith) belief about the name of the variety, and unsure traceability of some uncontrolled (sometimes contraband) movements of seeds. However, the fact that *C. arabica* can self-pollinate does not mean that it always self-pollinates. Various studies have established that the share of cross-pollination in *C. arabica* can reach 10 to 15% (37–39). Furthermore, recent works in Ethiopian forests showed that this share could go up to more than 50% (40). Several studies corroborate that coffee pollen can be transported by wind or insects over a distance up to 2 km (41–43).

To the best of our knowledge, Charrier (44) published the only study where the distance of transport of pollen within a plot was established, using pollen that had been radio-activated with phosphorus (^{32}P) and sulphur (^{35}S). In a plot with coffee trees of 3–4 m in height and a density of 1111 trees per hectare, during a typical good blossom, the pollen could be transported up to 42 m.

Hence, the genetic drift of an originally pure variety is very likely to happen through cross pollination if no specific measures are taken to isolate the plot and/or to harvest seeds from only the inner trees of the plot. Genetic drift can be significantly enhanced by the hybrid vigor phenomena, now well established in *C. arabica* (25) and *C. canephora* (45). For example, if there is a fixed-line variety, say a Bourbon, and just nearby there is a Gesha variety, and the seeds of the Gesha plot are harvested to plant a new plot even if there is a low contamination of Bourbon pollen in that Gesha plot, the few seeds resulting from a cross involving the Bourbon pollen will give rise to a F1 hybrid tree between Bourbon and Gesha. In the new plot, those F1 trees will be the more vigorous and high yielding and thus more likely to be selected by the farmer to plant a new generation of plots, because he/she will harvest the seeds from the most vigorous trees. Then, starting from what could have been less than 10% contamination of pollen, may end up after two generations in a Gesha plot that is in fact a F2 segregating population of an accidental Bourbon x Gesha F1 cross and the original true Gesha has been lost.

Through analysis of a large data set of coffee variety samples from farms, nurseries, and seed lots around the world, we have shown that the genetic conformity of coffee material is frequently questionable. We believe this is primarily due to the lack of order and good practices in the coffee seed sector. This is

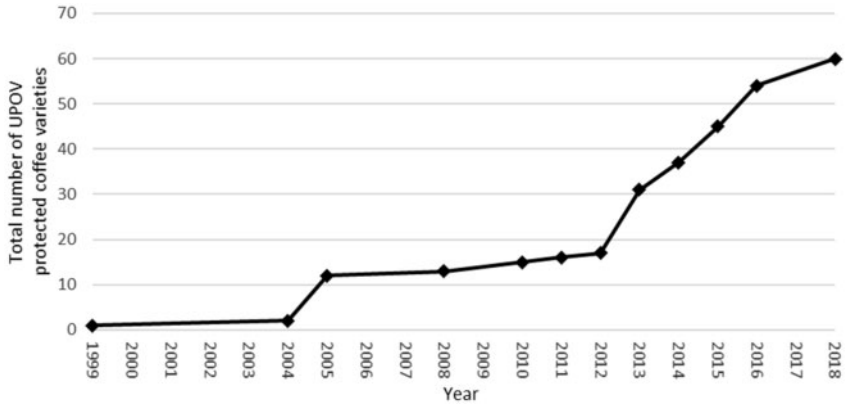

Figure 4. Evolution of the number of coffee protected varieties under International Union for the Protection of New Varieties of Plants (UPOV). Source: https://www.upov.int/pluto/en/.

a concern when improved varieties are increasingly acknowledged to be essential for a sustainable coffee industry, namely for disease resistance, adaptation to climate change and market demand for high quality coffees. The ability to have variety conformity (and authentication) is very likely to be an essential precondition to spurring increased investments in coffee breeding, and ensuring that genetic improvement reaches farmers.

The International Union for the Protection of New Varieties of Plants (UPOV) is the worldwide recognized institution to register new varieties and protect breeders' rights. The evolution of the number of UPOV registered coffee varieties (Figure 4) shows that there has been a major increase since 2012. Hence, even if the total number of registered varieties is still low (60 to be compared to more than 700 varieties for watermelon for instance), there is clearly a trend towards more variety innovation. This is an interesting trend as the first condition to have a genetic conformity is to have well-described reference varieties, which is an obligation when the variety is protected under UPOV.

Conclusions

For the first time we have demonstrated on a large scale the robustness of variety authentication of Arabica coffee varieties through a set of SSR markers. In addition to the 100% repeatability of the allelic phenotype of an individual, the vast majority of the varieties could be discriminated, including in the only situation when the varieties are known to be Full Sibs (Centroamericano and Milenio clones). Only when two varieties are different because of a single mutation (dwarf Caturra from tall Bourbon for instance) was it impossible to discriminate the two with the eight markers. However, in this case, a visual observation of the trees with the same "Bourbon" genetic background is enough to decide if it is a Bourbon or a Caturra. This DNA fingerprinting method provides nurseries, farmers and the whole coffee industry with a unique opportunity to increase knowledge about the genetic identify of trees that are planted or seeds that are traded. Our results show that most varieties can be easily identified through their allelic phenotype. In the field or in nurseries, recently bred and released varieties such as Marsellesa show an acceptable genetic conformity (91% for Marsellesa). However, we also show that highly sought after varieties whose seeds have not moved through formal pathways have much less genetic conformity (39% for Gesha). Genetic drift most likely related to the contamination of pollen from other varieties is significant.

Authentication of coffee varieties is acknowledged as a key part of coffee sector sustainability in the context of both disease pressure and climate change effects, and also growing demand for high-quality coffees. The coffee seed sector in general needs to be more organized and professional to ensure the genetic conformity that the coffee stakeholders deserve—and increasingly, demand—from farm to cup. In the last 5 years, WCR has produced several tools to contribute to the professionalization of the coffee seed sector, including a catalog of Arabica coffee varieties (46), WCR Verified^{SM}, a seed lots and nursery verification program (47), and the DNA fingerprinting tool presented here. As the coffee sector becomes increasingly professionalized, the genetic conformity of planting material will increase, and this may be monitored using the DNA fingerprinting tool described here. Meanwhile, the many actors in the coffee sector are now able to identify the varieties they are dealing with.

Acknowledgments

WCR thanks all its members and partners for making this study. The list of WCR members and partners is available here: https://worldcoffeeresearch.org/about/.

Guest edited as a special report on "Green and Roasted Coffee Authentication: Species, Origin and Diluent Methods of Analysis" by Brian T. Schaneberg.

Conflict of interest: World Coffee Research and RD2 operate DNA fingerprinting as a fee-based service.

References

1. Browning, D. (2018). *How many coffee farms are there in the world?* ASIC Conference Portland, 16–20 September 2018, https://www.youtube.com/watch?v=vKaeDkpqPSg (accessed August 24, 2019)
2. Montagnon, C., Marraccini, P., & Bertrand B. (2019) in *Specialty Coffee - Managing Quality*, 2nd Ed., T. Oberthür, P. Läderach, H.A.J. Pohlan, P. Lee Tan & J. Cock (Eds), Cropster, Innsbruck, Austria. 109 –143
3. Avelino, J., Cristancho, M., Georgiou, S., Imbach, P., Aguilar, L., Bornemann, G., Läderach, P., Anzueto, F., Hruska, A.J., & Morales, C. (2015) *Food Secur*. **7**, 303–321. doi:10.1007/s12571-015-0446-9
4. Bunn, C., Läderach, P., Jimenez, J.G.P., Montagnon, C., & Schilling, T. (2015) *PLoS One* **10**, e0140490
5. Morris, J. (2019) *Coffee: A Global History*, Reaktion Books Ltd, London, UK
6. Montagnon C., Leroy T., & Eskes, A.B. (1998) *Plant. Rech. Dév*. **5**, 18–33
7. Montagnon C., Leroy T., & Eskes, A.B. (1998) *Plant. Rech. Dév*. **5**, 89–98
8. Van der Vossen, H., Bertrand, B., & Charrier, A. (2015) *Euphytica* **204**, 243–256. doi:10.1007/s10681-015-1398-z
9. Guerreiro Filho, O., Ramalho, M.A.P., & Andrade, V.T. (2018) *Crop Breed. Appl. Biotechnol*. **18**, 460–466. doi:10.1590/1984-70332018v18n4p69
10. Etienne, H., Breton, D., Breitler, J.C., Bertrand, B., Déchamp, E., Awada, R., Marraccini, P., Léran, S., Alpizar, E., Campa, C., & Courtel, P. (2018) *Front. Plant Sci.* **9**, 21. doi: 10.3389/fpls.2018.01630
11. Louwaars, N.P., De Boef, W.S., & Edeme, J. (2013) *J. Crop Improv*. **27**, 186–214. doi:10.1080/15427528.2012.751472
12. Munyi, P., & De Jonge, B. (2015) *J. Sustain. Dev*. **8**, 161–173. doi:10.5539/jsd.v8n2p161
13. Sisay, D.T., Verhees, F.J., & van Trijp, H.C. (2017) *J. Crop Improv*. **31**, 323–355. doi:10.1080/15427528.2017.1303800
14. Garavito, A., Montagnon, C., Guyot, R., & Bertrand, B. (2016) *BMC Plant Biol*. **16**: 242. doi:10.1186/s12870-016-0933-y
15. Merot-L'anthoene, V., Tournebize, R., Darracq, O., Rattina, V., Lepelley, M., Bellanger, L., Tranchant-Dubreuil, C., Coulée, M., Pégard, M., Metairon S., Fournier C., Stoffelen, P., Janssens, S.B., Kiwula, C., Musoli, P., Sumirat, U., Legnate, H., Kambale, J.L., da Costa Neto, J.F., Revel, C., de Kochko, A., Descombes, P., Crouzillat, D., & Poncet, V. (2018) *Plant Biotechnol. J*. **17**, 1418–1430. doi:10.1111/pbi.13066
16. Teressa, A., Crouzillat, D., Petiard, V., & Brouhan, P. (2010) *EJAST* **1**, 63–79
17. Vieira, E.S., Von Pinho, É.V.D.R., Carvalho, M.G., Esselink, D.G., & Vosman, B. (2010) *Genet. Mol. Biol*. **33**, 507–514. doi:10.1590/S1415-47572010005000055

18. Motta, L.B., Soares, T.C.B., Ferrão, M.A.G., Caixeta, E.T., Lorenzoni, R.M., & Souza Neto, J.D.D. (2014) *Braz. Arch. Biol. Technol.* **57**, 728–735. doi:10.1590/S1516-8913201402071
19. Sousa, T.V., Caixeta, E.T., Alkimim, E.R., de Oliveira, A.C.B., Pereira, A.A., Zambolim, L., & Sakiyama, N.S. (2017) *Euphytica* **213**, 75. doi:10.1007/s10681-017-1865-9
20. da Silva, B.S.R., Sant'Ana, G.C., Chaves, C.L., Androcioli, L.G., Ferreira, R.V., Sera, G.H., Charmetant, P., Leroy, T., Pot, D., Domingues, D.S., & Pereira, L.F.P. (2019) *Genetica* **147**, 205–216. doi:10.1007/s10709-019-00064-4
21. Geleta, M., Herrera, I., Monzón, A., & Bryngelsson, T. (2012) *Sci. World J.* **2012**, 11. doi:10.1100/2012/939820
22. Zhou, L., Vega, F.E., Tan, H., Lluch, A.E.R., Meinhardt, L.W., Fang, W., Mischke, S., Irish, B., & Zhang, D. (2016) *Trop. Plant Biol.* **9**, 82–95. doi:10.1007/s12042-016-9167-2
23. Haarer, A.E. (1958) *Modern Coffee Production*, Leonard Hill, UK
24. Kushalappa, A.C., & Eskes, A.B. (1989) *Annu. Rev. Phytopathol.* **27**, 503–531. doi:10.1146/annurev.py.27.090189.002443
25. Bertrand, B., Alpizar, E., Lara, L., Santacreo, R., Hidalgo, M., Quijano, J.M., Montagnon, C., Georget, F., & Etienne, H. (2011) *Euphytica* **181**, 147–158. doi:10.1007/s10681-011-0372-7
26. FAO (1968) *FAO Coffee Mission to Ethiopia: 1964–1965*. FAO, Rome
27. Charrier, A. (1978) Etude de la structure et de la variabilité génétique des caféiers: Résultats des études et des expérimentations réalisées au Cameroun, en Côte d'Ivoire et à Madagascar sur l'espèce *Coffea arabica* L. collectée en Ethiopie par une mission Orstom en 1966. Bulletin IFCC no 14, Paris, France
28. Bramel, P., Krishnan, S., Horna, D., Lainoff, B., & Montagnon, C. (2017) *Global Conservation Strategy for Coffee Genetic Resources*, Crop Trust and World Coffee Research, Portland, OR, 72 p
29. Combes, M.C., Andrzejewski, S., Anthony, F., Bertrand, B., Rovelli, P., Graziosi, G., & Lashermes, P. (2000) *Mol. Ecol.* **9**, 1178–1180. doi:10.1046/j.1365-294x.2000.00954-5.x
30. Perrier, X., & Jacquemoud-Collet, J.P. (2006) DARwin Software, http://darwin.cirad.fr/darwin (accessed August 24, 2019)
31. Saitou, N., & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425. doi:10.1093/oxfordjournals.molbev.a040454
32. Georget, F., Alpizar, E., Courtel, P., Hidalgo, J.M., Dechamp, E., Poncon, C., Etienne, H., & Bertrand, B. (2014) *Development of a Male Sterility Based Reproductive System to Ensure a Cost Effective and Massive Propagation of New Outstanding F1 Arabica Hybrids by Seed*. 25th International Conference on Coffee Science, ASIC 2014, Armenia, Colombia, 8–13 September 2014
33. Smith, J.S.C., & Smith, O.S. (1992) *Adv. Agron.* **47**, 85–140. doi:10.1016/S0065-2113(08)60489-7
34. Turnbull, C.J., Butler, D.R., Cryer, N.C., Zhang, D., Lanaud, C., Daymond, A.J., Ford, C.S, Wilkinson, M.J & Hadley, P. (2004) *Ingenic Newsl.* **9**, 8–11
35. Motilal, L.A., Zhang, D., Umaharan, P., Mischke, S., Pinney, S., & Meinhardt, L.W. (2011) *Plant Genet. Resour.* **9**, 430–438. doi:10.1017/S147926211100058X
36. Saha, T., Roy, C.B., & Nazeer, M.A. (2005) *Plant Breed.* **124**, 86–92. doi:10.1111/j.1439-0523.2004.01053.x
37. Carvalho A., & Krug, C.A. (1949) *Bragantia* **9**, 11–24
38. Castillo-Zapata J. (1976) *Cenicafe* **27**, 78–88
39. Anthony, F., Bertrand, B., Quiros, O., Wilches, A., Lashermes, P., Berthaud, J., & Charrier, A. (2001) *Euphytica* **118**, 53–65. doi:10.1023/A:1004013815166
40. Berecha, G., Aerts, R., Vandepitte, K., Van Glabeke, S., Muys, B., Roldán-Ruiz, I., & Honnay, O. (2014) *Biol. J. Linn. Soc.* **112**, 76–88. doi:10.1111/bij.12274
41. Portères R. (1937) *Ann. l'Afrique Occidentale Française Etrangère* **1**, 68–91
42. Berthaud J. (1985) in *Genetic Differentiation and Dispersal in Plants. Nato Asi Series, Series G: Ecological Sciences*, Vol. **5**, P. Jacquart, G. Heim, & J. Antonovics (Eds), Springer, Berlin, pp 355–366
43. Jha, S., & Dick, C.W. (2010) *Proc. Natl. Acad. Sci. USA* **107**, 13760–13764. doi:10.1073/pnas.1002490107
44. Charrier A. (1971) *Café Cacao Thé* **15**, 181–190
45. Leroy, T., Montagnon, C., Charrier, A., & Eskes, A.B. (1993) *Euphytica* **67**, 113–125. doi:10.1007/BF00022734
46. World Coffee Research (2019) Arabica Coffee Varieties. 75 p, https://varieties.worldcoffeeresearch.org/varieties (accessed August 24, 2019)
47. World Coffee Research (2019) WCR Verified^{SM} Nurseries, https://varieties.worldcoffeeresearch.org/info/verified/about (accessed August 24, 2019)