# Genome-wide admixture mapping identifies wild ancestry-of-origin segments in cultivated Robusta coffee

Tram Vi[1,2], Yves Vigouroux[1], Philippe Cubry[1], Pierre Marraccini[1,3], Ha Viet Phan[4], Giang Ngan Khong[2], Valerie Poncet[1*]

1 UMR DIADE, Univ Montpellier, IRD, CIRAD, Montpellier, France

2 National Key Laboratory of Plant Cellular Biotechnology, Agricultural Genetics Institute, Hanoi, Vietnam

3 CIRAD, UMR DIADE, Montpellier, France

4 Western Highlands Agriculture & Forestry Science Institute, Buon Ma Thuot, Vietnam

*Corresponding authors: Tram Vi, DIADE, Univ Montpellier, IRD, CIRAD, Montpellier, France, National Key Laboratory of Plant Cellular Biotechnology, Agricultural Genetics Institute, Hanoi, Vietnam, vbt576@gmail.com

1

# Abstract

Humans have had a major influence on the dissemination of crops beyond their native range, thereby offering new hybridization opportunities. Characterizing admixed genomes with mosaic origins generates valuable insight into the adaptive history of crops and the impact on current varietal diversity. We applied the ELAI tool—an efficient local ancestry inference method based on a two-layer hidden Markov model to track segments of wild origin in cultivated accessions in the case of multiway admixtures. Source populations—which may actually be limited and partially admixed—must be generally specified when using such inference models. We thus developed a framework to identify local ancestry with admixed source populations. Using sequencing data for wild and cultivated *Coffea canephora* (commonly called Robusta), our approach was found to be highly efficient and accurate on simulated hybrids. Application of the method to assess elite Robusta varieties from Vietnam led to the identification of an accession derived from a likely backcross between two genetic groups from the Congo Basin and the western coastal region of Central Africa. Admixtures resulting from crop hybridization and diffusion could thus lead to the generation of elite high-yielding varieties. Our methods should be widely applicable to gain insight into the role of hybridization during plant and animal evolutionary history.

# Keywords

# Significance

Local ancestry inference (LAI) has been widely investigated in humans to decipher genomic and evolutionary history, yet it is less commonly used in crops. Here we applied this approach to study mosaic genome origins in cultivated *Coffea canephora* (Robusta) accessions. We have proposed a new method to derive source populations for this analysis based on ancestral genotype frequencies estimated from native populations. Validation using simulated hybrids and ancestry deconvolution of the cultivated accessions revealed this approach to be promising for genomic studies of *C. canephora* as well as other crops.

# Introduction

Human history and migrations have markedly impacted crop dispersal patterns worldwide (Khoury et al. 2016). Cultivated plants have gradually, over time, undergone diffusion beyond their centers of origin (Meyer et al. 2012), while exchanging genetic material from their original wild relatives via hybridization. For example, Australian wheat cultivars are the result of multiway admixture of lineages of European, African and American origins (Joukhadar et al. 2017). Intraspecific admixture is a key factor in population diversity, adaptation and evolution (Goetz et al. 2014; Rius & Darling 2014). Cultivated individuals may therefore express novel phenotypic variability, e.g. beneficial ear traits in domesticated emmers (Nave et al. 2019; Oliveira et al. 2020), or higher antioxidant activity in a new litchi cultivar (Zhao et al. 2020; Hu et al. 2022).

Population genetics tools have been widely developed to gain insight into admixture (Pritchard et al. 2000; Padhukasahasram 2014). STRUCTURE (Pritchard et al. 2000), ADMIXTURE

1 (Alexander et al. 2009), or other tools like sNMF (Frichot et al. 2014) are commonly used to

2 infer global admixture proportions and ancestry per individual in a population. Yet individuals

3 presenting the same global ancestry might differ with respect to admixture patterns at the

4 chromosome level. More recent advances in sequencing, and application of statistical and

5 computing technologies have also enhanced ancestry inference at this chromosome level based

6 on local ancestry inference (LAI) approaches. At a fine genome scale, the local ancestry

7 provides better information on the mosaic genetic origin of admixed individuals, i.e. enhancing

8 knowledge of demographic histories, facilitating detection of adaptive introgression,

9 deciphering complex traits, and mapping underlying genes in admixed populations (Shriner

10 2017; Padhukasahasram 2014; Thornton & Bermejo 2014; Mani 2017; Geza et al. 2018). For

11 instance, admixture mapping based on LAI of a three-way admixed human population

12 identified a genomic region in native American ancestry linked to Alzheimer's disease (Norris

13 et al. 2020; Horimoto et al. 2021).

14

15 Numerous LAI tools have been developed. In most of them, genotypes from putative ancestral

16 populations (so-called 'source populations') are used as reference to infer the local ancestry of

17 admixed individuals or tested individuals (Sankararaman et al. 2008). Most LAI models are

18 based on hidden Markov models (HMM), including SABER (Tang et al. 2006), LAMP-LD

19 (Baran et al. 2012) and ELAI (Guan 2014), which predict the classification of ancestries in

20 hidden states by monitoring genotypes of the source populations (Baran et al., 2012). Other

21 methods implement strategies based on principal component analysis, e.g. PCADMIX (Brisbin

22 et al. 2012), Markov chain Monte Carlo (Chromopainter, Lawson et al. 2012), random forest

23 (RFMix, Maples et al. 2013), K-means (EILA, Yang et al. 2013) and other clustering methods

24 (Wu et al. 2021). LAI models can also be categorized into linkage disequilibrium (LD)-based

1 and non-LD-based models (Geza et al. 2018). Most LAI software requires phased genotypes

2 (Wu et al. 2021), which is not always accurately estimated with short-read data obtained via

3 next generation sequencing (NGS) technologies (Garg et al. 2021). Some tools need biological

4 information such as genetic and physical mapping data, recombination rates and admixture

5 generations, or statistical parameters such as hidden states and misfitting probabilities (Geza et

6 al. 2018).

7

8 According to previous studies comparing the performance of several commonly-used software

9 programs (Cottin et al. 2019; Schubert et al. 2020; Molinaro et al. 2021), ELAI (Guan 2014)

10 has proven to be one of the most efficient ancestry inference tools for dealing with unphased

11 data. This two-layer HMM and LD-based method uses source populations to predict the

12 classification in two hidden state layers—haplotypes in the lower-layer and ancestries in the

13 upper-layer. Therefore it does not require haplotype phasing, but only prior assumptions

14 regarding the number of haplotype clusters and of admixture generations, which are needed for

15 hidden state modeling. ELAI has been applied in studies of different plants. For instance: in

16 perennial plants such as aspen, local ancestry signals obtained using ELAI have generated

17 insight into the local adaptation and demographic history of European varieties (Rendón-Anaya

18 et al. 2021),. In annual crops such as wheat, ELAI has also been applied to analyze gene flow

19 from wild emmers to bread wheat (Zhou et al. 2020).

20

21 LAI tools perform more accurately when using source populations with a large sample size and

22 high differentiation level (Cottin et al. 2019), whereas small, unbalanced structure or admixed

23 source populations might cause erroneous ancestry assignment (Shringarpure & Xing 2014;

24 Molinaro et al. 2021). In practice, it is often challenging to have a perfect sampling design

1   across source populations (Hübner & Kantar 2021). Here we propose a solution to this problem

2   of unbalanced and admixed individuals from source populations.

3

4   *C. canephora* Pierre ex A.Froehner, or so-called Robusta coffee, is an allogamous diploid

5   species with high genetic and phenotypic differentiation (Berthaud 1986; Montagnon et al.

6   1992; Gomez et al. 2009; Cubry et al. 2013; Mérot-L'Anthoëne et al. 2019; Kiwuka et al. 2021;

7   de Aquino et al. 2022). This species originates from central and western Africa (Davis et al.

8   2006), corresponding to two major genetic groups, i.e. Congolese and Guinean groups,

9   respectively (Montagnon et al. 1992, 1998a; Cubry et al. 2013). The Congolese group consists

10  of five well-described subgroups: group A in Benin and Gabon, group E in the Democratic

11  Republic of the Congo (DRC), group C in Cameroon and the western Central Africa region,

12  group B in eastern Central African Republic (CAR), and group O in Uganda; and two recently

13  described groups, G in Angola and R in southern DRC (Mérot-L'Anthoëne et al. 2019). The

14  Guinean group corresponds to group D (Mérot-L'Anthoëne et al. 2019).

15

16  While most of crop species were domesticated during the Neolithic period, over the past 12,000

17  years (Harlan 1971; Larson et al. 2014), Robusta coffee cultivation and diffusion is much more

18  recent. Robusta coffee has attracted interest as a potential cash crop since the late 19th century

19  (Berthaud 1986), and the species has only become globally widespread since 1900 (Montagnon

20  et al. 1998b). Coffee research stations and breeding centers have been set up since the early

21  20th century, in Java, Indonesia (1900–1930), and then DRC (1930–1960), and Central Africa

22  (1960 onwards) (Cramer 1957; Montagnon et al. 1998b). *C. canephora* breeding programs for

23  varietal improvement have mainly been based on heterosis of crosses between the Congolese

1    subgroup E and subgroup A or the Guinean group D (Leroy et al. 1993; Montagnon et al.

2    1998a; Oliveira et al. 2018).

3

4    Due to the gametophytic self-incompatibility of Robusta and its intense breeding history,

5    cultivated populations might have experienced a high level of admixture, resulting in hybrids

6    with complex mosaic genomes. The local ancestry deconvolution approach to cultivated

7    Robusta varieties could help gain insight into their genetic makeup and trace back their origins.

8    This novel approach would facilitate the development of admixture mapping—i.e. associating

9    the phenotype with ancestry segments in coffee—a technique that is sometimes more powerful

10    than conventional genome-wide association studies (GWAS) (Horimoto et al. 2021). Several

11    key traits for coffee breeding, such as drought tolerance, i.e. a polygenic trait that is also

12    considered to be a feature in the Congolese group A (Marraccini et al. 2012; Vieira et al. 2013).

13

14    Robusta coffee was first introduced to Vietnam in the early 20th century, probably from the

15    Congo via France (Nogent-sur-Marne acclimation gardens) or via Java (coffee breeding centre)

16    (Vanden Abeele et al. 2021). In cultivated coffee, it takes 5-8 years to reach the maximum

17    productive stage (Wrigley 1988), and it has been estimated to be even up to 20 years (Moat et

18    al. 2019, Nab & Maslin 2020). Therefore, given 100 years of cultivation, the number of

19    generations is not expected to be higher than 20. Even though the current Vietnamese Robusta

20    varieties are mostly supposed to have originated from Java, their ancestral genetic groups are

21    still largely unclear. Since materials of the Javanese breeding program mainly came from DRC,

22    Uganda and Gabon (Montagnon et al. 1998b), the accessions historically introduced in

23    Vietnam may have Congolese origins and putatively some extent of intergroup admixture.

24

1   In this study, we implemented the ELAI approach on cultivated *C. canephora* with unbalanced

2   and admixed native reference populations. We inferred ancestral genotypic frequencies for

3   these native populations to build perfect source populations for ELAI. We assessed and

4   validated our new approaches with simulated hybrids. We finally applied an optimal

5   framework to a set of elite accessions cultivated in the Central Highlands of Vietnam to

6   determine their mosaic genome origins.

7

# 8   **Results**

## 9   **Characterization of genetic groups**

10  A total of 11,919,576 high-quality biallelic SNPs were obtained in all of the 55 African and 10

11  Vietnamese individuals. We assessed the structure of African native groups by performing

12  genetic structure analysis using sNMF on a set of 1,191,957 randomly picked SNPs.

13  The African individuals were classified into five groups (Supplementary Figures S1 and S2)

14  that could be linked to geographical origins: a West African group with accessions from

15  Guinea, Ghana and Côte d'Ivoire (group D), a group with accessions from Cameroon (group

16  C), a group with accessions from Gabon and Angola (group AG), an East African group with

17  accessions from Uganda and CAR (group OB), and finally the last group consisted of

18  Democratic Republic of the Congo accessions (group ER). Most pairwise $F_{ST}$ values between

19  the five genetic groups were high and ranged from 0.39 to 0.55, except for the $F_{ST} = 0.22$

20  between ER and OB (Supplementary Table S1). This strong structuring was also confirmed by

21  PCA analysis (Supplementary Figure S3).

22

23

1   **ELAI accuracy assessment**

2   The size of the African reference set was small, with an unequal number of individuals

3   (unbalanced structure), while 15 individuals presented some extent of admixture (> 20%

4   admixture) (Supplementary Figure S2). We built near perfect source populations for the five

5   groups based on ancestry genotype frequencies. All of them had perfect ancestry coefficients

6   (> 97%) relative to their respective groups, as expected (Supplementary Figure S4). The

7   artificial source populations were then used for the assessment of ELAI performance on

8   detecting simulated hybrids (Supplementary Figure S5).

9   Using simulated hybrids, we found that our approach achieved accurate inferences with high

10  correlations ($r^2$) ranging from 0.859 to 0.997 (Figure 1), regardless of the set of parameters

11  used. The lowest squared correlation ($r^2 = 0.859$) corresponded to ELAI runs using all SNPs

12  with c (number of lower clusters) = 5 and mg (number of admixture generations) = 30. All

13  other ELAI runs had $r^2$ values > 0.9. The number of lower clusters and of admixture generations

14  did not have marked impacts on the ELAI accuracy, except when all SNPs were used.

15  Conversely, the use of higher parameter values and SNP numbers increased the ELAI run time

16  and memory usage (Supplementary Table S2).

17

18  Among the four SNP sets tested, the overall accuracy was higher for ELAI runs with the 100K

19  SNP set, while slightly decreasing with the other SNP sets, 10K SNPs, evenly distributed 1

20  SNP/5 kb SNPs, and all SNPs (when mg = 20 and 30), respectively. These results were in line

21  with the fact that ELAI accounts for the background LD when detecting the haplotype structure,

22  so a higher number of SNPs provides more haplotype information and thus greater ancestry

23  assessment accuracy. However, using whole-chromosome SNPs did not improve but instead

9

1    slightly reduced the ELAI accuracy as it might cause background noise or false inference with

2    short segment lengths (< 500 kb).

3

4    Despite the high $r^2$ values, some false dosages were observed in the simulated introgression

5    segments where the ancestry switched on both alleles compared to the flanking sequence. We

6    computed the RMSE between true and estimated dosages in the homozygous introgressed

7    regions, and compared it among a range of introgression sizes (0.05 , 0.5, 1 , 2 , and 5 Mb).

8

9    Compared with the larger introgression tracts, RMSE values were higher for introgression

10   tracts < 1 Mb , except for some runs using 100K SNPs and the whole SNP set (Figure 2 and

11   Supplementary Figure S6).

12   RMSE values of around 0.5 in some cases indicated that it was likely that only one haplotype

13   had been correctly assigned along the introgression tracts. For introgression sizes 500 kb, 1

14   Mb, and especially 2 Mb, ELAI runs using whole-chromosome SNPs had the lowest RMSE

15   values (< 0.03), thereby indicating that the estimation was highly accurate. We observed an

16   RMSE increase in larger introgressions (5 Mb) associated with false inferences of small

17   fragments inside longer introgressed fragments. For 5 Mb admixture tracts, the least erroneous

18   inferences (RMSE ranging from 0.01 to 0.04) were obtained in runs using the even SNPs set

19   with c = 5 or 25, and in all SNP sets with c = 15 (Supplementary Figure S6).

20

21   In summary, the ELAI method was highly accurate in assessing the ancestry deconvolution of

22   the artificial hybrids using the simulated source populations, with a good confidence for

23   admixture tracts of > 1 Mb length. The required parameters (number of lower clusters and of

24   admixture generations) only had a minor effect on the detection, while the SNPs chosen for

1    analysis had more marked impacts on the admixture segment inference accuracy. The

2    validation enabled us to define the parameters for application of the method on cultivated

3    Robusta individuals.

4

## Optimized framework and application for inference of the local ancestry of the tested Vietnamese Robusta cultivated accessions

7    Based on our validation and optimization results using simulated hybrids, we developed an

8    LAI framework that encompassed ELAI to efficiently study the admixture origin in Robusta

9    coffee (Figure 3). For each chromosome, we performed ELAI using two SNP datasets (a set of

10   evenly distributed SNPs, and another of whole-chromosome SNPs), with simulated ancestral

11   groups serving as source populations. The lower cluster number was set at five as this factor

12   did not influence the detection but did reduce the run time and memory usage (Supplementary

13   Table S2). We set the number of admixture generations at 20, which reflected the maximum

14   possible number of generations of the cultivated accessions. Common results between the two

15   datasets (evenly distributed SNPs and whole-chromosome SNPs) were then considered as the

16   final local ancestry inference.

17

18   The results of the two SNP sets were pooled in 3 steps. Firstly, as the theoretical dosage of a

19   given ancestry at an SNP locus is either 0, 0.5 or 1 but the dosage inferred by ELAI can be any

20   value between 0 and 1, we approximated the ELAI-inferred dosage at each SNP with respect

21   to the theoretical values, i.e. the dosage was set at 0 if the inferred dosage was in the [0, 0.1)

22   range, at 0.5 if it was in the (0.4, 0.6) range, at 1 if it was in the (0.9, 1] range, or classified as

23   'undetermined' otherwise. Secondly, the approximated dosages in the two SNP sets were

24   compared locus-wise and also classified as 'undetermined' if they were not equal. Lastly,

11

1 ancestry blocks were determined if contiguous positions had the same dosage and the distance

2 between adjacent positions was not > 1 Mb; and < 1 Mb segments were also classified as

3 'undetermined'.

4

5 This framework was then applied for local ancestry inference of the Vietnamese accessions.

6 A total of 94% to > 100% of the genome could be assigned (Figure S7) for all of the 10 tested

7 Vietnamese accessions. Undetermined regions were due to disagreement between the results

8 of the two datasets, or the uncertainty in the ELAI inferred dosage (ancestry dosages within 0.1

9 - 0.4 or 0.6 - 0.9 ranges, or ancestry tracts < 1 Mb were treated as uncertainties). Some

10 accessions represented the same undetermined region on chromosome 10 (Supplementary

11 Figure S7) because this region was assigned with a dosage of 1 to the ER group by the even

12 SNPs set but with a dosage of 0.5 for both ER and AG groups by the whole-chromosome SNP

13 set.

14

15 Based on these ancestry blocks, the global ancestry inference of the tested individuals could be

16 estimated as follows: for each ancestry, the overall proportion was the sum of all block

17 dosages/genome assembly size (≈ 585 Mb), with the block dosage being the length of an

18 ancestry block x ancestry inference for the block. The global ancestry coefficients detected by

19 our ELAI framework were generally similar to the results estimated by sNMF (Supplementary

20 Figure S2). Nine accessions presented > 99% ancestry in the DRC-native ER group

21 (Supplementary Figure S7). Two of these accessions, i.e. TR5 and TR15, had minor admixture

22 proportions in one haplotype, i.e. 1.6 Mb of group AG on chromosome 4 and 1.2 Mb of group

23 C on chromosome 10, respectively.

24

1  The TR6 accession genome consisted of segments from two ancestral groups, i.e. ER and AG,

2  with 72% and 22% of the genome, respectively (Supplementary Figure S7 and Figure 4). The

3  admixture patterns varied in different parts of genome: chromosomes 3, 7, 8, and 11 were

4  completely heterozygous (excluding the undetermined segments); chromosomes 4, 9, and 10

5  had a single admixture tract at the terminal end of the chromosome; chromosome 1, 2 and 5

6  had admixture segments separately distributed along the chromosome; while, exceptionally,

7  chromosome 6 did not present any admixture signal. A small haplotype fragment of 2.8 Mb

8  length on chromosome 8 was assigned to group OB, which accounted for only 0.2% of the

9  genome. The local ancestry pattern suggested that TR6 resulted from recombination events

10  between individuals of the two main Congolese AG and ER subgroups, backcrossed with the

11  ER group.

12

13  # Discussion

14  ## An optimized framework to infer local ancestry in Robusta coffee

15  In this study, we developed a LAI framework implementing the ELAI method, that performed

16  with high accuracy for ancestry deconvolution of admixed individuals using derived source

17  populations and with a good confidence for admixture tracts of > 1 Mb length.

18

19  In particular, we implemented a simple method to overcome the lack of or bias in native

20  reference individual sampling, which was efficient for ELAI on both simulated and real data

21  inference. Our method was based on ancestral genotype frequencies, which can be directly

22  obtained from unphased genotypic data using sNMF. To our knowledge, this is the first time

23  that such an approach has been proposed for unphased data. Another method combining

13

1     WINC-ChromoPainter with the non-negative least squares approach has been developed to

2     analyze LAI when there are very few reference individuals and no source population

3     simulations (Molinaro et al. 2021). A test of this method on real data showed that WINC was

4     comparable to or could outperform ELAI in certain admixture scenarios when only two

5     individuals per source population were used. However, WINC requires phased data and

6     recombination maps, which are not always available for other datasets, especially our dataset.

7     ELAI was later adapted to use a large number of admixed samples (a cohort set) to compensate

8     for the lack of a pure source population (Zhou et al. 2016). The cohort set was down-weighted

9     to not outweigh other training source sets. This method could be applicable if a large cohort

10     sample size is available. We acknowledge that our simulation method did not preserve LD in

11     the native populations, which might explain the misassignment or uncertainty in the inference

12     of small ancestry tracts in admixed individuals. However, the uncertainty only accounted for a

13     minor portion of the genome (< 10%) and the global ancestry findings of our approach were

14     close to the global ancestry obtained by sNMF.

15

16     We also assessed the number of haplotype clusters and admixture generations required by

17     ELAI to predict the ancestry model. The Robusta coffee breeding program was launched about

18     100 years ago, which could be considered as relatively recent compared to most other crops,

19     and therefore the maximum number of Robusta generations out of Africa and especially in

20     Vietnam was estimated at most 20. We tested the methods with 5, 10, 20 and 30 generations.

21     Guan (2014) found that a higher number of admixture generations improved the inference

22     smoothness. In many studies using ELAI, the number of haplotype clusters is often set at 5-

23     fold the number of source populations. These parameters were shown to have an impact on

24     inference in human genomes (Guan 2014), but we did not clearly observe such an effect in our

14

1     simulations. Even when the number of lower clusters was set at the number of source

2     populations, so each source population was linked to only one haplotype in the higher clusters,

3     we could still quite accurately infer the ancestry source. Our approach to generate ideal source

4     populations will only keep LD linked to population structure. As we observed high genetic

5     differentiation in *C. canephora*, we certainly had high LD linked to the population structure.

6     Lower differentiation between ancestral populations might lead to lower performance of the

7     approach we have proposed here.

8

9     ELAI has been shown to be a highly robust LAI tool (Cottin et al. 2019; Schubert et al. 2020;

10    Molinaro et al. 2021). Indeed, our results obtained using simulated *C. canephora* hybrids also

11    illustrated its high accuracy for detecting > 1 Mb ancestry blocks, even with a small portion

12    (1%) of whole-genome SNPs, i.e. 10K SNPs out of the 1.1M whole-genome SNPs on

13    chromosome 1. Yet in our framework we combined inferences obtained with SNP sets of two

14    densities (whole-chromosome SNPs and evenly distributed SNPs, i.e. 203 and 1.8 SNPs per 10

15    kb, respectively) to enhance the inferred ancestry confidence.

16

## 17   **Robusta origin, diffusion and local ancestry inference**

18    The native African Robusta individuals were classified into five groups that could be linked to

19    geographical origins, and this genetic structure was perfectly congruent with previous study

20    findings (Tournebize et al. 2022). The latest classification of *C. canephora* using 8.5K SNP

21    arrays (Mérot-L'Anthoëne et al. 2019) led to an eight-group classification, but the

22    differentiation between groups O and B, E and R, and A and G was low, so they were clustered

23    in this study and our previous study (Tournebize et al. 2022). Clustering of individuals of

15

1   closely related groups was also due to bias towards one group when the other contained a small

2   number of individuals.

3

4   Using source populations derived from these native groups, we ran our optimized framework

5   method on a sample of 10 elite cultivars from Vietnam. Inference of these test Robusta

6   accessions revealed that all of them originated from the Congolese groups ER and AG. Nine

7   accessions shared a common ancestry of group ER, and one likely came from a hybrid between

8   the two Congolese ER and AG groups, backcrossed with ER. Previous studies on other Robusta

9   accessions in Vietnam also identified their Congolese origin (Garavito et al. 2016; Akpertey et

10  al. 2021). Garavito et al. (2016) used DArTseq SNPs and found six Vietnamese accessions

11  from the Congolese E group (included in our ER group). Akpertey et al. (2021) used KASP

12  SNPs and found 33 Vietnamese accessions distinguished from Côte d'Ivoire and Togo

13  accessions (putatively the Guinean group), but no reference for the Congolese groups was used

14  in that study. These inferences are in line with historical coffee breeding data.

15

16  These accessions, which are recognized elite Robusta accessions in Vietnam, could serve as

17  potential breeding materials for varietal improvement. The TR6 accession genome was found

18  to be composed of two ancestral ER and AG groups, accounting for 72% and 22% of the

19  genome, respectively. The Congolese genetic group E was previously found to present

20  advantageous phenotypic characteristics such as good aroma and low acidity, high leaf rust

21  resistance, but with susceptibility to drought and twig borers (Montagnon et al. 1998a). In

22  contrast to group E, genetic group A has very high twig borer resistance and drought tolerance,

23  but is only moderately resistant to leaf rust, and sometimes exhibits lower cup quality

16

1   (Montagnon et al. 1998a). Hybridization of these two groups might produce accessions with

2   heterosis characteristics combining these advantageous agronomic traits.

3

4   Inference of wild ancestry segments in the cultivated accessions could also enable downstream

5   analyses such as admixture mapping of important traits, or genomic selection for breeding

6   programs. Breeding strategies could now also be tailored for different purposes . Reciprocal

7   recurrent selection between the Congolese group and Guinean group (group D) has been used

8   to improve yield and vegetative vigor (Leroy et al. 1993; Montagnon et al. 1998b; Oliveira et

9   al. 2018), while recurrent selection within hybrid populations was more effective for enhancing

10  disease resistance (Alkimim et al. 2021). Therefore, studies on the genetic origin, especially

11  the local ancestry inference of Robusta materials available in collections, could also boost the

12  efficiency of coffee breeding programs.

13

14  This approach could also be adapted to other species when studying admixed populations with

15  a low number of reference individuals.

16

# 17 **Materials and Methods**

## 18 **Materials**

19  We used two sets of accessions: (1) 55 previously sequenced wild *C. canephora* accessions

20  from Africa (Tournebize 2017; Tournebize et al. 2022); and (2) 10 newly sequenced cultivated

21  *C. canephora* accessions from Vietnam. The wild African samples are representative of the

22  native range of *C. canephora*. The cultivated samples from Vietnam are recognized as elite

23  plants and conserved in the germplasm bank of the Western Highlands Agriculture & Forestry

24  Science Institute (WASI).

17

## Sequencing, mapping and SNP calling

The 10 Vietnamese individuals were sequenced using Illumina Hiseq X Ten PE 150 bp. The 55 samples from Africa were obtained from GenBank and analyzed with the new 10 Vietnamese genomes. Variant calling was performed according to GATK Best Practices recommendations for germline short variant discovery using the TOGGLE framework (Monat et al. 2015). The reads were first mapped against the v1.8 reference genome (ACGC, 2018) using BWA mem 0.7.2 (Li 2013), then sorted using Picard Tools 1.83 (https://broadinstitute.github.io/picard/) and SAMtools 0.1.3 (Danecek et al. 2021). Variants per sample were called in individual GVCFs using GATK HaplotypeCaller 3.6, and consolidated using GATK CombineGVCFs, and then final variant calling was jointly performed in the whole GVCF set using GATK GenotypeGVCFs (Poplin et al. 2018). High-quality biallelic SNPs were obtained by applying the following filtering criteria: remove indels, consider only biallelic SNPs, remove clusters of at least 4 SNPs in 10 bp sliding windows, remove SNPs with QUAL <200, MQ0 >4 & MQ0/DP >0.1, mean depth >100 or <10, and missing data >15%, by using GATK 4.0.0.0, BCFtools 1.9 (Danecek et al. 2021) and VCFtools 0.1.16 (Danecek et al. 2011).

## Characterization of population structure

We characterized the genetic groups present in the native African Robusta coffee populations via genetic structure analysis with sNMF (Frichot et al. 2014), using the R package LEA (Frichot & François 2015). 10% randomly chosen SNPs were used to assess the number of genetic groups. The optimal number of ancestral groups (K) was determined

1 using cross-entropy criteria over 10 iterations with K ranging from 1 to 10. Individuals

2 were assigned to a given cluster at an 80% ancestry threshold.

3 Pairwise $F_{ST}$ between the genetic groups (restricted to individuals with > 80% ancestry)

4 were computed using the R package StAMPP (Pembleton et al. 2013). 100 bootstraps

5 across loci were performed to assess the significance. We also performed a principal

6 component analysis (PCA) in the R package LEA (Frichot & François 2015).

7

## Inference of local ancestry

9 We inferred the local ancestry of cultivated Robusta using ELAI (Guan 2014), which was

10 suitable for our unphased data and did not require any additional biological information. The

11 source populations must be specified when using ELAI. To overcome limitations due to the

12 unbalanced structure of wild populations that include admixed individuals, we tested an

13 alternative method using population structure analysis. This new method was validated by

14 simulation of known hybrids, while the ELAI accuracy was assessed and optimal parameters

15 determined. Based on these results, we built a framework to detect the local ancestry in

16 Vietnamese cultivated Robusta accessions.

17

18 **Source population**

19 Genotypes from ancestral groups of wild accessions must be defined for the purpose of ELAI

20 analysis (Guan 2014). Given the small and unbalanced sample size of the wild populations and

21 some evidence of admixed ancestry between groups in the wild accessions, we generated new

22 synthetic populations exhibiting genotypic frequencies similar to those estimated in the

23 ancestral pools. To this end, we used ancestral genotypic frequencies inferred using the sNMF

24 analysis (Frichot et al. 2014). For each chromosome, we ran sNMF on whole-chromosome

19

1  SNPs of all the African and Vietnamese individuals, with 10 iterations and the optimized K

2  value. The best run over the 10 iterations was chosen on the basis of cross-entropy criteria. We

3  applied the G function (R package LEA) on the result of the best run to retrieve a G matrix

4  containing genotype frequencies inferred in each different ancestral group for all diploid SNPs.

5  This G matrix was then used as a probability matrix for randomly choosing genotypes at each

6  site by groups using the sample function in R (R Core Team 2022).

7  We checked if the generated genotypes were representative of the ancestral groups by

8  performing a further sNMF analysis jointly on simulated and real accessions using random

9  100K SNPs on chromosome 1 (with optimal K and 10 iterations). The ancestral proportion of

10  each simulated individual was obtained from the run with the lowest cross entropy.

11

12  **Simulated hybrids**

13  We simulated hybrids with known admixture levels in order to determine optimal parameters

14  and assess the accuracy of the ELAI approach with the simulated source populations in our *C.*

15  *canephora* model. We chose two accessions from the wild African set (BGQ07 and 20738)

16  representative of two divergent genetic groups to simulate three different hybrids with different

17  admixed segment sizes. Each hybrid had admixture segments of different lengths (50 kb, 500

18  kb, 1 Mb, 2 Mb and 5 Mb), and in homozygous (both alleles originating from one of the

19  ancestral groups) and/or heterozygous form (one allele from each of the two ancestries). In the

20  homozygous regions, the hybrid genotypes were copied from one respective progenitor, while

21  at heterozygous loci each of two alleles was drawn randomly from alleles of the two parents.

22  Simulations were based on chromosome 1 SNPs.

23

24

1 **Sets of ELAI parameters**

2 We determined the optimal ELAI parameters by running the software with varying parameter

3 values, including: the number of haplotype clusters (c), i.e. lower clusters, the number of

4 admixture generations (mg), and the set of SNPs used for the analysis. We tested 3 values (5,

5 15, and 25) for the number of haplotype clusters, and 4 values (5, 10, 20, and 30) for the number

6 of admixture generations. We also used 4 different sets of SNPs: (1) randomly selected 10K

7 SNPs, (2) randomly selected 100K SNPs, (3) about 11K SNPs resulting from randomly

8 selecting one SNP in every non-overlapping 5 kb window (referred to as the 'even SNP set'),

9 and (4) whole-chromosome SNPs, i.e. the 'all SNP set' comprising 1,133,736 SNPs. The

10 average SNP densities in the 4 datasets were 1.5, 15.1, 1.8 and 203.0 SNPs per 10 kb,

11 respectively.

12 In summary, we performed 48 ELAI runs with different combinations of parameters, and each

13 run used 20 expectation maximization (EM) steps (Guan 2014). To reduce the computational

14 cost, for the run with whole-chromosome SNPs, the analysis was performed by splitting the

15 chromosome in consecutive SNP chunks so that each subset contained a maximum of 100K

16 SNPs. The ELAI results obtained on the SNP subsets were then concatenated.

17

18 **Assessment of the ELAI inference accuracy based on simulated hybrids**

19 Each ancestry group was defined for the simulated hybrids as: ancestry dosage = 1, where the

20 two alleles were copied from the first parent (first genetic group); ancestry dosage = 0, where

21 both alleles were from the second parent (second genetic group); and ancestry dosage = 0.5,

22 where the alleles were derived from both parents (50% admixture). As we knew the true allelic

23 dosages of the simulated hybrids, we could compare them to those inferred via ELAI.

1 The ELAI performance was assessed by correlation and root mean square error (RMSE)

2 metrics. Correlation was the average Pearson's correlation between the estimated and true

3 dosages. The RMSE between the estimated and true dosages of each admixture tract was

4 calculated and averaged for different segment lengths (50 and 500 kb, and 1, 2 and 5 Mb).

5

6 **Framework to infer local ancestry in Vietnamese cultivated Robusta coffee**

7 Based on the ELAI validation data, a workflow (Figure 3) was developed to detect local

8 ancestry in cultivated Robusta accessions and applied to the 10 Vietnamese Robusta

9 accessions. ELAI was performed for each chromosome, with the simulated ancestral

10 populations as the ancestry source using the whole-chromosome SNP set and the evenly

11 distributed SNP set. Three independent ELAI runs with 20 EM steps were conducted for each

12 individual and SNP set to obtain average results. The ELAI inferences of these sets were then

13 combined to obtain the final ELAI.

14

# Supplementary Material

16 Supplementary data are available at Genome Biology and Evolution online.

17

# Acknowledgments

providing HPC resources that contributed to the research results reported in this paper. URL:

https://bioinfo.ird.fr/- http://www.southgreen.fr.

## Author Contributions

V.P., Y.V., P.C., N.G.K., and T.V. designed the approach, V.P., P.M. and V.H.P. selected the Vietnamese coffee materials, P.C. and T.V. performed the mapping and SNP calling, T.V. performed the genetic structure and local ancestry inference analyses, all co-authors interpreted the results, T.V. wrote the first draft of the manuscript, V.P., Y.V., P.M., and P.C. commented and edited the manuscript, while all co-authors approved the manuscript.

## Data Availability

The raw sequencing data of the African accessions were taken from NCBI SRA database under project accession number PRJNA803612 (Tournebize et al. 2022). The raw sequencing data of the Vietnamese accessions are available in NCBI SRA database under project accession number PRJNA950219 (present study). The software used for this study was downloaded from: https://github.com/haplotype/elai, https://github.com/bcm-uga/LEA, and other packages were available via IRD i-Trop HPC (South Green Platform).

## References

Akpertey A, Padi FK, Meinhardt L, Zhang D. 2021. Effectiveness of single nucleotide polymorphism markers in genotyping germplasm collections of *Coffea canephora* using KASP assay. Front. Plant Sci. 11:612593. doi: 10.3389/fpls.2020.612593.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in

23

1  unrelated individuals. Genome Res. 19:1655–1664. doi: 10.1101/gr.094052.109.

2  Alkimim ER et al. 2021. Designing the best breeding strategy for *Coffea canephora*: Genetic

3  evaluation of pure and hybrid individuals aiming to select for productivity and disease

4  resistance traits. PLoS ONE. 16:e0260997. doi: 10.1371/journal.pone.0260997.

5  Baran Y et al. 2012. Fast and accurate inference of local ancestry in Latino populations.

6  Bioinform. 28:1359–1367. doi: 10.1093/bioinformatics/bts144.

7  Berthaud J. 1986. *Les ressources genetiques pour l'amelioration des cafeires africains*

8  *diploides*. ORSTOM: Paris.

9  Brisbin A et al. 2012. PCAdmix: Principal components-based assignment of ancestry along

10  each chromosome in individuals with admixed ancestry from two or more populations.

11  Hum. Biol. 84:343–364. doi: 10.3378/027.084.0401.

12  Cottin A, Penaud B, Glaszmann J-C, Yahiaoui N, Gautier M. 2019. Simulation-based

13  evaluation of three methods for local ancestry deconvolution of non-model crop species

14  genomes. G3 (Bethesda). 10:569–579. doi: 10.1534/g3.119.400873.

15  Cramer PJS. 1957. *A review of literature of coffee research in Indonesia*. SIC Editorial, Inter

16  American Institute of Agricultural Sciences: Turrialba, Costa Rica.

17  Cubry P, De Bellis F, Pot D, Musoli P, Leroy T. 2013. Global analysis of *Coffea canephora*

18  Pierre ex Froehner (Rubiaceae) from the Guineo-Congolese region reveals impacts from

19  climatic refuges and migration effects. Genet. Resour. Crop. Evol. 60:483–501. doi:

20  10.1007/s10722-012-9851-5.

21  Danecek P et al. 2011. The variant call format and VCFtools. Bioinform. 27:2156–2158. doi:

22  10.1093/bioinformatics/btr330.

23  Danecek P et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience. 10:giab008.

24  doi: 10.1093/gigascience/giab008.

24

1 Davis AP, Govaerts R, Bridson DM, Stoffelen P. 2006. An annotated taxonomic conspectus of

2    the genus *Coffea* (Rubiaceae). Bot. J. Linn. Soc. 152:465–512. doi: 10.1111/j.1095-

3    8339.2006.00584.x.

4 de Aquino SO et al. 2022. Adaptive potential of *Coffea canephora* from Uganda in response to

5    climate change. Mol. Ecol. 31:1800–1819. doi: 10.1111/mec.16360.

6 Frichot E, François O. 2015. LEA: An R package for landscape and ecological association

7    studies. Methods Ecol. Evol. 6:925–929. doi: 10.1111/2041-210X.12382.

8 Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and efficient estimation

9    of individual ancestry coefficients. Genet. 196:973–983. doi:

10    10.1534/genetics.113.160572.

11 Garavito A, Montagnon C, Guyot R, Bertrand B. 2016. Identification by the DArTseq method

12    of the genetic origin of the *Coffea canephora* cultivated in Vietnam and Mexico. BMC

13    Plant Biol. 16:242. doi: 10.1186/s12870-016-0933-y.

14 Garg S et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. Nat

15    Biotechnol. 39:309–312. doi: 10.1038/s41587-020-0711-0.

16 Geza E et al. 2018. A comprehensive survey of models for dissecting local ancestry

17    deconvolution in human genome. Brief Bioinform. 20:1709–1724. doi:

18    10.1093/bib/bby044.

19 Gomez C et al. 2009. Current genetic differentiation of *Coffea canephora* Pierre ex A. Froehn

20    in the Guineo-Congolian African zone: cumulative impact of ancient climatic changes and

21    recent human activities. BMC Evol Biol. 9:167. doi: 10.1186/1471-2148-9-167.

22 Goetz LH, Uribe-Bruce L, Quarless D, Libiger O, Schork NJ. 2014. Admixture and clinical

23    phenotypic variation. Hum. Hered. 77:73–86. doi: 10.1159/000362233.

24 Guan Y. 2014. Detecting structure of haplotypes and local ancestry. Genet. 196:625–642. doi:

1    10.1534/genetics.113.160697.

2    Harlan JR. 1971. Agricultural origins: centers and noncenters. Sci. 174:468–474. doi:
3        10.1126/science.174.4008.468.

4    Horimoto ARVR, Xue D, Thornton TA, Blue EE. 2021. Admixture mapping reveals the
5        association between native American ancestry at 3q13.11 and reduced risk of Alzheimer's
6        disease in Caribbean Hispanics. Alzheimers Res. Ther. 13:122. doi: 10.1186/s13195-021-
7        00866-9.

8    Hu G et al. 2022. Two divergent haplotypes from a highly heterozygous lychee genome suggest
9        independent domestication events for early and late-maturing cultivars. Nat. Genet. 54:73–
10       83. doi: 10.1038/s41588-021-00971-3.

11   Hübner S, Kantar MB. 2021. Tapping diversity from the wild: from sampling to
12       implementation. Front. Plant. Sci. 12. doi: 10.3389/fpls.2021.626565.

13   Joukhadar R, Daetwyler HD, Bansal UK, Gendall AR, Hayden MJ. 2017. Genetic diversity,
14       population structure and ancestral origin of Australian wheat. Front. Plant. Sci. 8. doi:
15       10.3389/fpls.2017.02115.

16   Khoury CK et al. 2016. Origins of food crops connect countries worldwide. Proc. Royal Soc.
17       B. 283. doi: 10.1098/rspb.2016.0792.

18   Kiwuka C et al. 2021. Genetic diversity of native and cultivated Ugandan Robusta coffee
19       (*Coffea canephora* Pierre ex A. Froehner): Climate influences, breeding potential and
20       diversity conservation. PLoS ONE. 16:e0245965. doi: 10.1371/journal.pone.0245965.

21   Larson G et al. 2014. Current perspectives and the future of domestication studies. Proc. Natl.
22       Acad. Sci. U.S.A. 111:6139–6146. doi: 10.1073/pnas.1323964111.

23   Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using
24       dense haplotype data. PLoS Genet. 8:e1002453. doi: 10.1371/journal.pgen.1002453.

26

1  Leroy T, Montagnon C, Charrier A, Eskes AB. 1993. Reciprocal recurrent selection applied to

2  *Coffea canephora* Pierre. I. Characterization and evaluation of breeding populations and

3  value of intergroup hybrids. Euphytica. 67:113–125. doi: 10.1007/BF00022734.

4  Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

5  doi: 10.48550/arXiv.1303.3997.

6  Mani A. 2017. Local ancestry association, admixture mapping, and ongoing challenges. Circ.

7  Cardiovasc. Genet. 10:e001747. doi: 10.1161/CIRCGENETICS.117.001747.

8  Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: A discriminative modeling

9  approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93:278–288.

10  doi: 10.1016/j.ajhg.2013.06.020.

11  Marraccini P et al. 2012. Differentially expressed genes and proteins upon drought acclimation

12  in tolerant and sensitive genotypes of *Coffea canephora*. J. Exp. Bot. 63:4191–212. doi:

13  10.1093/jxb/ers103.

14  Mérot-L'Anthoëne V et al. 2019. Development and evaluation of a genome-wide Coffee 8.5K

15  SNP array and its application for high-density genetic mapping and for investigating the

16  origin of *Coffea arabica* L. Plant Biotechnol. J. 17:1418–1430. doi: 10.1111/pbi.13066.

17  Meyer RS, DuVal AE, Jensen HR. 2012. Patterns and processes in crop domestication: an

18  historical review and quantitative analysis of 203 global food crops. New Phytol. 196:29–

19  48. doi: 10.1111/j.1469-8137.2012.04253.x.

20  Moat J, Gole TW, Davis AP. 2019. Least concern to endangered: Applying climate change

21  projections profoundly influences the extinction risk assessment for wild Arabica coffee.

22  Glob. Chang. Biol. 25:390–403. doi: 10.1111/gcb.14341.

23  Molinaro L et al. 2021. A chromosome-painting-based pipeline to infer local ancestry under

24  limited source availability. Genome Biol. Evol. 13:evab025. doi: 10.1093/gbe/evab025.

27

1  Monat C et al. 2015. TOGGLE: toolbox for generic NGS analyses. BMC Bioinform. 16:374.

2     doi: 10.1186/s12859-015-0795-6.

3  Montagnon C, Leroy T, Eskes A. 1998a. Amélioration variétale de *Coffea canephora*. 1 :

4     critères et méthodes de sélection. Plantations, Recherche, Développement. 89–98.

5  Montagnon C, Leroy T, Eskes A. 1998b. Amélioration variétale de *Coffea canephora*. 2 : les

6     programmes de sélection et leurs résultats. Plantations, Recherche, Développement.

7     http://agritrop.cirad.fr/390311/ (Accessed June 19, 2019).

8  Montagnon C, Leroy T, Yapo A. 1992. Diversité génotypique et phénotypique de quelques

9     groupes de caféiers (*Coffea canephora* Pierre) en collection. Conséquences sur leur

10    utilisation en sélection. Café, Cacao, Thé. 36:187–198.

11 Musoli P et al. 2009. Genetic differentiation of wild and cultivated populations: diversity of

12    *Coffea canephora* Pierre in Uganda. Genome. 52:634–646. doi: 10.1139/G09-037.

13 Nab C, Maslin M. 2020. Life cycle assessment synthesis of the carbon footprint of Arabica

14    coffee: Case study of Brazil and Vietnam conventional and sustainable coffee production

15    and export to the United Kingdom. Geo: Geogr. Environ. 7:e00096. doi: 10.1002/geo2.96.

16 Nave M et al. 2019. Wheat domestication in light of haplotype analyses of the Brittle rachis 1

17    genes    (BTR1-A    and    BTR1-B).    Plant    Sci.    285:193–199.    doi:

18    10.1016/j.plantsci.2019.05.012.

19 Noirot M et al. 2003. Genome size variations in diploid african *Coffea* species. Ann. Bot.

20    92:709–714. doi: 10.1093/aob/mcg183.

21 Norris ET et al. 2020. Admixture-enabled selection for rapid adaptive evolution in the

22    Americas. Genome Biol. 21:29. doi: 10.1186/s13059-020-1946-2.

23 Oliveira HR, Jacocks L, Czajkowska BI, Kennedy SL, Brown TA. 2020. Multiregional origins

24    of   the   domesticated   tetraploid   wheats.   PLoS   ONE.   15:e0227148.   doi:

10.1371/journal.pone.0227148.

Oliveira LNL et al. 2018. Selection of *Coffea canephora* parents from the botanical varieties Conilon and Robusta for the production of intervarietal hybrids. Cienc. Rural. 48. doi: 10.1590/0103-8478CR20170444.

Padhukasahasram B. 2014. Inferring ancestry from population genomic data and its applications. Front. Genet. 5. https://www.frontiersin.org/article/10.3389/fgene.2014.00204 (Accessed March 14, 2022).

Pembleton LW, Cogan NOI, Forster JW. 2013. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Mol. Ecol. Resour. 13:946–952. doi: 10.1111/1755-0998.12129.

Poplin R et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of sample. doi: 10.1101/201178.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959. doi: 10.1093/genetics/155.2.945.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Rendón-Anaya M et al. 2021. Adaptive introgression facilitates adaptation to high latitudes in European aspen (*Populus tremula* L.). Mol. Biol. Evol. 38:5034–5050. doi: 10.1093/molbev/msab229.

Rius M, Darling JA. 2014. How important is intraspecific genetic admixture to the success of colonising populations? Trends Ecol. Evol. 29:233–242. doi: 10.1016/j.tree.2014.02.003.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating local ancestry in admixed populations. Am. J. Hum. Genet. 82:290–303. doi: 10.1016/j.ajhg.2007.09.022.

29

1    Schubert R, Andaleon A, Wheeler HE. 2020. Comparing local ancestry inference models in

2        populations of two- and three-way admixture. PeerJ. 8:e10090. doi: 10.7717/peerj.10090.

3    Shriner D. 2017. Overview of admixture mapping. In: Current Protocols in Human Genetics.

4        94(1):1–23. http://onlinelibrary.wiley.com/doi/abs/10.1002/cphg.44 (Accessed February

5        18, 2022).

6    Shringarpure S, Xing EP. 2014. Effects of sample selection bias on the accuracy of population

7        structure and ancestry inference. G3 (Bethesda). 4:901–911. doi: 10.1534/g3.113.007633.

8    Tang H, Coram M, Wang P, Zhu X, Risch N. 2006. Reconstructing genetic ancestry blocks in

9        admixed individuals. Am. J. Hum. Genet. 79:1–12. doi: 10.1086/504302.

10    Thornton TA, Bermejo JL. 2014. Local and global ancestry inference, and applications to

11        genetic association analysis for admixed populations. Genet. Epidemiol. 38:S5–S12. doi:

12        10.1002/gepi.21819.

13    Tournebize R et al. 2022. Ecological and genomic vulnerability to climate change across native

14        populations of Robusta coffee (*Coffea canephora*). Glob. Change Biol. 28:4124–4142.

15        doi: 10.1111/gcb.16191.

16    Tournebize R. 2017. Influence des variations spatio-temporelles de l'environnement sur la

17        distribution actuelle de la diversité génétique des populations. University of Montpellier:

18        Montpellier, France.

19    Vanden Abeele S et al. 2021. Genetic diversity of wild and cultivated *Coffea canephora* in

20        northeastern DR Congo and the implications for conservation. Am. J. Bot. 108:2425–2434.

21        doi: 10.1002/ajb2.1769.

22    Vieira NG et al. 2013. Different molecular mechanisms account for drought tolerance in *Coffea*

23        *canephora* var. Conilon. Trop. Plant Biol. 6:181–190. doi: 10.1007/s12042-013-9126-0.

24    Wrigley G. 1988. Coffee–tropical agriculture series. Harlow, UK: Longman Scientific &

1     Technical.

2 Wu J, Liu Y, Zhao Y. 2021. Systematic review on local ancestor inference from a mathematical

3     and algorithmic perspective. Front. Genet. 12:639877.

4 Yang JJ, Li J, Buu A, Williams LK. 2013. Efficient inference of local ancestry. Bioinform.

5     29:2750–2756. doi: 10.1093/bioinformatics/btt488.

6 Zhao L, Wang Kun, Wang Kai, Zhu J, Hu Z. 2020. Nutrient components, health benefits, and

7     safety of litchi (*Litchi chinensis* Sonn.): A review. Compr. Rev. Food. Sci. Food Saf.

8     19:2139–2163. doi: 10.1111/1541-4337.12590.

9 Zhou Q, Zhao L, Guan Y. 2016. Strong Selection at MHC in Mexicans since Admixture. PLoS

10     Genet. 12:e1005847. doi: 10.1371/journal.pgen.1005847.

11 Zhou Y et al. 2020. *Triticum* population sequencing provides insights into wheat adaptation.

12     Nat. Genet. 52:1412–1422. doi: 10.1038/s41588-020-00722-w.

13

## 14 Figure legends

15 **Fig. 1: ELAI accuracy in inferring local ancestry in simulated hybrids.** The plot shows

16 correlations between the true local ancestry dosage and ELAI dosages with different parameter

17 numbers: number of lower clusters (c = 5, 15, and 25), number of admixture generations (mg

18 = 5, 10, 20, and 30), and different SNP sets (10K SNPs, 100K SNPs, evenly distributed SNPs

19 – 1 SNP/5 kb, and all SNPs - 1M SNPs, represented in yellow, orange, violet and black points,

20 respectively). Each point represents an average of correlations computed for the three simulated

21 hybrids, and error bars represent the standard deviation.

22

23 **Fig. 2: Error in detecting simulated introgression tracts of different lengths in simulated**

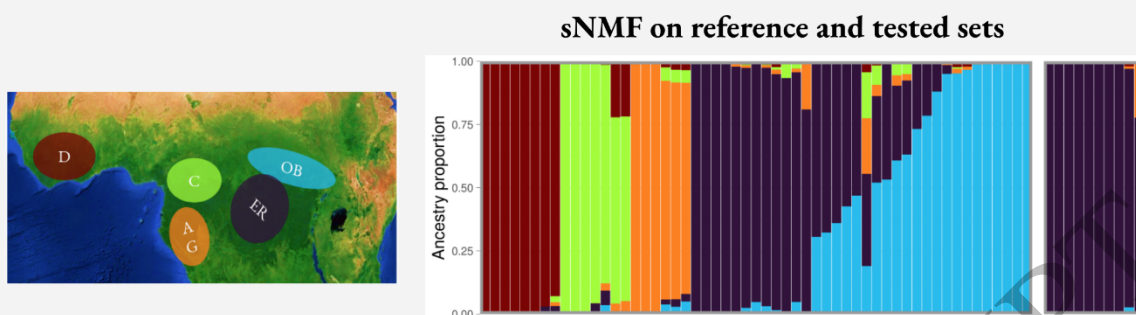24 **hybrids.** RMSE was calculated between the true dosage and ELAI dosages of the simulated

31

1 hybrids, for different homozygous introgressed segment lengths (0.05, 0.5, 1, 2, and 5 Mb).

2 Each panel shows, for each tested length, the RMSE values (y-axis) for ELAI runs with the

3 numbers of lower clusters (c = 5), different numbers of generations (mg = 5, 10, 20 and 30) on

4 the x-axis, and 4 different SNP sets (10K SNPs, 100K SNPs, evenly distributed SNPs – 1

5 SNP/5 kb, and all SNPs - 1M SNPs, represented in yellow, orange, violet and black points,

6 respectively).

7

8 **Fig. 3: Framework for local ancestry inference of cultivated *C. canephora*.** ELAI was

9 performed for each individual chromosome and involved three main steps. Step 1: analyzing

10 the genetic structure of the ancestral group, by performing sNMF on the reference set and tested

11 set. Step 2: simulating source populations based on sNMF-estimated  ancestral genotypic

12 frequencies. Step 3: running ELAI on the tested individuals using two marker sets, i.e. whole-

13 chromosome SNPs and evenly distributed SNPs. Step 4: merging the ancestry dosages inferred

14 in the two SNP sets to determine the final consensus inference of the target chromosome.

15

16 **Fig. 4: Local ancestry inference of the Vietnamese TR6 accession.** Each bar presents the

17 consensus local ancestry dosage (y-axis) along the positions (x-axis) of each chromosome. The

18 x-axis labels are in the Mb unit. The inferred ancestral groups - ER (from Democratic Republic

19 of the Congo), AG (from Benin, Gabon, and Angola), and OB (from Uganda and Central

20 Africa), are denoted by dark purple, orange and blue, respectively. The gray portions are
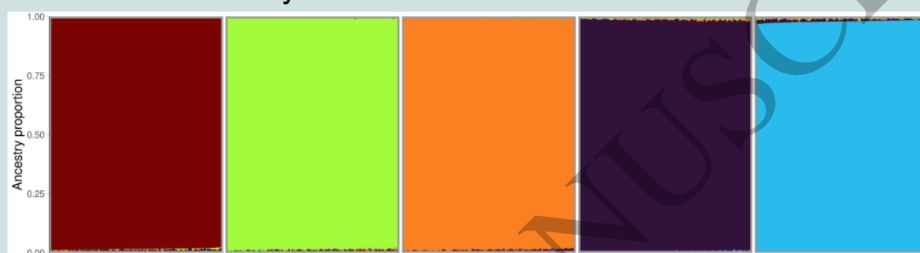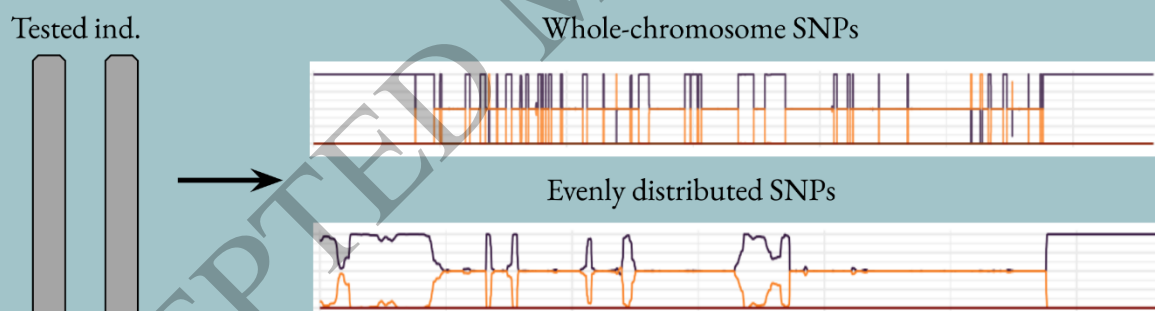
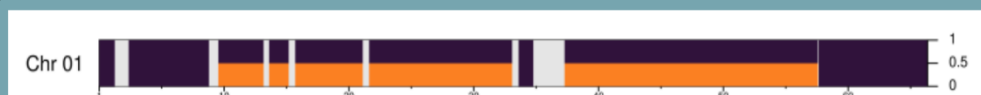21 undetermined regions.

*Figure 1*
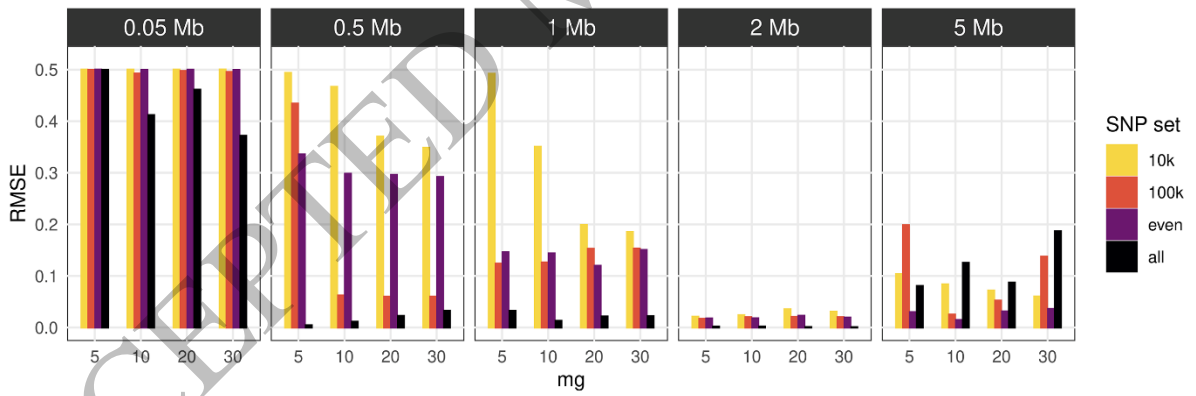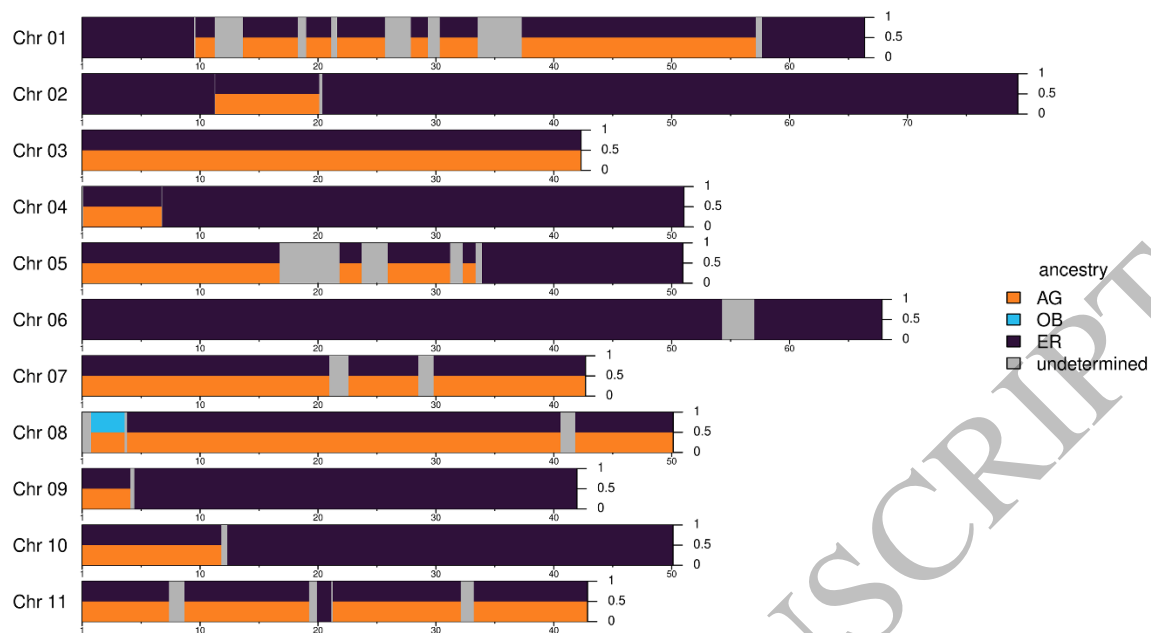*159x208 mm ( x  DPI)*
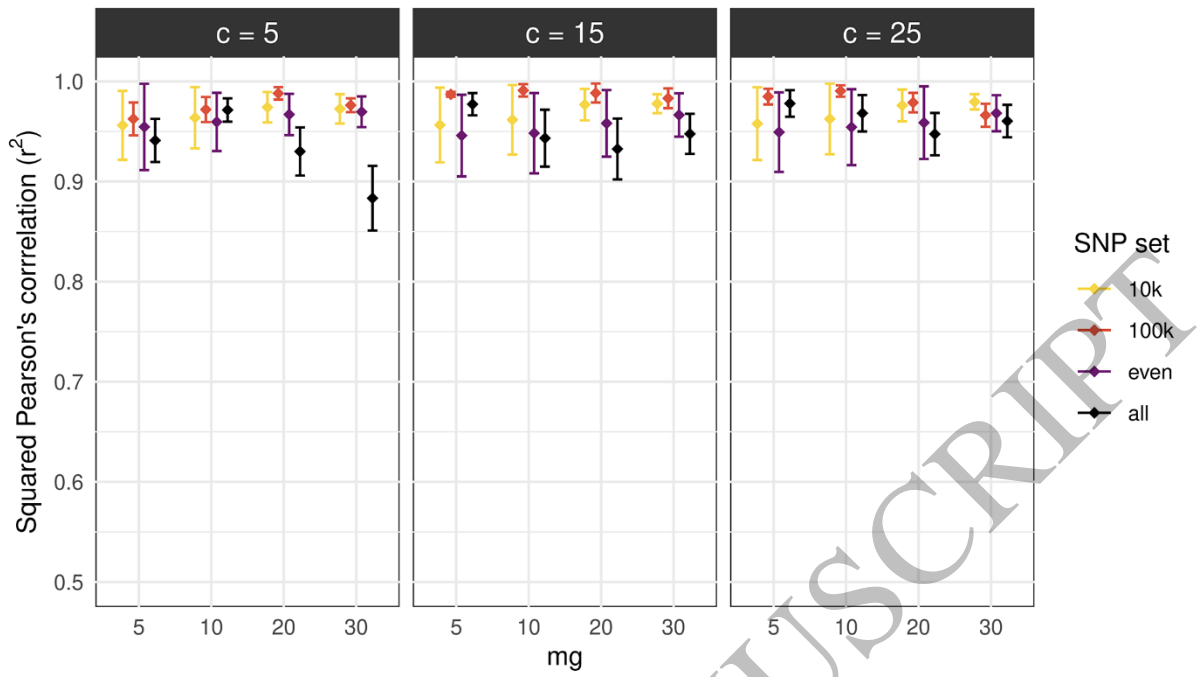
*Figure 2*
*159x91 mm ( x  DPI)*

1

2



*Figure 3*
*159x53 mm ( x  DPI)*

3

*Figure 4*
*159x87 mm ( x  DPI)*