Annals of Botany **XX**: 1–13, 2024
https://doi.org/10.1093/aob/mcae034, available online at www.academic.oup.com/aob

ANNALS OF BOTANY
Founded 1887

RESEARCH IN CONTEXT

Crop-to-wild gene flow in wild coffee species: the case of *Coffea canephora* in the Democratic Republic of the Congo

Lauren Verleysen[1,2,†], Jonas Depecker[1,3,4,*,†], Robrecht Bollen[1,3], Justin Asimonyio[5], Yves Hatangi[3,6,7], Jean-Léon Kambale[5], Ithe Mwanga Mwanga[8], Thsimi Ebele[9], Benoit Dhed'a[6], Piet Stoffelen[3], Tom Ruttink[2,10], Filip Vandelook[1,3] and Olivier Honnay[1,4]

[1]Division of Ecology, Evolution and Biodiversity Conservation, KU Leuven, Leuven, Belgium, [2]Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium, [3]Meise Botanic Garden, Meise, Belgium, [4]KU Leuven Plant Institute, Leuven, Belgium, [5]Centre de Surveillance de la Biodiversité et Université de Kisangani, Kisangani, DR Congo, [6]Université de Kisangani, Kisangani, DR Congo, [7]Liège University, Gembloux Agro-Bio Tech, Gembloux, Belgium, [8]Centre de Recherche en Science Naturelles, Lwiro, DR Congo, [9]Institut National des Etudes et Recherches Agronomique, Yangambi, DR Congo and [10]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium
[†]These authors contributed equally to this work.
[*]For correspondence: E-mail jonas.depecker@kuleuven.be

Received: 22 January 2024 Editorial decision: 28 February 2024 Accepted: 1 March 2024

• **Background and Aims** Plant breeders are increasingly turning to crop wild relatives (CWRs) to ensure food security in a rapidly changing environment. However, CWR populations are confronted with various human-induced threats, including hybridization with their nearby cultivated crops. This might be a particular problem for wild coffee species, which often occur near coffee cultivation areas. Here, we briefly review the evidence for wild *Coffea arabica* (cultivated as Arabica coffee) and *Coffea canephora* (cultivated as Robusta coffee) and then focused on *C. canephora* in the Yangambi region in the Democratic Republic of the Congo. There, we examined the geographical distribution of cultivated *C. canephora* and the incidence of hybridization between cultivated and wild individuals within the rainforest.
• **Methods** We collected 71 *C. canephora* individuals from home gardens and 12 *C. canephora* individuals from the tropical rainforest in the Yangambi region and genotyped them using genotyping-by-sequencing (GBS). We compared the fingerprints with existing GBS data from 388 *C. canephora* individuals from natural tropical rainforests and the INERA Coffee Collection, a Robusta coffee field gene bank and the most probable source of cultivated genotypes in the area. We then established robust diagnostic fingerprints that genetically differentiate cultivated from wild coffee, identified cultivated–wild hybrids and mapped their geographical position in the rainforest.
• **Key Results** We identified cultivated genotypes and cultivated–wild hybrids in zones with clear anthropogenic activity, and where cultivated *C. canephora* in home gardens may serve as a source for crop-to-wild gene flow. We found relatively few hybrids and backcrosses in the rainforests.
• **Conclusions** The cultivation of *C. canephora* in close proximity to its wild gene pool has led to cultivated genotypes and cultivated–wild hybrids appearing within the natural habitats of *C. canephora*. Yet, given the high genetic similarity between the cultivated and wild gene pool, together with the relatively low incidence of hybridization, our results indicate that the overall impact in terms of risk of introgression remains limited so far.

Key words: *Coffea canephora*, Robusta coffee, crop wild relatives (CWRs), cultivated–wild hybridization, Congo Basin, field gene bank, gene flow, introgression.

INTRODUCTION

A growing human population in an increasingly warmer world may jeopardize the food security of many households (FAO, 2018; Bohra *et al*., 2022). To cope with this future challenge, innovations in plant breeding and agricultural systems are an important part of the solution (Zhang *et al*., 2017; Bohra *et al*., 2022). Crop improvement should focus on increasing yields, enhancing crop climate change resilience and providing nutritional security (Brozynska *et al*., 2016). To do so, plant breeders can fall back on crop wild relatives (CWRs), which harbour genetic diversity that is not present in their cultivated relatives, and may underlie desirable traits, including resistance to pests and diseases, and tolerance to abiotic stresses (Heywood *et al*., 2007; Zhang *et al*., 2017; Saeed and Fatima, 2021). Conservation of the genetic resources in populations of

CWRs is thus of utmost importance and can be realized both *in situ* and *ex situ*. Solely relying on *ex situ* conservation of CWRs is, however, problematic for several reasons (Meilleur and Hodgkin, 2004). First, it remains challenging to cover all extant CWR species and their genetic diversity, and place duplicates in conservation repositories (Castañeda-Álvarez *et al*., 2016; Wambugu and Henry, 2022). Second, evolutionary adaptive processes are strongly impeded in *ex situ* germplasm collections. More so, artificial selection during *ex situ* conservation can drive populations away from their phenotypic optimum in nature (Ensslin and Godefroid, 2020). As a consequence, there is a considerable risk of maladaptation when the CWRs are reintroduced into their original environments (Schoen and Brown, 2001; Heywood, 2015; Ensslin *et al*., 2023). Finally, for some species *ex situ* conservation can be a very costly option due to seed recalcitrance (Mertens *et al*., 2022).

In situ conservation of CWRs, on the other hand, is increasingly compromised by multiple anthropogenic processes. One conspicuous global threat is habitat degradation and habitat loss, the latter mainly caused by the conversion of natural habitat into agricultural land (Balvanera *et al*., 2019; Jaureguiberry *et al*., 2022). Agricultural encroachment has furthermore resulted in the introduction of cultivated crops in or near the natural habitats of their wild relatives, increasing the chance of genetic exchange (Ellstrand *et al*., 1999; Kareiva *et al*., 2007; Hufford *et al*., 2013). Such genetic exchange has already been observed in multiple CWR species, including wild apple (*Malus sieversii*) in Kazakhstan (Ha *et al*., 2021), Macademia trees in Australia (O'Connor *et al*., 2015) and Japanese chestnut (*Castanea crenata*) (Nishio *et al*., 2021). The transfer of cultivated genetic material can lead to cultivated–wild hybrids and subsequent backcrosses (Kwit *et al*., 2011). Hybridization can be detrimental to the wild populations through genetic swamping if it happens at a large scale (Todesco *et al*., 2016; Macková *et al*., 2018). Ultimately, crop-to-wild gene flow can result in introgression, i.e. the stable incorporation of alleles of the cultivar gene pool into the CWR gene pool at a relatively high frequency, which can cause the loss of genetic variation and even local extinction of the original populations (Anderson and Hubricht, 1938; Ellstrand, 2003; Laikre *et al*., 2010). The extent of these processes is, however, expected to vary among species, populations, pollinating system and pollinating agents, and distances between the crop and wild populations, and thus needs to be carefully studied across systems (Arriola and Ellstrand, 1996; Ellstrand *et al*., 1999; O'Connor *et al*., 2015).

Crop-to-wild gene flow can be expected to be a particularly important emerging threat for wild coffee species in their region of origin, where coffee cultivation areas are close to, or even mixed with, wild *Coffea* populations. Coffee is an immensely important crop, providing livelihoods to millions of people throughout the world (Guido *et al*., 2020). Of the 131 *Coffea* species, only two, *Coffea arabica* and *Coffea canephora,* gave rise to most of the currently cultivated coffee varieties (Davis and Rakotonasolo, 2021; Stoffelen *et al*., 2021). Presently, Arabica coffee (*C. arabica*) accounts for ±56 % of the global coffee market share, whereas Robusta coffee (*C. canephora*) accounts for the remaining share (ICO, 2023). The importance of Robusta coffee is, however, growing in the coffee sector presumably thanks to its high disease resistance and broad climatic range fit for cultivation (Craparo *et al*., 2015; Davis *et al*., 2019). The coffee CWRs are native to the African continent; whereas *C. arabica* originated in the Ethiopian highlands, *C. canephora* arose in the lowlands of West and Central Africa (Hamon *et al*., 2017; Charr *et al*., 2020; Bawin *et al*., 2021). For both species, the genetic diversity still present in the wild populations is key for future coffee breeding efforts. For example, in the 1930s *C. arabica* var. Geisha was first discovered in the Afromontane region of western Ethiopia and exhibits coffee leaf rust resistance. This variety has since demonstrated its significance in coffee breeding (Boot, 2013; Krishnan, 2014) and is a key example of the importance of the wild gene pool of both *C. arabica* and *C. canephora*. The improvement of Arabica coffee through crossbreeding with Robusta coffee is a long-standing practice dating back to the early 20th century in Indonesia (Cramer, 1957; Rodrigues *et al*., 1975). This is mostly based on cultivated material; for example, Castro Caicedo *et al*. (2013) derived Arabica genotypes with high levels of resistance to pathogens by crossing *C. arabica* cultivars with cultivated *C. canephora* material. The *C. canephora* wild gene pool may contain additional desirable, yet undiscovered genetic diversity. Furthermore, previous studies have shown that the genetic diversity of wild coffee populations is significantly higher than material that is currently available for cultivation (Krishnan, 2013; Leroy *et al*., 2014; Scalabrin *et al*., 2020; Vanden Abeele *et al*., 2021). Despite the importance of the wild gene pool, its genetic integrity is increasingly comprised by anthropogenic factors including rainforest disturbance and agricultural encroachment (Depecker *et al*., 2023). Moreover, the spatial distance between the cultivated gene pool and the wild gene pool of coffee species is progressively diminishing, as plantations and cultivated coffee are advancing further into the habitats of wild coffee populations. In the south-western highlands of Ethiopia, where *C. arabica* grows naturally in the Afromontane rainforest and where Arabica coffee is cultivated in close proximity in the coffee-producing agricultural landscape, the presence of alleles from coffee berry disease-resistant cultivars have been reported to be present in the gene pool of wild coffee populations (Aerts *et al*., 2013). This suggests the possibility of cultivated–wild introgression (Aerts *et al*., 2013). In the same region in south-west Ethiopia, Zewdie *et al*. (2022) observed the spread of these coffee berry disease-resistant cultivars across the landscape, highlighting the elevated risk of hybridization and introgression of genetic material from cultivars into wild *C. arabica* plants. Likewise in Uganda, Robusta coffee is cultivated near populations of its CWR, and putative crop-to-wild gene flow has been identified by two separate studies, both employing microsatellite (simple sequence repeat, SSR) markers (Musoli *et al*., 2009; Kiwuka *et al*., 2021). Musoli *et al*. (2009) attributed the high genetic similarity between wild samples and cultivated material to either their close genetic origin or to pollen flow between wild individuals and plantations. Kiwuka *et al*. (2021) detected wild accessions exhibiting signals of hybridization and introgression in fragmented natural forests. Within the Congo Basin, from which many of the currently cultivated Robusta coffee genotypes originate and which harbours genetically highly diverse wild *C. canephora* populations (Cubry *et al*., 2013; Ferrão *et al*., 2019; Merot-L'anthoene *et al*., 2019; Depecker *et al*., 2023), Vanden Abeele *et al*. (2021) identified two putative cultivated–wild hybrids in the rainforests. All previous studies used a limited number of SSR markers, yet a

more comprehensive assessment, with genome-wide molecular markers, of the occurrence of cultivated–wild hybrids and the potential impact on the genetic composition of wild populations of *C. canephora* remains to be done. This requires the development of robust diagnostic fingerprints that genetically differentiate cultivated from wild coffee.

In the history of Robusta coffee cultivation, the Democratic Republic of the Congo (DR Congo) has played a key role in terms of coffee breeding and commercialization. As such, Robusta coffee that was first bred in Java at the start of the 20th century originated from the DR Congo (Coste *et al*., 1955; Montagnon *et al*., 1998*a*, *b*; Ferrão *et al*., 2019). Along the colonial routes, the crop eventually returned to the DR Congo with the establishment of the Lula Coffee Research Station. Later, in 1927, a second Congolese research station was founded in Yangambi. Currently, in Yangambi, cultivated coffee is grown in home gardens and in the field gene bank of the INERA (Institut National des Etudes et Recherches Agronomiques) Coffee Collection, whereas wild populations of *C. canephora* grow in the understorey of the surrounding rainforests. Verleysen *et al*. (2023) recently catalogued genetic fingerprints of the accessions of the INERA Coffee Collection and showed that most accessions of the germplasm collection were highly similar to 'Lula' cultivars, whereas some accessions were more similar to the Congolese subgroup A (as previously described by Labouisse *et al*., 2020; Tournebize *et al*., 2022; Vi *et al*., 2023) or to local wild genotypes. Here, we refer to 'Lula' cultivars as populations and clones selected at the Coffee Research Station in Lula in the DR Congo. In local villages along the main road through the area, cultivated *C. canephora* is predominantly cultivated in home garden systems, which are often located very close (sometimes <1 km) to wild *C. canephora* populations. These materials were presumably obtained from the INERA Coffee Collection in Yangambi (Vanden Abeele *et al*., 2021), and probably represent 'Lula' cultivars. This close proximity of the cultivated *C. canephora* in the small-scale home garden systems to the wild *C. canephora* populations may provide ample opportunities for cultivated–wild hybridization.

Here, we aimed to assess the extent of the threat to the genetic integrity of wild *C. canephora* by investigating the overlap in geographical distribution of cultivated and wild *C. canephora*, and the occurrence of cultivated–wild hybridization in the Yangambi region in the DR Congo. We therefore collected 71 *C. canephora* individuals from home gardens in the Yangambi region and an additional 12 *C. canephora* individuals from tropical rainforests, and genotyped them using genotyping-by-sequencing (GBS). Next, we compared those to two previously published GBS datasets of *C. canephora* individuals collected from nearby natural tropical rainforests (Depecker *et al*., 2023) and the INERA Coffee Collection in Yangambi (Verleysen *et al*., 2023). These home gardens are relatively small, containing one to several cultivated *C. canephora* plants (age 3–50 years) situated near the local residents' homes and in some cases located very close (<1 km) to the natural rainforest where the wild *C. canephora* populations grow. We used GBS data of all 471 *C. canephora* individuals sampled across the Yangambi region. Our specific objectives were to: (1) compare the genetic diversity within and between cultivated *C. canephora* and local wild populations, and establish genetic fingerprints that can discriminate between both groups; (2) assign individuals to a cultivated or wild origin, identify clonal material and identify individuals derived from cultivated–wild hybridization events; (3) map the geographical position of the cultivated–wild hybrids in the landscape; and (4) discuss scenarios that may affect the genetic integrity of wild *C. canephora* populations.

MATERIALS AND METHODS

Sampling

We collected leaf material of 71 cultivated *C. canephora* individuals from 21 home gardens spread across villages in the Yangambi region. We complemented this with samples previously obtained by Depecker *et al*. (2023) and Verleysen *et al*. (2023). Depecker *et al*. (2023) established 24 survey plots in the understorey of the rainforests in the Yangambi region, covering an area of ~50 × 20 km. From this study, the corresponding GBS data of 249 putatively wild *C. canephora* individuals were retrieved from the Sequence Read Archive (SRA, see below), and 12 additional individuals, including individuals from one additional survey plot, were included here to create novel GBS data. Additionally, we used GBS data from 139 individuals from Verleysen *et al*. (2023), covering cultivated and local wild material now maintained in the field gene bank of the INERA Coffee Collection in Yangambi. Metadata on the sampling site and classification based on the genetic analysis of all 471 individuals are available in Supplementary Data Table S1. Leaf material of the 71 home garden individuals and the 12 additional rainforest individuals was dried with silica gel and genomic DNA was extracted from 20–30 mg dried leaf material using an optimized cetyltrimethylammonium bromide (CTAB) protocol adapted from Doyle and Doyle (1987). DNA quantities were measured with the Quantifluor dsDNA system on a Promega Quantus Fluorometer (Promega, Madison, WI, USA).

GBS and read data processing

DNA extracts of the 83 individuals were subjected to GBS. Following Depecker *et al*. (2023), GBS libraries were prepared using a double-enzyme GBS protocol adapted from Elshire (2011) and Poland *et al*. (2012). In short, 100 ng of genomic DNA was digested with PstI and MseI restriction enzymes (New England Biolabs, Ipswich, MA, USA), and barcoded and common adapters were ligated with T4 ligase (New England Biolabs) in a final volume of 35 µL. Ligation products were purified with 1.6× MagNA magnetic beads (GE Healthcare Europe, Machelen, Belgium) and eluted in 30 µL TE buffer. Of the purified DNA eluate, 3 µL was used for amplification with Taq 2× Master Mix (New England Biolabs) using an 18-cycle PCR protocol. PCR products were bead-purified with 1.6× MagNA, and their DNA concentrations were quantified using a Quantus Fluorometer. The library quality and fragment size distributions were assessed using a QIAxcel system (Qiagen, Venlo, the Netherlands). Equimolar amounts of the GBS libraries were pooled, bead-purified and 150-bp paired-end sequenced on an Illumina HiSeq-X instrument by Admera Health (South Plainfield, NJ, USA).

Reads were processed with a customized script available on Gitlab (https://gitlab.com/ilvo/GBprocesS). The quality of

sequence data was validated with FastQC 0.11 (Andrews, 2010) and reads were demultiplexed using Cutadapt 2.10 (Martin, 2011), allowing zero mismatches in the barcode-restriction site remnant combination. The 3′ restriction site remnant and the common adapter sequence of forward reads, and the 3′ restriction site remnant, the barcode and the barcode adapter sequence of reverse reads were removed based on sequence-specific pattern recognition and positional trimming using Cutadapt. After trimming the 5′ restriction site remnant of forward and reverse reads using positional trimming in Cutadapt, forward and reverse reads with a minimum read length of 60 bp and a minimum overlap of 10 bp were merged using PEAR 0.9.11 (Zhang *et al*., 2014). Merged reads with a mean base quality <25 or with >5 % of the nucleotides uncalled and reads containing internal restriction sites were discarded using GBprocesS.

The GBS data of the 83 newly sampled individuals were complemented with GBS data derived from SRA (Bioproject PRJNA901681) of the 249 wild *C. canephora* individuals studied by Depecker *et al*. (2023), and the 139 cultivated *C. canephora* individuals studied by Verleysen *et al*. (2023). GBS data of all 471 *C. canephora* individuals were mapped onto the *C. canephora* reference genome sequence (Denoeud *et al*., 2014) with the BWA-mem algorithm in BWA 0.7.17 with default parameters (Li and Durbin, 2009). Alignments were sorted, indexed and filtered based on mapping quality >20 with SAMtools 1.10 (Li *et al*., 2009). Next, high-quality GBS loci and Stack Mapping Anchor Points (SMAPs) were identified in the mapped reads using the SMAP *delineate* module within the SMAP package v4.4.0 (manual available at: https://ngs-smap.readthedocs.io/en/latest/; source code available at: https://gitlab.ilvo.be/genomics/smap-package/smap) with parameters: *mapping_orientation* ignore, *min_stack_depth* 4, *max_stack_depth* 400, *min_cluster_depth* 8, *max_cluster_depth* 400, *completeness* 80 and *min_mapping_quality* 20.

SNP and haplotype calling

Single nucleotide polymorphisms (SNPs) within the high-quality GBS loci were called with GATK (Genome Analysis Toolkit) Unified Genotyper 3.7.0 (McKenna *et al*., 2010). Multi-allelic SNPs were removed with GATK, and the remaining SNPs were filtered using the following parameters: *min-meanDP* 30, *mac* 4 and *minQ* 20. The remaining SNPs were then subjected to further filtering using VCFtools 0.1.16 with the following parameters: *minDP* 10, *minGQ* 30, *max-missing* 0.7, *mac* 3, *minQ* 30, *min-alleles* 2, *max-alleles* 2 and *maf* 0.05 (Danecek *et al*., 2011).

Read-backed haplotyping was conducted based on the combined variation in SMAPs and SNPs using the SMAP *haplotype-sites* module within the SMAP package v4.4.0 with parameters: *mapping_orientation* ignore, *partial* include, *no_indels*, *min_read_count* 10, *min_distinct_haplotypes* 2, *min_haplotype_frequency* 5, *discrete_calls* dosage, *frequency_interval_bounds* 10 10 90 90 and *dosage_filter* 2.

Genetic structure

To investigate the genetic structure among all 471 individuals, a principal component analysis (PCA) was performed using the R package *ADEGENET* (Jombart, 2008; Rstudio Team, 2016). Additionally, a Bayesian clustering implemented in fastSTRUCTURE v1.0 (Raj *et al*., 2014) was run. In total, 100 iterations were run for each expected cluster setting *K*, ranging from 2 to 9. The StructureSelector software (Li and Liu, 2018) was used to determine the optimum number of *K*, by first plotting the mean log probability of each successive *K* and then using the Delta *K* method following Evanno *et al*. (2005).

Genetic similarity

The genetic similarity between all 471 individuals was quantified with the SMAP *grm* module within the SMAP package v4.4.0, using the Jaccard Inversed Distance (Jaccard, 1912) that was calculated based on the discrete dosage haplotype calls in high-quality GBS loci. SMAP *grm* was run with parameters: *locus_completeness* 0.1, *similarity_coefficient* Jaccard, *distance_method* Euclidean, *locus_information_content* shared and *partial* FALSE, creating a pairwise Jaccard Inversed Distance (JID) matrix. The procedure of Verleysen *et al*. (2023) was used to calculate the minimal JID as a threshold to identify all pairs of genetically identical individuals (i.e. clones).

Genetic diversity

Cluster setting $K = 2$ separated the cultivated from wild genotypes. The estimated admixture proportions, Bayesian *Q*-value (Pritchard *et al*., 2000), for each individual for cluster setting $K = 2$ were used to establish a ‘wild reference group’, containing all individuals collected from the Yangambi rainforest with a proportion of wild genotype >0.9 ($n = 249$), and a ‘cultivated reference group’, containing all individuals collected from the INERA Coffee Collection and home gardens with a proportion of cultivated genotype >0.9 ($n = 152$). The reasons for including samples (with a clear genetic fingerprint assigned to cultivated material) from the home gardens in the ‘cultivated reference group’ were three-fold. First, materials distributed by the INERA Coffee Collection were propagated both clonally and seed-based. Seed-propagated individuals derived from cultivated materials and found in the home gardens are closely related to known cultivated genotypes in the INERA Coffee Collection, yet genetically unique. Second, sampling of the INERA Coffee Collection (Verleysen *et al*., 2023) did not cover all possible cultivated genotypes. Any additional information on allele diversity in the cultivated material is thus helpful to genetically differentiate cultivated and wild material. Last, the home gardens are located closest to the wild genotypes in the rainforest, and may be the source of gene flow of cultivated material. Including their genetic fingerprints into the reference set includes the sensitivity of identifying crop–wild hybridization as these individuals are the most likely source of exchanged alleles. These reference groups were used to calculate genetic diversity. Allelic richness (A_r) was calculated according to El Mousadik and Petit (1996) using the *allelic.richness* function of the R package *hierfstat* (Goudet, 2013). The observed and expected heterozygosity (H_o and H_e, respectively) and inbreeding coefficient (F_{IS}) were calculated using the *gl.report.heterozygosity* function of the R package *dartR* (Gruber *et al*., 2018). All genetic diversity indices were compared between the

wild and cultivated reference group using a Mann–Whitney U test. Genetic differentiation between the two reference groups was further quantified with a pairwise F_{ST} according to Weir and Cockerham, (1984).

Hybridization analysis

For the hybridization analysis, we only retained SNPs with large differences in allele frequency ($F_{ST} > 0.8$) between the 'wild' and 'cultivated' reference groups. Hybridization levels were tested for all individuals not included in the two established reference groups ($n = 70$). First, heterozygosity (H) within individuals was compared to their proportion of ancestry (S) from either reference group using the R package *Hiest* (Fitzpatrick, 2012). Next, likelihoods for six early generation hybrid classes (wild genotype, cultivated genotype, F_1 hybrid, F_2 hybrid, hybrid–wild backcross, hybrid–cultivated backcross) were calculated using the *Hiclass* function. The best fit of these hybrid classes was compared to the maximum likelihood genotype described by ancestry (S) and individual heterozygosity (H). We accepted a putative classification as credible if the log-likelihood of the best fit was within 2 units of the maximum log-likelihood.

RESULTS

The 471 coffee individuals collected from rainforests, the INERA Coffee Collection and home gardens (Supplementary Data Fig. S1) yielded a total of 19 678 bi-allelic SNPs within 14 251 high-quality GBS loci with a completeness of at least 80 % across all 471 individuals. Of these, 8131 SNPs with a minimum minor allele count of 3 and a minimum minor allele frequency of 0.05 were used for all further analyses.

Genetic identities

The PCA performed on the 8131 SNPs showed that individuals collected from the rainforests and the INERA Coffee Collection were separated along the PC1-axis (Fig. 1A). Along

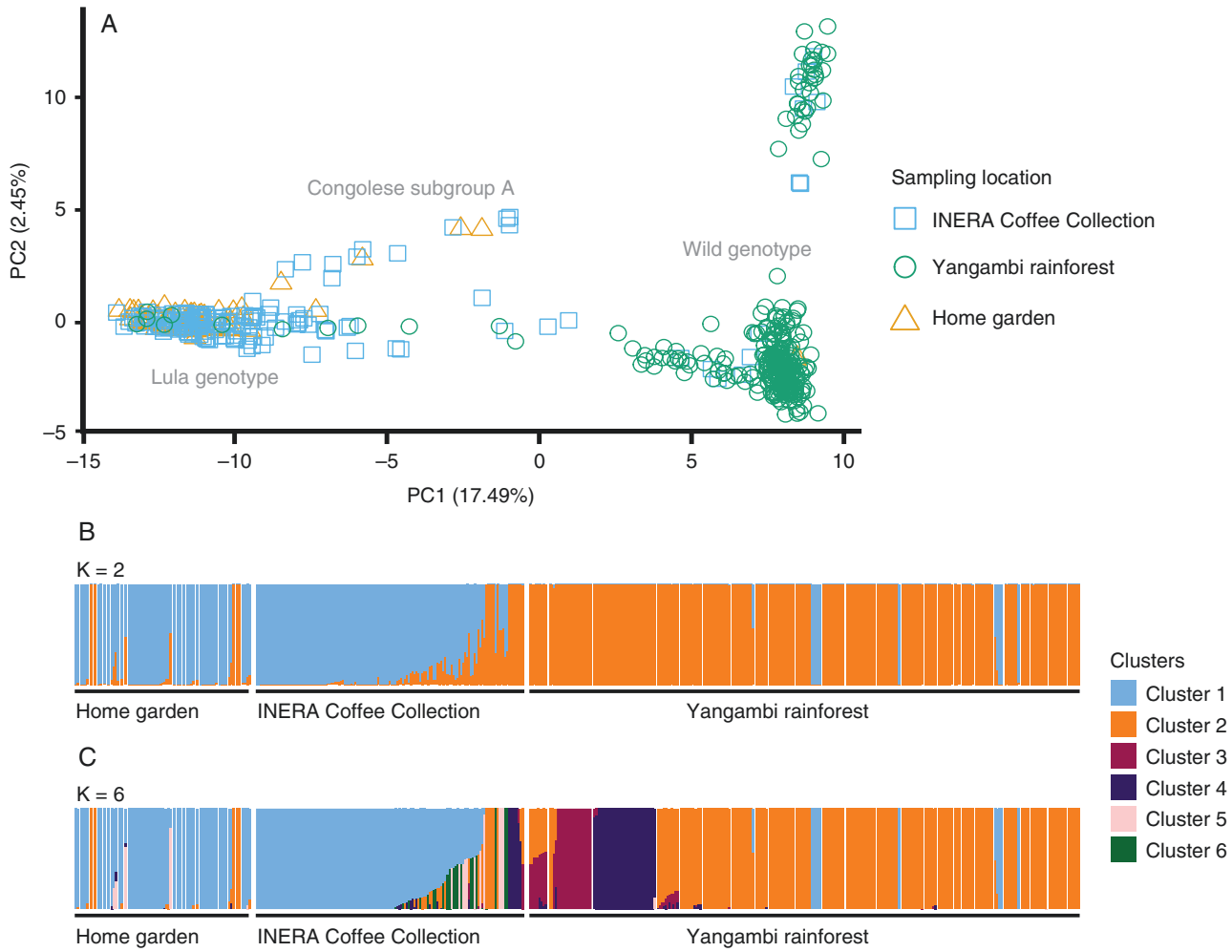


FIG. 1. Population genetic structure within the *Coffea canephora* sample set. (A) Principal component analysis using 8131 SNPs indicating individuals collected from the INERA Coffee Collection, the Yangambi rainforest and the home gardens in the Yangambi region. (B) fastSTRUCTURE bar plot representing two ($K = 2$) and six clusters ($K = 6$). Colours define subpopulations: blue ('Lula' cultivars), orange (wild genotype), dark pink (wild genotypes), purple (wild genotypes), light pink (Congolese subgroup A) and green (unknown origin). Individuals are shown by thin vertical lines, which are divided into K coloured segments representing the estimated membership probabilities (Q) of each individual.

the PC2-axis, 32 rainforest individuals were separated from all other individuals collected in the rainforest. Likewise, 15 individuals belonging to the Congolese subgroup A were separated from all the other individuals collected from the INERA Coffee Collection along the PC2-axis.

FastSTRUCTURE revealed that the cultivated individuals collected in the home gardens and INERA Coffee Collection (Cluster 1) were separated from wild individuals collected in the rainforest (Cluster 2) for a minimal number of clusters ($K = 2$) (Fig. 1B). The optimum number of clusters for wild and cultivated individuals together was six ($K = 6$) (Fig. 1C). Three genetic clusters were present in the Yangambi rainforest (Clusters 2, 3 and 4). All three wild clusters were also present in the INERA Coffee Collection, along with three cultivated clusters. Based on Verleysen *et al*. (2023), Cluster 1 represents the 'Lula' cultivar genotype and Cluster 5 represents the 'Congolese subgroup A'. The origin of Cluster 6 remained unknown in the frame of this analysis. Most of the individuals ($n = 61$) cultivated in home gardens belonged to Cluster 1 and were classified as 'Lula' cultivars, but also wild ($n = 5$) and one Congolese subgroup A genotype were present in the home gardens. From the individuals collected in the rainforest, three were positioned between the cultivated and the wild group in the PCA and were assigned an admixed wild–Lula genotype by fastSTRUCTURE, and nine individuals were positioned on the negative PC1-axis and were assigned a 'Lula' genotype by fastSTRUCTURE.

The pairwise JID calculated for all 471 individuals based on 41 522 haplotypes within 10 045 polymorphic high-quality GBS loci showed a group of replicates with pairwise JID values >0.979 (Supplementary Data Table S2, Fig. S1). Using the minimal JID values of 0.979 to identify pairs of clones, one rainforest individual was genetically identical to a 'Lula' genotype from the INERA Coffee Collection. Within the home gardens, one individual collected was genetically identical to an individual collected from another home garden, one individual was genetically identical to a 'Lula' genotype collected in the INERA Coffee Collection and another was genetically identical to an individual collected from the rainforest.

Genetic diversity and differentiation within and between reference groups

To calculate the genetic diversity and differentiation within and between wild and cultivated groups, 249 individuals collected from the Yangambi rainforest with a proportion of wild genotype >0.9 for $K = 2$ were used as a 'wild reference group', which includes all three wild genetic clusters, and 152 individuals collected from the INERA Coffee Collection and home gardens with a proportion of cultivated genotype >0.9 were used as a 'cultivated reference group', which includes the 'Lula' genetic group, Congolese subgroup A and Cluster 6.

No significant differences were found in allelic richness (A_r) ($P = 0.5$) and observed heterozygosity (H_o) ($P = 0.35$) between wild and cultivated reference groups (Table 1). Expected heterozygosity (H_e) ($P = 0.00074$) and inbreeding coefficient (F_{IS}) ($P < 2.2e^{-16}$) were significantly lower in the cultivated reference group than in the wild reference group. Genetic differentiation as measured by overall F_{ST} between both groups was 0.142.

Table 1. *Genetic diversity estimates for the wild and cultivated reference group.*

	N	A_r	H_o	H_e	F_{IS}
Wild reference group	249	1.90	0.38	0.30	−0.28
Cultivated reference group	152	1.90	0.38	0.28	0.38

Admixture and hybridization

To explore putative cultivated–wild hybridization, the wild and cultivated reference groups were used to identify SNPs with large differences in allele frequency between wild and cultivated material, resulting in 24 SNPs with $F_{ST} > 0.8$, here considered as discriminatory SNPs (Fig. 2A; detailed information on the 24 SNPs is provided in Supplementary Data Table S3). The proportion of ancestry (S) and individual heterozygosity (H) were calculated on the 24 SNPs for all 70 individuals not included in the two reference groups (Fig. 2C). Using the ancestry–heterozygosity ratio (Fig. 2B), 40 individuals were assigned to six different hybrid classes based on statistical support (Fig. 2D): two individuals were assigned as F_1 hybrids (one in the INERA Coffee Collection and one in a home garden), four as F_2 hybrids (one in the rainforest, one in a home garden and two in the INERA Coffee Collection), 16 as hybrid–cultivated backcrosses (one in the rainforest, one in a home garden and 14 in the INERA Coffee Collection), one as a hybrid–wild backcross in the INERA Coffee Collection, ten as cultivated genotypes (nine in the rainforest and one in the INERA Coffee Collection) and seven as wild genotypes (two in the home gardens and five in the INERA Coffee Collection) (Table S1).

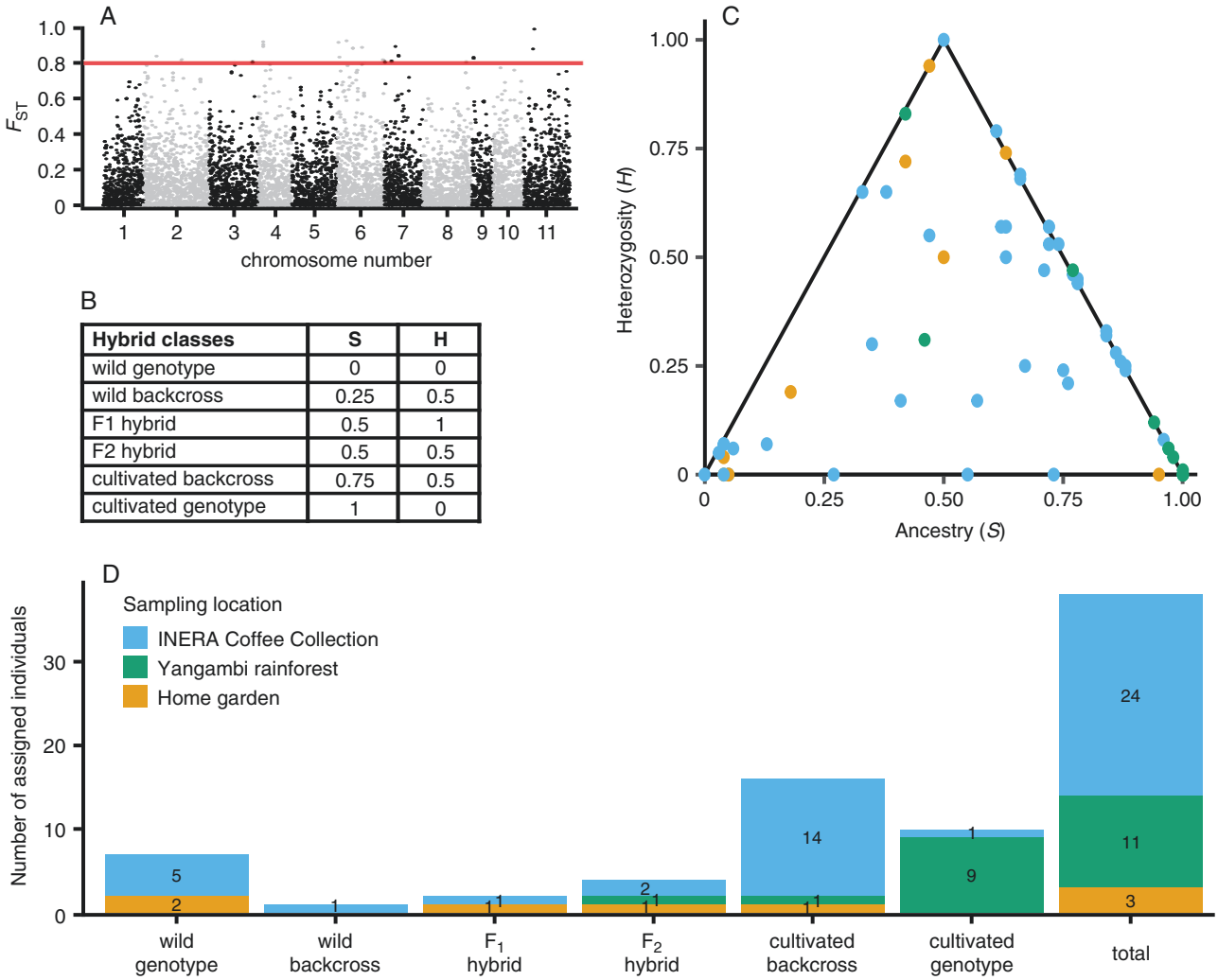
Wild, cultivated and cultivated–wild genotypes in the Yangambi rainforest

To assess the risk of cultivated–wild hybridization, the fastSTRUCTURE results ($K = 6$) were placed on the landscape map, revealing four different situations (Fig. 3).

Situation I: four plots (Plots 16–19) from undisturbed old-growth rainforest more than 16 km away from home gardens contained only wild genotypes (Cluster 2).

Situation II: Plot 3 established in disturbed old-growth rainforest was located in close vicinity (<1 to 4 km) of 14 home gardens. All individuals collected from this plot were assigned a wild (Cluster 4) genotype and we did not detect any cultivated or cultivated–wild genotypes. The home gardens in this locality contained individuals assigned to a 'Lula' genotype and individuals with an admixed Lula–Congolese subgroup A genotype, of which one individual was identified as an F_1 hybrid and one individual as a hybrid–cultivated backcross.

Situation III: five plots were located in disturbed old-growth and four plots in regrowth rainforest, which were surrounded by home gardens. Almost all individuals collected in these nine rainforest plots had a wild genotype, except for one individual from Plot 13 that was assigned a 'Lula' genotype and one individual from Plot 7 that showed indications of an admixed



Hybrid classes	S	H
wild genotype	0	0
wild backcross	0.25	0.5
F1 hybrid	0.5	1
F2 hybrid	0.5	0.5
cultivated backcross	0.75	0.5
cultivated genotype	1	0

FIG. 2. Estimation of hybridization within all 70 individuals, not included in the wild and cultivated reference groups. (A) F_{ST} of all 8131 SNPs calculated between the wild and cultivated reference group. Red line indicates the threshold to identify differentiating SNPs (F_{ST} = = 0.8). (B) Theoretical hybrid assignments based on the ratio between the proportion of ancestry (*S*) and individual heterozygosity (*H*). (C) The proportion of ancestry (*S*) and individual heterozygosity (*H*) calculated based on the 24 SNPs for all 70 individuals coloured by their sampling location. (D) All 38 statistically supported hybrid assignments into six hybrid classes coloured by their sampling location.

wild–Lula genotype, although this was not statistically supported. Four individuals collected from three home gardens were assigned a wild genotype. In one home garden, all four individuals were identified as a 'Lula' genotype, and another home garden contained one wild genotype, one 'Lula' genotype and one admixed wild–Lula genotype that was statistically assigned as an F_2 hybrid.

Situation IV: one plot in regrowth rainforest, two plots in disturbed old-growth rainforest and six plots in undisturbed old-growth rainforest (Depecker *et al.*, 2022, 2023) had no home gardens nearby (>7 km). Plots 11, 12, 22, 23, 24 and 25 contained only individuals with wild genotypes. Plot 10 in disturbed old-growth rainforest, on the other hand, contained nine individuals with a wild genotype and five individuals with a 'Lula' genotype but no admixed wild–Lula genotypes. Plot 21 in undisturbed old-growth rainforest had seven individuals with a wild genotype and one individual with a 'Lula' genotype, but no admixed wild–Lula genotypes. Plot 20 in undisturbed old-growth rainforest contained no individuals with a wild genotype, two individuals with a 'Lula' genotype, one F_2 hybrid and one hybrid–cultivated backcross.

DISCUSSION

The coffee cultivars grown globally at present are a result of the selection and the breeding of *C. arabica* and *C. canephora*, of which natural populations occur in the highlands of Ethiopia and the lowlands of the Congo Basin, respectively. Because natural populations often occur near to coffee cultivation areas, the wild coffee gene pool might be especially prone to hybridization and introgression. Previous studies have already provided some evidence for this in *C. arabica* in Ethiopia (Aerts *et al.*, 2013) and *C. canephora* in the DR Congo (Vanden Abeele *et al.*, 2021) and Uganda (Musoli *et al.*, 2009; Kiwuka *et al.*, 2021), yet those studies were based on a limited number of SSR

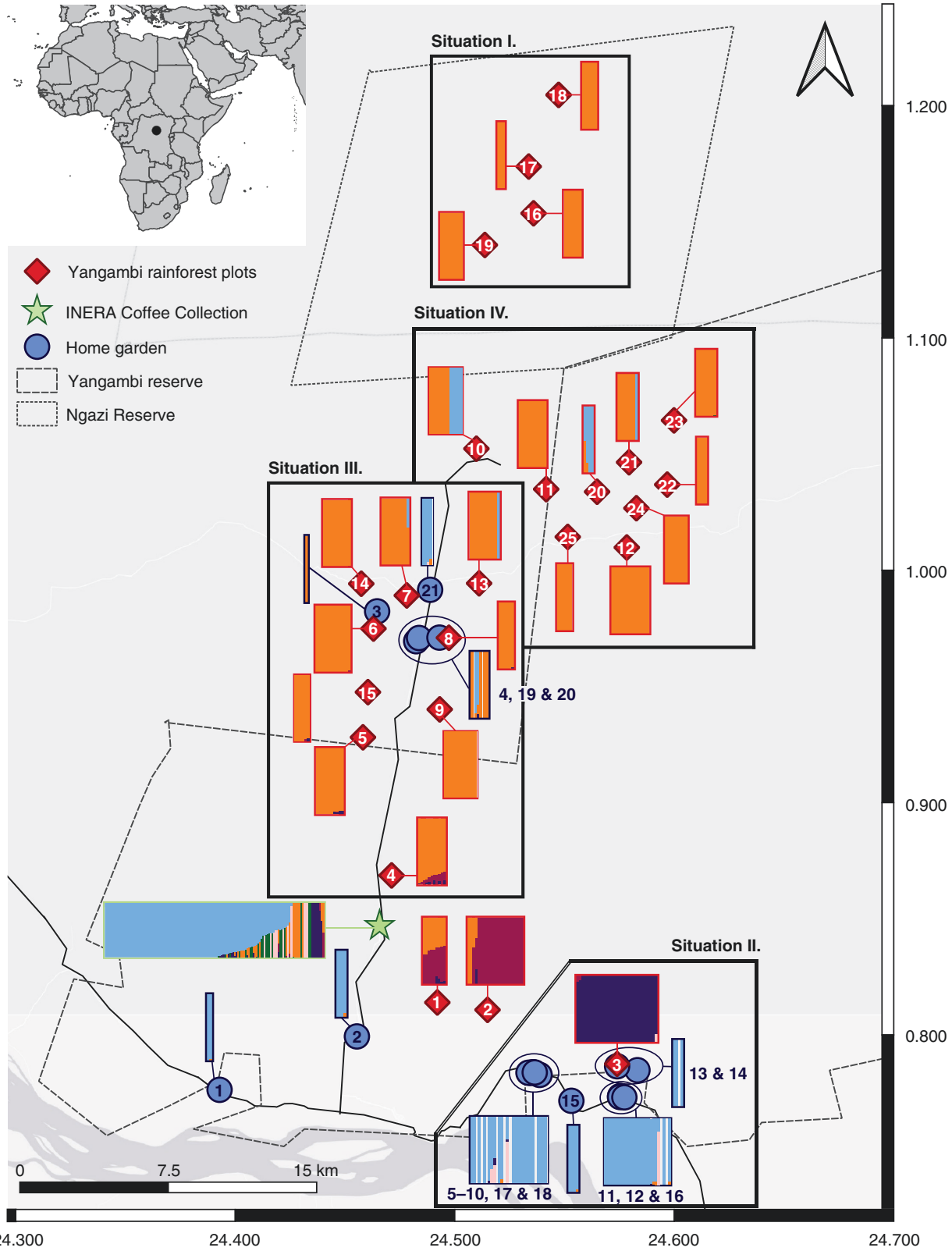


FIG 3. Map of the Yangambi region showing the location of each Yangambi rainforest plot (red rhombuses), each home garden (blue rhombuses) and the INERA Coffee Collection (green star), as well as the *K* coloured segments (*K* = 6) representing the estimated membership probabilities (*Q*) of each individual within each location. Individuals are shown by thin vertical lines and grouped together based on sample location. Red border: all individuals collected in the Yangambi rainforest; blue border: all individuals collected in the home gardens.

markers. We established robust diagnostic fingerprints based on genome-wide genetic markers to attain a more comprehensive assessment of the occurrence of cultivated–wild hybrids and the impact on the genetic composition of wild *C. canephora* populations in the Yangambi area of the Congo Basin.

We combined novel GBS genotyping data of coffee trees from home gardens and rainforests with GBS data from previous studies on the genetic diversity in wild populations (Depecker *et al*., 2023), and cultivated source material of the INERA Coffee Collection (Verleysen *et al*., 2023) to investigate whether planting of selected cultivated materials in home gardens in close vicinity to the wild populations may lead to crop-to-wild gene flow, and whether cultivated–wild hybridization could lead to changes in the genetic composition of the wild populations.

INERA Coffee Collection germplasm has a broad genetic base

Our results suggest that cultivated materials found in home gardens in the area were probably created and distributed by the Yangambi INERA station. Notably, the cultivated materials present in the INERA Coffee Collection and distributed in the Yangambi area should be considered as one or more heterogeneous 'genetic groups', and not as 'pure' cultivars with a narrow genetic base. The INERA Coffee Collection that is used for the breeding and distribution of germplasm for cultivation encompasses genetic material from several different origins, mostly 'Lula' material, Congolese subgroup A (see Labouisse *et al*., 2020; Tournebize *et al*., 2022; Vi *et al*., 2023) and wild material from local wild populations, as previously described in detail in Verleysen *et al*. (2023). In addition to Verleysen *et al*. (2023), a novel group with unknown origin was found (Cluster 6). The creation of new genetic diversity by crossings between materials of different genetic origin and between historically cultivated materials and local wild materials further adds to the broad genetic base (Verleysen *et al*., 2023). Furthermore, most of this material is distributed after seed-based propagation, produced by open-pollination, and it appears that only a fraction of the germplasm is multiplied clonally for distribution, as illustrated by a relatively low proportion of clonal pairs between the INERA Coffee Collection, the wild populations and the home gardens. Our data also showed that most cultivated coffees in the home gardens belong to the genetic group of 'Lula' cultivars, confirming an earlier study (Vanden Abeele *et al*., 2021).

We found a relatively high allelic richness and observed heterozygosity in the cultivated reference group that was similar to that observed in the wild reference group, which was in contrast to our expectations. Typically, CWRs are considered to exhibit greater genetic diversity than their cultivated counterparts (Cubry *et al*., 2013; Zhang *et al*., 2017). For instance, Vanden Abeele *et al*. (2021) previously reported higher levels of observed heterozygosity in cultivated materials in the Yangambi region, as compared to the wild populations of *C. canephora* in the region, and equivalent or higher levels of observed heterozygosity were found in cultivated *C. canephora* elsewhere in the Afrotropical region compared to wild *C. canephora* populations (Musoli *et al*., 2009; Kiwuka *et al*., 2021). However, such comparisons rely strongly on the composition and structure of the cultivated germplasm (i.e. the breeding history), the geographical area of wild populations covered and their population structure (previously described for the Yangambi region in Depecker *et al*., 2023), the number of individuals sampled per group, and the type and number of molecular markers. For instance, Vanden Abeele *et al*. (2021) analysed a substantially lower number of individuals with 18 multi-allelic SSRs compared to our study with hundreds of individuals and thousands of bi-allelic SNP markers. Not only was *within*-genetic group allelic richness and observed heterozygosity similar in the cultivar and wild reference groups in our study, but also *between*-reference group genetic diversity comparisons showed a relatively high overlap in allele composition, albeit at different allele frequency per group (no strictly private alleles and only 24 out of 8131 SNPs with an $F_{ST} > 0.8$ in our study). The relatively close genetic relatedness between cultivated and wild *C. canephora* was also highlighted by Cubry *et al*. (2013), and Kiwuka *et al*. (2021) found a low number of private alleles in cultivated *C. canephora* populations. This is probably related to the long generation time of coffee, limited breeding selection and a broad genetic base (Stoffelen, 1998; Cubry *et al*., 2013; Gomez *et al*., 2016), and because 'Lula' cultivars probably came from sources relatively close to the Yangambi region (Verleysen *et al*., 2023).

Differentiation of cultivated, wild and cultivated–wild hybrids, and reconstruction of spreading patterns

Despite the relatively high proportion of common alleles in cultivated materials and local wild populations, genetic analyses delineated 24 'diagnostic' molecular markers that distinguish between the groups of cultivated and wild genotypes. Our analyses are thus consistent with previous studies that genetically separated cultivated and wild *C. canephora* in the Congo Basin (Vanden Abeele *et al*., 2021; Verleysen *et al*., 2023) and other studies in Africa comparing wild with cultivated *C. canephora* (Musoli *et al*., 2009; Kiwuka *et al*., 2021). The genetic distinction between cultivated and wild was instrumental in our study to assign genotypes to cultivar or wild origins and identify cultivated–wild hybridization events, and placing them onto the landscape map revealed their distribution pattern. Genetic analyses also help to disentangle the different ways cultivated materials spread in the area, for instance by distinguishing between the different cultivated genetic groups and first- or second-generation cultivated–wild hybrids and/or backcrosses. As expected, the first and most prominent route of spreading cultivated materials is the intentional planting of cultivars in home gardens. The second type of location where cultivars were found was the regrowth rainforest, disturbed old-growth rainforest and even presumed undisturbed old-growth rainforest. Cultivar individuals may end up there via different routes, either being planted or as remnants of abandoned plantations (based historical land-use maps; see Depecker *et al*., 2022, 2023), which would be expected to yield clusters of cultivars in a given small area (e.g. Plot 20). Alternatively, seed flow may partly account for the dispersal of berries or seeds from within-cultivar-group pollinated trees, which may be expected to yield more sporadic instances of cultivar individuals in otherwise predominantly wild populations, such as Plot 7 or Plot 10. Seeds of *Coffea* species are predominantly dispersed by birds

and mammals that can cover long distances (Stoffelen, 1998; Noirot *et al*., 2016). Conversely, cultivated–wild hybrids may be derived from pollination between cultivated and wild materials. Such materials are known to be generated in the INERA Coffee Collection (Verleysen *et al*., 2023), and may be distributed and planted, just like the cultivated 'Lula' cultivar material. Alternatively, cultivated–wild hybrids may result from the natural process of cross-pollination by pollinators in the rainforest. Crops grown in the vicinity of the rainforest have been shown to be visited by a rich pollinator community, as compared to more isolated crop fields (Klein *et al*., 2008, 2009). Such spread of cultivar genetic material would then also lead to more sporadic patterns of F_1, F_2 or backcross materials. The frequency of cultivated–wild hybrid occurrence in nature would be determined by a combination of the density of cultivars as source material and pollen transport over distance in that area, followed by seed dispersal from the mother plant. The relatively short geographical distance between cultivated *C. canephora* genotypes and the wild *C. canephora* populations creates opportunities for gene flow, while isolation-by-distance (IBD) may limit gene flow. Interestingly, IBD may not be homogeneous across the landscape in the study system (Depecker *et al*., 2023), with Plot 3 genetically distinct from the closest neighbouring Plot 1 and Plot 2, while the population in the northern part displayed more genetic similarity (i.e. connectivity) between plots at a comparatively longer distance. Taking the human and natural factors that potentially facilitate cultivar spread and cultivated–wild gene flow together with the locations at which cultivars and cultivated–wild hybrids were found in the landscape, we were able to distinguish four different situations that revealed heterogeneity of potential cultivated–wild interactions at the landscape level within the relatively small study area.

First, in the most northern part of the Yangambi region (Fig. 3, Situation I), where rainforests have been classified as undisturbed old-growth (Depecker *et al*., 2022, 2023), genetic analysis confirmed the presence of exclusively wild individuals. This remote area contained no cultivated materials, thus representing a pristine rainforest fit for *in situ* conservation as genetic exchange between cultivated and wild material is still absent.

Second, in the south-eastern part (Fig. 3, Situation II), wild individuals were collected in disturbed old-growth rainforest (Depecker *et al*., 2022, 2023) in close vicinity of several home gardens containing cultivated coffee genotypes. Despite the substantial level of anthropogenic activity, in terms of both disturbance and agricultural encroachment, no signs of admixture and hybridization between cultivated and wild *C. canephora* were detected. As reported by Kearsley *et al*. (2017), a monodominant *Gilbertiodendron dewevrei* forest isolates this part of the rainforest from the surrounding environment, and Depecker *et al*. (2023) coined this to be a natural barrier explaining the high genetic differentiation between the *C. canephora* population in this area and the other populations in the Yangambi region, observing IBD even at relatively small distances.

Third, in the centre of the Yangambi region (Fig. 3, Situation III), disturbed old-growth and regrowth rainforests are patched together (Depecker *et al*., 2022, 2023), with a strip of village development with home gardens along the road that crosses these rainforests. Here, cultivated *C. canephora* and wild *C. canephora* populations can be found in very close proximity, in some occasions even less than 1 km, without an apparent natural barrier as observed in situation II. A cultivated genotype and a cultivated–wild hybrid were present within the wild population, indicating a low level of spreading of cultivated material and crop-to-wild gene flow. Such spurious individuals are consistent with several scenarios of spreading; in regrowth rainforest, they could be founders originating from neighbouring populations (Depecker *et al*., 2023), while in disturbed old-growth or regrowth rainforest cultivars might be occasionally planted, or cultivated–wild hybridization may arise from cross-pollination and seed flow from neighbouring fields or gardens.

Finally, in the fourth situation, just north-east of the centre of the region (Fig. 3, Situation IV), wild genotypes of *C. canephora* were found in disturbed and undisturbed old-growth rainforests, without any home gardens in close proximity (Depecker *et al*., 2022, 2023). Nevertheless, multiple cultivar genotypes and cultivated–wild hybrid genotypes were also identified in three different plots. The high level of grouping of the cultivated genotypes raises questions about whether these individuals have emerged naturally following pollen and/or seed dispersal. For instance, the cultivated coffees in Plot 10 might have been a remnant of a former coffee plantation based on the historical land-use maps (see Depecker *et al*., 2022, 2023). In addition to the home gardens, these cultivated genotypes should also be considered as sources for crop-to-wild gene flow, and we indeed found hybrids in close proximity to these cultivated genotypes. Notably, in Plot 20, in an area of the rainforest that was previously classified as undisturbed old-growth rainforest based on canopy structure, previous land-use maps and absence of logging, an individual was found that was genetically identical to a cultivated genotype of the INERA Coffee Collection, suggesting that this individual was derived from clonal propagation and was dispersed by human intervention, rather than by seed dispersal.

In conclusion, we have identified cultivated genotypes and cultivated–wild hybrids in zones with clear anthropogenic activity, and where *C. canephora* cultivated in the home gardens may serve as a source for crop-to-wild gene flow. Our data further show that the current distribution area of the cultivated genotypes is more extensive than previously believed. The presence of clonally propagated cultivated genetic material in presumed undisturbed locations is a sign of past or present human activity and expands the region of the Yangambi rainforest that is subject to disturbance with possible consequences for habitat integrity. Our study further illustrates that genetic analyses may uncover sites of human activities in parallel with field observations of forest canopy structure, vegetation composition or logging, and therefore may be of complementary value in landscape-level monitoring of anthropogenic activities and habitat disturbance. In other African regions, hybridization between cultivated and wild *C. canephora* has also been reported, for example in Uganda (Musoli *et al*., 2009). In *C. arabica*, hybridization and introgression were reported in montane rainforests in south-western Ethiopia (Aerts *et al*., 2013). Introgression, however, will only occur when cultivated–wild hybrids form backcrosses with the wild population for many generations and at a large scale (Ridley, 2004; Verónica *et al*., 2017). In our study, we found only relatively few F_1 and F_2 hybrids and backcrosses in the

wild, which is probably not sufficient for the stable incorporation of alleles of the cultivated gene pool into the CWR gene pool (Anderson and Hubricht, 1938; Ellstrand, 2003; Laikre *et al*., 2010). Furthermore, the cultivated material distributed in the Yangambi region generally has a broad genetic base, is derived from several origins and is genetically closely related to the local wild populations. Therefore, if hybridization occurs, mostly genetic material is exchanged that was already part of the wild gene pool. Nevertheless, it is important to continue monitoring habitat integrity and cultivated–wild gene flow to safeguard the wild gene pool of *C. canephora*, as the *in situ* conservation of CWRs is important to guarantee future food security.

SUPPLEMENTARY DATA

Supplementary data are available at *Annals of Botany* online and consist of the following.

Table S1: Source data of all 471 individuals collected from the Yangambi rainforest, the home gardens and the INERA Coffee Collection in Yangambi. Table S2: Pairwise Jaccard Inversed Distance (JID) calculated for all 471 individuals based on 41 522 haplotypes within 10 045 polymorphic high-quality GBS loci. The JID matrix was arranged based on the sample location. Red-shaded JID values were above the minimal JID and indicate genetically identical individuals (pairs of clones). Table S3: Detailed information on the 24 diagnostic SNPs. Figure S1: Distribution of all pairwise Jaccard Inversed Distance (JID) calculated for all 471 individuals based on 41 522 haplotypes within 10 045 polymorphic high-quality GBS loci.

ACKNOWLEDGMENTS

We are grateful for the field assistance by the Institut National pour l'Étude et la Recherche Agronomiques (INERA) and the FORETS project, which is financed by the 11th European Development Fund. We would also like to thank the Ministère de L'Environnement et Développement Durable (MEDD) for their indispensable help with obtaining research and export permits in accordance with the Nagoya regulations of the DR Congo (No. 003/ANCCB-RDC/SG-EDD/BTB/02/2020, No. 008/ANCCB-RDC/SG-EDD/BTB/11/2020, No. 001/ANCCB-RDC/SG-EDD/BTB/01/2021, No. 004/ANCCB-RDC/SG-EDD/BTB/2021, No. 014/ANCCB-RDC/SG-EDD/BTB/11/2021, No. 025/ANCCB-RDC/SG-EDD/BTB/11/2022).

FUNDING

This work was supported by Research Foundation-Flanders, via a research mandate granted to J.D. (FWO; 1125221N), a research project to O.H. (FWO; G090719N), and by the Belgian Science Policy Office (BELSPO) under contract No. 632 B2/191/P1/COFFEEBRIDGE (CoffeeBridge Project) of the Belgian Research Action through 633 Interdisciplinary Networks (BRAIN-be 2.0). Additional funding was granted to J.D. and Y.H. through the Foundation for the promotion of biodiversity research in Africa (SBBOA, www.sbboa.be).

DATA AVAILABILITY

The data that support the findings of this study are available on request (curator@plantentuinmeise.be). In accordance with DR Congo and international regulations, restrictions apply on the availability of these data, which were used under licence for this study.

AUTHOR CONTRIBUTIONS

O.H., F.V., T.R., J.D. and L.V. designed this study. J.D., J.A., Y.H., J.-L.K., I.M.M., T.E. and R.B. participated in fieldwork. L.V. executed the lab work. L.V. and J.D. analysed the data. J.D., L.V., F.V., T.R. and O.H. wrote the manuscript. All authors contributed to finalizing the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Aerts R, **Berecha G**, **Gijbels P**, ***et al***. **2013**. Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in south-western Ethiopian montane rainforests. *Evolutionary Applications* **6**: 243–252.

Anderson E, **Hubricht L. 1938**. Hybridization in Tradescantia. III. The evidence for introgressive hybridization. *American Journal of Botany* **25**: 396–402.

Andrews S. 2010. *FastQC: a quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arriola PE, **Ellstrand NC. 1996**. Crop-to-weed gene flow in the genus *Sorghum* (Poaceae): spontaneous interspecific hybridization between johnsongrass, *Sorghum halepense*, and crop *Sorghum, S. bicolor*. *American Journal of Botany* **83**: 1153–1159.

Balvanera P, **Pfaff A**, **Viña A**, ***et al.*** **2019**. *Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-policy Platform on Biodiversity and Ecosystem Services*, In: **Brondizio ES**, **Settele J**, **Nho HT**, Díaz eds, Secretariat of the Intergovernmental Science-Policy Platform for Biodiversity And Ecosystem Services, Bonn, Germany. 54–200.

Bawin Y, **Ruttink T**, **Staelens A**, ***et al***. **2021**. Phylogenomic analysis clarifies the evolutionary origin of *Coffea arabica*. *Journal of Systematics and Evolution* **59**: 953–963.

Bohra A, **Kilian B**, **Sivasankar S**, ***et al***. **2022**. Reap the crop wild relatives for breeding future crops. *Trends in Biotechnology* **40**: 412–431.

Boot W. 2013. Exploring the holy grail: Geisha Coffee, 10 years on. *Roast Magazine* **May/June 2013**: 39–49.

Brozynska M, **Furtado A**, **Henry RJ. 2016**. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnology Journal* **14**: 1070–1085.

Castañeda-Álvarez NP, **Khoury CK**, **Achicanoy HA**, ***et al***. **2016**. Global conservation priorities for crop wild relatives. *Nature Plants* **2**: 16022.

Castro Caicedo BL, **Cortina Guerrero HA**, **Roux J**, **Wingfield MJ. 2013**. New coffee (*Coffea arabica*) genotypes derived from *Coffea canephora* exhibiting high levels of resistance to leaf rust and *Ceratocystis* canker. *Tropical Plant Pathology* **38**: 485–494.

Charr J-C, **Garavito A**, **Guyeux C**, ***et al***. **2020**. Complex evolutionary history of coffees revealed by full plastid genomes and 28,800 nuclear SNP analyses, with particular emphasis on *Coffea canephora* (Robusta coffee). *Molecular Phylogenetics and Evolution* **151**: 106906.

Coste R. 1955. *Les caféiers et les cafés dans le monde*. Paris: Editions Larose.

Cramer PJS. 1957. *A review of literature of coffee research in Indonesia*. SIC Editorial. Inter Turrialba: American Institute of Agricultural Sciences.

Craparo ACW, **Van Asten PJA**, **Läderach P**, **Jassogne LTP**, **Grab SW. 2015**. *Coffea arabica* yields in decline in Tanzania due to climate change: global implications. *Agricultural and Forest Meteorology* **207**: 1–10.

Cubry P, **De Bellis F**, **Pot D**, **Musoli P**, **Leroy T. 2013**. Global analysis of *Coffea canephora* Pierre ex Froehner (Rubiaceae) from the Guino-Congolese region reveals impacts from climatic refuges and migration effects. *Genetic Resources and Crop Evolution* **60**: 483–501.

Danecek P, **Auton A**, **Abecasis G**, ***et al***; 1000 Genomes Project Analysis Group. **2011**. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Davis AP, **Rakotonasolo F. 2021**. Six new species of coffee (*Coffea*) from northern Madagascar. *Kew Bulletin* **76**: 497–511.

Davis AP, **Chadburn H**, **Moat J**, **O'Sullivan R**, **Hargreaves S**, **Lughadha EN. 2019**. High extinction risk for wild coffee species and implications for coffee sector sustainability. *Science Advances* **5**: 3473–3489.

Denoeud F, **Carretero-Paulet L**, **Dereeper A**, ***et al***. **2014**. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**: 1181–1184.

Depecker J, **Asimonyio JA**, **Miteho R**, ***et al***. **2022**. The association between rainforest disturbance and recovery, tree community composition, and community traits in the Yangambi area in the Democratic Republic of the Congo. *Journal of Tropical Ecology* **38**: 426–436.

Depecker J, **Verleysen L**, **Asimonyio JA**, ***et al***. **2023**. Genetic diversity and structure in wild Robusta coffee (*Coffea canephora* A. Froehner) populations in Yangambi (DR Congo) and their relation to forest disturbance. *Heredity* **130**: 145–153.

Doyle JJ, **Doyle JL. 1987**. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**: 11–15.

Ellstrand NC. 2003. Current knowledge of gene flow in plants: implications for transgene flow. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **358**: 1163–1170.

Ellstrand NC, **Prentice HC**, **Hancock JF. 1999**. Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics* **30**: 539–563.

El Mousadik A, **Petit RJ. 1996**. High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics* **92**: 832–839.

Elshire RJ, **Glaubitz JC**, **Sun Q**, ***et al***. **2011**. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**: e19379.

Ennslin A, **Godefroid S. 2020**. Ex situ cultivation impacts on plant traits and drought stress response in a multi-species experiment. *Biological Conservation* **248**: 108630.

Ennslin A, **Sandner TM**, **Godefroid S. 2023**. Does the reduction of seed dormancy during ex situ cultivation affect the germination and establishment of plants reintroduced into the wild? *Journal of Applied Ecology* **60**: 685–695.

Evanno G, **Regnaut S**, **Goudet J. 2005**. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**: 2611–2620.

FAO. **2018**. *The future of food and agriculture - Alternative pathways to 2050*. Rome: FAO.

Ferrão MAG, **Ferrão RG**, **da Fonseca AFA**, **Filho ACVF**, **Volpi PS. 2019**. Origin, geographical dispersion, taxonomy and genetic diversity of *Coffea canephora*. In: **Ferrão RG**, **da Fonseca AFA**, **Ferrão MAG**, **De Mune LH**. eds, *Conilon Coffee: the Coffea canephora produced in Brazil*. Bento Ferreira: Incaper, 85–109.

Fitzpatrick BM. 2012. Estimating ancestry and heterozygosity of hybrids using molecular markers. *BMC Evolutionary Biology* **12**: 131.

Gomez C, **Despinoy M**, **Hamon S**, ***et al***. **2016**. Shift in precipitation regime promotes interspecific hybridization of introduced Coffea species. *Ecology and Evolution* **6**: 3240–3255.

Goudet J. 2013. *hierfstat: estimation and tests of hierarchical F-statistics*. R Package version 0:04–10. http://CRAN.R-project.org/package=hierfstat

Gruber B, **Unmack PJ**, **Berry OF**, **Georges A. 2018**. DARTR: an R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources* **18**: 691–699.

Guido Z, **Knudson C**, **Rhiney K. 2020**. Will COVID-19 be one shock too many for smallholder coffee livelihoods? *World Development* **136**: 105172.

Ha Y-H, **Oh S-H**, **Lee S-R. 2021**. Genetic admixture in the population of wild apple (*Malus sieversii*) from the Tien Shan Mountains, Kazakhstan. *Genes* **12**: 104.

Hamon P, **Grover CE**, **Davis AP**, ***et al***. **2017**. Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (*Coffea*) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. *Molecular Phylogenetics and Evolution* **109**: 351–361.

Heywood V. 2015. In situ conservation of plant species – an unattainable goal? *Israel Journal of Plant Sciences* **63**: 211–231.

Heywood V, **Casas A**, **Ford-Lloyd B**, **Kell S**, **Maxted N. 2007**. Conservation and sustainable use of crop wild relatives. *Agriculture, Ecosystems & Environment* **121**: 245–255.

Hufford MB, **Lubinsky P**, **Pyhäjärvi T**, **Devengenzo MT**, **Ellstrand NC**, **Ross-Ibarra J. 2013**. The genomic signature of crop-wild introgression in maize. *PLoS Genetics* **9**: e1003477.

ICO **2023**. *Coffee market data per year*. London, United Kingdom: Data provided by ICO statistical service.

Jaccard P. 1912. The distribution of the flora in the alpine zone. 1. *New Phytologist* **11**: 37–50.

Jaureguiberry P, **Titeux N**, **Wiemers M**, ***et al***. **2022**. The direct drivers of recent global anthropogenic biodiversity loss. *Science Advances* **8**: 1–12.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Kareiva P, **Watts S**, **McDonald R**, **Boucher T. 2007**. Domesticated nature: shaping landscapes and ecosystems for human welfare. *Science* **316**: 1866–1869.

Kearsley E, **Verbeeck H**, **Hufkens K**, ***et al***. **2017**. Functional community structure of African monodominant *Gilbertiodendron dewevrei* forest influenced by local environmental filtering. *Ecology and Evolution* **7**: 295–304.

Kiwuka C, **Goudsmit E**, **Tournebize R**, ***et al***. **2021**. Genetic diversity of native and cultivated Ugandan Robusta coffee (*Coffea canephora* Pierre ex A. Froehner): Climate influences, breeding potential and diversity conservation. *PLoS One* **16**: e0245965.

Klein A-M. 2009. Nearby rainforest promotes coffee pollination by increasing spatio-temporal stability in bee species richness. *Forest Ecology and Management* **258**: 1838–1845.

Klein A-M, **Cunningham SA**, **Bos M**, **Steffan-Dewenter I. 2008**. Advances in pollination ecology from tropical plantation crops. *Ecology* **89**: 935–943.

Krishnan S. 2013. Current status of coffee genetic resources and implications for conservation. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources* **8**: 1–9.

Krishnan S. 2014. *Genetic Characterization of Geisha Coffee*. Final Report. Denver, CO: Denver Botanic Gardens.

Kwit C, **Moon HS**, **Warwick SI**, **Stewart CN Jr**. **2011**. Transgene introgression in crop relatives: molecular evidence and mitigation strategies. *Trends in Biotechnology* **29**: 284–293.

Labouisse J-P, **Cubry P**, **Austerlitz F**, **Rivallan R**, **Nguyen HA. 2020**. New insights on spatial genetic structure and diversity of *Coffea canephora* (Rubiaceae) in Upper Guinea based on old herbaria. *Plant Ecology and Evolution* **153**: 82–100.

Laikre L, **Schwartz MK**, **Waples RS**, **Ryman N**; GeM Working Group. **2010**. Comprising genetic diversity in the wild: unmonitored large-scale release of plants and animals. *Trends in Ecology & Evolution* **25**: 520–529.

Leroy T, **De Bellis F**, **Legnate H**, ***et al***. **2014**. Developing core collections to optimize the management and the exploitation of diversity of the coffee *Coffea canephora*. *Genetica* **142**: 185–199.

Li H, **Durbin R. 2009**. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li YL, **Liu JX. 2018**. StructureSelector: a web-based software to select and visualize the optimal number of clusters using multiple methods. *Molecular Ecology Resources* **18**: 176–177.

Li H, **Handsaker B**, **Wysoker A**, ***et al***; 1000 Genome Project Data Processing Subgroup. **2009**. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Macková L, **Vít P**, **Urfus T. 2018**. Crop-to-wild hybridization in cherries – Empirical evidence from *Prunus fruticose*. *Evolutionary Applications* **11**: 1748–1759.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMB Net Journal* **17**: 10–12.

McKenna A, **Hanna M**, **Banks E**, ***et al***. **2010**. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Meilleur BA, **Hodgkin T. 2004**. In situ conservation of crop wild relatives: status and trends. *Biodiversity and Conservation* **13**: 663–684.

Merot-L'anthoene V, **Tournebize R**, **Darracq O**, ***et al***. **2019**. Development and evaluation of a genome-wide Coffee 8.5K SNP array and its application for high-density genetic mapping and for investigating the origin of *Coffea arabica* L. *Plant Biotechnology Journal* **17**: 1418–1430.

Mertens A, **Bawin Y**, **Vanden Abeele S**, ***et al***. **2022**. Phylogeography and conservation gaps of *Musa balbisiana* Colla genetic diversity revealed by microsatellite markers. *Genetic Resources and Crop Evolution* **69**: 2515–2534.

Montagnon C, **Leroy T**, **Eskes AB. 1998***a*. Amélioration variétale de Coffea canephora. I. Critères et méthodes de selection. *Plantations, Recherche, Developpement* **5**: 18–33.

Montagnon C, **Leroy T**, **Eskes AB. 1998***b*. Amélioration variétale de Coffea canephora. II. Les programmes de selection et leurs résultats. *Plantations, Recherche, Developpement* **5**: 89–95.

Musoli P, **Cubry P**, **Aluka P**, ***et al***. **2009**. Genetic differentiation of wild and cultivated populations: diversity of *Coffea canephora* Pierre in Uganda. *Genome* **52**: 634–646.

Nishio S, **Takada N**, **Terakami S**, ***et al***. **2021**. Genetic structure analysis of cultivated and wild chestnut populations reveals gene flow from cultivars to natural stands. *Scientific Reports* **11**: 240.

Noirot M, **Charrier A**, **Stoffelen P**, **Anthony F. 2016**. Reproductive isolation, gene flow and speciation in the former *Coffea* subgenus: a review. *Trees* **30**: 597–608.

O'Connor K, **Powell M**, **Nock C**, **Shapcott A. 2015**. Crop to wild gene flow and genetic diversity in a vulnerable *Macademia* (Proteaceae) species in New South Wales, Australia. *Biological Conservation* **191**: 504–511.

Poland JA, **Rife TW. 2012**. Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome* **5**: 92–102.

Pritchard JK, **Stephens M**, **Donnelly P. 2000**. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Raj A, **Stephens M**, **Pritschard JK. 2014**. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Ridley M. 2004. *Evolution*. Hoboken: Blackwell Publishing.

Rodrigues CJ Jr, **Bettencourt AJ**, **Rijo L. 1975**. Races of the pathogen and resistance to coffee rust. *Annual Review of Phytopathology* **13**: 49–70.

R Studio Team **2016**. *RStudio: Integrated Development for R*. Boston, MA: RStudio Inc.

Saeed A, **Fatima N. 2021**. Wild germplasm: shaping future tomato breeding. In: **Azhar, MT.**, **Wani, SH.** eds. *Wild germplasm for genetic improvement in crop plants*. Cambridge: Academic Press, 201–2014.

Scalabrin S, **Toniutti L**, **Di Gaspero G**, ***et al***. **2020**. A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Scientific Reports* **10**: 4642.

Schoen DJ, **Brown AHD. 2001**. The conservation of wild plant species in seed banks: attention to both taxonomic coverage and population biology will improve the role of seed banks as conservation tools. *BioScience* **51**: 960–966.

Stoffelen P. 1998. *Coffea and Psilanthus (Rubiaceae) in tropical Africa: a systematic and palynological study, including a revision of the West and Central African species*. PhD Thesis, KU Leuven, Belgium.

Stoffelen P, **Anthony F**, **Janssens S**, **Noirot M. 2021**. A new coffee species from South-West Cameroon, the principal hotspot of diversity for *Coffea L.* (Coffeeae, Ixoroideae, Rubiaceae) in Africa. *Adansonia* **43**: 277–285.

Todesco M, **Pascual MA**, **Owens GL**, ***et al***. **2016**. Hybridization and extinction. *Evolutionary Applications* **9**: 892–908.

Tournebize R, **Borner L**, **Manel S**, ***et al***. **2022**. Ecological and genomic vulnerability to climate change across native populations of Robusta coffee (*Coffea canephora*). *Global Change Biology* **28**: 4124–4142.

Vanden Abeele S, **Janssens SB**, **Asimonyio JA**, ***et al***. **2021**. Genetic diversity of wild and cultivated *Coffea canephora* in northeastern DR Congo and the implications for conservation. *American Journal of Botany* **108**: 2425–2434.

Verleysen L, **Bollen R**, **Kambale J-L**, ***et al***. **2023**. Characterization of the genetic composition and establishment of a core collection for the INERA Robusta coffee (*Coffea canephora*) field genebank from the Democratic Republic of Congo. *Frontiers in Sustainable Food Systems* **7**: 1–33.

Verónica EM, **Georgina S**, **Alejandro A**, **Leonardo G. 2017**. Pattern of natural introgression in a *Nothofagus* hybrid zone from South American temperate forests. *Tree Genetics and Genomes* **13**: 49.

Vi T, **Vigouroux Y**, **Cubry P**, ***et al***. **2023**. Genome-wide admixture mapping identifies wild ancestry-of-origin segments in cultivated Robusta coffee. *Genome Biology and Evolution* **15**: 1–12.

Wambugu PW, **Henry R. 2022**. Supporting in situ conservation of the genetic diversity of crop wild relatives using genomics technologies. *Molecular Ecology* **31**: 2207–2222.

Weir BS, **Cockerham CC. 1984**. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Zewdie B, **Bawin Y**, **Tack AJM**, ***et al***. **2022**. Genetic composition and diversity of Arabica coffee in the crop's centre of origin and its impact on four major fungal diseases. *Molecular Ecology* **32**: 2484–2503.

Zhang J, **Kobert K**, **Flouri T**, **Stamatakis A. 2014**. PEAR: a fast and accurate Illumina Paired-End read mergeR. *Bioinformatics* **30**: 614–620.

Zhang H, **Mittal N**, **Leamy LJ**, **Barazani O**, **Song B. 2017**. Back into the wild - Apply untapped genetic diversity of wild relatives for crop improvement. *Evolutionary Applications* **10**: 5–24.