the geneticssociety

www.nature.com/hdy

ARTICLE

Genetic diversity and structure in wild Robusta coffee (*Coffea canephora* A. Froehner) populations in Yangambi (DR Congo) and their relation to forest disturbance

Jonas Depecker[1,2,3,9], Lauren Verleysen[1,4,9], Justin A. Asimonyio[5], Yves Hatangi[2,6], Jean-Léon Kambale[5], Ithe Mwanga Mwanga[7], Tshimi Ebele[8], Benoit Dhed'a[6], Yves Bawin[1,4], Ariane Staelens[4], Piet Stoffelen[2], Tom Ruttink[4], Filip Vandelook[2,3] and Olivier Honnay[1,3]

Degradation and regeneration of tropical forests can strongly affect gene flow in understorey species, resulting in genetic erosion and changes in genetic structure. Yet, these processes remain poorly studied in tropical Africa. *Coffea canephora* is an economically important species, found in the understorey of tropical rainforests of Central and West Africa, and the genetic diversity harboured in its wild populations is vital for sustainable coffee production worldwide. Here, we aimed to quantify genetic diversity, genetic structure, and pedigree relations in wild *C. canephora* populations, and we investigated associations between these descriptors and forest disturbance and regeneration. Therefore, we sampled 256 *C. canephora* individuals within 24 plots across three forest categories in Yangambi (DR Congo), and used genotyping-by-sequencing to identify 18,894 SNPs. Overall, we found high genetic diversity, and no evidence of genetic erosion in *C. canephora* in disturbed old-growth forest, as compared to undisturbed old-growth forest. In addition, an overall heterozygosity excess was found in all populations, which was expected for a self-incompatible species. Genetic structure was mainly a result of isolation-by-distance, reflecting geographical location, with low to moderate relatedness at finer scales. Populations in regrowth forest had lower allelic richness than populations in old-growth forest and were characterised by a lower inter-individual relatedness and a lack of isolation-by-distance, suggesting that they originated from different neighbouring populations and were subject to founder effects. Wild Robusta coffee populations in the study area still harbour high levels of genetic diversity, yet careful monitoring of their response to ongoing forest degradation remains required.

Heredity; https://doi.org/10.1038/s41437-022-00588-0

INTRODUCTION

Tropical rainforests cover only 7% of the Earth's land surface but represent the world's richest reservoir of terrestrial biodiversity (Kier et al. 2005; Kreft and Jetz 2007). Over the past decades, human activities such as industrial logging and the encroachment of agriculture and infrastructure have negatively impacted tropical forest cover and resulted in the loss of biodiversity, jeopardising the provisioning of important ecosystem services such as carbon sequestration and climate regulation (Gardner et al. 2009; Curtis et al. 2018; Edwards et al. 2019). Less conspicuous than the loss of tropical forest cover is the ongoing degradation of tropical forests. Forest degradation refers to within-forest disturbance under a more or less intact canopy, and is mainly caused by selective logging and the removal of the understorey vegetation (Sasaki and Putz 2009; Tyukavina et al. 2018). Degradation of tropical forests may be as detrimental to biodiversity as forest cover loss due to the large spatial scales at which it occurs (Barlow et al. 2016). In the Congo Basin, for example, the rate of forest degradation has been estimated at 317,000 ha per year between 2000 and 2005 (Ernst et al. 2013), whereas Shapiro et al. (2021) reported that 23 million ha of forest has been degraded between 2000 and 2016 in this region.

Forest degradation may compromise the resilience and long-term stability of tropical rainforests because it can negatively affect the regeneration of the remaining woody plant species (Norden et al. 2009). Plant regeneration and fitness depend on multiple processes, including pollination, seed dispersal, germination and seedling establishment (Barret and Eckert 1990). Crucial aspects of gene flow early in the regeneration cycle, such as pollination and seed dispersal, can become strongly jeopardised through ongoing large-scale anthropogenic disturbance of tropical forests (Neushulz et al. 2016). Because many tropical canopy trees and understorey shrubs typically occur in population densities of less than one individual per ha, and due to widespread dioecy and self-incompatibility (SI) (Bawa et al. 1985; Hubbell and Foster 1986), pollen flow and successful pollination

[1]Division of Ecology, Evolution and Biodiversity Conservation, KU Leuven, Leuven, Belgium. [2]Meise Botanic Garden, Meise, Belgium. [3]KU Leuven Plant Institute, Leuven, Belgium. [4]Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food, Melle, Belgium. [5]Centre de Surveillance de la Biodiversité et Université de Kisangani, Kisangani, Democratic Republic of the Congo. [6]Université de Kisangani, Kisangani, Democratic Republic of the Congo. [7]Centre de Recherche en Science Naturelles, Lwiro, Democratic Republic of the Congo. [8]Institut National des Etudes et Recherches, Agronomique, Democratic Republic of the Congo. [9]These authors contributed equally: Jonas Depecker, Lauren Verleysen. Associate editor: Pår Ingvarsson. ✉email: Jonas.depecker@kuleuven.be; Lauren.verleysen@ilvo.vlaanderen.be

Received: 7 October 2022 Revised: 14 December 2022 Accepted: 14 December 2022
Published online: 03 January 2023

and reproduction can be expected to be particularly susceptible to changes in the understorey plant species density and composition (Aguilar et al. 2019; Chiriboga-Arroyo et al. 2021). Reduced gene flow may not only result in decreased reproductive capacity, but also in changes in the genetic structure and in genetic erosion of the remaining populations through increased genetic drift and inbreeding (Vranckx et al. 2012; Ismail et al. 2017; Campbell et al. 2018). This process can be exacerbated by the disappearance of large frugivores (by hunting, or as a result of habitat loss) from disturbed tropical rainforests (Bello et al. 2015), hampering seed dispersal and recruitment. Ultimately, reduced pollen flow and seed dispersal may even result in the local extinction of shrub and tree species (da Silva and Tabarelli 2000).

Apart from tropical forest disturbance, tropical forest regeneration on abandoned agricultural land may also significantly impact genetic diversity and structure of the recolonising woody species. Such regrowth forests make up an increasing fraction of the forested area throughout the tropics (FAO and UNEP 2020; Poorter et al. 2021). Recolonisation of abandoned agricultural fields by tropical woody species almost entirely depends on seed dispersal, as these species usually do not have a persistent soil seed bank (Sezen et al. 2007). These colonisation events are expected to be prone to founder effects, in which the newly founded population represents only a subsample from one or a few neighbouring source populations (Wright 1932; Mayr 1954; Widmer and Lexer 2001). These founder effects can result in major genetic changes, including loss of genetic diversity and increased genetic differentiation among populations, with associated fitness consequences (Born et al. 2008; Vandepitte et al. 2012). Whereas ample research has already been done on species diversity and community composition in regrowth tropical forests (e.g., Oberleitner et al. 2021; Makelele et al. 2021; Depecker et al. 2022), studies on the genetic diversity of tropical tree species in regrowth forests in the tropics are still scarce, especially in Africa.

Coffea canephora (Robusta coffee) is an understorey tree from the lowland tropical rainforests of Central and West Africa. The conservation of its genetic diversity is of utmost importance for future sustainable coffee production worldwide as wild populations carry useful traits for coffee breeding, such as disease resistance (Silva et al. 2006; Lashermes et al. 2010), tolerance to climate change (Davis et al. 2012) and drought tolerance (Cramer and Wellman 1957). Robusta coffee currently accounts for more than 40% of the global coffee production (ICO 2022) but is gaining commercial importance thanks to its higher disease resistance (Leroy et al. 2005), higher productivity (Wellman 1961) and its assumedly lower susceptibility to climate change than Arabica coffee (Craparo et al. 2015; Davis et al. 2012). *Coffea canephora* is a self-incompatible species, without a persistent soil seed bank (Oryem-Origa 1999; Nowak et al. 2011). Natural populations of *C. canephora* are usually disconnected, with 10–20 individuals per ha, and few offspring scattered across the forest floor (Musoli et al. 2009; Cubry et al. 2013; Depecker and Vandelook pers. obs.). Such characteristics can be expected to render the genetic diversity and structure of this species very susceptible to both the processes of forest disturbance and forest regrowth. Yet, research on the genetic diversity of wild *C. canephora* is still rare (but see Musoli et al. 2009; Kiwuka et al. 2021; Vanden Abeele et al. 2021 at the nationwide scale; and Nyakaana 2007 at the population scale).

In this study, we aimed to quantify the association between rainforest disturbance and regrowth on the one side, and genetic diversity and genetic structure of wild *C. canephora* on the other side, focusing on the Yangambi area (DR Congo) in the Congo Basin, an important Robusta coffee genetic diversity hotspot (Ferrão et al. 2019; Merot-L'anthoene et al. 2019). Therefore, we surveyed 24 inventory plots across undisturbed old-growth forest, disturbed old-growth forest, and regrowth forest, in which a total of 256 *C. canephora* individuals were sampled, and genotyped using genotyping-by-sequencing (GBS). We hypothesised to find: (i) lower genetic diversity in disturbed old-growth forest and regrowth forest, as compared to undisturbed old-growth forest; (ii) more pronounced genetic structure and pedigree relations in disturbed old-growth forest; and (iii) that populations in regrowth forests have emerged through colonisation via seed dispersal from multiple neighbouring coffee populations in old-growth forest, resulting in strongly admixed populations.

MATERIAL AND METHODS

Study population and sampling

The Yangambi region is located in the Tshopo province in North-Eastern DR Congo, approximately 100 km west of Kisangani. The Yangambi landscape consists of a mosaic of land tenures, typical for the Congo Basin: the Yangambi Man and Biosphere Reserve; the Ngazi Forest Reserve; a logging concession; and customary land (van Vliet et al. 2018).

Previously, Depecker et al. (2022) established 25 forest inventory plots of 125 × 125 m (1.56 ha), covering an area of ca. 50-by-20 km, just North of the Congo River (Fig. 1A). We adopted their classification of the plots into three different forest categories: (i) plots in regrowth forest (8 plots) located on historical agricultural land. Depecker et al. (2022) estimated that these agricultural lands were abandoned somewhere between 1962 and 1980, and since then overgrown; (ii) plots in disturbed old-growth forest (7 plots), with clear indications of small-scale selective logging through the presence of tree stumps and (iii) plots in undisturbed old-growth forest (10 plots) without signs of disturbance.

In this study, these plots were systematically surveyed for *C. canephora* by multiple Afrotropical plant experts. A leaf sample of all *C. canephora* individuals was collected and silica-dried, yielding a total of 256 samples. One survey plot (plot #20) from Depecker et al. (2022) was omitted, because there were too few *C. canephora* individuals to adequately analyse.

Genomic DNA extraction and genotyping-by-sequencing (GBS)

A total of 20–30 mg dried leaf material was homogenised with a Retsch TissueLyser II (Mixer Mill MM 500 Nano; Retsch®). Genomic DNA was extracted from the dried leaf material using an optimised 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, USA).

GBS libraries were prepared using a double-enzyme GBS protocol adapted from Elshire et al. (2011) and Poland and Rife (2012). In short, 100 ng of genomic DNA was digested with *Pst*I and *Mse*I restriction enzymes (New England Biolabs, Ipswich, USA), and barcoded and common adapter constructs were ligated with T4 ligase (New England Biolabs, Ipswich, USA) in a final volume of 35 µL. Ligation products were purified with 1.6× MagNA magnetic beads (GE Healthcare Europe, Machelen, BE) and eluted in 30 µL TE. Of the purified DNA eluate, 3 µL was used for amplification with *Taq* 2× Master Mix (New England Biolabs, Ipswich, USA) using a 18 cycles 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, NL). 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, USA).

Data processing

Reads were processed with a customised 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 barcodes or 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 below 25 or with more than 5% of the nucleotides

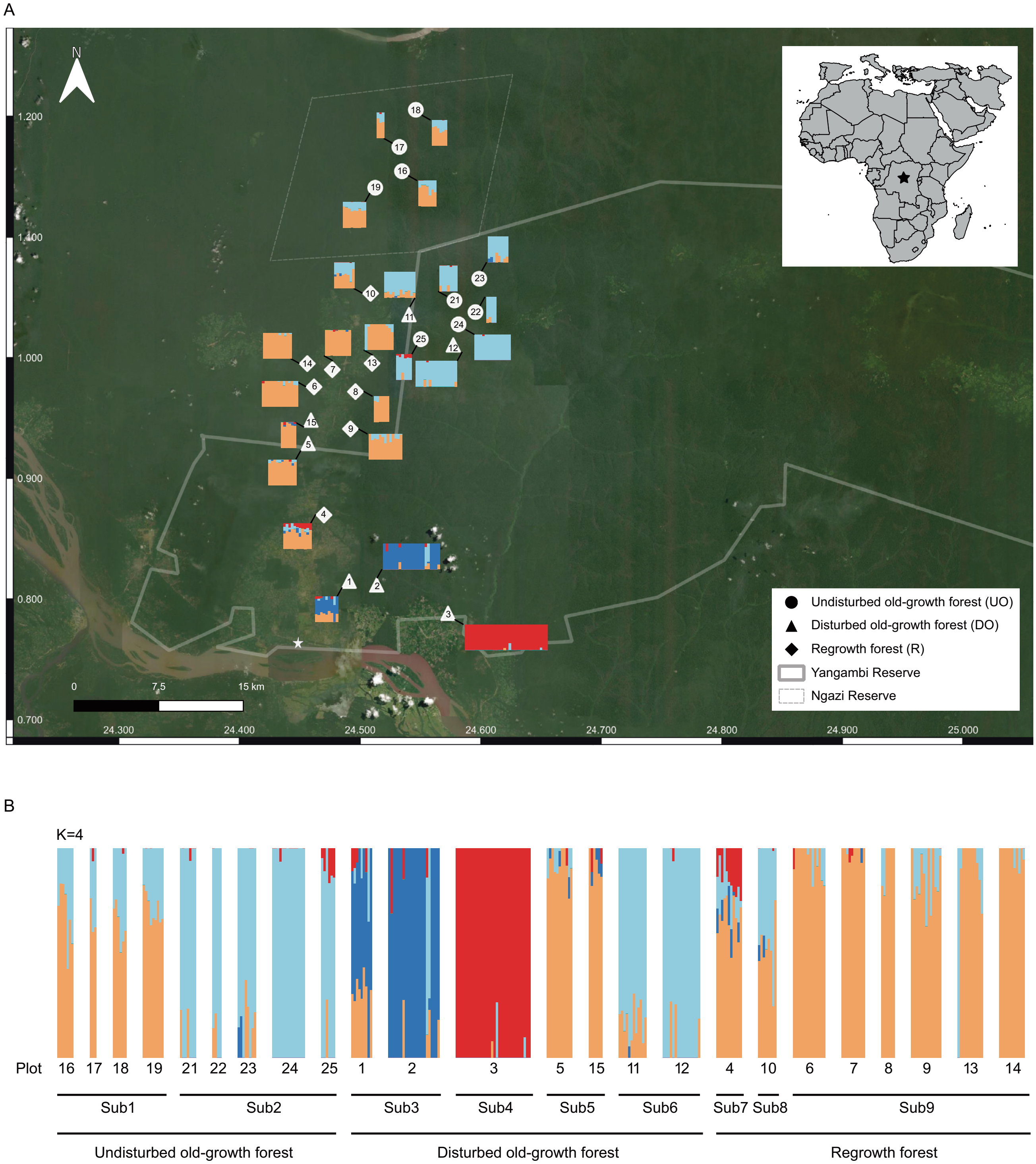

Fig. 1 Estimation of subpopulations using 256 wild *C. canephora* individuals and 794 SNPs. A Map of the Yangambi region showing the location of each plot as well as the K coloured segments (K = 4) of each individual within each plot. The white star on the map locates the commune of Yangambi. **B** The entire population of *C. canephora* was divided into 4 clusters (K = 4) using fastSTRUCTURE. Individuals are shown by thin vertical lines, which are divided into K coloured segments representing the estimated membership probabilities (Q) of each individual. Plots were grouped together in subpopulations according to DAPC.

uncalled and reads containing internal restriction sites were discarded using GBprocesS. Merged reads were aligned to the *C. canephora* reference genome sequence (Denoeud et al. 2014) with the BWA-mem algorithm in BWA 0.7.17 with default parameters. Alignments were sorted, indexed, and filtered on mapping quality above 20 with SAMtools 1.10 (Li et al. 2009).

Single-nucleotide polymorphisms (SNPs) 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, minQ 20 and minimal 80% completeness. The remaining SNPs were then subjected to further filtering with the following parameters: minDP10,

Table 1. Genetic diversity estimates for all *C. canephora* sampling plots across the three forest categories in the Yangambi region in DRC.

Plot	*N*	N_e	A_r	H_e	H_o	F_{is}
Plot_16	7	1.47	1.30	0.32	0.41	−0.27
Plot_17	3	1.46	1.31	0.32	0.41	−0.27
Plot_18	6	1.48	1.30	0.33	0.43	−0.30
Plot_19	9	1.47	1.29	0.32	0.41	−0.27
Plot_21	7	1.49	1.30	0.32	0.40	−0.25
Plot_22	4	1.47	1.30	0.32	0.40	−0.24
Plot_23	8	1.49	1.30	0.32	0.40	−0.25
Plot_24	14	1.49	1.30	0.32	0.41	−0.27
Plot_25	6	1.46	1.32	0.32	0.41	−0.28
Undisturbed old-growth forest	64	1.52	1.94	0.32	0.41	−0.27
Plot_01	9	1.50	1.30	0.32	0.41	−0.26
Plot_02	22	1.50	1.30	0.32	0.41	−0.27
Plot_03	32	1.50	1.30	0.32	0.41	−0.26
Plot_05	11	1.51	1.31	0.32	0.42	−0.31
Plot_11	12	1.48	1.29	0.32	0.41	−0.26
Plot_12	16	1.50	1.30	0.32	0.40	−0.25
Plot_15	6	1.50	1.30	0.32	0.41	−0.26
Disturbed old-growth forest	108	1.53	1.95	0.32	0.41	−0.27
Plot_04	11	1.50	1.31	0.32	0.42	−0.29
Plot_06	14	1.52	1.30	0.32	0.42	−0.29
Plot_07	10	1.50	1.30	0.32	0.41	−0.27
Plot_08	6	1.48	1.30	0.32	0.42	−0.29
Plot_09	13	1.50	1.30	0.32	0.40	−0.24
Plot_10	8	1.48	1.30	0.32	0.41	−0.27
Plot_13	11	1.49	1.31	0.32	0.41	−0.28
Plot_14	11	1.50	1.30	0.32	0.42	−0.29
Regrowth forest	84	1.53	1.94	0.32	0.41	−0.28
Total	256	1.57	1.97	0.32	0.41	−0.27

N sample size, N_e effective number of alleles, A_r allelic richness, H_o observed heterozygosity, H_e expected heterozygosity, F_{IS} inbreeding coefficient.

minGQ 30, max-missing 0.7, mac 3, minQ 30, min-alleles 2, max-alleles 2 and maf 0.05 using VCFtools 0.1.16 (Danecek et al. 2011).

Genetic diversity and structure analysis

The number of effective alleles (N_e) was calculated in GenAlEx 6.5 (Peakall and Smouse 2012) for each plot and forest category. Allelic richness (A_r) was calculated according to El Mousadik and Petit (1996) using the *allelic.richness* function of R package *hierfstat* (Goudet 2013). The observed and expected heterozygosity (H_o and H_e, respectively) and inbreeding coefficient (F_{IS}) were calculated in VCFtools, for each plot and forest category. All genetic diversity indices were compared between the three forest categories using non-parametric Kruskal–Wallis rank sum tests, followed by Dunn's Multiple Comparison tests. To estimate the genetic distance between all coffee plots, pairwise F_{ST} values (Weir and Cockerham 1984) were calculated using PLINK 1.9 (Chang et al. 2015). Mantel tests (Podani 2000) were performed in RStudio to test the correlation between genetic (F_{ST}) and geographical distances across all plots, and across plots within each of the three forest categories.

To comply with the assumptions for a genetic structure analysis, SNPs were filtered, using VCFtools based on Hardy–Weinberg Equilibrium (hwe 0.01), minor allele frequencies (maf 0.05) and linkage disequilibrium (indep-pairwise 50 10 0.5). Discriminant analysis of principal components (DAPC) (Jombart and Collins 2015) was then conducted using the R package ADEGENET (Jombart 2008; RStudio Team 2016). First, the *find.clusters* function, which runs successive K-means clustering with increasing number of clusters (k), was used to assess the number of clusters that maximises between-group variance and minimises within-group variance. The Bayesian Information Criterion (BIC) was applied to select the most optimal value of k. DAPC was then performed on the most optimal number of clusters (k) using the *dapc* function. Second, a Bayesian clustering implemented in fastSTRUCTURE 1.0 (Raj et al. 2014) was run to assess genetic structure in the *C. canephora* individuals given the most optimal number of genetic clusters (K). Hundred 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 most optimal 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). Graphical representation of the fastSTRUCTURE results was done in RStudio.

Relatedness and relationships

To reveal patterns of historical gene flow, we analysed the relatedness and relationships among the 256 individuals using multi-allelic short haplotypes (created via read-backed phasing of the SNPs with the SMAP package), which were preferred over single bi-allelic SNPs because of their increased information content (Schaumont et al. 2022). Furthermore, the use of haplotypes reduces the optimum number of SNPs needed to achieve high assignment success rates in complex scenarios (García-Fernández et al. 2018). Read-backed haplotyping of the SNPs was done with SMAP haplotype-sites using the optimal parameter settings for diploid individuals and double-enzyme GBS merged reads with: no-indels, min-haplotype-frequency 5, discrete calls dosage, dosage-filter 2, min-read-count 10, min-distinct-haplotypes 2, max-distinct-haplotypes 10, frequency-interval-bounds default for diploids, and locus-correctness 90%. Subsequently, SMAPapp-Matrix (https://gitlab.com/ybawin/smapapps) was used to calculate the locus information content (LIC). The criteria were set so that all loci with at least one unique haplotype were considered. Based on this criterion, we selected

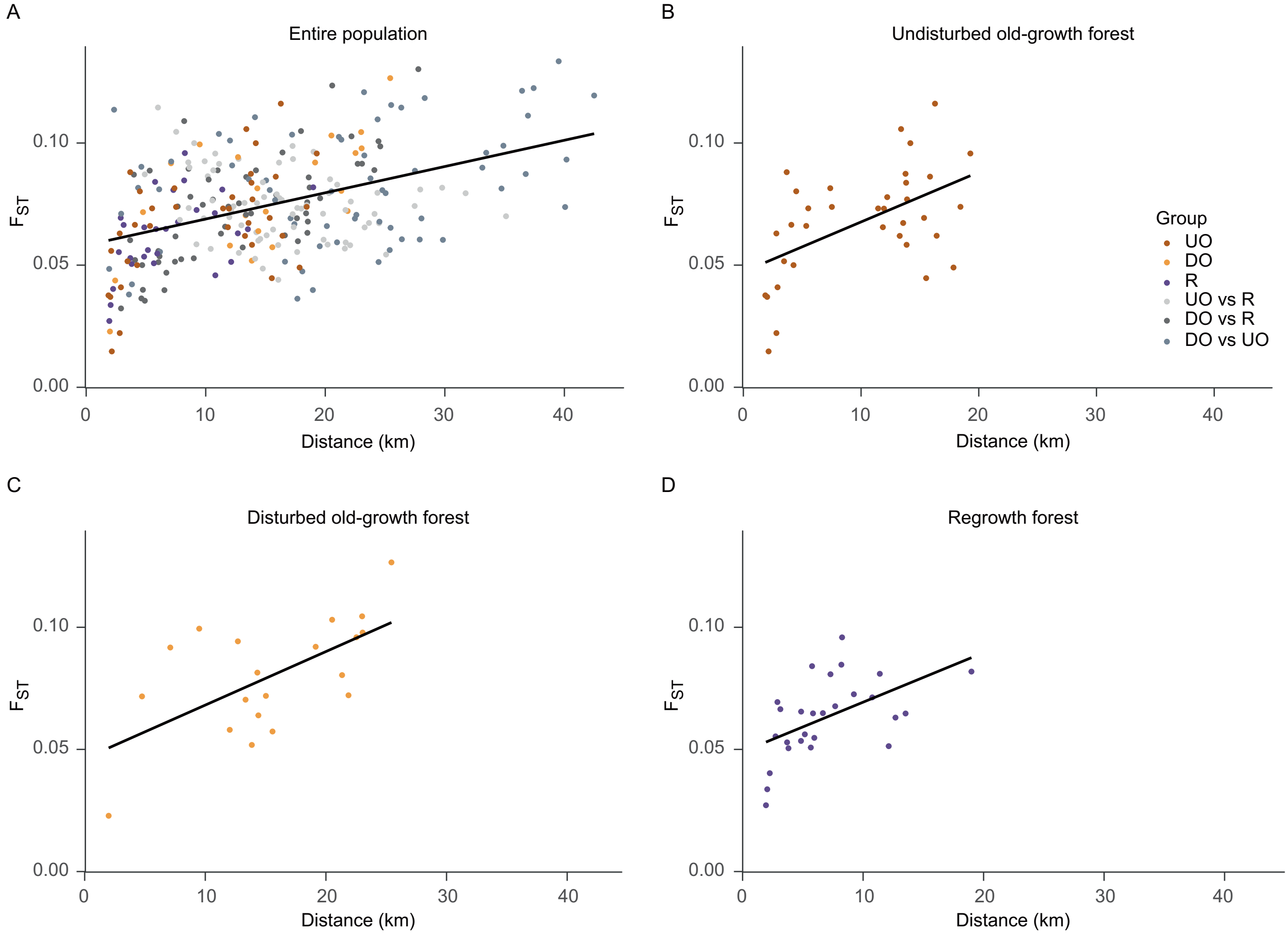

Fig. 2 Relationship between geographic (km) and genetic (F_{ST}) distances between *C. canephora* populations in different forest categories of the Yangambi region. The lines show the positive relationship between both variables. Relationship shown over **A** the entire population of *C. canephora*; **B** all individuals located in undisturbed old-growth forest (UO); **C** all individuals located in disturbed old-growth forest (DO); and **D** all individuals located in regrowth forest (R).

the haploset with the 100 most informative loci. The strength of the LIC is that informative loci with more than two haplotypes are retained even if the minor haplotype frequency is low, and thus LIC is a better criterion for the discriminatory power of all haplotypes per locus.

The relatedness (*r*), based on maximum likelihood, was calculated among all pairs of individuals using ML-Relate (Kalinowski et al. 2006). The relatedness between individual pairs represents the overall identity-by-descent in a continuous measure, ranging between zero and one (Blouin 2003). Mean relatedness was calculated afterwards between pairs of individuals at both the plot level and the forest category level. To evaluate differences between forest categories, a Kruskal–Wallis rank sum test and Dunn's Multiple Comparisons test was used.

In addition, and also using ML-Relate, pairs of individuals were classified into four pedigree relations using maximum likelihood estimates: unrelated (U), half-siblings (HS), full-siblings (FS), and parent-offspring (PO) (Jones et al. 2010). Log-likelihoods were calculated for all four relationships and the one with the highest value was assigned to the corresponding tested pair of individuals. Afterwards, the frequency of each assigned relationship was counted at both plot level and forest category level. Differences in frequencies of relationships at the category level were tested using Pearson's χ^2 test and pairwise Pearson's χ^2 tests with simulated *p* values based on 9999 replicates, using the chisq.test function in the *stats* package (RStudio Team 2016).

RESULTS

SNP discovery and selection

A total of 18,894 bi-allelic SNPs with a completeness of at least 80% were identified across all individuals in undisturbed old-growth forest ($n = 64$ coffee samples), disturbed old-growth forest ($n = 108$), and regrowth forest ($n = 84$) plots. Of these, 3212 SNPs with a minimum minor allele count of 3 and a minimum minor allele frequency of 0.05 were used for the genetic diversity analysis. We used 794 SNPs for the genetic structure analysis, after filtering based on the Hardy–Weinberg and linkage disequilibrium criteria.

Genetic diversity

Genetic diversity measures for all plots and forest categories are presented in Table 1. The number of effective alleles (N_e) was significantly lower in plots in undisturbed old-growth forest than in plots in disturbed old-growth forest ($p = 0.006$), but N_e in plots in regrowth forest was not significantly different from plots in disturbed and undisturbed old-growth forest ($p = 0.10$ and $p = 0.11$, respectively). Allelic richness (A_r) was significantly lower in plots in undisturbed old-growth forest compared to plots in disturbed old-growth forest ($p < 0.001$) and significantly lower in plots in regrowth forest compared to both plots in disturbed and undisturbed old-growth forest ($p = 0.03$ and $p < 0.001$, respectively). No significant differences were found in observed heterozygosity (H_o), expected heterozygosity (H_e), and inbreeding coefficient (F_{IS}) between plots from the three forest categories ($p = 0.16$, $p = 0.85$, $p = 0.15$, respectively).

Pairwise genetic differentiation (F_{ST}) was highest among plots in disturbed and plots in undisturbed old-growth forest ($F_{ST} = 0.025$).

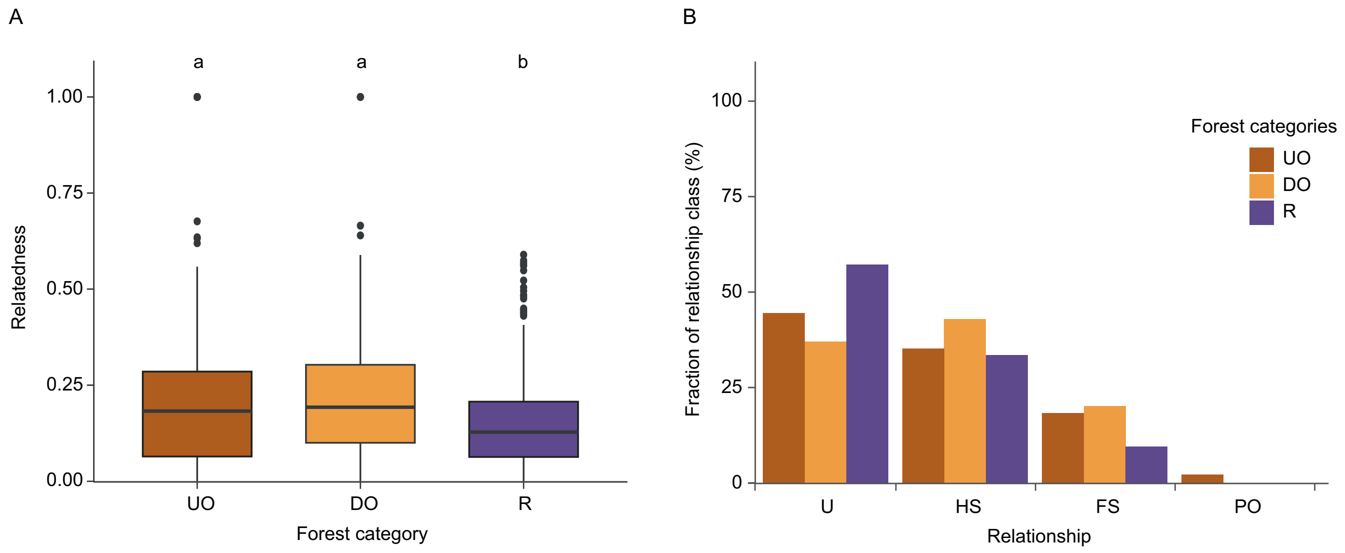

Fig. 3 Determination of genetic relatedness and relationships at the forest category level. A Estimated relatedness between forest categories undisturbed old-growth forest (UO), disturbed old-growth forest (DO), and regrowth forest (R). **B** Estimated frequencies of unrelated individuals (U), half-siblings (HS), full-siblings (FS), and parent-offspring (PO) relationships between forest categories undisturbed old-growth forest (UO), disturbed old-growth forest (DO), and regrowth forest (R). Letters code for significantly different forest categories.

Genetic differentiation was higher among plots in undisturbed old-growth forest and plots in regrowth forest ($F_{ST} = 0.025$) than among plots in disturbed old-growth forest and plots in regrowth forest ($F_{ST} = 0.017$). Isolation-by-distance was found across all plots (Mantel *r*-statistic = 0.24, $p = 0.01$; Fig. 2A) and across the plots within disturbed and undisturbed old-growth forest separately (undisturbed old-growth forest: Mantel *r*-statistic = 0.48, $p = 0.024$, Fig. 2B; disturbed old-growth forest: Mantel *r*-statistic = 0.47, $p = 0.036$, Fig. 2C). No significant isolation-by-distance was found across plots in regrowth forest (Mantel *r*-statistic = 0.24, $p = 0.2$; Fig. 2D).

Genetic structure analysis

Four different clusters were identified using the DAPC analysis, performed on the first hundred PCs of the PCA and three discriminant eigenvalues. Cluster 1, containing samples from plot 3, was separated from the other clusters according to Linear Discriminant 1 (LD1) (Supplementary Fig. S1B).

The fastSTRUCTURE analysis showed that the plots in undisturbed old-growth forest were divided into two subpopulations (Fig. 1B), namely plot 16–19 and plot 21–25, with individuals in plots 16–19 showing a mixture of DAPC cluster 2 and 4 (Fig. 1A). Plots in disturbed old-growth forest were divided into four subpopulations (Fig. 1B), namely: plots 1 and 2; plot 3; plots 5 and 15; plots 11 and 12 (Fig. 1A). Plots in regrowth forest were divided into three subpopulations (Fig. 1A), namely: plot 4; plot 10; plots 6–9, 13 and 14. Individuals in plot 10 showed to be a mixture of DAPC clusters 2 and 4, whereas individuals in plot 4 showed a mixture of all DAPC clusters. Overall, clustering of the samples from old-growth forest was consistently differentiated according to the geographical location of the plots (Fig. 1A).

Relatedness and relationships

The average relatedness values per plot ranged between 0.074 and 0.304, with an average value of 0.184. The relatedness between pairs of individuals was significantly different among forest categories ($\chi^2 = 65.27$, $p < 0.001$) (Fig. 3A). Specifically, it was lower in regrowth forest than in both undisturbed old-growth forest and disturbed old-growth forest ($p < 0.001$). There were no significant differences between undisturbed old-growth forest and disturbed old-growth forest.

The frequency distribution of the relationship (unrelated, half-siblings, full-siblings, parent-offspring) of pairs of individuals within plots was significantly different among forest categories ($\chi^2 = 86.543$, $p < 0.05$). Specifically, significant differences in the frequency distribution were found among plots in undisturbed old-growth forest and disturbed old-growth forest ($\chi^2 = 27.49$, $p < 0.001$), plots in undisturbed old-growth forest and regrowth forest ($\chi^2 = 22.83$, $p < 0.001$), and plots in disturbed old-growth forest and regrowth forest ($\chi^2 = 54.69$, $p < 0.001$) (Fig. 3B). Parent-offspring pairs were only found in undisturbed old-growth forest, and even then, only at low incidence, namely: two pairs in plot 21; one pair in plot 22; one pair in plot 23; one pair in plot 24.

DISCUSSION

Understanding the genetic variation and its relation with anthropogenic disturbance is of major importance for the conservation of *C. canephora* genetic resources in the Congo Basin. Our study encompasses the most densely sampled set of wild individuals of *C. canephora* so far. By using GBS-derived SNP markers, we were able to quantify genetic diversity, map genetic structure, and determine pedigree relations in *C. canephora* populations and compare these indicators among undisturbed old-growth forest, disturbed old-growth forest, and regrowth forest.

Genetic diversity

A high genetic diversity, both in terms of allelic diversity and heterozygosity, was found in all 24 sampling plots across the three forest categories. Our findings are in line with other studies that have used SSR markers, such as Vanden Abeele et al. (2021), who found high heterozygosity ($H_o = 0.48$) in wild *C. canephora* populations in the Tshopo Province of DR Congo (Yangambi and Yoko, Kisangani). Likewise, Nyakaana (2007) found a high mean observed heterozygosity ($H_o = 0.46$) across five localities in the Kibale National Park in Uganda. Elsewhere in Uganda, Musoli et al. (2009) found a mean H_o of 0.37 over two separate regions, while Kiwuka et al. (2021) found a mean H_o of 0.51 over seven distinct regions. All our sampling plots harboured a higher H_o than reported over the whole Guineo-Congolian region, where H_o ranged between 0.27 and 0.38 (Gomez et al. 2009; Cubry et al.

2013). This suggests that the Yangambi area is key for the conservation of *C. canephora* genetic resources.

We hypothesised that anthropogenic disturbance leads to decreased population genetic diversity, specifically due to selective logging (Depecker et al. 2022). However, we found no evidence of reduced genetic diversity in plots in disturbed old-growth forest, as compared to plots in undisturbed old-growth forest. On the contrary, the number of effective alleles and allelic richness were significantly lower in plots in undisturbed old-growth forest, compared to plots in disturbed old-growth forest. One explanation could be historical spatial variation in genetic diversity. Unfortunately, there is no information on the historical genetic diversity of wild *C. canephora* in the Yangambi area. An alternative explanation for the higher genetic diversity could be increased levels of gene flow in disturbed areas, contrary to our expectations. Several studies have found enhanced pollinator activity through disturbance, promoting gene flow (Dick et al. 2003 and references therein). It is possible that forest disturbance altered the pollinating insect communities, with for instance a higher abundance of *Apis mellifera*, which can govern long-distance pollination as has been observed in the tropical Amazonian tree species *Dinizia excelsa* (Dick et al. 2003). Furthermore, competition for space and resources is likely to be reduced in disturbed areas, possibly resulting in higher seed survival and germination rates, which can lead to better seedling establishment (Olsson et al. 2019).

In addition, we found that, in general, observed heterozygosity was higher than expected heterozygosity, indicating an excess of heterozygotes. Such negative values of F_{IS} are in accordance with obligate outcrossing in self-incompatible plant species like *C. canephora* (Mateu-Andrés and De Paco 2006).

Genetic structure

Limited genetic connectivity between plots may have resulted in relatively high genetic differentiation among our sampled plots at relatively short geographical distance. Likewise, in the tropical rainforest of West Uganda, Nyakaana (2007) detected strong genetic differentiation between five *C. canephora* populations, which were only separated by short geographical distances. Similar patterns were detected in other tropical woody plant species, including several *Psychotria* species (Theim et al. 2014), *Paypayrola blanchetiana* (Braun et al. 2020), and *Theobroma cacao* (Lachenaud and Zhang 2008).

The genetic differentiation was also reflected in the significantly genetically diverged clusters, which were demonstrated by DAPC analysis and further supported by the fastSTRUCTURE analysis. It is remarkable that the *C. canephora* individuals sampled in plot 3 were markedly genetically separated from the individuals sampled in all other plots. This may be explained by the monodominant and species-poor *Gilbertiodendron dewevrei* forest that forms a natural barrier and isolates plot 3 from the other ones (Kearsley et al. 2017). In general, this type of monodominant forests significantly alters the understorey environment, making it difficult for other species to establish and survive (Torti et al. 2001). *Coffea canephora* has never been observed in the *G. dewevrei* forest understorey (Asimonyio and Kambale, pers. obs.), but more sampling in the vicinity of plot 3 is needed to confirm this hypothesis.

Geographic location, rather than anthropogenic disturbance, appears to be the main driver of genetic structure across the whole study area. The identified subclusters can be attributed to two different gradients in terms of geographic distance, and which are clearly subjected to isolation-by-distance. Firstly, a more or less east-west gradient separating plots 11 and 21–25 from the other plots. Secondly, a north-south gradient, visible in the more continuously sampled area. Within this north-south gradient, plots in the south are clearly more differentiated and admixed than plots in the north. Within the Yangambi region, anthropogenic activity is high close to the Congo River. Nevertheless, further research is necessary to test the association between the anthropogenic activities and the higher rate of differentiation in plots in the south. Plots in regrowth forest follow the same patterns of gene flow and show high levels of admixture but are not subjected to isolation-by-distance. These findings suggest that after agricultural abandonment, *C. canephora* individuals coming from neighbouring old-growth forests recolonised this area. This is further supported by the lower allelic diversity, which hints at founder effects, possibly due to a limited number of migrants.

Pedigree relations

By assessing the relatedness and relationships between pairs of individuals, we were able to reveal putative patterns of dispersal and recruitment. Overall, we found a low to moderate relatedness among pairs of individuals within one plot. Furthermore, most pairs of individuals in the majority of the plots were unrelated or half-siblings. Combined with the observed genetic structure, we hypothesise that gene flow is limited at larger distances (between plots), as also indicated by the significant isolation-by-distance and the observation that no parent-offspring pairs were found between plots, with a minimal distance of two km. This distance-dependent decay of gene flow is consistent with the theory of isolation-by-distance models (Vekemans and Hardy 2004).

A lower relatedness was detected within plots in regrowth forest as compared to within plots in undisturbed old-growth forest and in disturbed old-growth forest. This confirms that these regrowth areas were colonised by *C. canephora* migrants from multiple neighbouring sources, thus, lowering the average relatedness among individuals within plots in regrowth forests. This pattern is supported by the common observation that relatedness decreases as migration increases (Jones and Wang 2012).

CONCLUSION

We found that the wild *C. canephora* populations in the Yangambi region harbour both a high allelic diversity and heterozygosity, thereby pointing at the importance of the wild *C. canephora* populations in the Congo Basin as hotspots of genetic diversity. Because local studies on the genetic diversity of wild *C. canephora*, and by extension other rainforest understorey species, are very rare in the Congo Basin, our study can be used as a reference for future research, in which novel quantifications of genetic diversity can be compared with the values found in our work. Indeed, although we could not detect genetic erosion in disturbed forests, it is important to continue monitoring the effect of anthropogenic disturbance on the genetic diversity, genetic structure, and gene flow in wild populations of *C. canephora*, because the observations made in this study might be influenced by the historical distribution of genetic diversity. Conservation of the genetic diversity and actors governing gene flow in old-growth forests is crucial, although the populations in regrowth forests can aid in the maintenance of the genetic resources, which are important for the future of coffee cultivation.

DATA AVAILABILITY

FASTQ read files of all GBS libraries were deposited at the NCBI sequence read archive (SRA) in BioProject (PRJNA901681).

REFERENCES

Aguilar R, Cristóbal-Pérez ED, Balvino-Olvera FJ, Aguilar-Aguilar MDJ, Aguirre-Acosta N, Ashworth L et al. (2019) Habitat fragmentation reduces plant progeny quality: a global synthesis. Ecol Lett 22:1163–1173

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

Barlow J, Lennow GD, Ferreira J, Berenguer E, Lees AC, Nally RM et al. (2016) Anthropogenic disturbance in tropical forests can double biodiversity loss from deforestation. Nature 535:144–147

Barret SC, Eckert CG (1990) Current issues in plant reproductive ecology. Isr J Plant Sci 39:5–12

Bawa KS, Bullock SH, Perry DR, Coville RE, Grayum MH (1985) Reproductive biology of tropical lowland rain forest trees II. Pollination systems. Am J Bot 72:346–356

Bello C, Galetti M, Pizo MA, Magnago LFS, Roch MF, Lima RA, et al. (2015) Defaunation affects carbon storage in tropical forests. Sci Adv 1:e1501105. https://doi.org/10.1126/sciadv.1501105

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol 18:503–511

Born C, Kjellberg F, Chevallier M-H, Vignes H, Dikangadissi J-T, Sanguié J et al. (2008) Colonization processes and the maintenance of genetic diversity: insight from a pioneer rainforest tree, *Aucoumea Klaineana*. Proc R Soc B 275:2171–2179

Braun M, Dantas L, Esposito T, Pedrosa-Harand A (2020) Strong genetic differentiation on a small geographic scale in the Neotropical rainforest understory tree *Paypayrola blanchetiana* (Violaceae). Tree Genet Genomes. https://doi.org/10.1007/s11295-020-01477-5

Campbell AJ, Carvalheiro LG, Maués MM, Jaffé R, Giannini TC, Freitas MAB et al. (2018) Anthropogenic disturbance of tropical forests threatens pollination services to açai palm in the Amazon river delta. J Appl Ecol 55:1725–1736

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience https://doi.org/10.1186/s13742-015-0047-8

Chiriboga-Arroyo F, Jansen M, Bardales-Lozano R, Ismail SA, Thomas E, Garcia M et al. (2021) Genetic threats to the Forest Giants of the Amazon: Habitat degradation effects on the socio-economically important Brazil nut tree (*Bertholletia excelsa*). Plants People Planet 3:194–210

Cramer PJS, Wellman FL (1957) Review of literature of coffee research in Indonesia. SIC Editorial, Inter-American Institute of Agricultural Sciences

Craparo ACW, Van Asten PJ, Läderach P, Jassogne LT, Grab SW (2015) *Coffea arabica* yields decline in Tanzania due to climate change: Global implications. Agric Meteorol 207:1–10

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

Curtis PG, Slay CM, Harris NL, Tyukavina A, Hansen MC (2018) Classifying drivers of global forest loss. Science 361:1108–1111

Da Silva JMC, Tabarelli M (2000) Tree species impoverishment and the future flora of the Atlantic forest of northeast Brazil. Nature 404:72–74

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Davis AP, Gole TW, Baena S, Moat J (2012) The impact of climate change on indigenous arabica coffee (*Coffea arabica*): predicting future trends and identifying priorities. PLoS One. https://doi.org/10.1371/journal.pone.0047981

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M 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, Hatangi Y, Kambale J-L, Verleysen L, 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. J Trop Ecol. https://doi.org/10.1017/S0266467422000347

Dick CW, Etchelecu G, Austerlitz F (2003) Pollen dispersal of tropical trees (*Dinizia excelsa*: Fabaceae) by native insects and African honeybees in pristine and fragmented Amazonian rainforest. Mol Ecol 12:753–764

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

Edwards DP, Socolar JB, Mills SC, Burivalova Z, Koh LP, Wilcove DS (2019) Conservation of tropical forests in the Anthropocene. Curr Biol 29:R1008–R1020

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. Theor Appl Genet 92:832–839

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. https://doi.org/10.1371/journal.pone.0019379

Ernst C, Mayaux P, Verhegghen A, Bodart C, Christophe M, Defourny P (2013) National forest cover change in Congo Basin: deforestation, reforestation, degradation and regeneration for the years 1990, 2000 and 2005. Glob Chang Biol 19:1173–1187

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

FAO, UNEP (2020) The State of the World's Forests 2020. In Forests, bio-diversity and people. FAO and UNEP

Ferrão RG, da Fonseca AFA, Ferrão MAG, De Mune LH (2019) Conilon Coffee: the *Coffea canephora* produced in Brazil. Incaper, Vitória-ES, Brasil

Gardner TA, Barlow J, Chazdon R, Ewers RM, Harvey CA, Peres CA et al. (2009) Prospects for tropical forest biodiversity in a human-modified world. Ecol Lett 12:561–582

García-Fernández C, Sánchez JA, Blanco G (2018) SNP-haplotypes: An accurate approach for parentage and relatedness inference in gilthead sea bream (*Sparus aurata*). Aquaculture 495:582–591

Gomez C, Dussert S, Hamon P, Hamon S, De Kochko A, Poncert V (2009) Current genetic differentiation of *Coffea canephora* pierre ex a. Froehn in the guineo-Congolian african zone: Cumulative impact of ancient climatic changes and recent human activities. BMC Evol Biol 9:167

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

Hubbell SP, Foster RB (1986) Biology, chance and history and the structure of tropical rain forest tree communities. In: Diamond JM, Case TJ (eds) Community ecology. Harper and Row, New York, NY, p 314–329

ICO (2022) Coffee Market Report: August 2022. Donwloaded from International Coffee Organization https://www.ico.org/documents/cy2021-22/cmr-0822-e.pdf

Ismail SA, Ghazoul J, Ravikanth G, Kushalappa CG, Uma Shaanker R, Kettle CJ (2017) Evaluating realized seed dispersal across fragmented tropical landscapes: A two-fold approach using parentage analysis and the neighbourhood model. N Phytol 214:1307–1316

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

Jombart T, Collins C (2015) Analysing genome-wide SNP data using adegenet 2.0.0. https://adegenet.r-forge.r-project.org/files/tutorial-genomics.pdf

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. Mol Ecol Resour 10:6–30

Jones OR, Wang J (2012) A comparison of four methods for detecting weak genetic structures from maker data. Ecol Evol 2:1048–1055

Kalinowski ST, Wagner AP, Taper ML (2006) ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship. Mol Ecol Notes 6:576–579

Kearsley E, Verbeeck H, Hufkens K, Van, de Perre F, doetterl S, Baert G et al. (2017) Functional community structure of African monodominant *Gilbertiodendron dewevrei* forest influenced by local environmental filtering. Ecol Evol 7:295–304

Kier G, Mutke J, Dinerstein E, Ricketss TH, Küper W, Kreft H et al. (2005) Global patterns of plant diversity and floristic knowledge. J Biogeogr 32:1107–1116

Kiwuka C, Goudsmit E, Tournebize R, Oliveir de Aquino S, Douma JC, Bellanger L 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

Kreft H, Jetz W (2007) Global patterns and determinants of vascular plant diversity. Proc Natl Acad Sci USA 104:5925–5930

Lachenaud P, Zhang D (2008) Genetic diversity and population structure in wild stands of cacao trees (*Theobroma cacao* L.) in French Guiana. Ann For Sci. https://doi.org/10.1051/forest:2008011

Lashermes P, Combes MC, Ribas A, Cenci A, Mahé L, Etienne H (2010) Genetic and physical mapping of the SH3 region that confers resistance to leaf rust in coffee tree (*Coffea arabica* L.). Tree Genet Genomes 6:973–980

Leroy T, Marraccini P, Dufour M, Montagnon C, Lashermes P, Sabau X et al. (2005) Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes. Theor Appl Genet 111:1031–1041

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 14:2078–2079

Li YL, Liu JX (2018) StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. Mol Ecol Resour 18:176–177

Makelele IA, Verheyen K, Boeckx P, Ntaboba LC, Bazirake BM, Ewango C et al. (2021) Afrotropical secondary forests exhibit fast diversity and functional recovery, but slow compositional and carbon recovery after shifting cultivation. J Veg Sci 32:1–13

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12

Mateu-Andrés I, De Paco L (2006) Genetic diversity and the reproductive system in related species of *Antirrhinum*. Ann Bot 98:1053–1060

Mayr E (1954) Change of genetic environment and evolution. In: Huxley A, Hardy AC, Ford EB (eds) Evolution as a process. Allen and Unwin, London, p 157–180

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Merot-L'anthoene V, Tournebize R, Darracq O, Rattina V, Lepelley M, Bellanger L 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 Biotechnol J 17:1418–1430

Musoli P, Cubry P, Aluka P, Billot C, Dufour M, De Bellis F et al. (2009) Genetic differentiation of wild and cultivated populations: diversity of *Coffea canephora* Pierre in Uganda. Genome 52:634–646

Neushulz EL, Mueller T, Schleuning M, Böhning-Gaese K (2016) Pollination and seed dispersal are the most threatened processes of plant regeneration. Sci Rep 6:1–6

Norden N, Chazdon RL, Chao A, Jiang YH, Vilchez-Alvarado B (2009) Resilience of tropical rain forests: tree community reassembly in secondary forests. Ecol Lett 12:385–394

Nowak MD, Davis AP, Anthony F, Yoder AD (2011) Expression and trans-specific polymorphism of self-incompatibility RNases in *Coffea* (Rubiaceae). PLoS One. https://doi.org/10.1371/journal.pone.0021019

Nyakaana S (2007) Microgeographical genetic structure of forest robusta coffee (*Coffea canephora*, Pierre), in Kibale National Park, Uganda. Afr J Ecol 45:71–75

Oberleitner F, Egger C, Oberdorfer S, Dullinger S, Wanek W, Hietz P (2021) Recovery of aboveground biomass, species richness and composition in tropical secondary forests in SW Costa Rica. Ecol Manag 479:118580

Olsson O, Nuñez-Iturri G, Smith HG, Ottosson U, Effium EO (2019) Competition, seed dispersal and hunting: what drives germination and seedling survival in an Afrotropical forest? AoB Plants https://doi.org/10.1093/aobpla/plz018

Oryem-Origa H (1999) Fruit and seed ecology of wild Robusta coffee (*Coffea canephora* Froehner) in Kibale National Park. Uganda Afr J Ecol 37:439–448

Peakall R, Smouse RPP (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28:2537–2539

Podani J (2000) Introduction to the exploration of multivariate biological data. Backhuys Publishers, Kerkwere

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome. https://doi.org/10.3835/plantgenome2012.05.0005

Poorter L, Craven D, Jakovac CC, van der Sande MT, Amissah L, Bongers F et al. (2021) Multidimensional tropical forest recovery. Science 374:1370–1376

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573–589

RStudio Team (2016) RStudio: Integrated Development for R

Sasaki N, Putz FE (2009) Critical need for new definitions of "forest" and "forest degradation" in global climate change agreements. Conserv Lett 2:226–232

Sezen UU, Chazdon RL, Holsinger KE (2007) Multigenerational genetic analysis of tropical secondary regeneration in a canopy palm. Ecology 88:3065–3075

Schaumont D, Veeckman E, Van der Jeugt F, Haegeman A, van Glabeke S, Bawin Y et al. (2022) Stack Mapping Anchor Points (SMAP): a versatile suite of tools for read-backed haplotyping. Preprint at bioRxiv https://doi.org/10.1101/2022.03.10.483555

Shapiro AC, Grantham HS, Aguilar-Amuchastegui N, Murray NJ, Gond V, Bonfils D, et al. (2021) Forest condition in the Congo Basin for the assessment of ecosystem conservation status. Ecol Indic. https://doi.org/10.1016/j.ecolind.2020.107268

Silva MDC, Várzea V, Guerra-Guimarães L, Azinheira HG, Fernandez D, Petitot AS et al. (2006) Coffee resistance to the main diseases: leaf rust and coffee berry disease. Braz J Plant Physiol 18:119–147

Theim TJ, Shirk RY, Givnish TJ (2014) Spatial genetic structure in four understorey *Psychotria* species (Rubiaceae) and implications for tropical forest diversity. Am J Bot 101:1189–1199

Torti SD, Coley PD, Kursar TA (2001) Causes and consequences of monodominance in tropical lowland forests. Am Nat 157:141–153

Tyukavina A, Hansen MC, Potapov P, Parker D, Okpa C, Stehman SV, et al. (2018) Congo Basin forest loss dominated by increasing smallholder clearing. Sci Adv. https://doi.org/10.1126/sciadv.aat2993

Vanden Abeele S, Janssens SB, Asimonyio Anio J, Bawin Y, Depecker J, Kambale B et al. (2021) Genetic diversity of wild and cultivated *Coffea canephora* in northeastern DR Congo and the implications for conservation. Am J Bot 108:2425–2434

Vandepitte K, Gristina AS, De Hert K, Meekers T, Roldán-Ruiz I, Honnay O (2012) Recolonization after habitat restoration leads to decreased genetic variation in populations of a terrestrial orchid. Mol Ecol 21:4206–4215

Van Vliet N, Muhindo J, Kbale Nyumu J, Mushagalusa O, Nasi R (2018) Mammal depletion processes as evidenced from spatially explicit and temporal local ecological knowledge. Trop Conserv Sci 11:1–16

Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. Mol Ecol 13:921–935

Vranckx G, Jacquemyn H, Muys B, Honnay O (2012) Meta-analysis of susceptibility of woody plants to loss of genetic diversity through habitat fragmentation. Conserv Biol 26:228–237

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

Wellman FL (1961) Coffee. Botany, cultivation, and utilization. Leonard Hill, London

Widmer A, Lexer C (2001) Glacial refugia: sanctuaries for allelic richness, but not for gene diversity. Trends Ecol Evol 16:267–269

Wright S (1932) The role of mutation, inbreeding, crossbreeding and selection in evolution. In: Proceedings of the sixth international congress of genetics. pp 356–366.

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

ACKNOWLEDGEMENTS

We would like to thank the Institut National pour l'Étude et la Recherche Agronomiques (INERA) and the FORETS project, which is financed by the 11th European Development Fund, for facilitating the field mission. We would also like to express our sincere gratitude to the Ministère de L'Environnement et Développement Durable (MEDD) for their help with obtaining permits (N°008/ANCCB-RDC/SG-EDD/BTB/11/2020 and N°001/ANCCB-RDC/SG-EDD/BTB/01/2021).

AUTHOR CONTRIBUTIONS

OH, FV, TR, JD and LV designed this study. JD, JAA, YH, J-LK, IMM, LV and TE participated in fieldwork. LV and AS executed the lab work. JD, LV and YB analysed the data. JD, LV, FV, OH and TR wrote the manuscript. All authors contributed to finalising the manuscript.

FUNDING

This study was funded by Research Foundation-Flanders, research mandate granted to JD (FWO; 1125221N) and research project granted to OH (FWO; G090719N), and the Foundation for the promotion of biodiversity research in Africa, granted to JD and YH (SBBOA, www.sbboa.be).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-022-00588-0.

Correspondence and requests for materials should be addressed to Jonas Depecker or Lauren Verleysen.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.