THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Génétique et Génomique

École doctorale GAIA

Unité de recherche DIADE

Tracing *Coffea canephora* genetic diversity from wild African population to cultivated germplasm in Vietnam and its suitability to future climate

Présentée par Tram Bao VI
Le 08 decembre 2023

Sous la direction de Yves VIGOUROUX et Valérie PONCET
et co-encadrée par Philippe CUBRY, Ngan Giang KHONG, et Pierre MARRACCINI

Devant le jury composé de

Roberto PAPA, Professeur, Università Politecnica delle Marche, Ancona	Rapporteur
Karine ALIX, Professeur, AgroParisTech, Palaiseau	Rapportrice
Mathieu GAUTIER, Chargé de Recherche, INRAE, Montpellier	Examinateur
Judith BURSTIN, Directrice de Recherche, INRAE, Dijon	Examinatrice

UNIVERSITÉ DE MONTPELLIER

Acknowledgments

The three-year journey through my Ph.D. has been a paradox—long enough to test the limits of mental endurance, yet a little too short to fully explore the intricacies of an engaging project. Throughout this challenging expedition, I have had the privilege of interacting with outstanding individuals and institutions who have become integral to the narrative of my academic adventure.

Dr. Valerie Poncet, your determination, close guidance and continuous encouragement were the driving forces behind the commencement, completion, innovation and success of my Ph.D. Beyond the formalities of work, your supportive presence resembled that of a "friend," providing solace and comradeship throughout the journey.

Dr. Yves Vigrouroux, your strict guidance and constructive criticism were the compass that kept my project on a precise course, enhancing not only my scientific acumen but also my transferable skills. Your unwavering commitment significantly contributed to the success of the work.

To my co-supervisors, Dr. Philippe Cubry, Dr. Khong Ngan Giang, and Dr. Pierre Marraccini, your diverse scientific advice covering genetics, physiology, and agronomy was instrumental in the development of my project. The collaboration of these different aspects created a well-rounded and enriched scholarly experience.

The members of the Ph.D. committees and defense jury, your invaluable scientific feedback played a pivotal role in refining and improving the quality of my work.

Colleagues at IRD, I extend my gratitude for your collaboration in the project and the generous assistance with scientific and technical aspects. Special thanks to Mrs. Julie Orjuela for the collaborative efforts in bioinformatics that significantly accelerated my progress. Dr. Christine Tranchant-Dubreuil and Mr. Ndomassi Tando, your technical support in bioinformatics and clusters was indispensable. Dr. Francois Sabot, your motivational guidance inspired a crucial transition from a biomedicine major to bioinformatics, and Dr. Sebastien Cunnac, a year working with you has made a significant improvement of my bioinformatic skill, preparing me well for the PhD work.

To my fellow Ph.D. students and postdocs—Francis, Marine, Laura, Serafin, and all the others—you were more than just colleagues; you became friends who shared both the professional and personal aspects of this challenging journey. Your camaraderie, shared experiences, and encouragement were a source of strength, especially during the most challenging moments at the conclusion of my Ph.D.

Collaborators at AGI and WASI, your contributions were essential, and I express my gratitude to Ms. Le Thi Nhu, Dr. Phan Viet Ha, and Mrs. Dinh Thi Tieu Oanh for their significant roles in preparing coffee samples and providing valuable insights into the coffee in Vietnam.

To the funding supporters, including French Embassy in Vietnam, Ministry of Science and Technology in Vietnam, and IRD, your financial support laid the groundwork for this research endeavor. Special thanks to Mrs. Au Co Vu for efficiently managing my administrative procedures for grants in France and to Mrs. Gaelle Brule for overseeing my documents for grants in Vietnam.

Friends in France and Vietnam—Nguyen, Linh, Thao, Nga, and others—I appreciate your support in keeping me connected to the modern world during the Ph.D. journey. Your shared perspectives on life and emotional support were invaluable.

Finally, **To my family**, your unwavering support in accommodation, nutrition, and the roles you played as my

"co-workers" during my near 100% home office experience in Vietnam are deeply cherished. Your unconditional love provided the foundation upon which this academic endeavor stood.

List of publications

Articles related to the PhD thesis

— **Vi, T.**, Vigouroux, Y., Cubry, P., Marraccini, P., Phan, H. V., Khong, G. N., & Poncet, V. (2023). Genome-wide admixture mapping identifies wild ancestry-of-origin segments in cultivated Robusta coffee. Genome Biology and Evolution, 15(5), evad065.

— **Vi, T.**, Marraccini, P., De Kochko, A., Cubry, P., Khong, N. G., & Poncet, V. (2022). Sequencing-based molecular markers for wild and cultivated coffee diversity exploration and crop improvement.

Articles not related to the PhD thesis

— Orjuela, J., Comte, A., Ravel, S., Charriat, F., **Vi, T.**, Sabot, F., & Cunnac, S. (2022). CulebrONT : a streamlined long reads multi-assembler pipeline for prokaryotic and eukaryotic genomes. Peer Community Journal, 2.

— Ho, T. T., Pham, V. T., Nguyen, T. T., Trinh, V. T., **Vi, T.**, Lin, H. H., ... & Pham, M. D. (2021). Effects of Size and Surface Properties of Nanodiamonds on the Immunogenicity of Plant-Based H5 Protein of A/H5N1 Virus in Mice. Nanomaterials, 11(6), 1597.

Conferences (Posters)

— **Vi, T.**, Cubry P., Marraccini P., Dinh T. T. O., Phan V. H., Zhang D., Stoffelen P., Vigouroux Y., Poncet V., & Khong N. G. Which genetic diversity was brought to Vietnamese Robusta coffee (Coffea canephora) ? ASIC International Conference on Coffee Science. Hanoi, Vietnam. September 2023.

— Stoffelen, P., Léonard, G., Mwanga, I. M., ..., **Vi, T.**, Angbonda, D.-M. A., Vandelook. F. The Democratic Republic of the Congo, the cradle of cultivated Robusta coffee (Coffea canephora) : can we safeguard these coffee genetic resources of world importance ? ASIC International Conference on Coffee Science. Hanoi, Vietnam. September 2023.

— Le, T.N., Marraccini, P., Nguyen, V.T., Dinh, T. T. O., ..., **Vi, T.**, Poncet, V., Khong, N. G. Genetic evaluation of Coffea canephorain Vietnam,for pre-breeding and generation of new hybrids tolerant to drought. ASIC International Conference on Coffee Science. Hanoi, Vietnam. September 2023.

— Millet, P., Allinne, C., **Vi, T.**, Marraccini, P., ..., & Poncet, V. Unexpected coffee varietal diversity in haitian coffee agroforestry systems. ASIC International Conference on Coffee Science. Hanoi, Vietnam. September 2023.

— **Vi, T.**, Vigouroux, Y., Cubry, P., Marraccini, P., Phan, H. V., Khong, G. N., & Poncet, V. Genome-Wide Admixture Mapping Using Admixed Source Populations In Vietnamese Cultivated Robusta Coffee. SMBE Society for Molecular Biology and Evolution. Ferrara, Italy. July 2023.

— **Vi, T.**, Cubry, P., Marraccini, P., Dinh, T. T. O., Phan, V. H., Khong, N. G., & Poncet, V. Genomic characterization of 10 Vietnamese elite clones of Robusta (Coffea canephora). ASIC International Conference on Coffee Science. Montpellier, France. September 2021.

Résumé

Contexte

Domestication et adaptation des cultures

La domestication est un processus évolutif piloté par l'homme qui a commencé il y a environ 12 000 ans (Meyer et al., 2012 ; Zohary et Hopf, 2000). Le processus de domestication est associé à un échantillonnage génétique de la diversité sauvage, résultant souvent en une diversité plus faible dans les variétés cultivées (Allaby et al., 2019 ; R. Kumar et al., 2021 ; Meyer et Purugganan, 2013). On parle généralement de goulot d'étranglement associé à la domestication pour caractériser cette perte de diversité (Allaby et al., 2019 ; R. Kumar et al., 2021 ; Meyer et Purugganan, 2013). La perte de diversité génétique ne s'est pas limitée à la domestication, le développement de l'agriculture dans les 100 dernières années a promu la culture sur de grandes surfaces de cultivars dits élites (Gross et al., 2014) issus de la sélection moderne et réduisant encore la base génétique cultivée. En raison de ces pertes de diversité génétique, la question de la capacité d'adaptation de nos cultivars élites à un environnement changeant rapidement (température plus élevée, précipitations en hausse ou en baisse, augmentation des aléas climatiques) se pose.

Au cours des 50 dernières années, d'importants changements climatiques ont été observés. Pour adapter l'agriculture à ces changements environnementaux, soient les cultures migrent vers de nouveaux sites favorables , soient elles doivent s'adapter aux nouvelles conditions dans le site actuel (Aitken et al., 2008). Les espèces sauvages apparentées aux cultures, les variétés domestiquées ou cultivées et les variétés élites sont des ressources importantes qui peuvent être mobilisées pour cette adaptation (Flint-Garcia et al., 2023). Dans ce cadre, une compréhension plus fine des bases génétiques de l'adaptation des plantes et leur caractérisation dans les ressources génétiques peut aider à construire des stratégies de sélection efficaces.

L'espèce *Coffea canephora*

L'espèce *Coffea canephora* est diploïde (2n = 22), avec un génome d'environ 710 Mb (Noirot et al., 2003), et strictement allogame (Berthaud, 1986 ; Lashermes et al., 2000). Son aire de répartition recouvre l'Afrique de l'Ouest et l'Afrique centrale (A. P. Davis et al., 2006). Elle produit le café robusta qui représente environ 40 % de la production mondiale de café, derrière l'arabica, produit par l'espèce *C. arabica* (ICO, 2019) Les premiers programmes de sélection ont été lancés en Indonésie, en République démocratique du Congo et en Côte d'Ivoire au début du 20ème siècle (Montagnon et al., 1998a). Par la suite, le matériel sélectionné a été largement cultivé dans d'autres parties du monde, y compris au Vietnam.

La diversité de *Coffea canephora* a été récemment étudiée et huit groupes génétiques correspondant à différentes régions d'Afrique ont été proposés (Mérot-L'Anthoëne et al., 2019) (figure .0.1). Les groupes génétiques sont également caractérisés par la variabilité de nombreux caractères morphologiques et de résistance biotique et abiotique. *Coffea canephora* est donc un modèle intéressant pour étudier la diversité intraspécifique et l'adaptation locale dans la zone intertropicale. L'étude des relations entre climat, phénotype, et données génomiques chez le caféier *Coffea canephora* serait utile pour comprendre les bases génétiques de son adaptation. Elle peut également permettre de construire des scénarios sur l'impact du changement climatique sur la distribution et la productivité du Robusta à l'avenir.

Le caféier *C. canephora* au Vietnam

C. canephora a été initialement introduit au Vietnam en 1908 (ICO, 2019 ; Phan, 2017). Les premières variétés introduites provenaient d'Indonésie et d'Afrique centrale, mais leur origine génétique reste globalement incertaine. Depuis les années 2000, le Vietnam est devenu le plus important producteur mondial de café robusta. Le caféier *C. canephora* joue par conséquent un rôle majeur dans l'agriculture, l'écologie et l'économie sociale du pays.

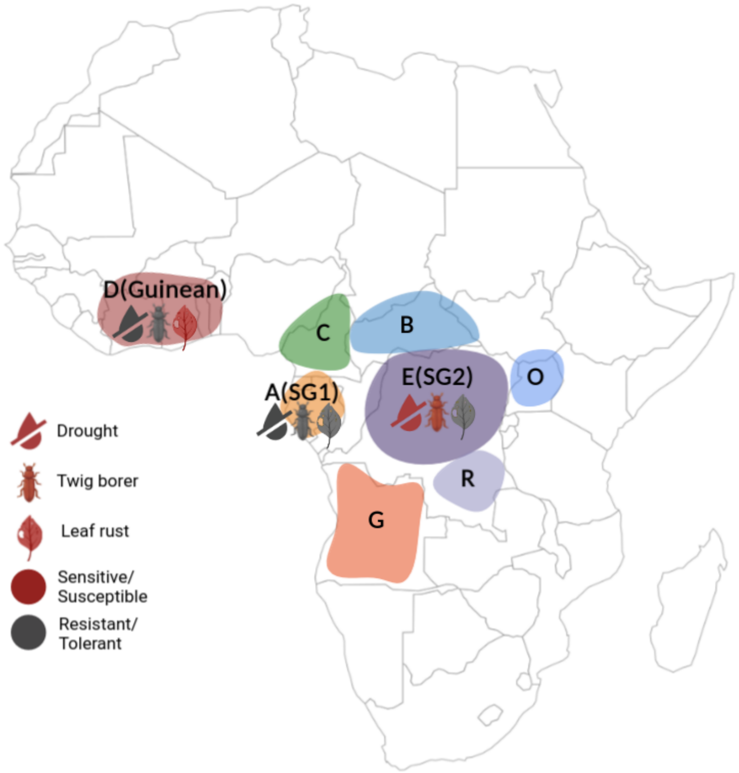


Figure .0.1: Geographical representation of the genetic groups of wild C. canephora in Africa and their characteristics. The group distribution is according to Mérot-L'Anthoëne et al. (2019), the biotic and abiotic stress behaviour is according to Montagnon et al. (1998b). These behaviours has been observed for some accessions, but might be variable within the group.

La production de robusta au Vietnam est concentrée dans les hauts plateaux du centre, situés dans la région centrale du sud. Des travaux de conservation et de recherche sur l'amélioration de *C. canephora* sont menés à l'Institut des sciences agricoles et forestières des hauts plateaux de l'Ouest (WASI), dans la province de Dak Lak. Certaines variétés élites ont été reconnues compatibles entre elles et sont couramment utilisées en mélange dans les plantations. Elles ont une productivité élevée (de 3,5 à 7 tonnes/hectare), une bonne qualité à la tasse et une résistance aux ravageurs et à la rouille des feuilles (ICO, 2019 ; Phan, 2017).

Cependant, la culture de *C. canephora* est confrontée à divers problèmes au Vietnam, principalement dus au changement climatique et au vieillissement des plantations (ICO, 2019). Une précédente étude a prévu que la production de café allait diminuée de 50 % par rapport à la production actuelle d'ici 2050 en raison des conséquences du changement climatique (ICO, 2019). Pour faire face à ces problèmes, l'amélioration de la tolérance des caféiers au stress abiotiques et biotiques peut être une des mesures durables. En 2020, le ministère de la science et de la technologie a lancé un projet (projet MOST) visant à améliorer la tolérance à la sécheresse et la résistance aux nématodes de *C. canephora* au Vietnam par le biais de programmes de sélection. La diversité génétique est essentielle pour l'amélioration des cultures (Swarup et al., 2021), mais elle est largement méconnue pour les variétés vietnamiennes de *C. canephora*. La compréhension de l'origine des caféiers vietnamien est également l'un des objectifs du projet MOST. Sur la base de l'étude de diversité génétique que nous allons mener, du matériel sera sélectionné et proposé comme base de nouveaux programmes de sélection afin de créer des variétés plus résistantes et durables.

Marqueurs basés sur le séquençage pour l'exploration de la diversité du café

Les marqueurs moléculaires sont indispensables pour de nombreuses études génétiques et génomiques visant à explorer la diversité des espèces. L'avènement du séquençage à haut débit (séquençage de nouvelle génération) à permis une détection efficace de différents marqueurs puissants tels que les polymorphismes nucléotidiques simples (SNP), les insertions et les délétions (Indels), la variation d'insertion des éléments transposables (TE), et plus largement les variants structuraux (SV). Ces marqueurs ont été utilisés dans les recherche sur les caféiers, comme la génomique évolutive, la diversité génétique et les études d'association à l'échelle du génome. Le développement rapide de la biologie moléculaire et de la bioinformatique améliore l'identification des marqueurs moléculaires et étend leur application, ce qui permet d'explorer de nouveaux aspects de la génomique du café et de révolutionner les programmes de sélection à venir.

Objectifs de la recherche

Originaire d'Afrique, le café *C. canephora* est l'une des cultures les plus importantes au Vietnam, propulsant le pays au premier rang mondial de la production de robusta au cours des 20 dernières années. À l'avenir, la production de café robusta vietnamien sera menacée par plusieurs facteurs, notamment le changement climatique. Des stratégies de conservation et des programmes de sélection sont nécessaires de toute urgence pour trouver des solutions durables afin d'améliorer la capacité d'adaptation du café au changement climatique au Vietnam. L'adaptation du café à de nouveaux environnements dépend fortement de la variabilité génétique disponible. La collection du WASI est la principale source disponible de variabilité pour l'amélioration de *C. canephora* au Vietnam. La compréhension de la diversité génétique disponible au Vietnam et de son adaptation au climat local pourrait permettre de prévoir la réaction du caféier *C. canephora* vietnamien au changement climatique à venir. Cette compréhension facilitera les plans futurs de conservation et d'amélioration.

Pour établir les futures stratégies de conservation et de sélection visant à faire face au changement climatique au Vietnam, il est nécessaire de bien comprendre quelle partie de la diversité génétique des populations sauvages a été transmise au café C. canephora vietnamien, et quelle diversité génétique de cette source est adaptée au climat local prévu au Vietnam à l'avenir. L'objectif de cette thèse était de répondre à ces questions et a été divisé en trois parties principales.

La première partie (chapitre II) s'est concentrée sur le développement d'une approche à l'échelle du génome pour identifier les segments d'ascendance d'origine sauvages dans le café *C. canephora* cultivé. La deuxième partie (chapitre III) a été consacrée à une étude sur la diversité génétique et l'origine des accessions de *C. canephora* cultivées dans les hauts plateaux du centre du Vietnam. La dernière partie (chapitre IV) a évalué l'adéquation des accessions sauvages de *C. canephora* aux conditions climatiques locales du Vietnam, à la fois dans le présent et dans le futur. L'ensemble des résultats de ces études contribuera à l'élaboration des stratégies de sélection et de conservation, en suggérant des matériels de sélection potentiels, des groupes de croisement, et en indiquant s'il convient d'introduire une plus grande diversité génétique que celle disponible au sein du germplasm vietnamien. Les principales conclusions, les limites, les perspectives et les implications du projet sont examinées dans le chapitre V.

La cartographie des mélanges à l'échelle du génome identifie des segments d'ascendance d'origine sauvage dans le café *C. canephora* cultivé

Les progrès du séquençage et l'application de technologies statistiques et informatiques ont permis d'améliorer l'inférence de l'ascendance au niveau des chromosomes sur la base d'approches d'inférence de l'ascendance locale (LAI). Ces méthodes fournissent de meilleures informations sur l'origine génétique mosaïque des individus mélangés, améliorant la connaissance des histoires démographiques et facilitant la détection de l'introgression adaptative (GEZA et al., 2019 ; MANI, 2017 ; PADHUKASAHASRAM, 2014 ; SHRINER, 2017 ; THORNTON et BERMEJO, 2014). Les outils LAI sont plus précis lorsqu'ils utilisent des populations sources ayant une taille d'échantillon importante et un niveau de différenciation élevé (COTTIN et al., 2019 ; MOLINARO et al., 2021 ; SHRINGARPURE et XING, 2014). Dans la pratique, il est souvent difficile d'avoir un plan d'échantillonnage parfait parmi les populations sources (HÜBNER et KANTAR, 2021).

Selon des études antérieures (Cottin et al., 2019 ; Molinaro et al., 2021 ; Schubert et al., 2020), ELAI (Y. Guan, 2014) s'est avéré être l'un des outils de LAI les plus efficaces pour traiter des données non phasées. Dans cette étude, nous avons mis en œuvre l'approche ELAI sur des C. canephora cultivés au Vietnam avec des populations de référence issues de la zone d'origine de ce caféier en Afrique déséquilibrées et mélangées. Nous avons déduit des fréquences génotypiques ancestrales pour ces populations natives afin de constituer des populations sources parfaites pour l'analyse ELAI. Nous avons évalué et validé notre approche avec des hybrides simulés. Enfin, nous avons appliqué cette méthode à un ensemble d'accessions élites cultivées dans les hauts plateaux du centre du Vietnam afin de déterminer l'origine de leur génome en mosaïque.

L'ensemble de référence africain (de petite taille) a été classé en cinq groupes, avec une structure déséquilibrée et un niveau élevé de mélange. Nous avons construit des populations sources presque parfaites pour les cinq groupes sur la base des fréquences génotypiques ancestrales. Toutes ces populations simulées avaient des coefficients d'ascendance parfaits (> 97 %) par rapport à leurs groupes respectifs, comme attendu. Les populations sources artificielles ont ensuite été utilisées pour évaluer les performances de l'analyse ELAI en matière de détection des hybrides simulés.

En utilisant des hybrides simulés, nous avons constaté que notre approche permettait des inférences précises, avec des corrélations élevées (r^2 = 0,859 - 0,997) entre les dosages d'ascendance déduits et réels, quel que soit l'ensemble des paramètres utilisés. Toutefois, nous avons constaté un taux d'erreur de détection plus élevé pour les tailles d'introgression inférieures à 1Mb. Notre validation a également montré que les SNPs choisis pour l'analyse avaient un impact marqué sur la précision de l'inférence des segments d'introgression.

Sur la base de nos résultats de validation et d'optimisation utilisant des hybrides simulés, nous avons développé une méthode afin d'étudier efficacement l'origine des mélanges dans le café *C. canephora* (figure .0.2). Dix accessions élites ont été analysées et attribués aux cinq groupes génétiques majeurs africains. Une de ces variétés élites est un hybride avec des schémas de mélange variés dans différentes parties du génome, ce qui suggère un rétro-croisement entre les groupes AG et ER.

L'inférence de segments d'ascendance sauvage dans les accessions cultivées pourrait également permettre des analyses en aval telles que la cartographie des mélanges de caractères importants ou la sélection génomique pour les programmes de sélection. Cette approche pourrait également être adaptée à d'autres espèces lorsque l'on étudie des populations mélangées avec un faible nombre d'individus de référence.

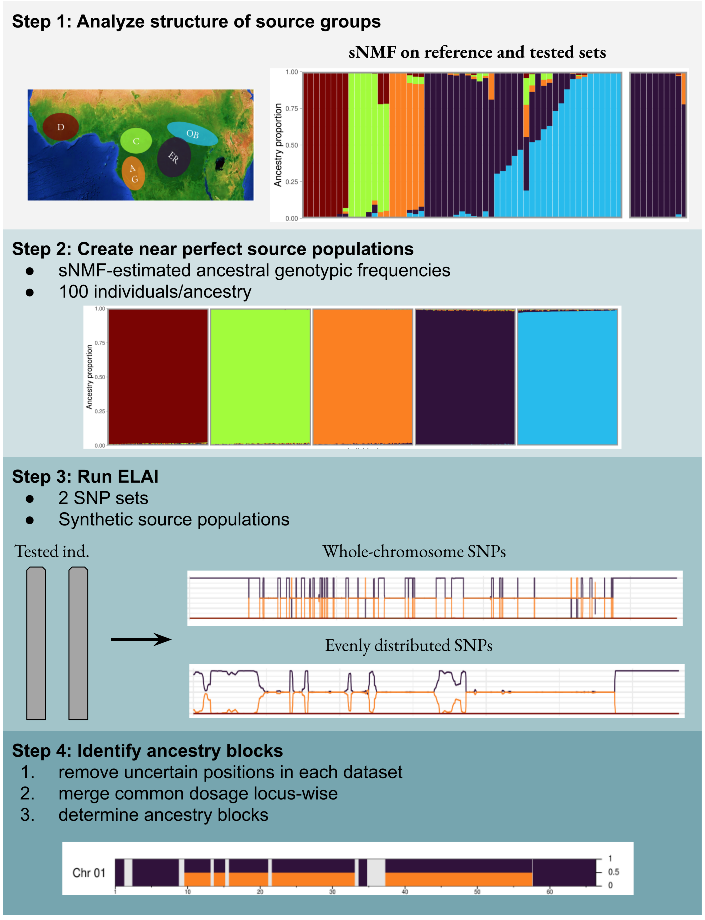

Step 1: Analyze structure of source groups
sNMF on reference and tested sets
D
C
OB
ER
A
G
Ancestry proportion
1.00
0.75
0.50
0.25
0.00
Step 2: Create near perfect source populations
● sNMF-estimated ancestral genotypic frequencies
● 100 individuals/ancestry
Step 3: Run ELAI
● 2 SNP sets
● Synthetic source populations
Tested ind.
Whole-chromosome SNPs
Evenly distributed SNPs
Step 4: Identify ancestry blocks
1. remove uncertain positions in each dataset
2. merge common dosage locus-wise
3. determine ancestry blocks
Chr 01
1
0.5
0
1
10
20
30
40
50
60

FIGURE .0.2: Cadre de travail pour le LAI de C. canephora cultivé. L'analyse ELAI a été réalisée pour chaque chromosome individuellement et a comporté trois étapes principales. Étape 1 : analyse de la structure génétique du groupe ancestral, par l'exécution de la méthode sNMF sur l'ensemble de référence et l'ensemble testé. Étape 2 : simulation des populations sources sur la base des fréquences génotypiques ancestrales estimées par la méthode sNMF. Étape 3 : exécution de l'analyse ELAI sur les individus testés à l'aide de deux ensembles de marqueurs, à savoir l'ensemble des SNPs d'un chromosome entier et un sous-ensemble de SNP à répartition uniforme. Étape 4 : fusion des dosages d'ascendance déduits dans les deux ensembles de SNP pour déterminer le consensus final d'inférence du chromosome cible.

La diffusion à partir du bassin du Congo et l'hybridation avec d'autres origines ont façonné la diversité du café *C. canephora* vietnamien

C. canephora (Robusta) est fortement structuré dans son habitat d'origine, les forêts tropicales d'Afrique. Seule une partie de cette diversité a contribué à la diffusion du robusta à travers le monde. Nous avons retracé ici l'origine africaine d'accessions de C. canephora cultivées dans les hauts plateaux centraux du Vietnam.

Une collection de 126 accessions de la collection WASI a été caractérisée, y compris des clones cultivés anciens, élites et locaux. Leur diversité génétique et leur origine ont été déduites par comparaison avec des échantillons sauvages de référence, à l'aide d'un nouvel ensemble de 261 SNP à l'échelle du génome. Les accessions maximisant la distance génétique et la richesse allélique ont été sélectionnées dans une core-collection. Des segments d'ascendance au niveau des chromosomes de chaque individu de la core-collection ont été détectés en utilisant des données de séquençage du génome entier.

L'ascendance par le groupe génétique congolais du bassin du Congo (groupe ER) était présent dans toutes les accessions vietnamiennes, à des proportions variables (figure .0.3). Une majorité d'entre elles 77 % (97 individus) présentaient une ascendance très forte > 90 % issue du groupe ER. Des signatures génétiques du groupe A (12 à 82 %, six individus) et des signatures du groupe D (12 à 38 %, 13 individus) ont aussi été observées. Dix accessions de la collection présentaient également 86 à 100 % du groupe ER, avec 10 ou 12 % du groupe OB dans deux accessions. Les résultats des accessions élites sont cohérents avec les résultats du chapitre précédent.

Une core-collection de 45 accessions, représentatif de la diversité génétique de la collection WASI, a été proposé. Cette collection présente une hétérozygotie attendue et une richesse allélique comparables à celles de l'ensemble de la collection. La plupart des individus étroitement apparentés ont été retirés, et la plupart des hybrides avec toutes les accessions élites ont été conservés. Cette core-collection a été séquencée au niveau du génome entier des individus, permettant une analyse approfondie de leur histoire évolutive.

L'ancestralité d'une partie du génome dans un groupe congolais du bassin du Congo (groupe ER) a été validée dans toutes les accessions vietnamiennes reséquencées (figure .0.3). Sur les 45 accessions resequencées, différents schémas de mélange et différentes longueurs de tract ont été détectés (figure .0.4). Des mélanges, avec des distributions différentes sur le génome, provenant d'au moins un autre groupe (D dans la région guinéenne, AG dans la région côtière atlantique de l'Afrique centrale et OB dans le bassin du Congo) ont été trouvés chez 31 individus.

Les caféiers *C. canephora* vietnamiens sont principalement issus de *C. canephora* congolais, mais il existe également une diversité d'autres sources génétiques identifiée dans des hybrides rétro-croisés. L'hybridation de ces groupes a été largement utilisée pour produire des clones élite. Les résultats de cette étude ont permis de mieux comprendre les ressources génétiques disponibles au Vietnam, ce qui sera utile pour l'établissement de stratégies de sélection durables.

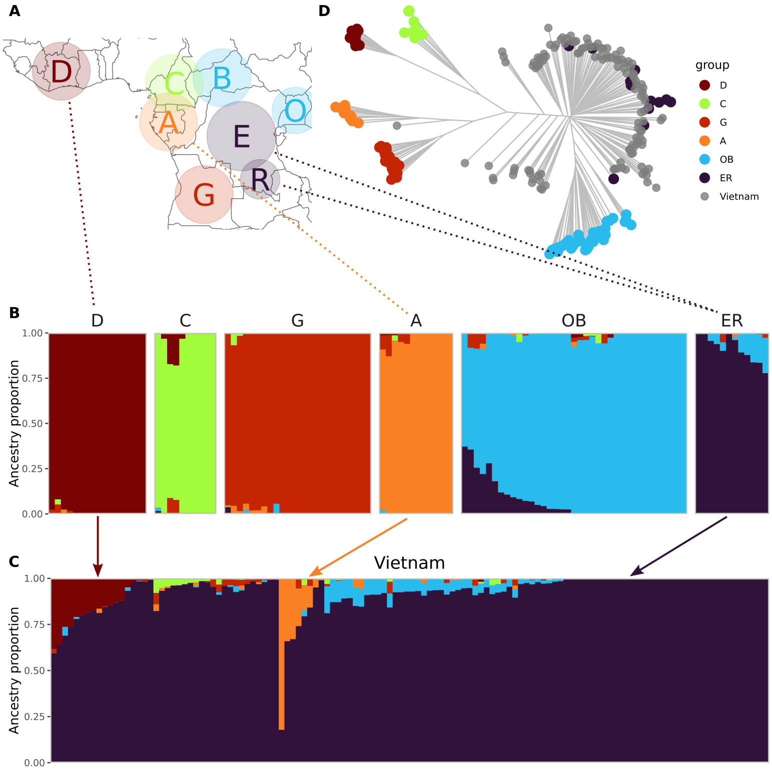


FIGURE .0.3: Origine, diversité et structure génétiques. (A) Répartition géographique des groupes génétiques sauvages en Afrique de l'Ouest et centrale (adapté de Mérot-L'Anthoëne et al. (2019)). Structure génétique de tous les individus, (B) pour 110 individus africains et (C) pour 126 individus vietnamiens. La proportion d'ascendance de tous les individus a été obtenue à partir de l'analyse sNMF avec K = 6 en utilisant 261 SNP. (D) Arbre de Neighbor-Joining des individus analysés basé sur les distances euclidiennes.

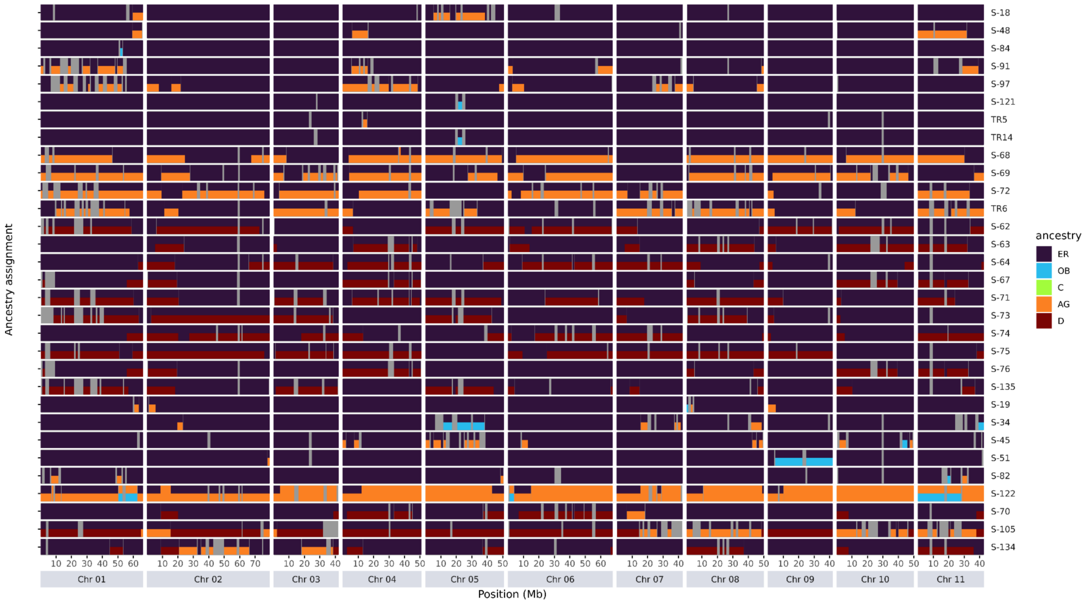


FIGURE .0.4: Introgression à l'échelle du génome chez 31 individus mélangés de la core-collection. Proportion d'ascendance locale (0, 50 %, 100 %) déduite le long des chromosomes. Les groupes d'ascendance sont représentés par des couleurs (AG - orange, OB - bleu, D - rouge, ER - violet foncé), et les régions indéterminées sont en gris.

Les sources génétiques de café *C. canephora* les plus adaptées au climat futur du Vietnam

Dans sa zone d'origine, *C. canephora* présente différents groupes génétiques (Mérot-L'Anthoëne et al., 2019) répartis dans une gamme relativement large de conditions environnementales (A. P. Davis et al., 2006). La dispersion mondiale de *C. canephora* et le changement climatique pourraient accroître l'inadéquation entre l'adaptation locale et la distribution à l'avenir. Selon les projections, le Vietnam, premier producteur et exportateur mondial de robusta depuis les années 2000, pourrait perdre 50 % de sa production actuelle d'ici 2050 (ICO, 2019). Il est donc important d'évaluer l'adéquation des populations sauvages au climat local du Vietnam afin d'identifier des sources potentielles d'adaptation aux conditions futures.

Pour prédire la potentielle adaptation d'une espèce à un nouvel environnement, on peut analyser les corrélations entre les variables bioclimatiques et les variants génétiques (Rellstab et al., 2021). En supposant que les populations sont adaptées à leur environnement local, ces relations peuvent être extrapolées dans l'espace et le temps, afin de prévoir les changements de composition génomique nécessaires dans des environnements nouveaux ou futurs (Bay et al., 2018 ; Capblancq et Forester, 2021 ; Fitzpatrick et Keller, 2015 ; Rellstab et al., 2016), ce que l'on appelle le "décalage génomique" ou "décalage génétique" (Rellstab et al., 2021). Dans cette étude, nous avons appliqué une méthode sans génome de référence (en utilisant des k-mers d'une longueur de 31 pb extraits directement des lectures de séquences) pour capturer toutes les variations génomiques dans la population.

Pour évaluer l'adéquation des variétés sauvages de *C. canephora* (60 individus couvrant tous les groupes génétiques) au climat local du Vietnam, nous avons tout d'abord évalué l'adéquation des groupes natifs au climat des zones plantées du Vietnam (640 occurrences), sur la base de la différence de 19 variables bioclimatiques (base de données WorldClim). Nous avons ensuite étudié les k-mers associés aux facteurs bioclimatiques, afin d'estimer les compensations génétiques pour le climat actuel et futur du Vietnam (prédit par trois modèles climatiques différents et deux scénarios de voies socio-économiques partagées).

La diversité génétique des populations sauvages et de dix variétés élites communément plantées au Vietnam a été évaluée sur la base d'un ensemble de données k-mers. Les populations sauvages ont été fortement différenciées en cinq groupes génétiques principaux, mais un seul d'entre eux (le groupe ER) était dominant dans les variétés élites cultivées au Vietnam, avec une proportion mineure du groupe AG du Gabon et de l'Angola. Une corrélation négative a été établie entre les distances climatiques par rapport à l'aire d'origine du groupe ER et les rendements de café dans les zones plantées. Sur la base de cette corrélation, une réduction du rendement dans environ un tiers des zones plantées a été prédite pour toutes les prévisions climatiques futures. La distance climatique suggère une meilleure adéquation du groupe génétique AG avec le climat du Vietnam.

Plus de 18 millions de k-mers ont été détectés en association avec les variables bioclimatiques. L'annotation fonctionnelle de ces k-mers candidats a permis d'identifier des protéines putatives liées à la régulation des gènes. En utilisant ces k-mers candidats, le décalage génétique a été estimé pour chaque individu de *C. canephora* natif, avec une correction pour les facteurs confondants (Gain et al., 2023). Les décalages génétiques ont montré des variations dans les différents groupes, et ont également suggéré les génotypes les mieux adaptés dans le groupe AG à la plupart des régions plantées au Vietnam (figure .0.5).

L'adéquation climatique et l'adéquation génomique fournissent ensemble de meilleures suggestions pour les futures stratégies de sélection. L'introduction d'un plus grand nombre de matériels du groupe AG et la production d'un plus grand nombre d'hybrides ERxAG peuvent être prometteuses pour l'amélioration de l'adaptabilité du café au changement climatique au Vietnam, en particulier dans les régions moyennes où l'on prévoit une forte perte de rendement.

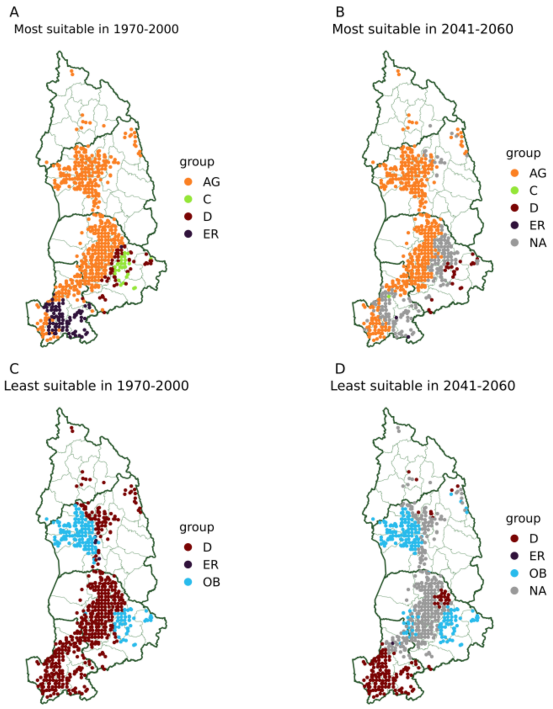

A
Most suitable in 1970-2000
group
AG
C
D
ER
B
Most suitable in 2041-2060
group
AG
C
D
ER
NA
C
Least suitable in 1970-2000
group
D
ER
OB
D
Least suitable in 2041-2060
group
D
ER
OB
NA

FIGURE .0.5: Matériel génétique le plus adapté et le moins adapté au climat au Vietnam. Les figures A et C montrent les groupes génétiques des génotypes ayant le décalage génétique le plus faible et le plus élevé dans le climat actuel dans 640 occurrences au Vietnam, respectivement. Les figures B et D montrent les groupes génétiques des génotypes les plus et les moins adaptés au climat futur. Pour chaque occurrence, le groupe génétique est présenté s'il a été trouvé dans au moins cinq des six scénarios futurs, sinon il est présenté par un point gris comme valeur NA.

Discussion générale

Principales conclusions

Les premières variétés introduites au Vietnam ont vraisemblablement été principalement des génotypes congolais (groupes E et R de la RDC), comme en témoignent les principaux groupes génétiques trouvés dans les accessions anciennes et actuelles des hauts plateaux centraux. Environ un quart des 126 génotypes étudiés ont deux ou trois sources mélangées, y compris le groupe ER (dans tous les hybrides), et impliquant tous les autres groupes à l'exception du groupe C. Cependant, bien qu'il y ait des preuves de mélanges récents et anciens, on ne sait pas si les croisements entre groupes ont eu lieu avant ou après la diffusion au Vietnam. La plupart des variétés élites proviennent uniquement du groupe congolais ER, à l'exception d'une variété issue d'un rétro-croisement entre les groupes ER et AG. Ces variétés élites pourraient avoir été plus communément cultivées au Vietnam (ICO, 2019), et utilisées en mélange dans les plantations (en raison de leur auto-incompatibilité). Par conséquent, la diversité des cultivars locaux vietnamiens pourrait être réduite et limitée au groupe ER.

Le groupe génétique ER pourrait ne pas être le mieux adapté au climat local des hauts-plateaux centraux, car les conditions bioclimatiques locales se sont révélées plus proches des régions d'origine du groupe AG. Cette hypothèse a été confirmée par l'étude de la relation génotype-environnement dans les populations sauvages de *C. canephora*. Les accessions sauvages du groupe AG semblaient également être les génotypes les plus adaptés au climat local de la plupart des régions des hauts plateaux centraux, dans le présent et d'ici à 2060. Les résultats indiquent une incohérence entre les génotypes largement cultivés au Vietnam et les génotypes prédits comme les mieux adaptés à l'environnement local.

Limites et perspectives

Dans ce projet, la diversité du café *C. canephora* vietnamien a été évaluée uniquement sur la base de la variation génétique, sans tenir compte de la variabilité phénotypique. Diverses caractéristiques morpho-agronomiques à forte héritabilité (Leroy et al., 1993b ; Montagnon et al., 1998a) peuvent être utilisées comme marqueurs pour identifier la diversité variétale et compléter la variabilité génétique, afin de mieux comprendre les changements évolutifs et l'adaptation locale chez *C. canephora* (Loor Solórzano et al., 2017). La compréhension de la génétique quantitative des caractères souhaitables pourrait aider à sélectionner efficacement le matériel parental approprié pour le croisement et la production de la vigueur hybride, et aider les étapes de sélection (M. A. G. Ferrão et al., 2023).

Le décalage génomique des populations sauvages de *C. canephora* au Vietnam a été estimé pour évaluer leur adéquation au futur environnement local. Toutefois, l'application du décalage génomique à l'établissement de programmes de sélection n'est pas simple et nécessite une validation plus poussée et une meilleure compréhension des limites de la méthode (Lind et al., 2023 ; Rellstab et al., 2021). En outre, sur la base des modèles génotype-environnement, il est possible d'estimer non seulement le décalage génomique des populations sauvages adaptées, mais aussi d'interpoler l'aptitude des nouveaux génotypes dans l'environnement existant ou nouveau. Cette approche, si elle est applicable, sera utile lorsque les populations sauvages ne sont pas accessibles dans un nouvel environnement, ou pour prédire l'adaptabilité des hybrides issus des programmes de sélection.

Implications pour les plans de sélection

Comme les génotypes de café *C. canephora* vietnamien sont redondants et étroitement liés à un groupe génétique (ER), l'introduction de diversité issue des autres groupes génétique sera utile. Le groupe génétique AG du Gabon et de l'Angola (type "Conilon"), en particulier, devrait faire l'objet d'une plus grande attention, car il pourrait avoir un potentiel d'adaptation plus élevé dans l'environnement local, et il est généralement croisé avec le groupe ER (type "Robusta") pour l'amélioration variétale. Les croisements entre groupes et la sélection au sein de la population hybride pourraient convenir aux stratégies de sélection au Vietnam, car ils se sont révélés efficaces pour améliorer toutes les gammes de caractères (Alkimim et al., 2021).

La génomique et la génétique jouent un rôle de plus en plus important dans la sélection des cultures. Une meilleure compréhension de la base génétique des caractères agronomiques, des phénotypes bénéfiques et de l'adaptation locale sera la clé de voûte de l'amélioration du café.

Contents

I General introduction

I.1 Crop domestication and adaptation

I.1.1 Side effects of domestication in diversity and adaptation

Domestication is an evolutionary process driven by humans that occurred in most plants about 12,000 years ago (Meyer et al., 2012; Zohary and Hopf, 2000). Domestication process is associated with genetic sampling from the wild diversity, often leading to lower diversity in cultivated varieties (Allaby et al., 2019; Meyer and Purugganan, 2013). Domestication is also associated with morphology and physiology changes in the cultivars compared to the wild relatives, and these differences define the domestication syndrome (Harlan, 1992). Domesticated plants carry desirable traits for human, such as increased fruit size in tomato (Bai and Lindhout, 2007), seedless fruit in banana (Heslop-Harrison and Schwarzacher, 2007), sweetness in watermelon (Paris, 2015), non-shattering in rice (Zheng et al., 2016), and other traits related to yield and metabolites in many other species (Denham et al., 2020). Domesticated plants play important roles in human's sources of fiber, medicines, and ornamental (Dirzo and Raven, 2003).

Domestication however has some adverse effects on plants. It can cause a drastic loss of genetic diversity - the domestication bottleneck (Figure I.1.1). For instance, genetic diversity is reduced by 70% in cultivated rice (Z.-M. Li et al., 2011), and up to 81% rare alleles were lost in soybean landraces (Hyten et al., 2006). Genetic bottleneck does not only occur during domestication, but also happens in modern crops, because elite cultivars are mostly selected from precedent domestication events for crop improvement (Gross et al., 2014).

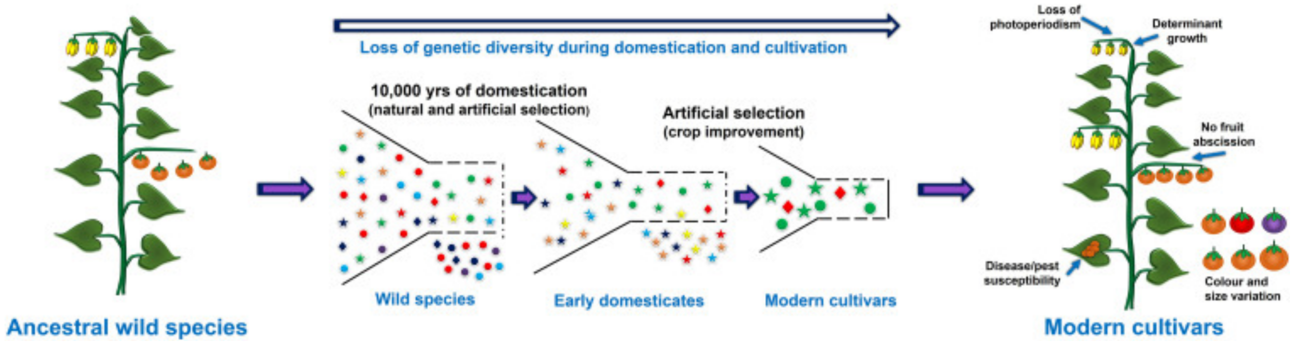


Figure I.1.1: An example of genetic loss during domestication and post-domestication (figure from R. Kumar et al., 2021)

Following domestication, crops are often diffused to other locations outside of their centers of origin. The new environment occupied is not necessarily optimal for their growth, necessitating high adaptability of the plants. Because of the initial genetic diversity loss, the adaptive potential of cultivated crops to new environments might be limited. For example, domesticated narrow-leaf lupin has high yield in eastern Australia but lower in western Australia, due to selection in vernalization response (Berger et al., 2012). Moreover, as a cost of gaining traits needed by humans, species under domestication may lose traits advantageous to themselves (Bergelson and Purrington, 1996; Pickersgill, 2007). Resistant genes or alleles might be lost during domestication, as observed in tomato or common beans, by direct or indirect (in genetic linkage with targeted alleles) selection (Singh et al., 2022).

I.1.2 Crops response to climate change

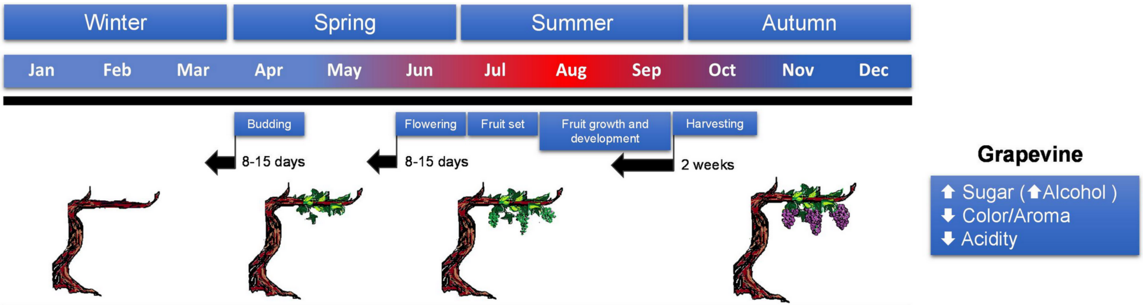
During the past century, severe climate changes have been observed, as a result of increasing greenhouse gas emission. Global mean surface temperature has risen by more than 1°C since the 1980s (IPCC report 2019). The frequency of extreme events, such as drought, heatwave, or storm, has increased worldwide and more severely in Asia (de Ruiter et al., 2020). Future climate change has been predicted using different global climate models (GCMs) based on representative concentration pathways (named as RCPy, with y = 1.9 to 8.5 forcing level W/m2), and/or shared socio-economic pathways (named as SSPx, with x = 1 to 5). RCPs propose different scenarios of emission concentration without societal factors, while SSPs derive emission scenarios based on levels of socio-economic activities without climate change impact (O'Neill et al., 2020). They are therefore combined to obtain more complete scenarios, which are named SSPx-y. While in the most sustainable and optimistic scenario (SSP1-1.9), global warming will be limited to 1.5°C by 2100, in the most pessimistic and economic optimism scenario (SSP5-5.8), global temperature will increase by 4°C (IPCC, 2019). If these predictions occur in the future, crops will be at significant risk.

Crops are and will continue to be affected by climate change in their morphology, physiology, phenology, production, distribution range, and species richness (Box I.1.2). Climate change not only has direct impacts with more extreme heat or drought stress, but also indirect impacts by increasing abiotic stress (e.g. soil nutrient, irrigation availability, flood, drought, etc) and biotic stress (e.g. facilitating growth of pests and diseases) (Raza et al., 2019). Numerous studies have shown and elucidated how abiotic stresses are associated to plant production, such as metabolite malfunction due to drought (Yadav et al., 2021) and heat (Gong et al., 1997; Griffin et al., 2004; Z. Z. Xu and Zhou, 2006), decrease in fertilization caused by water deficit (Suharyanti et al., 2020), shorter growing phases in response to increase in temperature (Moriondo and Bindi, 2007), or significant yield reduction proportional to temperature increase (Ray et al., 2015). Because of reduced climatic suitability, geographical distribution and abundance of crop species can shrink or become disrupted. Taking peanut (*Arachis*) as an example, it has been predicted that, its suitable areas will decrease by 89%, and half of the species will go extinction by 2055 (Jarvis et al., 2008). To cope with climate change, production areas of several crops in Europe, America, and Asia have shifted to more suitable environments (Sloat et al., 2020). Migration speed of agro-climatic zones (areas with similar climatic conditions suitable for a certain range of crops) in Europe has reached 100 km per 10 years for the past 40 years and will double in the next 30 years (Ceglar et al., 2019).

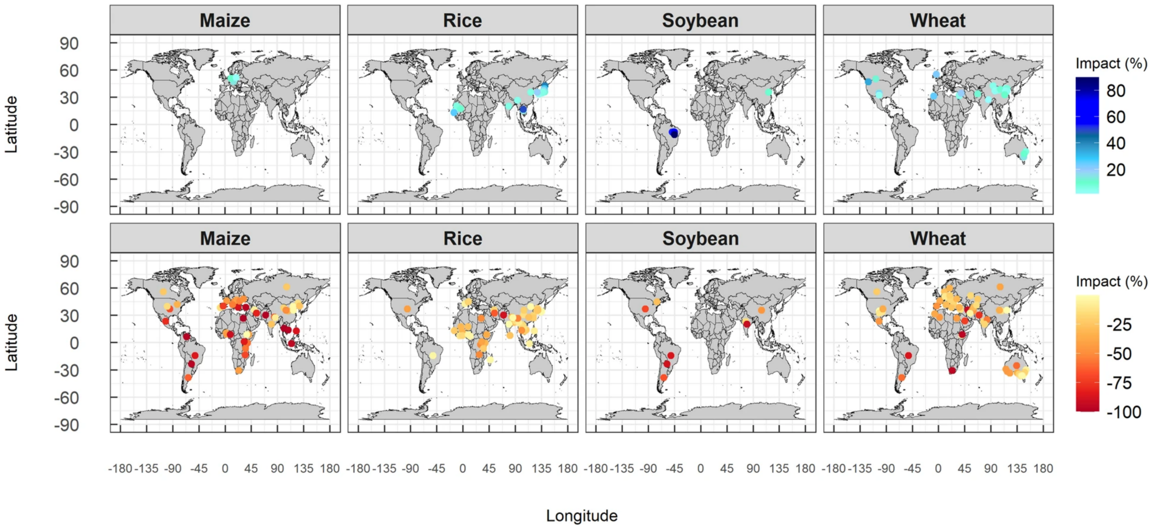
To mitigate the consequences of climate change on crops, farmers could improve cultural methodologies and management practices. In long terms, genetic and genomic strategies also play important role in developing new cultivars with higher resistance or tolerance to biotic and abiotic stress (Raza et al., 2019). Some authors suggested that enhancing resistance of crops might be more efficient than improving agricultural practices (K. Guan et al., 2017; Parkes et al., 2018). Both are certainly important and could contribute to a sustainable approach to mitigate or adapt climate change in agriculture.

Box 1: Examples of predicted climate change effects on crops

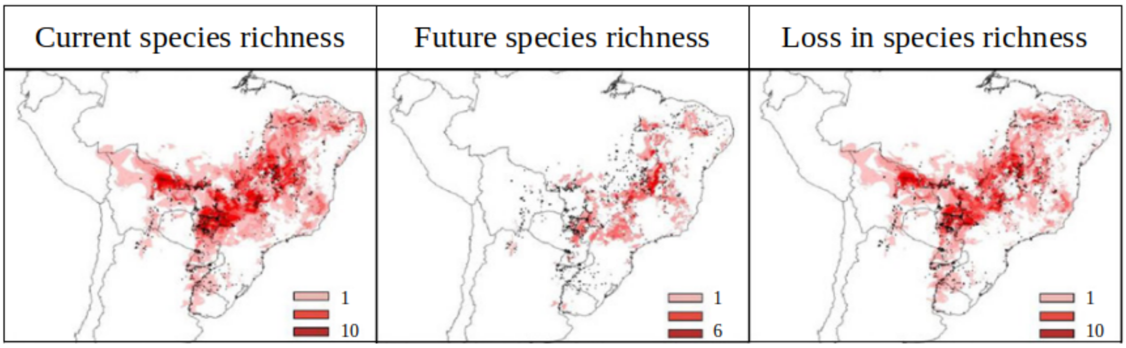
1. **Phenology**: increasing temperature and drought will cause earlier phenological events (e.g. budding, flowering, and harvesting) in grapevine, and changed fruit quality (De Ollas et al., 2019).



2. **Productivity**: under RCP8.5, until the end of the 21th century, yields of four main crops (maize, rice, soybean, and wheat) will drastically decrease in America, Africa and Asia, while slightly increase in only few regions (Hasegawa et al., 2022).



3. **Distribution and richness**: by 2055, peanuts will lose 24–31 of its 51 wild species and their distribution range will decrease by 89% of the current area (Jarvis et al., 2008).



I.1.3 Importance of genetic diversity in adaptation

To survive under environmental change, if crops cannot migrate to new locations with suitable conditions, they will have to adapt to the new conditions (Aitken et al., 2008). Adaptation to rapid climate change in population is facilitated by three main sources of genetic variation: standing genetic variation (preexisting in the population), adaptive introgression (from another population), and new mutation (Matuszewski et al., 2015; Molinaro et al., 2021; Tigano and Friesen, 2016) (Figure I.1.2). Wild crop relatives are massive resources for discovering such variants, and can be used directly in breeding programs (Flint-Garcia et al., 2023). Domesticated or cultivated varieties possess significantly lower genetic diversity, but may contain admixture or introgression from multiple wild sources in their genome. These admixed populations are useful for accessing some of the wild diversity and identifying variants from adaptive introgression (Burgarella et al., 2019). Elite varieties are also valuable sources of variation, as they already exhibit some advantageous traits (Kochevenko et al., 2018), and may present combinations of beneficial alleles (Badu-Apraku and Yallou, 2009). Genetic diversity in modern cultivars can be improved by crossing with wild populations (Smýkal et al., 2018). Finally, genetic variations can be newly created by induced mutations, which has been effectively applied in rice, barley and wheat (Sathee et al., 2022; Yali and Mitiku, 2022). The future of crop adaptation will rely mainly on their gene pools and genetic/genomic studies to understand the genetic basis of adaptation for effective breeding strategies.

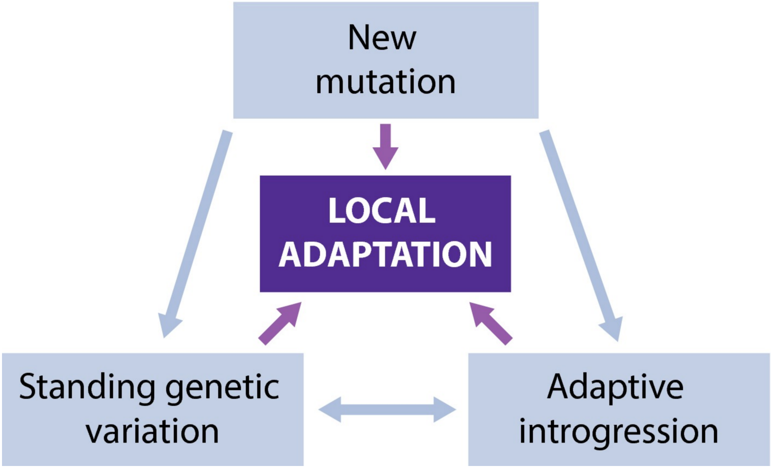
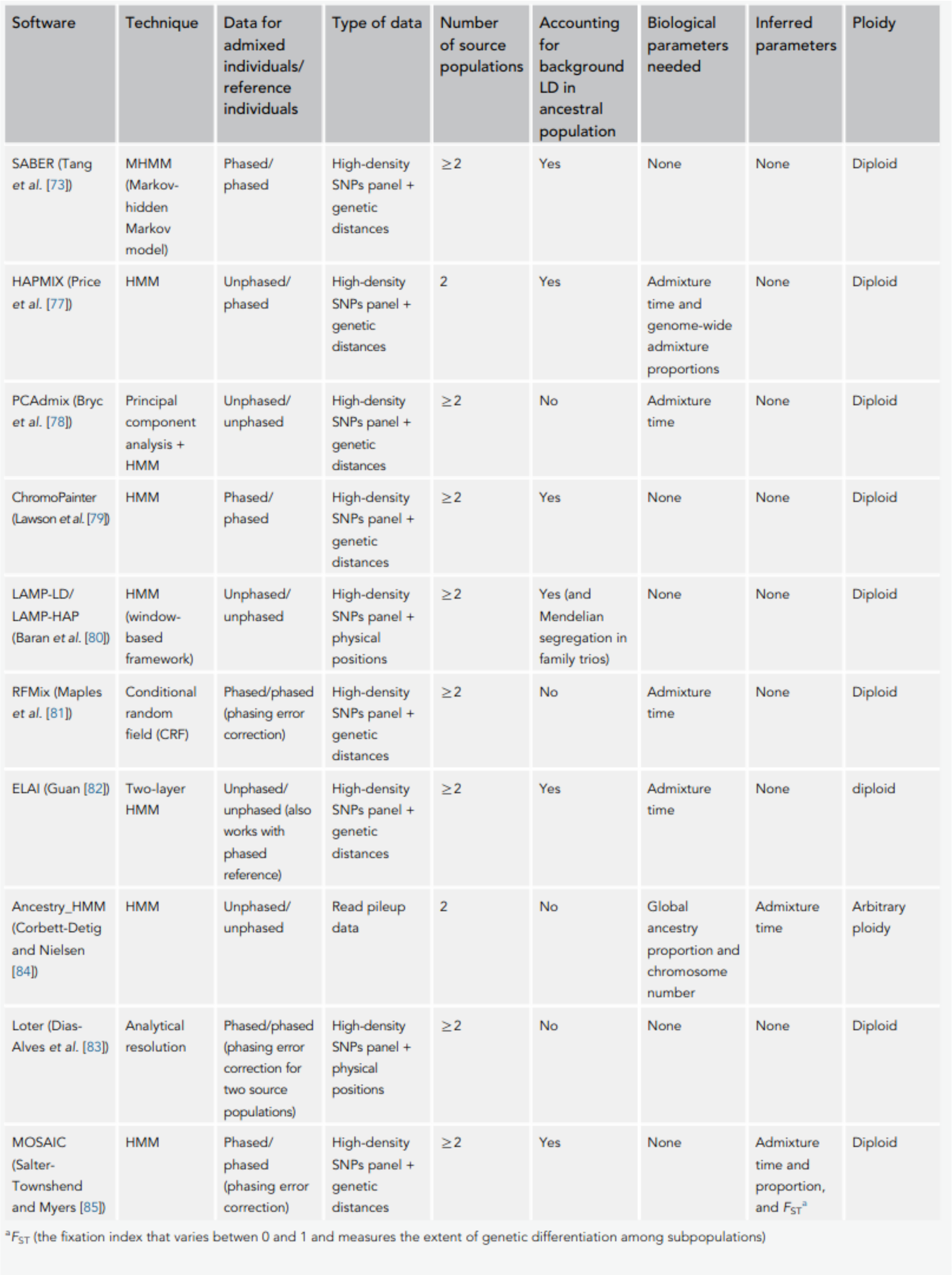


Figure I.1.2: Sources of genetic variation for local adaptation and their interactions (from Tigano and Friesen, 2016). Violet arrows indicate direct contribution of the sources to local adaptation, and light blue arrows indicate the dynamics between them. New mutation can provide new variation to standing genetic variation and adaptive introgression. Adaptive variation can be transferable between standing genetic variation and adaptive introgression, when they bring high fitness benefit and reach fixation.

Wild and cultivated gene pool diversity can be easily assessed by different types of molecular marker, such as single-nucleotide polymorphism (SNP) or microsatellite (reviewed in I.3). Detection of introgression, which has been widely applied on human research, has recently been used in plant genomics (Cottin et al., 2019; Rendón-Anaya et al., 2021; Y. Zhou et al., 2020). Many methods have been developed for inferring ancestry proportion at any given SNP (local ancestry inference), which are applicable for different ploidy levels and different types of data (Table I.1.1). These approaches can provide more insights into genetic structure at the chromosome level, and enable detection of adaptive introgression in different plant species (Leitwein et al., 2020).

Table I.1.1: Commonly used LAI software and their characteristics (from Leitwein et al., 2020)



Software	Technique	Data for admixed individuals/ reference individuals	Type of data	Number of source populations	Accounting for background LD in ancestral population	Biological parameters needed	Inferred parameters	Ploidy
SABER (Tang *et al.* [73])	MHMM (Markov-hidden Markov model)	Phased/ phased	High-density SNPs panel + genetic distances	≥2	Yes	None	None	Diploid
HAPMIX (Price *et al.* [77])	HMM	Unphased/ phased	High-density SNPs panel + genetic distances	2	Yes	Admixture time and genome-wide admixture proportions	None	Diploid
PCAdmix (Bryc *et al.* [78])	Principal component analysis + HMM	Unphased/ unphased	High-density SNPs panel + genetic distances	≥2	No	Admixture time	None	Diploid
ChromoPainter (Lawson *et al.* [79])	HMM	Phased/ phased	High-density SNPs panel + genetic distances	≥2	Yes	None	None	Diploid
LAMP-LD/ LAMP-HAP (Baran *et al.* [80])	HMM (window-based framework)	Unphased/ unphased	High-density SNPs panel + physical positions	≥2	Yes (and Mendelian segregation in family trios)	None	None	Diploid
RFMix (Maples *et al.* [81])	Conditional random field (CRF)	Phased/phased (phasing error correction)	High-density SNPs panel + genetic distances	≥2	No	Admixture time	None	Diploid
ELAI (Guan [82])	Two-layer HMM	Unphased/ unphased (also works with phased reference)	High-density SNPs panel + genetic distances	≥2	Yes	Admixture time	None	diploid
Ancestry_HMM (Corbett-Detig and Nielsen [84])	HMM	Unphased/ unphased	Read pileup data	2	No	Global ancestry proportion and chromosome number	Admixture time	Arbitrary ploidy
Loter (Dias-Alves *et al.* [83])	Analytical resolution	Phased/phased (phasing error correction for two source populations)	High-density SNPs panel + physical positions	≥2	No	None	None	Diploid
MOSAIC (Salter-Townshend and Myers [85])	HMM	Phased/ phased (phasing error correction)	High-density SNPs panel + genetic distances	≥2	Yes	None	Admixture time and proportion, and F_{ST}[a]	Diploid

[a] F_{ST} (the fixation index that varies betwen 0 and 1 and measures the extent of genetic differentiation among subpopulations)

The identification and characterization of adaptive variants is also facilitated by the high-throughput genotyping and sequencing. Approaches using phenotyping to discover quantitative trait loci (QTL mapping or genome-wide association studies - GWAS) or studying signature of selection associated with environmental factors (genotype-environment association – GEA) are now easier and easier to achieve. Using GWAS, a high number of genes or QTLs underlying several adaptive traits (e.g. flower time, metabolites, broad-spectrum resistance) related to multiple stresses (e.g. drought, heat, cold, salt, and multi-stress) have been identified in many crops (Zhu et al., 2022). GEA, which bridges evolutionary genomics and ecological data (e.g. temperature, precipitation), has provided more insights into geographical adaptation and prediction of genomic adaptability to new environments for several plant species (Cortés et al., 2022). However, further research is still needed to realize the application of these adaptive variations in practice.

I.2 *Coffea canephora* species

I.2.1 Origins and dispersal

Coffee is one of the most consumed beverages in the world (Neves et al., 2012). It is not only a hot drink but also contains several compounds with health benefits, such as caffeine (Farah, 2009). Coffee is produced from berries of species belonging to the *Coffea* genus in the Rubiaceae flowering-plant family. The *Coffea* genus is composed of 124 described species (A. P. Davis et al., 2006), but only two of them are mostly used for coffee production nowadays. *Coffea arabica* L. produces Arabica coffee, which has smooth, sweet and aromatic taste. *Coffea canephora* Pierre ex Froehner produces Robusta coffee with stronger and more bitter taste than Arabica coffee. Consequently, Robusta coffee is commonly used in instant coffee, espresso, or in blends of ground coffee as a filler. Robusta coffee tree is considered hardier than Arabica. It has greater resistance to common pests and diseases, such as coffee leaf rust caused by the fungus Hemileia vastatrix (Silva et al., 2006), twig borer caused by the beetle *Xylosandrus compactus* (Noir et al., 2003), and root damage caused by the nematode *Meloidogyne* species (Noir et al., 2003). It is also more tolerance to harsh conditions such as drought, heat or cold (DaMatta and Ramalho, 2006). In addition, *C. canephora* exhibits a wider environmental distribution in the wild than *C. arabica*, with higher genetic diversity (A. P. Davis et al., 2006; Lashermes et al., 2000; Mérot-L'Anthoëne et al., 2019).

With the widest native distribution range in the *Coffea* genus (A. P. Davis et al., 2006), and high genetic diversity (Cubry et al., 2013; Gomez et al., 2009; Mérot-L'Anthoëne et al., 2019), *C. canephora* is an interesting model for studying intraspecific diversity and local adaptation.

C. canephora is indigenous to West and Central Africa, with a relatively wide distribution in longitude scale ranging from Guinea to Uganda, as well as in latitude scale ranging from Cameroon and Central African Republic to Angola (Berthaud, 1986; A. P. Davis et al., 2006). It naturally grows in the understorey of evergreen forests (mainly), and can sometimes be found in seasonally dry tropical forest and gallery forest (A. P. Davis et al., 2006). Therefore, wild *C. canephora* also spans in (and may be adapted to) diverse environmental conditions such as elevation, precipitation or temperature. For example, in west Africa, the altitude ranges from sea level to 500 m and the annual precipitation is high at 2000 mm, while in the western coastal regions of central Africa (such as Gabon, Cameroon), the altitude is up to 1500 m and the annual precipitation is much lower at 1000 mm (Almazroui et al., 2020; Liebmann et al., 2012).

C. canephora was first cultivated locally in Africa, notably in Gabon, Uganda and Congo, in the late 19th century (Charrier and Berthaud, 1990; Chevalier, 1929), and has been widely cultivated in other parts of the world since the early 20th century. The first dispersal outside its native range was to Java (Indonesia), which was also the location of the first breeding program using unselected materials from Democratic Republic of Congo (DRC), Uganda and Gabon (Montagnon et al., 1998b). The breeding centers then moved back to Africa with selected materials from previous breeding programs (i.e. DRC in the 1930s, and Cote D'Ivoire in the 1960s) (Montagnon et al., 1998b). Selected materials from these breeding centers were also introduced to many other countries, such as Vietnam in 1908 (ICO, 2019), Brazil in 1912 (M. A. G. Ferrão et al., 2007; MERLO and Capixaba, 2012), Papua New Guinea in 1935 (Charmetant, 1994), Malaysia in 1980 (Montagnon et al., 1998b), etc.

Robusta coffee is currently grown in more than 50 countries, mostly near the equator in Asia, Africa and America, as these are tropical areas suitable for its growing. Production of Robusta (about 4.2 million tons in 2020) accounts for about 40% total coffee production worldwide; and Vietnam, Brazil, and Indonesia are the leading countries (according to ICO 2021 report).

I.2.2 Intraspecific diversity

C. canephora species is diploid (2n = 22), with a genome size of 7̃10 Mb (Noirot et al., 2003). It is strictly allogamous (Berthaud, 1986; Lashermes et al., 2000), resulting in high heterozygosity at the genomic level (Denoeud et al., 2014; Lashermes et al., 2000). A reference genome of *C. canephora* has been sequenced from a double haploid accession, covering at least 80% its genome size (Denoeud et al., 2014). This reference genome has enabled the construction of a high-density genetic map covering 6̃4% of the assembly, and annotation of more than 25k protein-coding genes (Denoeud et al., 2014). It could also allow further studies for better understanding the genomics, diversity, and evolution of the species (reviewed in I.3).

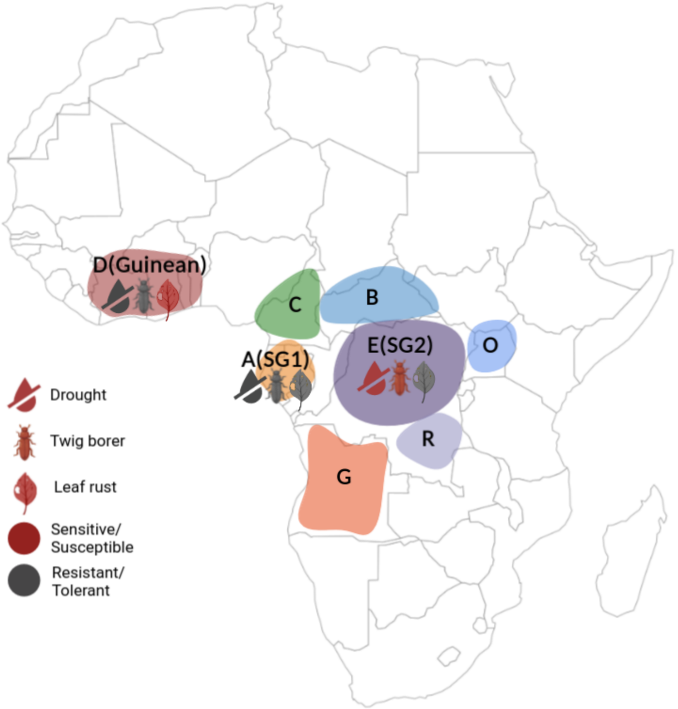


Figure I.2.1: Geographical representation of the genetic groups of wild C. canephora in Africa and their characteristics. The group distribution is according to Mérot-L'Anthoëne et al. (2019), the biotic and abiotic stress behaviour is according to Montagnon et al. (1998b). These behaviours has been observed for some accessions, but might be variable within the group.

Wild populations of *C. canephora* were initially classified into two diversity groups based on isozyme polymorphism by Berthaud (1986), which were named the "Guinean" group, corresponding to populations in Côte D'Ivoire, and the "Congolese" group, corresponding to populations in the Central African Republic (CAR) and Cameroon. Eventually, with the evolution in molecular biotechnology, such as the emergence of genomic markers (random amplified polymorphic DNA - RAPD, simple sequence repeat - SSR, SNP), and more studied samples from different origins, its genetic structure has been better characterized. *C. canephora* diversity was recently differentiated into eight genetic groups in Mérot-L'Anthoëne et al. (2019) by using an 8.5K SNPs array (Figure I.2.1 and Table I.2.1). The Guinean group is now known as group D, corresponding to the wild populations of Guinea and Côte d'Ivoire. This group has recently been revised and might comprise of up to five subgroups (Labouisse et al., 2020). The Congolese group has been further differentiated into five (sub-)groups: group A, previously referred to Congolese subgroup 1 (SG1) by Montagnon et al. (1992), distributed in Gabon and Togo, group B in southern Central African Republic (CAR), group C in Cameroon, group E, previously referred to Congolese subgroup 2 (SG2) by (Montagnon et al., 1992a), in the Democratic Republic of the Congo (DRC), and group O, initially described as UW by Musoli et al. (2009), in Uganda. Two additional groups, group G and R, were recently identified by (Mérot-L'Anthoëne et al., 2019), located in Angola and southern DRC, respectively. While group D and C appear to be more differentiated from others, closer genetic distances between the remaining groups were found, i.e. between group A and G, group E and R, and group O and B (Mérot-L'Anthoëne et al., 2019; Tournebize et al., 2022).

Table I.2.1: Summary table of the historical definition of C. canephora genetic groups with the references and the marker types in use (adapted from Mérot-L'Anthoëne et al., 2019)

Study	**Berthaud, 1986**	**Montagnon et al., 1992**	**Cubry et al., 2008** **Musoli et al. 2009**	**Dussert et al., 1999** **Gomez et al., 2009**	**Merot-L'anthoen et al., 2019**	**Labouisse et al, 2020**	**Geographic origin**	
Marker	**Isozymes**	**Isozymes**	**SSR**	**RFLP, SSR**	**SNP**	**SSR**	**Wild**	**Cultivated**
Genetic groups	Guinean	Guinean	Guinean	D	D	sgG1 (Maclaudii)	Guinea	
						sgG2 (Gamé)	Guinea	Guinea, Côte d'Ivoire
						sgG3	Guinea, Côte d'Ivoire	
						sgG4	Guinea, Côte d'Ivoire	
						sgG5	Guinea, Côte d'Ivoire	
	Congolese	SG1	SG1	A	A	Not included	North of Congo, South of Cameroon	« Niaouli », « Conilon » Togo, Côte d'Ivoire
		SG2	SG2	B	B	Not included	East of Central Afr. Rep.	
				E	E	Not included	Democr. Rep of the Congo, Cameroon	Guinea
		Not included	C	C	C	Not included	West of Central Afr. Rep., Cameroon	
	Not included		UW	Not included	O	Not included	Uganda, South Sudan	
			Not included	Not included	R	Not included	South of the Democr. Rep of the Congo	
					A	Not included	North and West of Angola	

The genetic groups are also characterized by variability in a number of agronomic traits, such as morphology of leaf and fruit, branching, ripeness, cup quality, biotic and abiotic resistance (Figure I.2.1 and Table I.2.2). Contrasting features have been described within the three main groups: Guinea, SG1 and SG2 (Berthaud, 1986; Leroy et al., 1993a; Montagnon et al., 1998a). Trees from the Congolese groups typically have bigger leaves, larger fruits, late ripeness, and better cup quality with good aroma and low acidity than the Guinean group. In terms of abiotic and biotic resistance, SG1 seems to be superior, with high resistance to twig borers and drought, and moderate resistance to leaf rust. SG2 is sensitive to drought, susceptible to twig borers but highly resistant to leaf rust, while in contrast, Guinean group is resistant to drought and twig borers yet susceptible to leaf rust. Because of their complementary characteristics and the hybrid vigor observed in the progeny of their crosses (heterosis), these three varietal groups have been commonly used in breeding programs.

Table I.2.2: Morphological description of the Congolese and Guinean group (adapted from Montagnon thesis 2000)

		Congolese group			**Guinea group**	
		SG1 (A)	**SG2 (E)**			
		Cultivated	**Cultivated**	**Wild**	**Cultivated**	**Wild**
Leaf	**Length (mm)**	180	203	221	170	176
	Width (mm)	71	82	93	68	75
	Acumen	Short to long	Long	Long	Long	Long
	Petiole	Short to long	Medium to long	Long	Short to long	Medium to long
	Domaties	Absent to clearly visible	Slighly to clearly visible	Slighly to clearly visible	Slighly to clearly visible	Slighly to clearly visible
Fruit	**Peduncle**	Short to medium	Short to medium	Short to long	Short to long	Short to long
	Size of disc	Small	Small	Small	Big	Small
	Disc relief	Flat to salient	Salient	Flat to salient	Flat	Flat
Others	**Intermode length**	Medium to long	Long	Long	Short	Short
	Ripeness	Very late	Moderately late	Moderately late	Early	Early
	Drought resistance	Good	Weak	Weak	Good	Good

I.2.3 Genetic adaptation to environment

As the genetic diversity of *C. canephora* has been shaped in a wide geographical distribution with high environmental variation, the wild populations might present genetic variation for local adaptation. Indeed, there is much evidence of local adaptation in Robusta coffee, detected by different relationships between genotype and climate, or genotype and fitness, or direct relationship between climate and fitness traits.

Recent studies have assessed the importance of genetic variation in the local adaptation of Robusta through the identification of adaptive variants linked with the environment. de Aquino et al., 2022, found in Ugandan populations 71 SNPs associated with temperature and precipitation factors, which were located in or near different genes involved in various functions such as caffeine synthesis (*DXMT* gene) or transcription factors controlling plant responses to abiotic stresses. Tournebize et al., 2022, studied wild populations across the whole geographical diversity, and identified 165 SNPs in association with mostly isothermality, seasonal temperature and precipitation, which were also found in genes involved in response to stress.

Relationships between genetic variation or gene expression and stress-related traits have also been characterized. Marraccini et al., 2012, identified more than 40 candidate genes up or down regulated under drought acclimation. Other genes in coffee were also found involved in different biological pathways in response to nitrogen starvation, salt stress, heat stress, cold stress, and humidity (e.g. *CaM6PR*, *CaPMI* and *CaMTD* in mannitol biosynthesis, *DREB* genes in abscisic acid synthesis) (de Carvalho et al., 2013; de Carvalho et al., 2014; Torres et al., 2019).

Direct relationship between environment and fitness traits was also observed: Anim-Kwapong and Boamah, 2010 showed significant correlations between soil types (sandy, clay, or humus) and climatic variables (rainfall and temperature -related variations) with agronomic traits such as fruit-set, number of fruits per node, number of flowering and fruiting nodes, girth and number of primary branches.

These various studies collectively provide a more comprehensive insights into the genetic mechanism underlying local adaptation in *C. canephora*. With high genetic diversity distributed across a wide range of environmental conditions, they have high potential of genetic adaptation to different environments. Indeed, Tournebize et al., 2022, observed different levels of genomic vulnerability to climate change in *C. canephora* across its wild distribution. Studying association between climate, genotype, and fitness phenotype in Robusta coffee will be useful in understanding how they adapt to the environment, and predicting the impact of climate change on the distribution and productivity of Robusta in the future.

I.3 Sequencing-based markers for coffee diversity exploration

Molecular markers are indispensable for many genetic and genomic studies to explore species diversity. High-throughput sequencing has enabled effective detection of different and powerful markers, such as single nucleotide polymorphism (SNP), insertions and deletions (Indels), transposable element (TE), structural variants (SV), and k-mers. Studies based on these markers has brought valuable insights into genetic diversity and evolutionary genomics of coffee species. We reviewed related studies in a published book chapter (reformatted below, with the addition of a new paragraph on k-mers):
Vi, T., Marraccini, P., De Kochko, A., Cubry, P., Khong, N. G., Poncet, V. (2022). Sequencing-based molecular markers for wild and cultivated coffee diversity exploration and crop improvement. In Coffee Science (pp. 213-220). CRC Press (DOI:0.1201/9781003043133-20).

I.3.1 Introduction

Before the emergence of sequencing technologies, several PCR-based approaches and methods that rely on restriction sites were applied to establish molecular markers for coffee genetic studies, such as random amplified polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), simple sequence repeat (SSR), and inter-simple sequence repeat (ISSR). Initial studies on the genetic diversity of Coffea canephora have defined two main groups (Congolese and Guinean) distributed in central and western Africa (Berthaud, 1986; Montagnon et al., 1992a), and later five and more divergent groups of wild population were determined using RFLP, AFLP, and SSR data (Dussert et al., 1999; L. F. V. Ferrão et al., 2013; Gomez et al., 2009; Kiwuka et al., 2021). Despite the fact that these markers, especially SSR, have high discriminant capacity in population genetics, they do not allow high-throughput production and only contribute to limited information in genetic and genomic studies. It is also difficult to link the different studies together because SSR markers are not easily transferable. The advancement of sequencing technologies has enabled not only high-throughput data analysis but also the availability of new molecular markers and thus has provided new insights at the whole-genome single-base level or structural level. Regarding sequencing methods, there are several, including whole-genome DNA sequencing, restriction-site associated DNA sequencing (RADSeq; Davey and Blaxter, 2010), genotyping-by-sequencing (GBS; J. He et al., 2014) and diversity arrays technology sequencing (DArTSeq). This chapter focuses on the molecular markers that can be detected by these sequencing approaches and their application and achievement on genetic and genomic studies of coffee species; also, their importance in coffee conservation and crop improvement will be further discussed.

I.3.2 Types of sequencing-based markers used in coffee

Earlier developed sequencing technologies, including Sanger, 454 sequencing, and especially next-generation sequencing (NGS), in combination with computational algorithms (bioinformatic tools), allow the detection of high-accuracy as well as high-density single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels). In coffee genetics studies, SNP markers (mostly biallelic SNPs) might be considered as one of the most commonly used and powerful types of molecular markers to date. Whole-genome DNA sequencing is used for detection of abundant genome-wide variations, whereas RADSeq and GBS could be alternatives to reduce the variation redundancy. RADSeq limits the read sequences from the restriction sites to their 300–400 bp downstream positions (Davey and Blaxter, 2010). GBS covers only the highly divergent genomic regions that are predetermined. GBS markers were used, for example, to reconstruct well-resolved phylogeny for the *Coffea* genus (P. Hamon et al., 2017). Complex reduced SNP arrays have been discovered and scored in several crop species using DArTseq (Huttner et al., 2005). DArT also allows detecting variations in specifically targeted regions. The use of these methods has several privileges: (1) lower sequencing cost, (2) simultaneous screening of multiple genotypes, and (3) preselection of variations that could be used across arrays. Despite the complexity reduction, RADSeq is capable of offering highly informative data (Davey and Blaxter, 2010). Genome-wide SNPs, besides having discrimination capacity, have been proved to be statistically associated with coffee quality traits such as caffeine content, aroma components (Sant'Ana et al., 2018; H. T. M. Tran et al., 2018; H. T. Tran et al., 2018), and resistance to diseases (Gimase et al., 2020). The wide application of SNP markers in genetic and genomic studies has led to various developments of robust bioinformatic methods for SNP detection, mostly based on Bayesian likelihood (S. Kumar et al., 2012; F. Xu et al., 2012). Long-read sequencing technologies, such as PacBio and Nanopore, because of the low coverage resulting in low-quality reads (Logsdon et al., 2020), are less effective in SNP calling.

Indels are less commonly used compared to SNPs, although they are relatively abundant along the genome. While SNPs are point mutations, Indels can concern a single base or a longer DNA sequence. They also play important roles in evolution and genetic differentiation. Indels were found to be the major type of variations in the chloroplast genome of Arabica (Min et al., 2019). Some Indels detected in the chloroplastic genome were supportive in classifying *C. arabica* and *C. canephora* (Nakagawa et al., 2019). Recently phylogenomic models have been developed to improve the use of Indels in phylogenetic analysis in species including plants (Donath and Stadler, 2018; Liu et al., 2019; Mahadani et al., 2018).

A large proportion of eukaryotic genomes is made of transposable elements (TEs) (Britten and Kohne, 1968), particularly in *C. canephora*, where the estimated proportion is over 50% (Denoeud et al., 2014). Studies of TEs in coffee species, however, have been mainly limited to their detection and classification, although some retrotransposons have shown abilities to resolve recent phylogenetic relationships (P. Hamon et al., 2011). There is currently numerous software for the identification and classification of TEs (Makałowski et al., 2019), but their accuracy is a constraint (Vendrell-Mir et al., 2019). Nonetheless, the improvement in read length and development in machine learning algorithms would leverage the efficiency in the discovery of TEs (Arango-López et al., 2017; Ewing, 2015).

Structural variation (SV), referring to Indels, duplications, interchromosomal/intrachromosomal translocations, and inversions that occur in approximately 50 bp or larger (Ho et al., 2020), is another type of molecular marker that can only be efficiently detected using whole-genome sequencing. Some of them are mediated by TEs. As is the case for TEs, long-read sequencing is more efficient in the identification of SVs compared to short-read sequencing (Mahmoud et al., 2019), but improvements are still needed in detection computational tools. SVs have been characterized and analyzed in evolutionary genetics for a number of plant species (Y. Zhou et al., 2020; Żmieńko et al., 2014); however, they are still poorly explored in coffee genomes.

Most of recent researches in coffee genomics, genetics, and phylogeny have been using SNPs because of their ease in detection and well-developed bioinformatic tools (Tranchant-Dubreuil et al., 2018), although other markers are evolutionally informative at certain levels.

I.3.3 Application in coffee phylogeny

Coffea is a genus belonging to the Rubiaceae family and strongly related with *Psilanthus* genus (A. P. Davis et al., 2006). These two genera have been assembled to include 124 described species (A. Davis et al., 2011), increasing the geographical range of *Coffea*. Before the inclusion of *Psilanthus*, *Coffea* was restricted to tropical Africa, Madagascar, the Comoros, and the Mascarenes. It also exists now in southern Asia, south tropical Asia, southeastern Asia, and Australasia.

The first nearly full-resolved phylogeny of the *Coffea* genus, including 81 species out of the 124 in the genus, was constructed based on 28,800 nuclear SNPs detected by GBS using Illumina sequencing (P. Hamon et al., 2017). The phylogeny analysis classified the studied species into two clades and four subclades which were geographical differentiated, and suggested that the biogeographical origin of the *Coffea* genus might be either Asia or Africa. This study also showed the evolution of caffeine in the *Coffea* genus and the transmission from free or low caffeine content in the ancestral species to moderate and high caffeine accumulators in west and central African species, especially *C. canephora*. Besides differentiated morphology from other coffee species, genetic distinction of *C. canephora* was confirmed by phylogenetic analysis of the 28,800 SNPs taken together with plastid phylogeny, which suggested that *C. canephora* might be still undergoing a sub-speciation process (Charr et al., 2020). Evolution of coffee has been also assessed by using TEs, for example, in *C. canephora* Pierre ex A. Froehner (Dereeper et al., 2013; P. Hamon et al., 2011), or in native *Coffea* species from Madagascar which provided results in accordance with analyses using SSR and/or SNPs (Charr et al., 2020; Razafinarivo et al., 2013; Roncal et al., 2016).

I.3.4 Application in genetic diversity analysis

Few *Coffea* species have been studied at the intra-specific level over its entire range, mainly the two cultivated species (*C. canephora*: Cubry et al., 2013, Gomez et al., 2009; *C. arabica*: Anthony et al., 2002, Scalabrin et al., 2020). Among studied *Coffea* species, *C. canephora* and *C. liberica* are the two species having widest genetic diversity, largely distributed in west and central Africa (A. P. Davis et al., 2006), and *C. arabica* has relatively low differentiation and a natural distribution centered on Ethiopia (Lashermes, 2018a).

SNP markers are widely used in genetic discrimination of populations within coffee species. Mérot-L'Anthoëne et al., 2019, have developed a high-density array of 8.5K SNPs that reveals a high level of polymorphism and was applicable for three species: *C. canephora*, *C. arabica*, and *C. eugenioides*. The SNP panel also allowed the discovery of two new *C. canephora* groups in the southern Democratic Republic of the Congo and Angola (Mérot-L'Anthoëne et al., 2019). Practically, by applying multiple SNPs selection methods such as filtering based on genome distribution, linkage disequilibrium, or minor allele frequency (ability to reveal polymorphism), only a reduced coreset of the 8.5K SNP array is sufficiently effective for the genetic diversity analysis of unidentified coffee collections (Akpertey et al., 2021; Zhang et al., 2021). To unravel the genetic origin and population structure of cultivated coffee species, one could also analyze complexity-reduced SNPs detected by DArTSeq (Garavito et al., 2016; Spinoso-Castillo et al., 2020). These reduced SNP arrays are powerful in genetic classification; however, they do not fully cover the variations of the large coffee genome and therefore a portion of high-resolved information might be lost. More than 15 million whole-genome SNPs were identified in 93 accessions of three species (*C. arabica*, *C. canephora*, and *C. excelsa* [*C. liberica* var. dewevrei]) by whole-genome NGS and mapping on the *C. canephora* reference genome (Huang et al., 2020). In addition to genetic evidence strongly supported by an abundance of variations, SNPs via whole-genome sequencing could also enable detection of genome regions responsible for the divergence (Huang et al., 2020). Some limitations in whole-genome SNPs detection are high consumption of time, cost, and computational resources, and possibly some background noises in close-related genomes.

Genetic diversity of wild and cultivated coffee accessions in different regions of the world have been evaluated using various selections of SNPs (Table I.3.1).

Table I.3.1: Recent studies on local genetic diversity in different regions using SNPs

Study	**Region**	**Species**	**Source of plants**	**Markers**
Garavito et al., 2016	Mexico and Vietnam	87 *C. canephora*	Wild and cultivated	4021 DArTSeq SNPs
Anagbogu et al., 2019	Nigeria	44 *C. canephora,* 1 *C. arabica*, 1 *"C. abeokutae"* (*C. liberica* var. *liberica*), 1 *C. liberica*, and 1 *C. stenophylla*	Cultivated	433,048 GBS SNPs
Garot et al., 2019	Reunion Island	25 *C. mauritiana*, 16 *C. canephora*, 1 *C. eugenioides*, 1 *C. liberica*, and 1 *C. bertrandii*	Wild	3953 DArTSeq SNPs
Spinoso-Castillo et al., 2020	Mexico	3 *C. canephora*, 83 *C. arabica*, and 1 *C. liberica*	Cultivated	1739 DArTSeq SNPs
Mérot-L'Anthoene et al., 2019	Africa	27 *C. canephora*, 17 *C. arabica*, and 6 *C. eugenioides*	Wild and cultivated	Coffee 8.5K SNP array (8580 SNPs)
Huang et al., 2020	China	48 *C. arabica*, 34 *C. canephora*, and 11 *"C. excelsa"* (*C. liberica* var. *dewevrei*)	Cultivated	15,367,960 whole-genome SNPs
Scalabrin et al., 2020	Ethiopia, Yemen, etc.	736 *C. arabica*	Wild and cultivated	698 GBS SNPs
Gimase et al., 2021	Kenya	91 *C. canephora*	Cultivated	2280 DArTSeq SNPs
Akpertey et al. 2021	Ghana	400 *C. canephora*	Cultivated	192 KASPar SNPs
Zhang et al. 2021	Ethiopia, Yemen, etc.	130 *C. arabica*	Wild and cultivated	96 KASPar SNPs

These studies have expanded the knowledge of the diversity of some species at the regional level, but they only provide a partial view. Coffee biodiversity has been partly preserved by in situ collections for wild species and ex situ collections for wild crop relatives (WCR; Bramel et al., 2017; Engelmann et al., 2007) (Figure I.3.1). Nevertheless, 60% of *Coffea* species are at high risk of extinction, and a large proportion of WCR is lacking ex situ collections (A. P. Davis et al., 2019). Therefore, understanding how much genetic diversity has been captured in present-day ex situ collections is indispensable to improve coffee conservation strategies. To determine genetic diversity and parentage within and between germplasms and compare them, a minimum set of key common markers and the inclusion of reference accessions would play an essential role to ensure the effective and efficient conservation and use of genetic diversity.

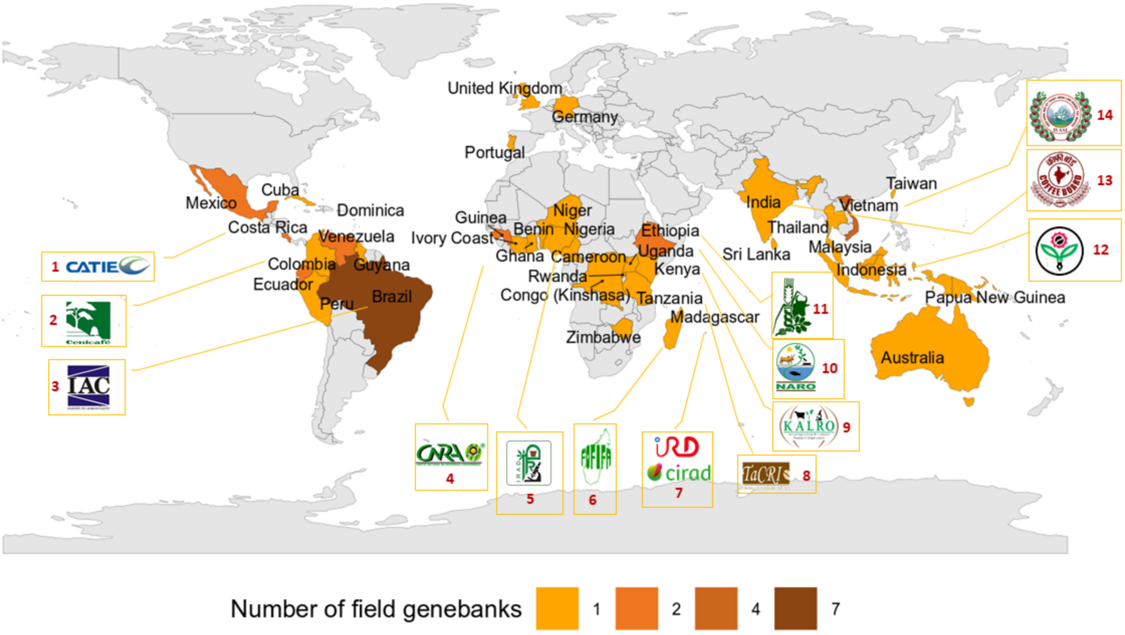


Figure I.3.1: Ex situ field collections in countries from 1986 to 2016 (using data in Bramel et al., 2017) and major coffee gene banks. (1) Centro Agronómico Tropical de Investigación y Enseñanza (CATIE); (2) Centro Nacional de Investigaciones de Café (CENICAFE); (3) Instituto Agronômico de Campinas (IAC); (4) Centre National de Recherche Agronomique (CNRA); (5) Institut de Recherche Agronomique et de Développement (IRAD); (6) Centre National de Recherche Appliquée au Développement (FOFIFA); (7) Institut de recherche pour le développement (IRD)/Centre de coopération internationale en recherche agronomique pour le développement (CIRAD); (8) Tanzania Coffee Research Institute (TACRI); (9) Kenya Agricultural and Livestock Research Organization (KALRO); (10) National Agricultural Research Organisation (NARO); (11) Ethiopian Biodiversity Institute (EBI); (12) Indonesian Coffee and Cocoa Research Institute (ICCRI); (13). Central Coffee Research Institute (CCRI); (14) Western Highlands Agriculture and Forestry Science Institute (WASI).

I.3.5 Application in genotype-phenotype association studies

The main traits of agronomic and organoleptic interest in coffee trees are resistance to pests and diseases (especially for the susceptible species *C. arabica*) and traits related to aroma and flavor in low cup quality species for *C. canephora*. Improvement in these traits is one of the keys in crop breeding and requires understanding of the relationship between phenotype and genotype.

Genome-wide association studies (GWAS), which are used to link molecular markers (mainly SNPs) to quantitative traits, have become routine in genomic analysis in plants (Purugganan and Jackson, 2021), including coffee. Several attempts have been done to identify molecular markers responsible for specific traits in coffee, such as caffeine content and aroma compounds biosynthesis. In *C. arabica*, 1444 nonsynonymous SNPs were found to be related to caffeine content, of which 11 SNPs were linked to genes involved in the caffeine biosynthesis pathway (H. T. Tran et al., 2018), and more than 20 SNPs exhibited association with aroma and flavor production (Sant'Ana et al., 2018). An equivalent number of caffeine-linked SNPs were identified in *C. canephora* and *C. eugenioides* (H. T. M. Tran et al., 2018). A number of Indels also showed high association with caffeine biosynthesis in tea plants (Liu et al., 2019), but that has not yet been evaluated in coffee species.

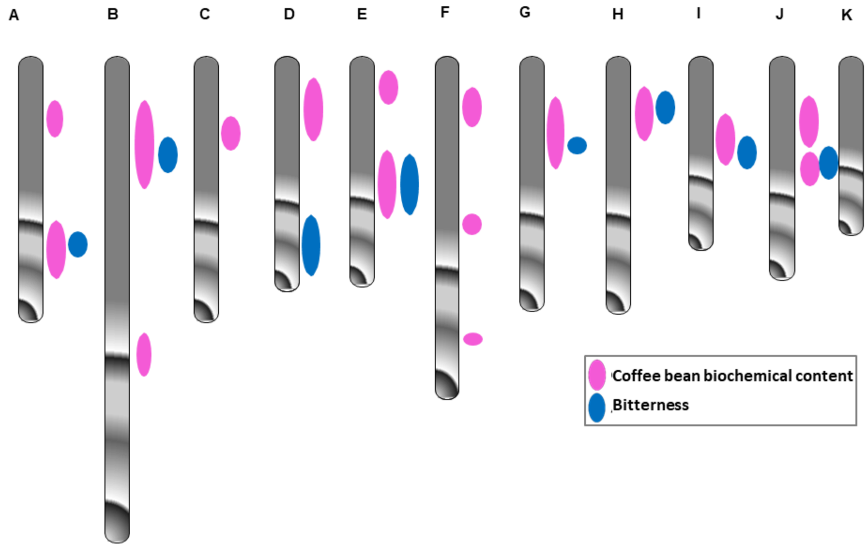


Figure I.3.2: QTLs linked to cup quality in C. canephora (Crouzillat et al., 2016). The genetic map is realized from the cross of two elite clones from ICCRI BP409 X Q121 on an F1 progeny of 92 individuals. Coffee biochemical bean content was assessed by NIR and bitterness from sensory panel analysis. Colocalization of QTLs from sensory and biochemical data could indicate the putative origin of the sensory trait analyzed.

SNPs are also robust markers for genetic linkage mapping and quantitative trait loci (QTL) mapping. Great efforts have been made, using a variety of molecular markers, to construct genetic maps of coffee species (Ky et al., 2000; Moncada et al., 2015; Paillard et al., 1996; Pearl et al., 2004), to locate QTL for productive traits (Leroy et al., 2011; Moncada et al., 2015), for resistance to diseases in *C. arabica* (Gimase et al., 2020; Romero et al., 2014), and for cup quality (Crouzillat et al., 2016) (Figure I.3.2). In addition, the availability of high-quality reference genomes could enable high-density genetic maps (Denoeud et al., 2014). The high-density SNPs can include markers that are in linkage disequilibrium with the QTL, and thus are useful in association studies of polygenic traits (Lashermes, 2018a; van der Vossen et al., 2015). However, major current QTL maps are based on SSRs and other PCR-based markers that are sparsely detected, while the advantage provided by high-density SNP markers has not yet been identified. Moreover, this approach is challenging in *C. arabica* because of its limited gene pools as a result of its probable recent single origin (Cui et al., 2020).

For genomics-assisted breeding (GAB), GWAS and molecular-marker-assisted selection (MAS) are useful in relating the phenotypes with the genetic markers. MAS underpinned by QTL mapping is only applicable in traits affected by strong-effect loci; however, with phenotypes such as resistance to diseases associated with a large number of minor-effect markers, GWAS gains more advantage as it takes all markers genome-wide into account (Lashermes, 2018b; Varshney et al., 2021).

I.3.6 Trending

Molecular markers potentially possess great information that has not been uncovered yet because of current limitations in mining and analysis tools.

Besides QTL mapping or linkage mapping, local ancestry inference and admixture mapping are further applications of genome-wide SNP identification, which are promising for association studies of polygenic traits and highly admixed populations such as in *C. canephora*. Such introgression analysis has been commonly used in tracking ancestral races in humans to study the complex admixture patterns in the Latino American population and their association with diseases (Mani, 2017; Shriner, 2013; Thornton and Bermejo, 2014), but this approach is still rarely applied in plants. A number of tools based on hidden Markov models were developed to assign the ancestry proportion at each locus by using haplotype-phased data (Leitwein et al., 2020), and some of them were evaluated to have high accuracy in plant species (Cottin et al., 2019; Dias-Alves et al., 2018). The investigation of mosaic genomes, in combination with linkage mapping, would enhance the knowledge of hybridization in self-incompatible coffee species and could be helpful in breeding programs. Most of GAB to improve adaptability in crops relies on introgression of QTL or resistance genes (Varshney et al., 2021). Introgression analysis might help to pinpoint QTL "hotspot" regions responsible for traits of interest (Figure I.3.3).

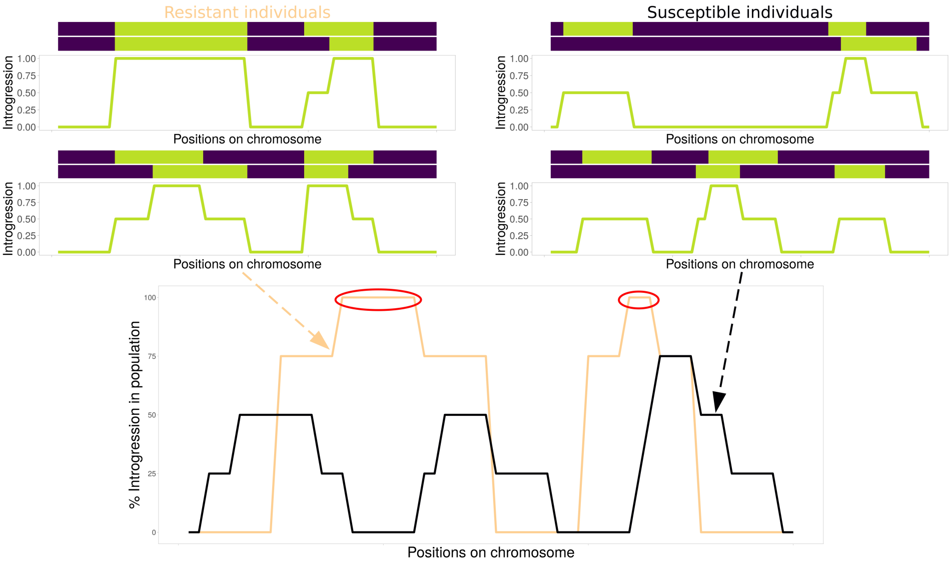


Figure I.3.3: Application of introgression detection and admixture mapping in identification of ancestry associated with disease resistance. Introgression is detected at each locus (green lines) in all individuals. Admixture mapping is constructed by using average local ancestry inference in susceptible population (black line) and resistant population (beige line). Regions where the two populations are discriminated by ancestry are spotted (red circles). This approach could be used to assess the relationship between introgression and disease resistance traits.

The combination of genotypic data and phenotypic data, moreover, could be used for genomic prediction, which is a useful tool for crop breeding. The first prediction models, using Bayesian linear regression, for coffee essential traits (bean production, leaf rust incidence, and green bean yield) were built in *C. canephora* using up to 60K SNPs detected by GBS (L. F. V. Ferrão et al., 2019). More recently, leaf rust resistance in *C. arabica* was predicted based on decision trees; however, this only used SSR and AFLP markers without the assistance of linkage maps (Sousa et al., 2020). Even though the accuracy of prediction was not high (< 70%), this is still a prospective approach in GAB and can be improved by using genome-wide dense SNP markers, increasing sample size and developing computational methods. Genomic prediction could be subsequent to genomic selection as a checkpoint to simulate and optimize the breeding efficacy (Cortés and López-Hernández, 2021). Moreover, using this approach, it would be possible to predict the adaptability of plants to different environments and climate change.

Furthermore, novel types of molecular markers could be possibly identified to fill the gap of the existing ones. Promisingly, development of a new category of molecular markers based on the sequencing of the methylome should allow a better understanding of the plant epigenetic response to biotic or abiotic stresses and the mechanisms involved in plant acclimatation and phenotypic plasticity (Bräutigam and Cronk, 2018; Corrêa et al., 2020; R. Li et al., 2020).

The most commonly-used markers have several limitations, for example, SNP markers are unable to capture variants that are not present on the reference genome (Rahman et al., 2018; Voichek and Weigel, 2020), or the identification of SVs is often challenging with short-read sequences (Ho et al., 2020). An effective alternate marker type can be k-mers, sub-strings of constant length k extracted from sequence reads (Karikari et al., 2023). Using k-mers allows reference-free and alignment-free approaches, which can easily capture variants outside of the reference genome. K-mers also cover a wide spectrum of other molecular markers, because SNPs, Indels, CNVs or SV all lead to differences in k-mer sequences (Rahman et al., 2018). Its application in GWAS has shown high statistical power and abilities to identify new genes important for adaptation and evolution in plants (Karikari et al., 2023). The use of k-mers is therefore a promising approach for other genetic and genomic studies to gain novel insights into crop genetics.

I.3.7 Conclusion

The ultimately main purpose of studying coffee genetics and genomics using molecular markers is to preserve and enhance the gene pool of high-quality coffee varieties and improve or produce varieties with desirable traits. Studying coffee evolution is important to understand their mechanism of adaptation to the changing environment and predict their divergent or convergent direction. Assessment of genetic diversity in wild and cultivated species plays a crucial role in preserving coffee species at high risk of extinction. Genomic selection is more robust with molecular marker assistance compared to conventional breeding, which is solely based on phenotyping; for example, localization of SNPs associated with disease resistance could help to optimize appropriate breeding strategies and produce prolonged resistant C. arabica varieties. However, there is still a big gap from genomic selection to breeding enhancement because phenotype is an output of complex genotype/environment interactions. Rapid development of biotechnology and bioinformatics is improving the identification of molecular markers and extending their application, allowing new aspects of the coffee genomics to be exposed and the revolution in breeding programs in the future.

I.4 Research objectives

Originating from Africa, Robusta coffee has been one of the most important crops in Vietnam, propelling the country to the world's leader in Robusta production for the last 20 years. In the future, production of Vietnamese Robusta coffee will be threatened by several factors, particularly climate change. Conservation strategies and breeding programs are urgently needed for sustainable solutions to improve coffee adaptive resilience to climate change in Vietnam. The adaptability to new environments in coffee strongly relies on genetic variation underlying local adaptation. Genetic diversity in the WASI germplasm is the main available source of climate-resilience variation for Vietnamese Robusta coffee. This genetic source might have been narrowed down from the wild gene pool, as a result of the diffusion and selection of Robusta coffee throughout the history of its cultivation. Understanding the genetic diversity available in Vietnam and its suitability to local climate will help to forecast the response of Vietnamese Robusta coffee to future climate change. This understanding will facilitate future plans for conservation and improvement.

For future conservation and breeding strategies to cope with climate change in Vietnam, it is necessary to fully understand What part of the genetic diversity from the wild source populations has been transmitted to Vietnamese Robusta coffee, and Which source genetic diversity is suitable to the predicted local climate in Vietnam in the future. The objectives of this thesis was to address these questions and was divided into three main parts.

The first part (chapter II) focused on the development of a genome-wide approach to identify wild ancestry-of-origin segments in cultivated Robusta coffee. Local ancestry inference is crucial in understanding the adaptive history and its impact on the current varietal diversity. Most of the current tools for local ancestry inference have been primarily developed and tailored for the human genome, generally requiring pure source populations, which is not always obtainable. We thus developed a framework to infer local ancestry with admixed source populations. The framework was tested on *C. canephora* whole-genome sequencing data using simulated hybrids, showing a high level of accuracy. We then applied it to analyze ten Vietnamese elite Robusta varieties. The application of this approach was further extended to a larger collection of Robusta varieties cultivated in Vietnam in the next chapter, to gain insights into the genomic make-up of cultivated Robusta and their diffusion history.

The second part (chapter III) delved into a study on genetic diversity and origin of *C. canephora* accessions cultivated in the Central Highlands. Robusta coffee was initially introduced from Africa to Vietnam through the breeding center in Indonesia, but its genetic and geographic origins remain largely unknown. Characterizing the genetic resources available in Vietnam is important in establishing effective breeding plans. The main objectives of this chapter was to trace the genetic origin of the Vietnamese cultivated accessions, assess the wild ancestor contributions to their genetic composition, and propose a core collection for potential use in future breeding. For this purpose, we studied a set of 126 accessions collected in WASI germplasm, including commonly planted accessions, elite varieties, and ancient accessions. We selected a new set of targeted SNP markers for efficient genotyping and genetic analysis. The genetic diversity of the WASI collection was compared to the diversity of wild samples from all the genetic groups in Africa. The source genetic groups contributing to the Vietnamese germplasm were identified, providing better insights into the genetic history and diffusion of Robusta coffee in Vietnam. A core set of individuals, representative of the whole collection, was chosen and re-sequenced. This sequencing data allowed local ancestry inference using the approach developed in chapter II to characterize the ancestral contributions along coffee chromosomes. Detection of ancestry segments in the core individuals enhanced our understanding of the admixture history in Vietnamese coffee. The results in this chapter offer important information for evaluating the potential uses of Vietnamese Robusta accessions in future breeding programs, and provide useful recommendations for conservation approaches.

The last part (chapter IV) evaluated the putative suitability of wild *C. canephora* accessions to local climatic conditions in Vietnam, both in the present and future. As *C. canephora* is indigenous to various regions in Africa, the wild populations might have reached an equilibrium with their local environment in their native distribution. Diffusion of Robusta to Vietnam, exposing them to new environments, can break this equilibrium and cause climatic mismatch. The main objective of this chapter was to predict which genetic materials from the wild diversity might be best suitable for the local climate of Vietnam. We first estimated the climatic shift between the native environments in Africa and those in Vietnam, using bioclimatic variables (e.g. annual precipitation, seasonality temperature). These climate differences allowed us to assess the climatic match/mismatch of the wild groups in Vietnam, according to their geographic origins. We also correlated these distances with coffee yields at the local areas to predict the effects of climate change on Vietnamese coffee production in the future. To gain more insights into the genetic drivers of local adaptation in the wild populations, we identified genetic variants associated with the bioclimatic variables, using a reference-free approach. This approach was based entirely upon the distribution of short k-mers, obtained from sequencing reads, across the sample genomes. The putative adaptive variants were then used to estimate the expected changes in allele frequencies of each sample to suit the present and future climate of Vietnam, indicating the genotypes best fitted. The results of this chapter provide a basis for proposing breeding materials likely to improve the adaptability of Vietnamese Robusta coffee to climate change.

Taken together, all the results of these studies will assist in shaping the breeding and conservation strategies, by suggesting potential breeding materials, interbreeding groups, and whether more genetic diversity should be introduced. The key findings, limitations, prospects, and implications of the project will be discussed in chapter V.

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

In this chapter, we developed a framework to infer genome segments of different genetic ancestries in cultivated Robusta coffee, and performed the analysis primarily on Vietnamese elite accessions. This chapter has been published (reformatted below):
Vi, T., Vigouroux, Y., Cubry, P., Marraccini, P., Phan, H. V., Khong, G. N., Poncet, V. (2023). Genome Biology and Evolution, 15(5), evad065.

II.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.

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.

II.2 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 a 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 and Darling, 2014). Cultivated individuals may, therefore, express novel phenotypic variability, for example, beneficial ear traits in domesticated emmers (Nave et al., 2019; H. R. Oliveira et al., 2020), or higher antioxidant activity in a new litchi cultivar (Hu et al., 2022; Zhao et al., 2020).

Population genetics tools have been widely developed to gain insight into admixture (Padhukasahasram, 2014; Pritchard et al., 2000). STRUCTURE (Pritchard et al., 2000), ADMIXTURE (Alexander et al., 2009), or other tools like sNMF (Frichot et al., 2014) are commonly used to infer global admixture proportions and ancestry per individual in a population. Yet individuals presenting the same global ancestry might differ with respect to admixture patterns at the chromosome level. More recent advances in sequencing and the application of statistical and computing technologies have also enhanced ancestry inference at this chromosome level based on LAI approaches. At a fine genome scale, the local ancestry provides better information on the mosaic genetic origin of admixed individuals, that is enhancing knowledge of demographic histories, facilitating detection of adaptive introgression, deciphering complex traits, and mapping underlying genes in admixed populations (Geza et al., 2019; Mani, 2017; Padhukasahasram, 2014; Shriner, 2013; Thornton and Bermejo, 2014). For instance, admixture mapping based on LAI of a three-way admixed human population identified a genomic region in native American ancestry linked to Alzheimer's disease (Horimoto et al., 2021; Norris et al., 2020).

Numerous LAI tools have been developed. In most of them, genotypes from putative ancestral populations (so-called "source populations") are used as a reference to infer the local ancestry of admixed individuals or tested individuals (Sankararaman et al., 2008). Most LAI models are based on hidden Markov models (HMM), including SABER (Tang et al., 2006), LAMP-LD (Baran et al., 2012), and ELAI (Y. Guan, 2014), which predict the classification of ancestries in hidden states by monitoring genotypes of the source populations (Baran et al., 2012). Other methods implement strategies based on principal component analysis, for example, PCADMIX (Brisbin et al., 2012), Markov chain Monte Carlo (Chromopainter, Lawson et al., 2012), random forest (RFMix, Maples et al., 2013), K-means (EILA, Yang et al., 2013), and other clustering methods (Wu et al., 2021). LAI models can also be categorized into linkage disequilibrium (LD)-based and non-LD-based models (Geza et al., 2019)). Most LAI software requires phased genotypes (Wu et al., 2021), which are not always accurately estimated with short-read data obtained via next-generation sequencing (NGS) technologies (Garg et al., 2021). Some tools need biological information such as genetic and physical mapping data, recombination rates, and admixture generations, or statistical parameters such as hidden states and misfitting probabilities (Geza et al., 2019).

According to previous studies comparing the performance of several commonly used software programs (Cottin et al., 2019; Molinaro et al., 2021; Schubert et al., 2020), ELAI (Y. Guan, 2014) has proven to be one of the most efficient ancestry inference tools for dealing with unphased data. This two-layer HMM and LD-based method use source populations to predict the classification in two hidden state layers—haplotypes in the lower-layer and ancestries in the upper-layer. Therefore it does not require haplotype phasing, but only prior assumptions regarding the number of haplotype clusters and admixture generations, which are needed for hidden state modeling. ELAI has been applied in studies of different plants. For instance: in perennial plants such as aspen, local ancestry signals obtained using ELAI have generated insight into the local adaptation and demographic history of European varieties (Rendón-Anaya et al., 2021). In annual crops such as wheat, ELAI has also been applied to analyze gene flow from wild emmers to bread wheat (Y. Zhou et al., 2020).

LAI tools perform more accurately when using source populations with a large sample size and high differentiation level (Cottin et al., 2019), whereas small, unbalanced structure or admixed source populations might cause erroneous ancestry assignment (Molinaro et al., 2021; Shringarpure and Xing, 2014). In practice, it is often challenging to have a perfect sampling design across source populations (Hübner and Kantar, 2021). Here we propose a solution to this problem of unbalanced and admixed individuals from source populations.

Coffea canephora Pierre ex A. Froehner, or so-called Robusta coffee, is an allogamous diploid species with high genetic and phenotypic differentiation (Berthaud, 1986; Cubry et al., 2013; de Aquino et al., 2022; Gomez et al., 2009; Kiwuka et al., 2021; Mérot-L'Anthoëne et al., 2019; Montagnon et al., 1992a). This species originates from central and western Africa (A. P. Davis et al., 2006), corresponding to two major genetic groups, that is Congolese and Guinean groups, respectively (Cubry et al., 2013; Montagnon et al., 1992a; Montagnon et al., 1998a). The Congolese group consists of five well-described subgroups: group A in Benin and Gabon, group E in the Democratic Republic of the Congo (DRC), group C in Cameroon and the western Central Africa region, group B in eastern Central African Republic (CAR), and group O in Uganda; and two recently described groups, G in Angola and R in southern DRC (Mérot-L'Anthoëne et al., 2019). The Guinean group corresponds to group D (Mérot-L'Anthoëne et al., 2019).

While most of the crop species were domesticated during the Neolithic period, over the past 12,000 years (Harlan, 1971; Larson et al., 2014), Robusta coffee cultivation and diffusion is much more recent. Robusta coffee has attracted interest as a potential cash crop since the late 19th century (Berthaud, 1986) and the species has only become globally widespread since 1900 (Montagnon et al., 1998b). Coffee research stations and breeding centers have been set up since the early 20th century, in Java, Indonesia (1900–1930), and then DRC (1930–1960), and Central Africa (1960 onward) (Cramer, 1957; Montagnon et al., 1998b). *C. canephora* breeding programs for varietal improvement have mainly been based on heterosis of crosses between the Congolese subgroup E and subgroup A or the Guinean group D (Leroy et al., 1993a; Montagnon et al., 1998a; Sant'Ana et al., 2018).

Due to the gametophytic self-incompatibility of Robusta and its intense breeding history, cultivated populations might have experienced a high level of admixture, resulting in hybrids with complex mosaic genomes. The local ancestry deconvolution approach to cultivated Robusta varieties could help gain insight into their genetic makeup and trace back their origins. This novel approach would facilitate the development of admixture mapping—that is, associating the phenotype with ancestry segments in coffee—a technique that is sometimes more powerful than conventional genome-wide association studies (GWAS) (Horimoto et al., 2021). Several key traits for coffee breeding, such as drought tolerance, that is, a polygenic trait that is also considered to be a feature in the Congolese group A (Marraccini et al., 2012; Vieira et al., 2013).

Robusta coffee was first introduced to Vietnam in the early 20th century, probably from the Congo via France (Nogent-sur-Marne acclimation gardens) or via Java (coffee breeding center) (Vanden Abeele et al., 2021). In cultivated coffee, it takes 5–8 years to reach the maximum productive stage (Wrigley, 1988), and it has been estimated to be even up to 20 years (Moat et al., 2019; Nab and Maslin, 2020). Therefore, given 100 years of cultivation, the number of generations is not expected to be higher than 20. Even though the current Vietnamese Robusta varieties are mostly supposed to have originated from Java, their ancestral genetic groups are still largely unclear. Since materials of the Javanese breeding program mainly come from DRC, Uganda, and Gabon (Montagnon et al., 1998b, the accessions historically introduced in Vietnam may have Congolese origins and putatively some extent of intergroup admixture.

In this study, we implemented the ELAI approach on cultivated *C. canephora* with unbalanced and admixed native reference populations. We inferred ancestral genotypic frequencies for these native populations to build perfect source populations for ELAI. We assessed and validated our new approaches with simulated hybrids. We finally applied an optimal framework to a set of elite accessions cultivated in the Central Highlands of Vietnam to determine their mosaic genome origins.

II.3 Materials and methods

II.3.1 Materials

We used two sets of accessions: (1) 55 previously sequenced wild *C. canephora* accessions from Africa (Tournebize, 2017; Tournebize et al., 2022); and (2) ten newly sequenced cultivated *C. canephora* accessions from Vietnam. The wild African samples are representative of the native range of *C. canephora*. The cultivated samples from Vietnam are recognized as elite plants and conserved in the germplasm bank of the Western Highlands Agriculture Forestry Science Institute (WASI).

II.3.2 Sequencing, mapping, and SNP calling

The ten Vietnamese individuals were sequenced using Illumina Hiseq X Ten PE 150bp. The 55 samples from Africa were obtained from GenBank and analyzed with the new ten 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 (De Kochko, 2018) using BWA mem 0.7.2 (H. 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 (Genomic Variant Call Format) 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., 2017). 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 10bp 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).

II.3.3 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 and François, 2015). 10% randomly chosen SNPs were used to assess the number of genetic groups. The optimal number of ancestral groups (K) was determined using cross-entropy criteria over ten iterations with K ranging from 1 to 10. Individuals were assigned to a given cluster at an 80% ancestry threshold.

Pairwise FST between the genetic groups (restricted to individuals with >80% ancestry) was computed using the R package StAMPP (Pembleton et al., 2013). Hundred bootstraps across loci were performed to assess the significance. We also performed a PCA in the R package *LEA* (Frichot and François, 2015).

II.3.4 Inference of local ancestry

We inferred the local ancestry of cultivated Robusta using ELAI (Y. Guan, 2014), which was suitable for our unphased data and did not require any additional biological information. The source populations must be specified when using ELAI. To overcome limitations due to the unbalanced structure of wild populations that include admixed individuals, we tested an alternative method using population structure analysis. This new method was validated by simulation of known hybrids, while the ELAI accuracy was assessed and optimal parameters determined. Based on these results, we built a framework to detect the local ancestry in Vietnamese-cultivated Robusta accessions.

II.3.5 Source population

Genotypes from ancestral groups of wild accessions must be defined for the purpose of ELAI analysis (Y. Guan, 2014). Given the small and unbalanced sample size of the wild populations and some evidence of admixed ancestry between groups in the wild accessions, we generated new synthetic populations exhibiting genotypic frequencies similar to those estimated in the ancestral pools. To this end, we used ancestral genotypic frequencies inferred using the sNMF analysis (Frichot et al., 2014). For each chromosome, we ran sNMF on whole-chromosome SNPs of all the African and Vietnamese individuals, with ten iterations and the optimized K value. The best run over the ten iterations was chosen based on cross-entropy criteria. We applied the *G* function (R package *LEA*) on the result of the best run to retrieve a G matrix containing genotype frequencies inferred in each different ancestral group for all diploid SNPs. This G matrix was then used as a probability matrix for randomly choosing genotypes at each site by groups using the sample function in R (R Core Team, 2022).

We checked if the generated genotypes were representative of the ancestral groups by performing a further sNMF analysis jointly on simulated and real accessions using random 100K SNPs on chromosome 1 (with optimal K and ten iterations). The ancestral proportion of each simulated individual was obtained from the run with the lowest cross entropy.

II.3.6 Simulated Hybrids

We simulated hybrids with known admixture levels in order to determine optimal parameters and assess the accuracy of the ELAI approach with the simulated source populations in our *C. canephora* model. We chose two accessions from the wild African set (BGQ07 and 20738) representative of two divergent genetic groups to simulate three different hybrids with different admixed segment sizes. Each hybrid had admixture segments of different lengths (50 kb, 500 kb, 1 Mb, 2 Mb, and 5 Mb), and in homozygous (both alleles originating from one of the ancestral groups) and/or heterozygous form (one allele from each of the two ancestries). In the homozygous regions, the hybrid genotypes were copied from one respective progenitor, while at heterozygous loci each of two alleles was drawn randomly from the alleles of the two parents. Simulations were based on chromosome 1 SNPs.

II.3.7 Sets of ELAI parameters

We determined the optimal ELAI parameters by running the software with varying parameter values, including the number of haplotype clusters (c), that is lower clusters, the number of admixture generations (mg), and the set of SNPs used for the analysis. We tested three values (5, 15, and 25) for the number of haplotype clusters, and four values (5, 10, 20, and 30) for the number of admixture generations. We also used four different sets of SNPs: (1) randomly selected 10 K SNPs, (2) randomly selected 100K SNPs, (3) about 11K SNPs resulting from randomly selecting one SNP in every nonoverlapping 5kb window (referred to as the “even SNP set”), and (4) whole-chromosome SNPs, that is the “all SNP set” comprising 1,133,736 SNPs. The average SNP densities in the four datasets were 1.5, 15.1, 1.8, and 203.0 SNPs per 10 kb, respectively.

In summary, we performed 48 ELAI runs with different combinations of parameters, and each run used 20 expectation maximization (EM) steps (Y. Guan, 2014). To reduce the computational cost, for the run with whole-chromosome SNPs, the analysis was performed by splitting the chromosome into consecutive SNP chunks so that each subset contained a maximum of 100K SNPs. The ELAI results obtained on the SNP subsets were then concatenated.

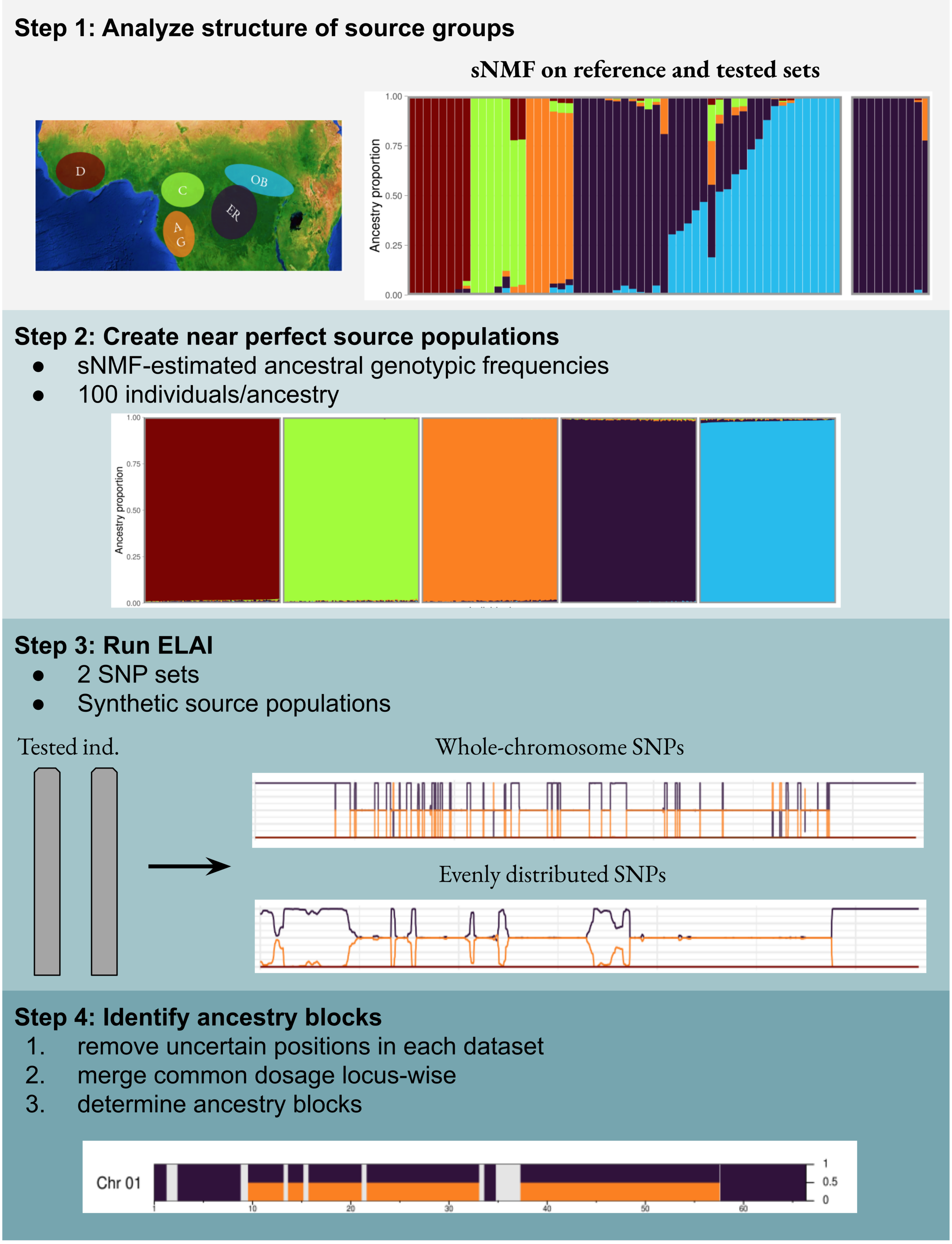
II.3.8 Assessment of the ELAI inference accuracy based on simulated hybrids

Each ancestry group was defined for the simulated hybrids as ancestry dosage = 1, where the two alleles were copied from the first parent (first genetic group); ancestry dosage = 0, where both alleles were from the second parent (second genetic group); and ancestry dosage = 0.5, where the alleles were derived from both parents (50% admixture). As we knew the true allelic dosages of the simulated hybrids, we could compare them to those inferred via ELAI.

The ELAI performance was assessed by correlation and root mean square error (RMSE) metrics. A correlation was the average Pearson's correlation between the estimated and true dosages. The RMSE between the estimated and true dosages of each admixture tract was calculated and averaged for different segment lengths (50 and 500 kb, and 1, 2, and 5 Mb).

II.3.9 Framework to infer local ancestry in Vietnamese cultivated Robusta coffee

Based on the ELAI validation data, a workflow (Figure II.3.1) was developed to detect local ancestry in cultivated Robusta accessions and applied to the ten Vietnamese Robusta accessions. ELAI was performed for each chromosome, with the simulated ancestral populations as the ancestry source using the whole-chromosome SNP set and the evenly distributed SNP set. Three independent ELAI runs with 20 EM steps were conducted for each individual and SNP was set to obtain average results. The ELAI inferences of these sets were then combined to obtain the final ELAI.



Step 1: Analyze structure of source groups

Step 2: Create near perfect source populations

- sNMF-estimated ancestral genotypic frequencies
- 100 individuals/ancestry

Step 3: Run ELAI

- 2 SNP sets
- Synthetic source populations

Step 4: Identify ancestry blocks

1. remove uncertain positions in each dataset
2. merge common dosage locus-wise
3. determine ancestry blocks

Figure II.3.1: Framework for LAI of cultivated C. canephora. ELAI was performed for each individual chromosome and involved three main steps. Step 1: analyzing the genetic structure of the ancestral group, by performing sNMF on the reference set and tested set. Step 2: simulating source populations based on sNMF-estimated ancestral genotypic frequencies. Step 3: running ELAI on the tested individuals using two marker sets, that is whole-chromosome SNPs and evenly distributed SNPs. Step 4: merging the ancestry dosages inferred in the two SNP sets to determine the final consensus inference of the target chromosome.

II.4 Results

II.4.1 Characterization of genetic groups

A total of 11,919,576 high-quality biallelic SNPs were obtained in all of the 55 African and 10 Vietnamese individuals. We assessed the structure of African native groups by performing genetic structure analysis using sNMF on a set of 1,191,957 randomly picked SNPs.

The African individuals were classified into five groups (Supplementary Figures S-II.9.1 and S-II.9.2) that could be linked to geographical origins: a West African group with accessions from Guinea, Ghana and Côte d'Ivoire (group D), a group with accessions from Cameroon (group C), a group with accessions from Gabon and Angola (group AG), an East African group with accessions from Uganda and CAR (group OB), and finally the last group consisted of Democratic Republic of the Congo accessions (group ER). Most pairwise F_{ST} values between the five genetic groups were high and ranged from 0.39 to 0.55, except for the F_{ST} = 0.22 between ER and OB (Supplementary Table S-II.9.1). This strong structuring was also confirmed by PCA analysis (Supplementary Figure S-II.9.3).

II.4.2 ELAI accuracy assessment

The size of the African reference set was small, with an unequal number of individuals (unbalanced structure), while 15 individuals presented some extent of admixture (>20% admixture) (Supplementary Figure S-II.9.2). We built near-perfect source populations for the five groups based on ancestry genotype frequencies. All of them had perfect ancestry coefficients (>97%) relative to their respective groups, as expected (Supplementary Figure S-II.9.4). The artificial source populations were then used for the assessment of ELAI performance in detecting simulated hybrids (Supplementary Figure S-II.9.5).

Using simulated hybrids, we found that our approach achieved accurate inferences with high correlations (r^2) ranging from 0.859 to 0.997 (Figure II.4.1), regardless of the set of parameters used. The lowest squared correlation (r^2 = 0.859) corresponded to ELAI runs using all SNPs with c (number of lower clusters) = 5 and mg (number of admixture generations) = 30. All other ELAI runs had r^2 values >0.9. The number of lower clusters and admixture generations did not have marked impacts on the ELAI accuracy, except when all SNPs were used. Conversely, the use of higher parameter values and SNP numbers increased the ELAI run time and memory usage (Supplementary Table S-II.9.2).

Among the four SNP sets tested, the overall accuracy was higher for ELAI runs with the 100K SNP set, while slightly decreasing with the other SNP sets, 10 K SNPs, evenly distributed 1 SNP/5 kb SNPs, and all SNPs (when mg = 20 and 30), respectively. These results were in line with the fact that ELAI accounts for the background LD when detecting the haplotype structure, so a higher number of SNPs provides more haplotype information and thus greater ancestry assessment accuracy. However, using whole-chromosome SNPs did not improve but instead slightly reduced the ELAI accuracy as it might cause background noise or false inference with short segment lengths (<500 kb).

Despite the high r^2 values, some false dosages were observed in the simulated introgression segments where the ancestry switched on both alleles compared to the flanking sequence. We computed the root mean square error (RMSE) between true and estimated dosages in the homozygous introgressed regions and compared it among a range of introgression sizes (0.05, 0.5, 1, 2, and 5 Mb).

Compared with the larger introgression tracts, RMSE values were higher for introgression tracts <1Mb, except for some runs using 100 K SNPs and the whole SNP set (Figure II.4.2 and Supplementary Figure S-II.9.6).

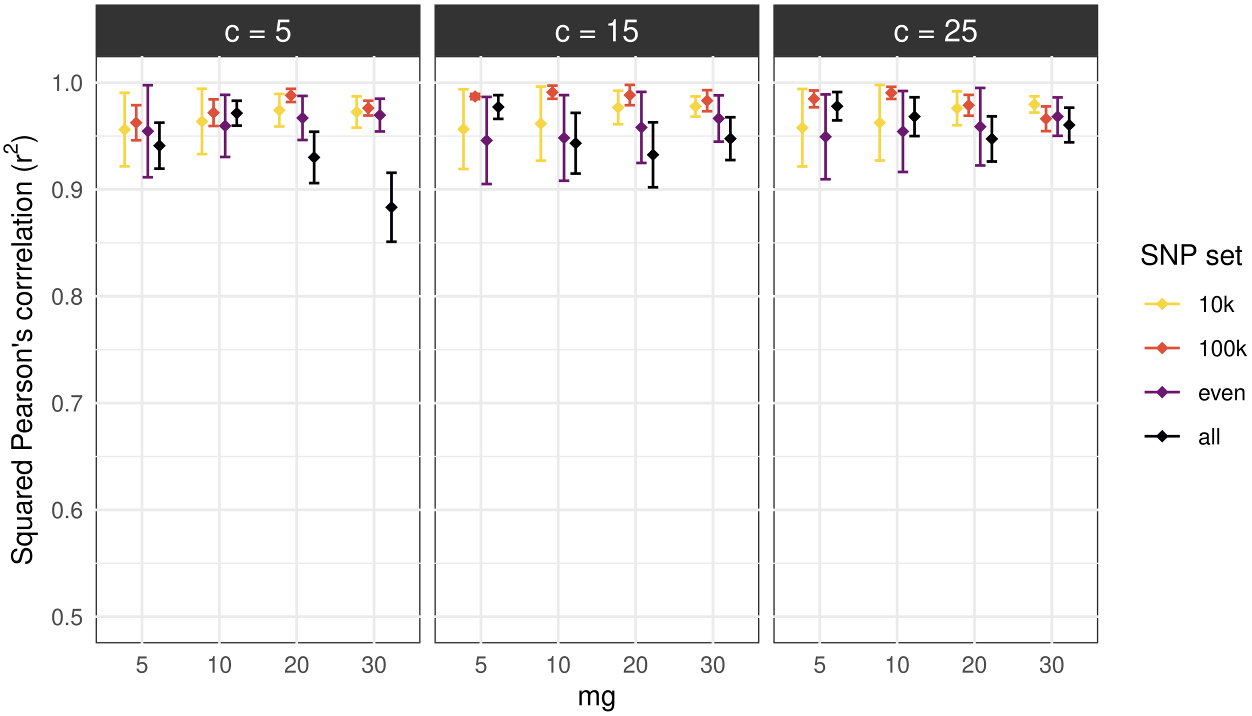


Figure II.4.1: ELAI accuracy in inferring local ancestry in simulated hybrids. The plot shows correlations between the true local ancestry dosage and ELAI dosages with different parameter numbers: number of lower clusters (c = 5, 15, and 25), number of admixture generations (mg = 5, 10, 20, and 30), and different SNP sets (10K SNPs, 100K SNPs, evenly distributed SNPs—1 SNP/5 kb, and all SNPs—1M SNPs). Each point represents an average of correlations computed for the three simulated hybrids, and error bars represent the standard deviation.

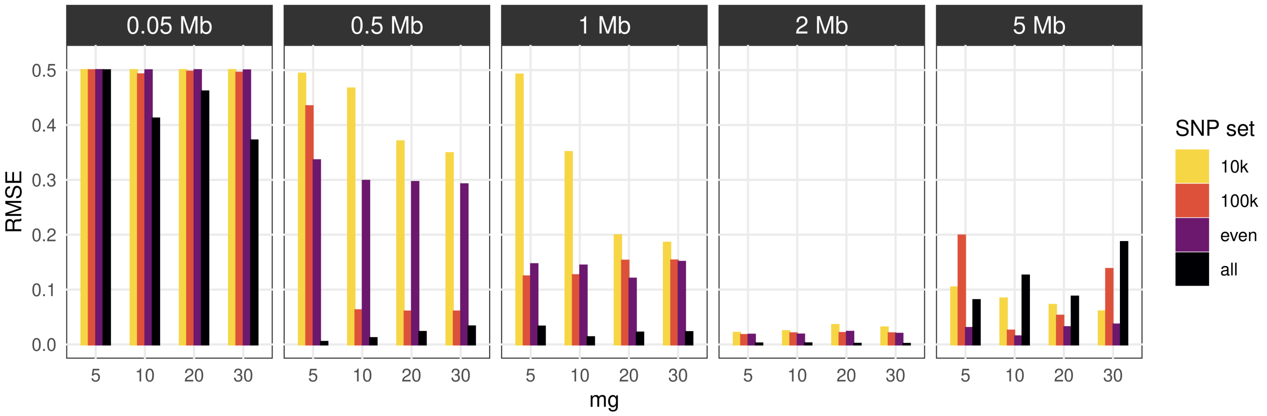


Figure II.4.2: Error in detecting simulated introgression tracts of different lengths in simulated hybrids. RMSE was calculated between the true dosage and ELAI dosages of the simulated hybrids, for different homozygous introgressed segment lengths (0.05, 0.5, 1, 2, and 5 Mb). Each panel shows, for each tested length, the RMSE values (y-axis) for ELAI runs with the numbers of lower clusters (c = 5), different numbers of generations (mg = 5, 10, 20, and 30) on the x-axis, and four different SNP sets (10K SNPs, 100K SNPs, evenly distributed SNPs—1 SNP/5 kb, and all SNPs—1M SNPs).

RMSE values of around 0.5 in some cases indicated that it was likely that only one haplotype had been correctly assigned along the introgression tracts. For introgression sizes 500 kb, 1 Mb, and especially 2 Mb, ELAI runs using whole-chromosome SNPs had the lowest RMSE values (<0.03), thereby indicating that the estimation was highly accurate. We observed an RMSE increase in larger introgressions (5 Mb) associated with false inferences of small fragments inside longer introgressed fragments. For 5 Mb admixture tracts, the least erroneous inferences (RMSE ranging from 0.01 to 0.04) were obtained in runs using the even SNPs set with c = 5 or 25, and in all, SNP sets with c = 15 (Supplementary Figure S-II.9.6).

In summary, the ELAI method was highly accurate in assessing the ancestry deconvolution of the artificial hybrids using the simulated source populations, with good confidence for admixture tracts of >1 Mb length. The required parameters (number of lower clusters and admixture generations) only had a minor effect on the detection, while the SNPs chosen for analysis had more marked impacts on the admixture segment inference accuracy. The validation enabled us to define the parameters for the application of the method on cultivated Robusta individuals.

II.4.3 Optimized framework and application for inference of the local ancestry of the tested Vietnamese Robusta cultivated accessions

Based on our validation and optimization results using simulated hybrids, we developed an LAI framework that encompassed ELAI to efficiently study the admixture origin in Robusta coffee (Figure II.3.1). For each chromosome, we performed ELAI using two SNP datasets (a set of evenly distributed SNPs, and another of whole-chromosome SNPs), with simulated ancestral groups serving as source populations. The lower cluster number was set at five as this factor did not influence the detection but did reduce the run time and memory usage (Supplementary Table S-II.9.2). We set the number of admixture generations at 20, which reflected the maximum possible number of generations of the cultivated accessions. Common results between the two datasets (evenly distributed SNPs and whole-chromosome SNPs) were then considered as the final LAI.

The results of the two SNP sets were pooled in three steps. First, as the theoretical dosage of a given ancestry at an SNP locus is either 0, 0.5, or 1 but the dosage inferred by ELAI can be any value between 0 and 1, we approximated the ELAI-inferred dosage at each SNP with respect to the theoretical values, that is the dosage was set at 0 if the inferred dosage was in the [0, 0.1) 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 “undetermined” otherwise. Second, the approximated dosages in the two SNP sets were compared locus-wise and also classified as “undetermined” if they were not equal. Finally, ancestry blocks were determined if contiguous positions had the same dosage and the distance between adjacent positions was not >1 Mb; and <1 Mb segments were also classified as “undetermined”.

This framework was then applied for LAI of the Vietnamese accessions.

A total of 94% to 100% of the genome could be assigned (Supplementary Figure S-II.9.7) for all of the ten tested Vietnamese accessions. Undetermined regions were due to disagreement between the results of the two datasets, or the uncertainty in the ELAI inferred dosage (ancestry dosages within 0.1–0.4 or 0.6–0.9 ranges, or ancestry tracts <1 Mb were treated as uncertainties). Some accessions represented the same undetermined region on chromosome 10 (Supplementary Figure S-II.9.7) because this region was assigned with a dosage of 1 to the ER group by the even SNPs set but with a dosage of 0.5 for both ER and AG groups by the whole-chromosome SNP set.

Based on these ancestry blocks, the global ancestry inference of the tested individuals could be estimated as follows: for each ancestry, the overall proportion was the sum of all block dosages/genome assembly size (585 Mb), with the block dosage being the length of an ancestry block × ancestry inference for the block. The global ancestry coefficients detected by our ELAI framework were generally similar to the results estimated by sNMF (Supplementary Figure S-II.9.2). Nine accessions presented >99% ancestry in the DRC-native ER group (Supplementary Figure S-II.9.7). Two of these accessions, that is TR5 and TR15, had minor admixture proportions in one haplotype, that is 1.6 Mb of group AG on chromosome 4 and 1.2 Mb of group C on chromosome 10, respectively.

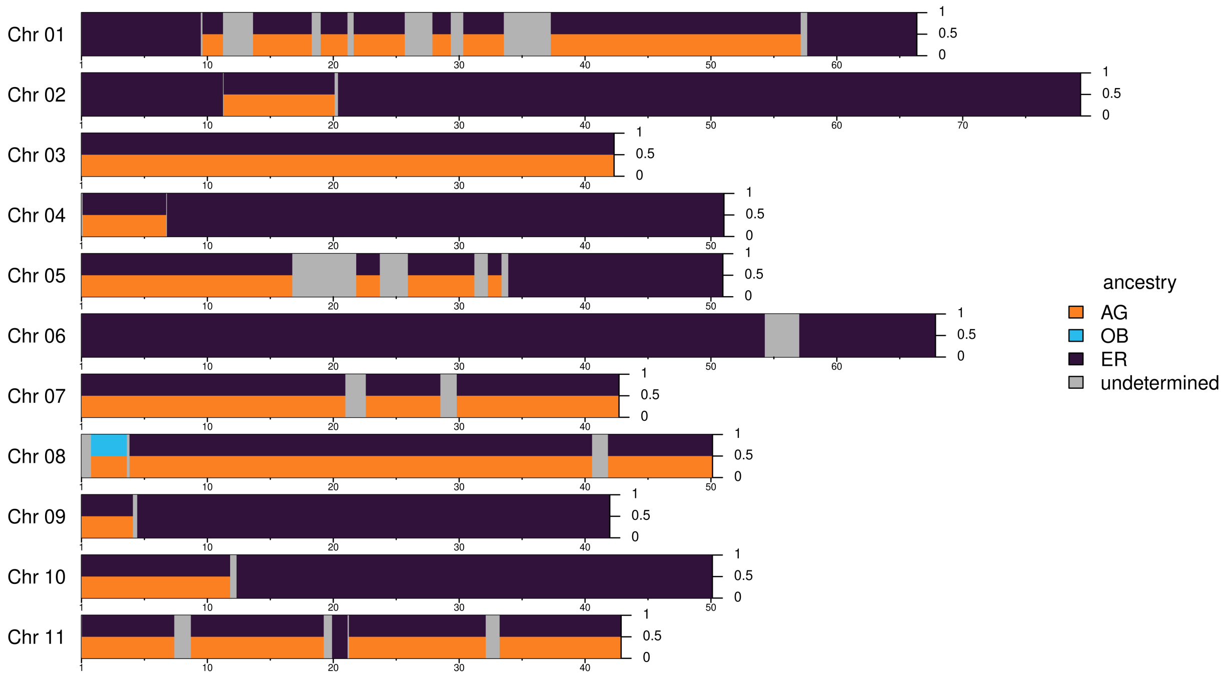


Figure II.4.3: LAI of the Vietnamese TR6 accession. Each bar presents the consensus local ancestry dosage (y-axis) along the positions (x-axis) of each chromosome. The x-axis labels are in the Mb unit. The inferred ancestral groups—ER (from DRC), AG (from Benin, Gabon, and Angola), and OB (from Uganda and Central Africa), are denoted by colors. The gray portions are undetermined regions.

The TR6 accession genome consisted of segments from two ancestral groups, that is, ER and AG, with 72% and 22% of the genome, respectively (Figure II.4.3 and Supplementary Figure S-II.9.7). The admixture patterns varied in different parts of genome: chromosomes 3, 7, 8, and 11 were completely heterozygous (excluding the undetermined segments); chromosomes 4, 9, and 10 had a single admixture tract at the terminal end of the chromosome; chromosome 1, 2, and 5 had admixture segments separately distributed along the chromosome; while, exceptionally, chromosome 6 did not present any admixture signal. A small haplotype fragment of 2.8 Mb length on chromosome 8 was assigned to group OB, which accounted for only 0.2% of the genome. The local ancestry pattern suggested that TR6 resulted from recombination events between individuals of the two main Congolese AG and ER subgroups, backcrossed with the ER group.

II.5 Discussion

An optimized framework to infer local ancestry in Robusta coffee

In this study, we developed an LAI framework implementing the ELAI method, which was performed with high accuracy for ancestry deconvolution of admixed individuals using derived source populations and with good confidence for admixture tracts of >1 Mb length.

In particular, we implemented a simple method to overcome the lack of bias in native reference individual sampling, which was efficient for ELAI on both simulated and real data inference. Our method was based on ancestral genotype frequencies, which can be directly obtained from unphased genotypic data using sNMF. To our knowledge, this is the first time that such an approach has been proposed for unphased data. Another method combining WINC-ChromoPainter with the non-negative least squares approach has been developed to analyze LAI when there are very few reference individuals and no source population simulations (Molinaro et al., 2021). A test of this method on real data showed that WINC was comparable to or could outperform ELAI in certain admixture scenarios when only two individuals per source population were used. However, WINC requires phased data and recombination maps, which are not always available for other datasets, especially our dataset. ELAI was later adapted to use a large number of admixed samples (a cohort set) to compensate for the lack of a pure source population (Q. Zhou et al., 2016). The cohort set was down-weighted to not outweigh other training source sets. This method could be applicable if a large cohort sample size is available. We acknowledge that our simulation method did not preserve LD in the native populations, which might explain the misassignment or uncertainty in the inference of small ancestry tracts in admixed individuals. However, the uncertainty only accounted for a minor portion of the genome (<10%) and the global ancestry findings of our approach were close to the global ancestry obtained by sNMF.

We also assessed the number of haplotype clusters and admixture generations required by ELAI to predict the ancestry model. The Robusta coffee breeding program was launched about 100 years ago, which could be considered as relatively recent compared to most other crops, and therefore the maximum number of Robusta generations out of Africa and especially in Vietnam was estimated at most 20. We tested the methods with 5, 10, 20, and 30 generations. Y. Guan, 2014, found that a higher number of admixture generations improved the inference smoothness. In many studies using ELAI, the number of haplotype clusters is often set at 5-fold the number of source populations. These parameters were shown to have an impact on inference in human genomes (Y. Guan, 2014), but we did not clearly observe such an effect in our simulations. Even when the number of lower clusters was set at the number of source populations, so each source population was linked to only one haplotype in the higher clusters, we could still quite accurately infer the ancestry source. Our approach to generate ideal source populations will only keep LD linked to population structure. As we observed high genetic differentiation in *C. canephora*, we certainly had high LD linked to the population structure. Lower differentiation between ancestral populations might lead to lower performance of the approach we have proposed here.

ELAI has been shown to be a highly robust LAI tool (Cottin et al., 2019; Molinaro et al., 2021; Schubert et al., 2020). Indeed, our results obtained using simulated *C. canephora* hybrids also illustrated its high accuracy for detecting >1 Mb ancestry blocks, even with a small portion (1%) of whole-genome SNPs, that is 10 K SNPs out of the 1.1 M whole-genome SNPs on chromosome 1. Yet in our framework, we combined inferences obtained with SNP sets of two densities (whole-chromosome SNPs and evenly distributed SNPs, i.e., 203 and 1.8 SNPs per 10 kb, respectively) to enhance the inferred ancestry confidence.

Robusta origin, diffusion, and LAI

The native African Robusta individuals were classified into five groups that could be linked to geographical origins, and this genetic structure was perfectly congruent with previous study findings (Tournebize et al., 2022). The latest classification of *C. canephora* using 8.5 K SNP arrays Mérot-L'Anthoëne et al., 2019) led to an eight-group classification, but the differentiation between groups O and B, E and R, and A and G was low, so they were clustered in this study and our previous study (Tournebize et al., 2022). The clustering of individuals of closely related groups was also due to bias toward one group when the other contained a small number of individuals.

Using source populations derived from these native groups, we ran our optimized framework method on a sample of ten elite cultivars from Vietnam. Inference of these test Robusta accessions revealed that all of them originated from the Congolese groups ER and AG. Nine accessions shared a common ancestry of group ER, and one likely came from a hybrid between the two Congolese ER and AG groups, backcrossed with ER. Previous studies on other Robusta accessions in Vietnam also identified their Congolese origin (Akpertey et al., 2021; Garavito et al., 2016). Garavito et al., 2016, used DArTseq SNPs and found six Vietnamese accessions from the Congolese E group (included in our ER group). Akpertey et al., 2021, used KASP (Kompetitive Allele Specific PCR) SNPs and found 33 Vietnamese accessions distinguished from Côte d'Ivoire and Togo accessions (putatively the Guinean group), but no reference for the Congolese groups was used in that study. These inferences are in line with historical coffee breeding data.

These accessions, which are recognized elite Robusta accessions in Vietnam, could serve as potential breeding materials for varietal improvement. The TR6 accession genome was found to be composed of two ancestral ER and AG groups, accounting for 72% and 22% of the genome, respectively. The Congolese genetic group E was previously found to present advantageous phenotypic characteristics such as good aroma and low acidity, high leaf rust resistance, but with susceptibility to drought and twig borers (Montagnon et al., 1998a). In contrast to group E, genetic group A has very high twig borer resistance and drought tolerance, but is only moderately resistant to leaf rust, and sometimes exhibits lower cup quality (Montagnon et al., 1998a). Hybridization of these two groups might produce accessions with heterosis characteristics combining these advantageous agronomic traits.

Inference of wild ancestry segments in the cultivated accessions could also enable downstream analyses such as admixture mapping of important traits, or genomic selection for breeding programs. Breeding strategies could now also be tailored for different purposes. Reciprocal recurrent selection between the Congolese group and Guinean group (group D) has been used to improve yield and vegetative vigor (Leroy et al., 1993a; Montagnon et al., 1998b; L. N. L. Oliveira et al., 2018), while recurrent selection within hybrid populations was more effective for enhancing disease resistance (Alkimim et al., 2021). Therefore, studies on the genetic origin, especially the LAI of Robusta materials available in collections, could also boost the efficiency of coffee breeding programs.

This approach could also be adapted to other species when studying admixed populations with a low number of reference individuals.

II.6 Acknowledgments

T.V. was funded by Ph.D. grants from the French Embassy in Vietnam and the ARTS program (IRD). This research was funded by the Ministry of Science and Technology of Vietnam (MOST) under grant number NVQG-2020/ĐT.15. The authors acknowledge the ISO 9001 certified IRD i-Trop High Performance Computing (member of the South Green Platform) at IRD Montpellier for providing HPC resources that contributed to the research results reported in this paper. URL: https://bioinfo.ird.fr/-http://www.southgreen.fr.

II.7 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 LAI 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.

II.8 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 (this 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).

II.9 Supplementary data

Table S-II.9.1: Pairwise F_{ST} between ancestral groups and Vietnamese set. The ancestral groups contained wild individuals with one ancestry accounting for > 80%. Pairwise F_{ST} was computed using 1.1M random genome-wide SNPs. All p-values were < 0.01.

Group	OB	ER	D	C	AG
ER	0.22				
D	0.55	0.54			
C	0.43	0.40	0.44		
AG	0.43	0.41	0.50	0.39	
Vietnam	0.23	0.01	0.54	0.41	0.41

Table S-II.9.2: Run time and memory usage of ELAI runs with different sets of parameters and SNP numbers.

		Time (hours)			Memory (Gb)		
		c			c		
snp	mg	5	15	25	5	15	25
10k	5	1.9	12.4	46.2	0.1	0.9	2.5
	10	1.9	7.2	32.6	0.1	0.9	2.5
	20	1.6	12.2	28.8	0.1	0.9	2.5
	30	1.3	7.6	27.7	0.1	0.9	2.5
100k	5	20.7	164	551	1.2	9.1	24.8
	10	19.2	155	327	1.2	9.1	24.8
	20	18	104	399	1.2	9.1	24.8
	30	16.8	103	309	1.2	9.1	24.8
even	5	1.7	8.7	41.5	0.1	1.1	2.9
	10	2	14.5	53.7	0.1	1.1	2.9
	20	1.6	13.4	45	0.1	1.1	2.9
	30	1.6	13.5	40.8	0.1	1.1	2.9
all	5	166.6	1372.8	4462.2	13.6	104	281.6
	10	164.5	1,020	3749.3	13.6	104	281.6
	20	151.4	1309.2	4396.4	13.6	104	281.6
	30	141.6	976.8	4270.8	13.6	104	281.6

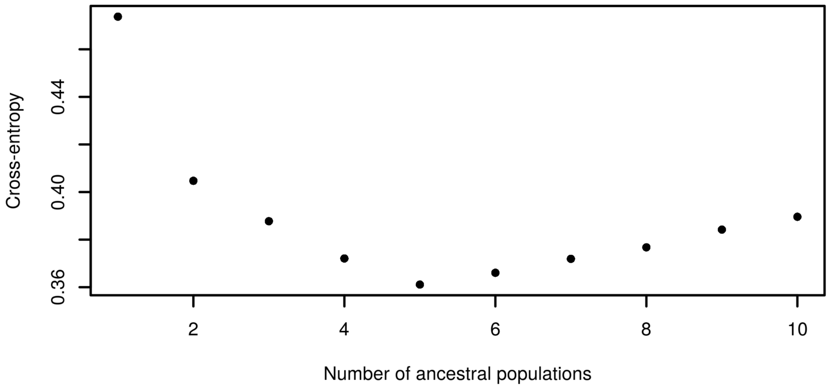


Figure S-II.9.1: Cross-entropy criterion for each tested number of ancestral groups (K) on the set of African references and Vietnamese accessions. Cross-entropy values of sNMF runs, with K = 1 to 10 and 10 iterations for each K, using 1.1M random genome-wide SNPs, were plotted. Lower cross-entropy implies a better prediction capacity.

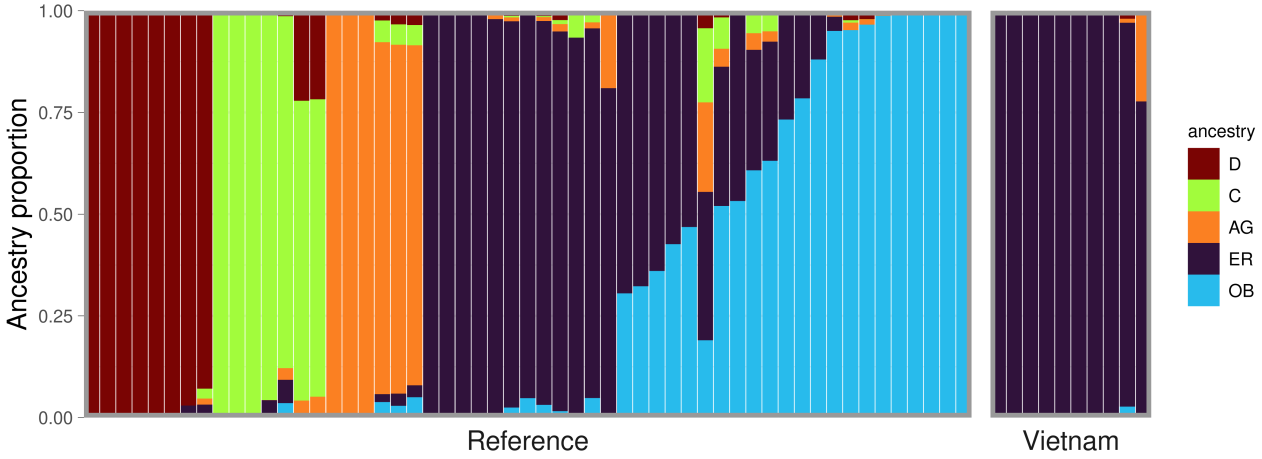


Figure S-II.9.2: Ancestry proportions for each African reference individual and tested Vietnamese individual. Five ancestral groups were estimated from the best sNMF run using 1.1M random genome-wide SNPs based on cross-entropy criteria. Proportions of ancestral groups (inferred by color) per individual are presented in each vertical bar. The ancestries were named according to the genetic groups associated with the geographical distribution of the African individuals which had > 80% ancestry proportions: group D (red) from Guinea, Ghana and Côte d'Ivoire, C (green): Cameroon, AG (orange): Gabon and Angola, OB (blue): Uganda and CAR, and ER (purple): DRC. The group names were as proposed previously in Tournebize et al., 2022.

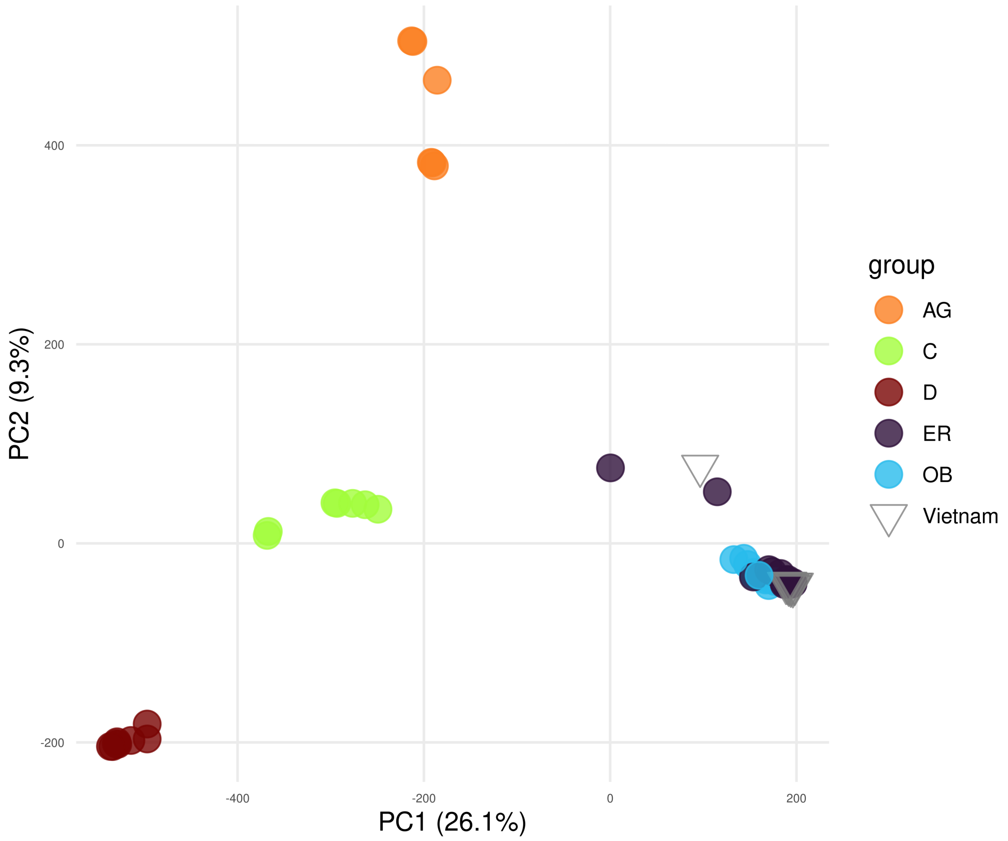


Figure S-II.9.3: Projection on PC axes 1 and 2 of 55 African wild accessions and 10 Vietnamese individuals using 1.1M random genome-wide SNPs. The African individuals were labeled by the most contributing ancestry based on the sNMF results. Group ER is in purple, OB in blue, C in green, AG in orange, and D in red), and Vietnamese individuals are in gray dots.

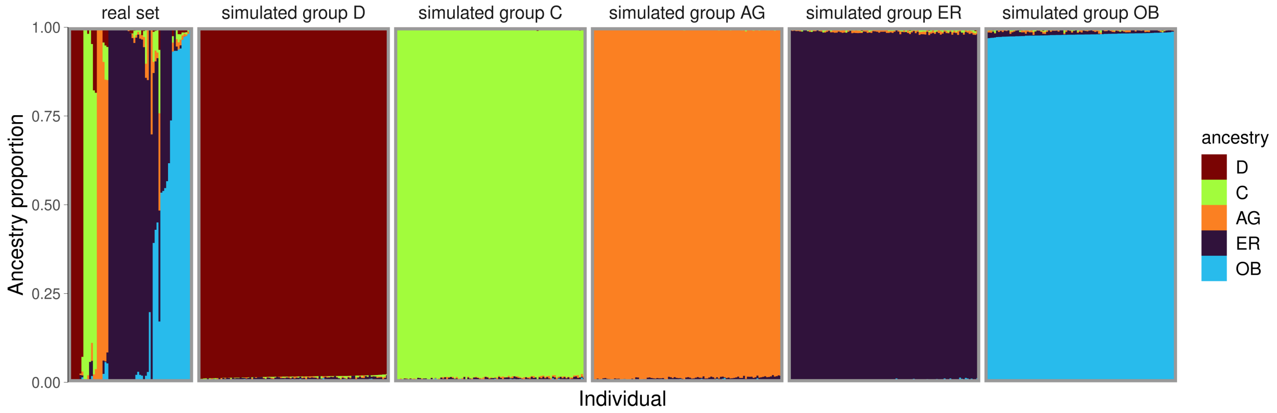
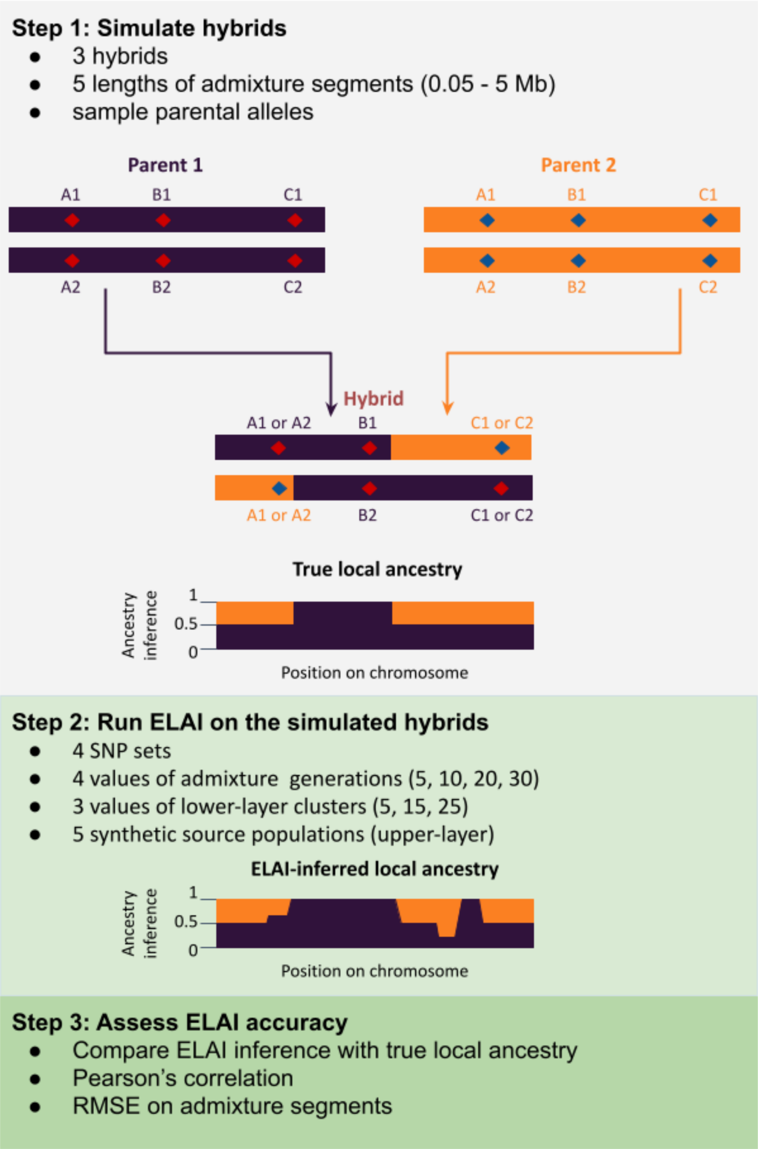


Figure S-II.9.4: Individual ancestry proportions of five simulated ancestral groups compared to the real dataset. Each of the five simulated groups consisted of 100 genotypes, corresponding to the ancestral groups as classified in Result 1. The barplot shows the proportion of the five ancestries (group ER in purple, OB in blue, C in green, AG in orange, and D in red) per real and simulated individual, estimated by sNMF using random 100K SNPs on chromosome 1. The average ancestral coefficient in the simulated individuals was 0.99 ± 0.004.



Step 1: Simulate hybrids

- 3 hybrids
- 5 lengths of admixture segments (0.05 - 5 Mb)
- sample parental alleles

Step 2: Run ELAI on the simulated hybrids

- 4 SNP sets
- 4 values of admixture generations (5, 10, 20, 30)
- 3 values of lower-layer clusters (5, 15, 25)
- 5 synthetic source populations (upper-layer)

Step 3: Assess ELAI accuracy

- Compare ELAI inference with true local ancestry
- Pearson's correlation
- RMSE on admixture segments

Figure S-II.9.5: Hybrid simulation steps and test ELAI performance. Step 1: creating 3 hybrids with different admixture patterns of lengths 50 kb, 500 kb, 1 Mb, 2 Mb and 5 Mb, by sampling alleles from two chosen parents. Step 2: running ELAI on the simulated hybrids with different sets of parameters and SNPs, using the synthetic source populations. Step 3: comparing the local ancestry inferred by ELAI and the true inference by Pearson's correlation and assessing the accuracy in known admixture segments by RMSE.

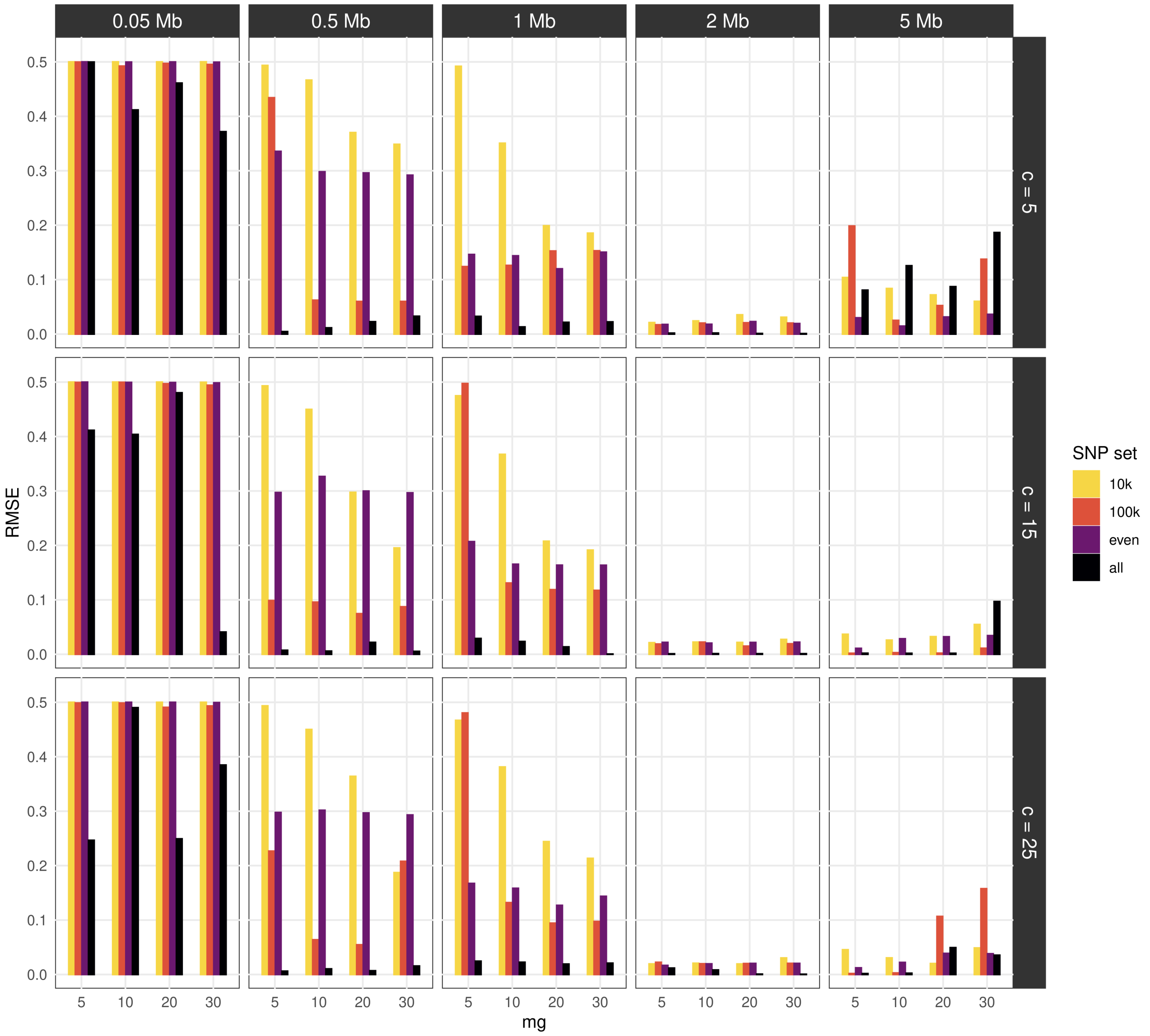


Figure S-II.9.6: ELAI errors in detecting true introgression tracts of different lengths. RMSE were computed between the true dosage and ELAI dosages of the simulated hybrids at introgression tracts differed by lengths (0.05, 0.5, 1, 2, and 5 Mb). The panel columns show results of different admixture tracts, and the rows show results in different numbers of lower clusters (5, 15, 25). Each plot presents the RMSE values on the y-axis for ELAI runs with different numbers of generations (5, 10, 20, 30) on the x-axis, and 4 SNP sets (10K SNPs, 100K SNPs, evenly distributed SNPs – 1 SNP/5 kb, and all SNPs - 1M SNPs, represented in yellow, orange, violet and black points, respectively).

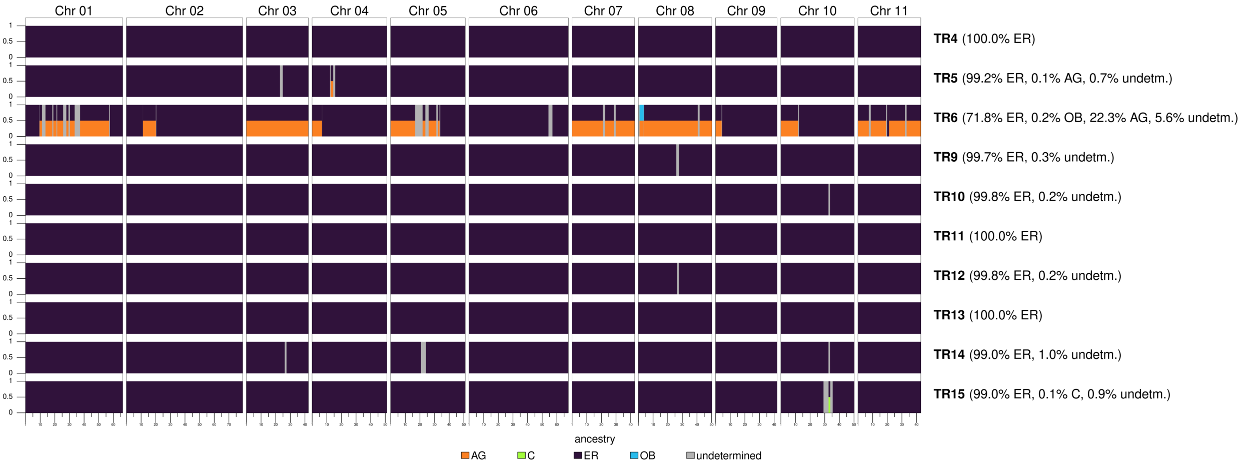


Figure S-II.9.7: Local ancestry inference of the 10 Vietnamese accessions. Each row presents the consensus local ancestry dosage (y-axis) along the positions (x-axis) in each genome. The x-axis labels are in the Mb unit. Each row presents the consensus local ancestry dosage (y-axis) along the positions (x-axis) in each genome. The inferred ancestral groups - ER (from Democratic Republic of Congo), OB (from Uganda and Central Africa), and AG (from Benin, Gabon, and Angola), are denoted by dark purple, blue, and orange, respectively. Uncolored (gray) portions are the undetermined regions. Global ancestry proportion, computed from the local ancestry, for each individual is shown on the label.

III Diffusion from the Congo basin and hybridization with other origins shaped the diversity of Vietnamese Robusta coffee

In the previous chapter, the genetic origin of some elite Vietnamese Robusta accessions has been revealed, showing the Congolese origins. However, there might be new diversity that have not been discovered yet. Therefore, in this chapter, we investigated in a large collection of *C. canephora* from a germplasm bank in the Central Highlands of Vietnam. The study would provide an overview of the genetic diversity of Vietnamese Robusta populations, and more insights into their breeding history, which might be useful for establishing future breeding programs.

Vi et al.

III.1 Abstract

— Coffea Canephora (Robusta) exhibits a genetic diversity strongly structured in its native habitat, the tropical forests of Africa. Only a part of this diversity contributed to the diffusion of Robusta across the world. In this study, we traced the African origin of Robusta accessions cultivated in the Central Highlands of Vietnam.

— A total of 126 accessions from the Vietnam germplasm collection were characterized, including ancient, elite and local cultivated clones. Their genetic diversity and origin was inferred through comparison with wild reference samples, using a new set of 261 genome-wide SNPs. Accessions maximizing genetic distance and allelic richness were selected in a core set. Ancestry segments at the chromosome level of each core individual were detected by using whole-genome sequencing data.

— Ancestral origin from a Congolese group of the Congo basin (group ER) presented in all the Vietnamese accessions, although in varying proportion. Admixtures, with different distributions on the genome, from at least one other groups (D in Guinean region, AG in Atlantic coastal region of central Africa, and OB in Congo basin) were found in 31 individuals.

— Vietnamese Robusta coffee was primarily derived from Congolese Robusta materials, but there was also a diversity from other genetic sources dispersed in backcrossed hybrids. These source groups have been widely used in crossbreeding to develop ”elite” clones. The results provided better understanding of the genetic resources available in Vietnam, which will be useful for establishing sustainable breeding strategies.

Societal Impact Statement

Robusta coffee was introduced from Africa to Vietnam in the 1900s, and in few decades has become one of the most important crops in Vietnam with a massive impact on the economy and life of local farmers. Robusta production in Vietnam is currently threatened by climate variation and aging of the trees, which urgently requires sustainable conservation approaches. Understanding the genetic diversity of Vietnamese cultivated varieties, compared with the wild diversity, is a prerequisite for maintaining the genetic resource and improving coffee production through breeding.

III.2 Introduction

Robusta coffee is the production of *Coffea canephora* Pierre ex A. Froehner, a diploid species with the widest native distribution in the *Coffea* genus (A. P. Davis et al., 2006). Its diversity has been structured up to 8 genetic groups corresponding to different regions in west and central Africa (Cubry et al., 2013; Gomez et al., 2009; Mérot-L'Anthoëne et al., 2019). The most recent study in classifying wild C. canephora has identified group D (also known as Guinean group) in Guinea and Côte D'Ivoire, A (also known as Congolese subgroup 1 (SG1), or Conilon) in Gabon and Togo, B in southern Central Africa Republic (CAR), C in Cameroon, E (also known as Congolese subgroup 2, SG2) in the Democratic Republic of the Congo (DRC), O in Uganda, G in Angola, and R in Sankuru region of DRC (Mérot-L'Anthoëne et al., 2019; Montagnon et al., 1992b; P. Musoli et al., 2009). However, the genetic differentiation between samples of some groups are sometimes low, for example between groups O and B, E and R, or A and G, which are often clumped together (Mérot-L'Anthoëne et al., 2019; Tournebize et al., 2022; Vi et al., 2023).

Most of the modern crops have undergone thousands of years of domestication since the Neolithic revolution (Meyer et al., 2012). But widespread cultivation of *C. canephora* has just started recently from its first breeding programs (Berthaud, 1986; Cramer, 1957; Montagnon et al., 1998a). *C. canephora* was initially cultivated at a small scale in the late 19th century in Gabon, Angola, Uganda, and the Sankuru region of the DRC (Chevalier, 1929; Durand et al., 1898; Montagnon et al., 1998b; Vanden Abeele et al., 2021). A first large-scale breeding program was established in Java, Indonesia, in the 1900s (L. F. V. Ferrão et al., 2019). Then, the next breeding centers returned to Africa from the 1930s (Montagnon et al., 1998b). The breeding programs were mostly based on parental accessions from three main groups, E, A and D, to produce elite varieties and disperse selected materials to other parts of the world (Montagnon et al., 1998a; Montagnon et al., 1998b). Therefore, even though cultivated Robusta populations might not have experienced drastic domestication bottlenecks or a strong selection process as other domesticated crops (Gaut et al., 2018), they are likely originating from a limited number of diversity sources. Indeed, many germplasm banks were found to have narrow genetic diversity compared to the wild, and high-quality "elite" varieties were mostly developed from crosses between Guinean and Congolese groups with limited parental materials (Anagbogu et al., 2019; M. A. G. Ferrão et al., 2021; Loor Solórzano et al., 2017; Montagnon et al., 1998b; Ramadiana et al., 2021; Teixeira Alexsandro et al., 2017). Late and limited sources of breeding make it easier to trace back to the origin of hybrids at the chromosome level. Recently, an approach to detect introgression locally at the chromosome level using SNP markers (local ancestry inference - LAI) has been adapted on cultivated *C. canephora* (Vi et al., 2023), allowing better understanding of the genetic diversity and breeding history.

Studying the genetic diversity of different resources, including cultivated and wild varieties, landraces, and germplasm collections containing elite and/or mutant varieties is important for crop improvement (Swarup et al., 2021). As wild crop relatives (WCRs) may present beneficial alleles that are missing in cultivated populations, they have been commonly used in breeding to enhance crop performance and adaptation for the past decades (Coyne et al., 2020; Dempewolf et al., 2017; Renzi et al., 2022). For instance, genes involved in blast, leaf blight resistance, and drought were identified in wild rice and successfully introgressed into cultivar lines (Babar et al., 2022; Hajjar and Hodgkin, 2007; Singh et al., 2022; Yamada et al., 2020). In *C. canephora*, the different wild origin groups also represent a great potential of genetic resources for coffee improvement. These groups have high genetic differentiation, and may present adaptive variants to a large range of environmental conditions due to their wide native range (Mérot-L'Anthoëne et al., 2019; Tournebize et al., 2022). In addition, wild coffee relatives can be screened for candidate loci and genes, such as those related to biotic stress (e.g. leaf rust; Nonato et al., 2021) and abiotic stress (e.g. drought, heat, and salt; de Carvalho et al., 2013; de Carvalho et al., 2014; Torres et al., 2019), which are useful for enhancing coffee resilience. They can also vary in agronomic traits across geographical locations and genetic groups (Kiwuka et al., 2023). For example, accessions from group E in the Congo basin were found to have high productivity, while accessions from group A in Gabon or group D in the Guinean region had high resistance to leaf rust and drought (Berthaud, 1986; Leroy et al., 1993a; Montagnon et al., 1992a; Montagnon, Leroy, Cilas, et al., 1993; Montagnon and Leroy, 1993; Moschetto et al., 1996).

Ex situ collections, such as germplasm banks, an approach to conserve the wild biodiversity, are also an important source for commercial breeding programs (Kjaer et al., 2001; Swarup et al., 2021). More than 50 ex situ Robusta collections over about 40 countries have been established, storing at least 30,000 accessions, with the CNRA collection in Côte d'Ivoire being the largest and most representative to our knowledge (Cubry et al., 2013; Dussert et al., 2003; Gomez et al., 2009; Labouisse et al., 2020; Montagnon et al., 1992a; Montagnon, Leroy, and Yapo, 1993). The genetic characterization of these worldwide collections have expanded the knowledge of diversity at the regional level, but they only provide a partial view. To estimate how much of the wild genetic diversity has been captured in present-day ex situ collections and to compare them, a minimum set of key common markers and the inclusion of reference accessions would play an essential role to ensure the effective and efficient conservation and use of genetic diversity (Vi et al., 2022). Large collections may contain redundant individuals, their diversity may not be fully utilized, and the management is not always efficient (Van Hintum et al., 2000). Therefore, the concept of representative core collections was established, in order to optimize the individuals representing the genetic diversity of larger populations with minimum redundancy (Brown, 1989; Frankel et al., 1984; Odong et al., 2013). In addition, core collections are useful for identifying elite lines, beneficial genes or quantitative trait loci, as applied in many crops such as rice, soybean, or cucumber (Gu et al., 2023). Sampling of core collections can be based on one or multiple features including passport, phenotypic, and genotypic data (Amalraj et al., 2006; Van Hintum et al., 2000), and various strategies, e.g. maximizing allelic richness, or maximizing phenotypic distance (De Beukelaer et al., 2018). Several Robusta core collections have been suggested based on genetic data. Gomez et al., 2009, defined two core sets of CNRA accessions that capture maximum diversity using the principal component score strategy (PCSS) (S. Hamon et al., 1998) on both SSR and RFLP markers. Leroy et al., 2014, proposed different core collections from Robusta accessions maintained in four field collections, by maximizing genetic diversity and/or allelic diversity at 13 microsatellite markers, and more recently, Verleysen et al., 2023, using SNPs, also developed different core collections by combining maximized genetic diversity with minimized entry-accession distance or maximized entry-to-nearest-entry distance, from 730 accessions in the INERA Yangambi Coffee Collection. These core collections would be valuable for conservation and breeding programs in Africa in the future.

Robusta coffee was first introduced to Vietnam in the 1900s (International Coffee Organization, ICO, 2019) but coffee production has only developed over the past several decades as a major export-oriented industry to make Vietnam the second largest producer of coffee in the world. Currently, Vietnam is the world's biggest Robusta producer (contributing 40% of global Robusta production), with more than 600,000 hectares cultivation concentrated in the Central Highlands (ICO, 2019). Robusta coffee is one of the country's major cash crops, contributing 10%-15% of agricultural gross domestic product in Vietnam (ICO, 2019). The largest germplasm collection is located at the Western Highlands Agriculture and Forestry Science Institute (WASI) in Dak Lak province, with nearly 200 accessions (Bramel et al., 2017). Since the 1990s, various local Robusta varieties have been selected from local farms and used to develop elite varieties at WASI (Phan, 2017). However, up to date, little is known about the origin and genetic diversity of this germplasm bank. Only a few Robusta accessions at WASI have been genotyped using SNP arrays and whole-genome sequencing (WGS), and most of them were found to be closely related to the Congolese (Akpertey et al., 2021; Vi et al., 2023) or a Congo-Uganda group (Garavito et al., 2016).

Vietnamese Robusta coffee is facing the risk of production decrease in the future. First, it is predicted that climate change will reduce the most suitable existing coffee lands in the Central Highlands by more than 50% by 2050 (Bunn et al., 2015; ICO, 2019). Consequences of climate change, such as rising temperatures, are also associated with the growth of pests and diseases in coffee (Baker, 2016). Secondly, the majority of Robusta trees in Vietnam are currently more than 15 and 20 years old (ICO, 2019), and will become economically unviable when they reach 25 years of age (Moat et al., 2019). Moreover, other abiotic and biotic factors also negatively affect Robusta coffee in the Central Highlands, e.g. polluted and exhausted soil due to inefficient irrigation and intensive cultivation (ICO 2019), and increasing nematode infection (P. Q. Trinh et al., 2009; Q. P. Trinh et al., 2019). Therefore, there is an urgent need to develop new Robusta plants with better adaptation to climate change, and resistance to biotic and abiotic stresses.

Understanding the genetic diversity of the WASI collection and developing core collections are the fundamental steps for improving Robusta varieties in Vietnam. In this study, 126 Robusta accessions were collected in the WASI germplasm bank, including elite varieties, ancient accessions, and other local accessions. A set of wild African accessions covering the species distribution range were used as genetic reference. Our study aimed to (1) understand the origin and genetic diversity of Vietnamese Robusta coffee in relation to the wild diversity, using a new Kaspar SNP genotyping set, then (2) suggest a representative core collection to maximize genetic diversity and minimize genetic redundancy, and finally (3) detect admixture segments, using whole-genome sequencing data in the core individuals, to assess the diversity at the chromosome level. The results were useful for highlighting the potential use of current germplasm varieties and suggesting future breeding strategies.

III.3 Materials and Methods

III.3.1 Plant material

A total of 126 Vietnamese Robusta accessions were collected in WASI, including 10 elite accessions created and recognized by WASI (previously sequenced in Vi et al. 2023), 11 ancient accessions which were introduced early in Vietnam, and other accessions collected from local gardens. The elite accessions were designated as clones TR4, TR5, TR6, TR9, TR10, TR11, TR12, TR13, TR14 and TR15. Some of these clones are the most favored and widely cultivated by Vietnamese coffee growers in recent years, especially clones TR14 and TR15 which were found to have late ripeness and adapt very well to changes in climate (Phan, 2017). The ancient accessions “Gx”, were collected from the Dak Lak Museum in 2020, probably aged about 100 years old but no related information could be found.

A set of 127 African Robusta accessions with known genetic groups, including 17 hybrids, were used as reference. They were obtained from previous studies (Mérot-L’Anthoëne et al., 2019; Tournebize et al., 2022) or the present study.

III.3.2 KASPar genotyping data

A new set of SNPs developed for KASPar genotyping was derived from the 8.5K SNP arrays (Mérot-L’Anthoëne et al., 2019), using the African reference set. In this dataset, 4 individuals were obtained from the 8.5K SNPs array data in Mérot-L’Anthoëne et al., 2019, 68 individuals were genotyped using the method in Mérot-L’Anthoëne et al., 2019, and 55 individuals were obtained from the resequencing data in Tournebize et al., 2022. By combining their genotypes at the 8.5K SNPs (Mérot-L’Anthoëne et al., 2019), biallelic SNPs with <5% missing data, and minor allele frequencies >0.05 were retained. The remaining SNPs were thinned by 40kb distance using vcftools 0.1.16 (Danecek et al., 2021).

The selected SNPs were submitted to LGC Biosearch Technologies for KASPar genotyping (Cuppen, 2007) of the Vietnamese accessions (excluding the sequenced elite ones). Genomic DNA was extracted from leaves using the sbeadexTM mini plant kit. Flanking sequences of 120 bp both upstream and downstream of the SNPs were used for primer design.

A total of 261 SNPs were obtained by KASPar genotyping in 116 WASI accessions, with 29 duplicates for quality control. For the remaining accessions, the genotypes at the same loci were extracted from their previous genotyping or sequencing data.

III.3.3 Genotyping analysis

To assess the genetic diversity of the Vietnamese accessions and their relationship with the wild accessions, we performed PCA and neighbor-joining clustering based on the Euclidean genetic distance of all the individuals using the 261 SNPs. Descriptive statistics of the African and Vietnamese groups, including allelic richness, observed heterozygosity, expected heterozygosity, and inbreeding coefficient, were computed using the R packages *adegenet* (Jombart, 2008; Jombart and Ahmed, 2011) and *hierfstat* (Goudet, 2005). To analyze the genetic structure, we performed sNMF (Frichot et al., 2014) for K ranging from 1 to 10 with 100 iterations. All the analyses were performed in R (R Core Team, 2022).

III.3.4 Core collection

To obtain an optimal representativeness as well as eliminate redundancy of the Vietnamese collection, we selected a representative set with the size of 45 individuals, using the R package *Core Hunter 3* (De Beukelaer et al., 2018). As 10 accessions have been recognized as elite varieties and used mainly in cultivation in Vietnam (Phan, 2017), they were forced to be included in the core set. Selection of the remaining 35 individuals were based on maximizing both Euclidean genetic distance and allelic diversity in the core.

To evaluate the core set representativeness, we generated 1000 sets of 45 random individuals from the whole collection, and compared them with the core set for expected heterozygosity and allelic richness.

III.3.5 Sequencing and SNP calling

The sequencing data included 55 African reference samples obtained from Tournebize et al., 2022, 10 elite accessions obtained from Vi et al., 2023, and 35 core individuals newly sequenced using DNBseq PE 150 (DNA extraction, library construction, and sequencing performed by BGI Hong Kong). All the raw sequencing data were cleaned by trimming read bases Q20 using cutadapt 3.1 (Martin, 2011). Variant calling was performed following GATK Best Practices recommendations for germline short variant discovery. Variants were then filtered by quality, depth, missing data, singletons and doubletons. Only biallelic SNPs were retained for further analysis. Details of SNP calling and software used are described in Supplementary Figure S-III.8.1.

III.3.6 Whole-genome SNP analysis

To assess the genetic structure of both the Vietnamese and reference sequenced samples, we performed sNMF (Frichot et al., 2014) on a set of genome-wide SNPs, which were filtered by minor allele frequency (MAF) > 0.05 and thinned by distance of 5 kb. sNMF was run for K ranging from 1 to 10 with 10 iterations, and the optimal K was determined based on the cross-entropy criterion.

To track admixture segments along the genome, we inferred the local wild ancestry-of-origin at the chromosome level for each individual in the Vietnamese core set. We used the genetic groups classified by the sNMF analysis as sources of ancestry and followed the approach developed in Vi et al., 2023. Briefly, this method applies the ELAI tool, an efficient local ancestry inference method based on a two-layer hidden Markov model (Y. Guan, 2014), to track genome segments of different ancestral origins in the case of multiway admixture.

III.4 Results

III.4.1 Genetic diversity and origin

From the 8.5K array SNPs (Mérot-L'Anthoëne et al., 2019), 268 biallelic SNPs distributed across all the 11 chromosomes were chosen for genotyping. These SNPs allowed us to efficiently discrimiate genotypes from different sources. For the Vietnamese Robusta accessions, 116 genotypes were obtained by KASPar genotyping, and 10 genotypes were obtained from re-sequencing data (Vi et al., 2023). For the African reference samples, 127 genotypes at the 268 loci were composed of genotypes extracted from re-sequencing data (Tournebize et al., 2022), and genotyping data using 8.5K array data (Mérot-L'Anthoëne et al., 2019). The quality of the KASPar SNPs was assessed using 29 duplicated samples, which showed agreement (percentage of samples with exactly replicated genotypes at a locus) at 99.5% SNP alleles on average. Finally, we obtained genotypes of all the Vietnamese and African individuals at 261 SNPs with missing data <5%.

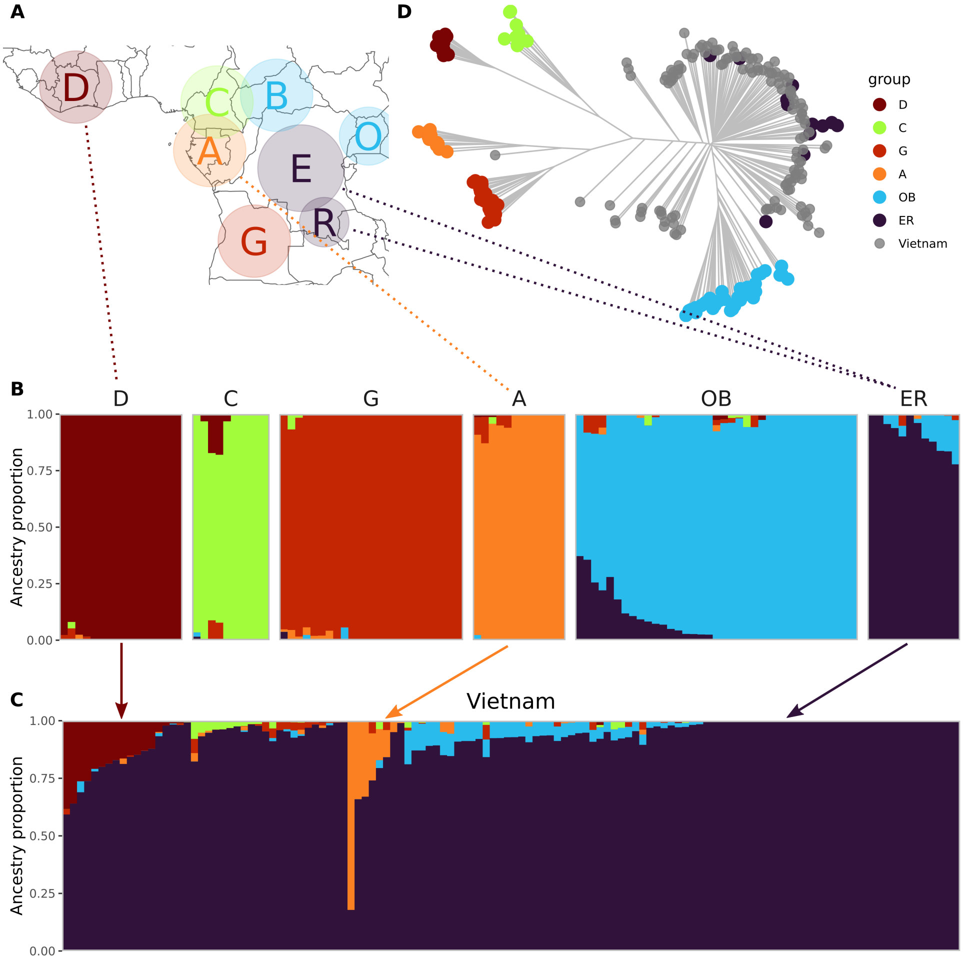


Figure III.4.1: Genetic origin, diversity and structure. (A) Geographical distribution of the wild genetic groups in west and central Africa (adapted from Mérot-L'Anthoëne et al., 2019). Genetic structure of all the individuals, (B) for 110 African individuals and (C) for 126 Vietnamese individuals. Ancestry proportion of all the individuals were obtained from sNMF analysis with K = 6 using 261 SNPs. (D) Neighbor joining clustering of all the individuals based on Euclidean distance.

Using these 261 genome-wide SNPs, we analyzed the genetic diversity of the Vietnamese Robusta collection (126 individuals) and the African groups (110 individuals). The genetic structure of all the individuals was analyzed using sNMF with K ranging from 1 to 10. The cross-entropy criterion did not show a clear optimum value for K (Supplementary Figure S-III.8.2). With K = 6, the African groups were well structured. The six groups were named by the genetic groups corresponding to their origin (Mérot-L'Anthoëne et al., 2019): group D in west Africa, group C in Cameroon, group G in Angola, group A in Gabon and Benin, group OB composed of 2 previously described groups in Uganda and CAR, and group ER composed of 2 previously described groups in DRC (Figure III.4.1A and B). All of the Vietnamese accessions exhibited a proportion of the merged group ER at certain levels (Figure III.4.1C). The majority of them (97 individuals) presented > 90% ancestry from group ER, while six of them were composed of 12-82% of group A, and thirteen individuals with 12-38% of group D. Ten ancient accessions in the collection also had 86-100% of group ER, with 10% or 12% of group OB in two accessions.

Table III.4.1: Summary statistics of genetic diversity of the Vietnamese Robusta collection and the genetic groups in Africa. The African groups were classified based on sNMF results, and the individuals with >= 70% ancestry proportion were assigned to the corresponding group. N = number of individuals; AR = average allelic richness; Ho = observed heterozygosity; He = expected heterozygosity; Fis = inbreeding coefficient

Statistics	D	C	G	A	OB	ER	Vietnam
N	16	10	24	12	32	12	126
AR	1.18	1.45	1.28	1.31	1.42	1.51	1.62
Ho	0.04	0.11	0.08	0.1	0.12	0.17	0.18
He	0.05	0.14	0.08	0.1	0.15	0.18	0.19
F_{IS}	0.08	0.19	0.02	0.06	0.19	0.06	0.03

The relationship between the Vietnamese cultivated individuals and the wild accessions was assessed based on their pairwise Euclidean distances (Figure III.4.1D). African individuals from the six groups defined by sNMF also grouped into long branches. Groups ER and OB were at closer distance, which was congruent with the structure analysis as they shared some ancestry proportions. Most of the Vietnamese genotypes were closely clustered together and with the accessions originating from the DRC, the group ER, while some others were spread at closer distances to the remaining groups. These results were also consistent with PCA results. The first two PCs, which explained 43.6% of the total variance, also showed the clustering of Vietnamese individuals with the wild groups from the center of Africa (Supplementary Figure S-III.8.3).

Table III.4.2: Differentiation coefficient (F_{ST}) between the wild and Vietnamese groups. All the Fst values are statistically significant (p-value = 0.000).

	D	C	G	A	OB	ER
C	0.687					
G	0.814	0.679				
A	0.835	0.705	0.536			
OB	0.753	0.592	0.677	0.708		
ER	0.774	0.573	0.67	0.684	0.339	
Vietnam	0.654	0.515	0.558	0.612	0.309	0.035

The genetic diversity of the Vietnamese population and the six African groups was assessed based on some descriptive statistics (Table III.4.1 and Table III.4.2). African individuals presenting < 70% ancestry proportion were excluded from the analysis. The mean allelic richness (AR) of the Vietnamese group was 1.62, which was higher than most of the wild genetic groups (AR = 1.18 –1.51). The observed heterozygosity (Ho = 0.18) and expected heterozygosity (He = 0.19) were similar to the Congolese group ER, which were higher than the other groups, and lower than those of the whole African populations (Ho = 0.36, He = 0.28). The inbreeding coefficient among the Vietnamese individuals was low - F_{IS} = 0.03, while it ranged from 0.06 to 0.38 in the wild groups. However, there were three pairs of identical genotypes in the Vietnamese collection, suggesting putative clones or possibility of mislabeling. The differentiation coefficients, F_{ST}, were also high between the African groups, ranging from 0.339 (OB and ER) to 0.814 (D and G). The F_{ST} values between the Vietnamese group and most of the wild groups were high, ranging from 0.309 (group OB) to 0.654 (group D), except for group ER with F_{ST} = 0.035 which again confirmed their close relationship.

III.4.2 Core set construction

As the Vietnamese Robusta genotypes presented a high level of redundancy, we selected a core set containing the most representative individuals, i.e. maximizing genetic distance and allelic richness. The core set of 45 accessions had expected heterozygosity (He = 0.22) and allelic richness (AR = 1.91) markedly higher than randomly selected sets of the same size (Figure III.4.2A and B), and comparable to the whole collection (He = 0.21, and AR = 1.41). Most of the closely related individuals were removed from the core, and most of the hybrids were retained (Figure III.4.2C and D). This core collection was sequenced for the whole genome to be more deeply analyzed.

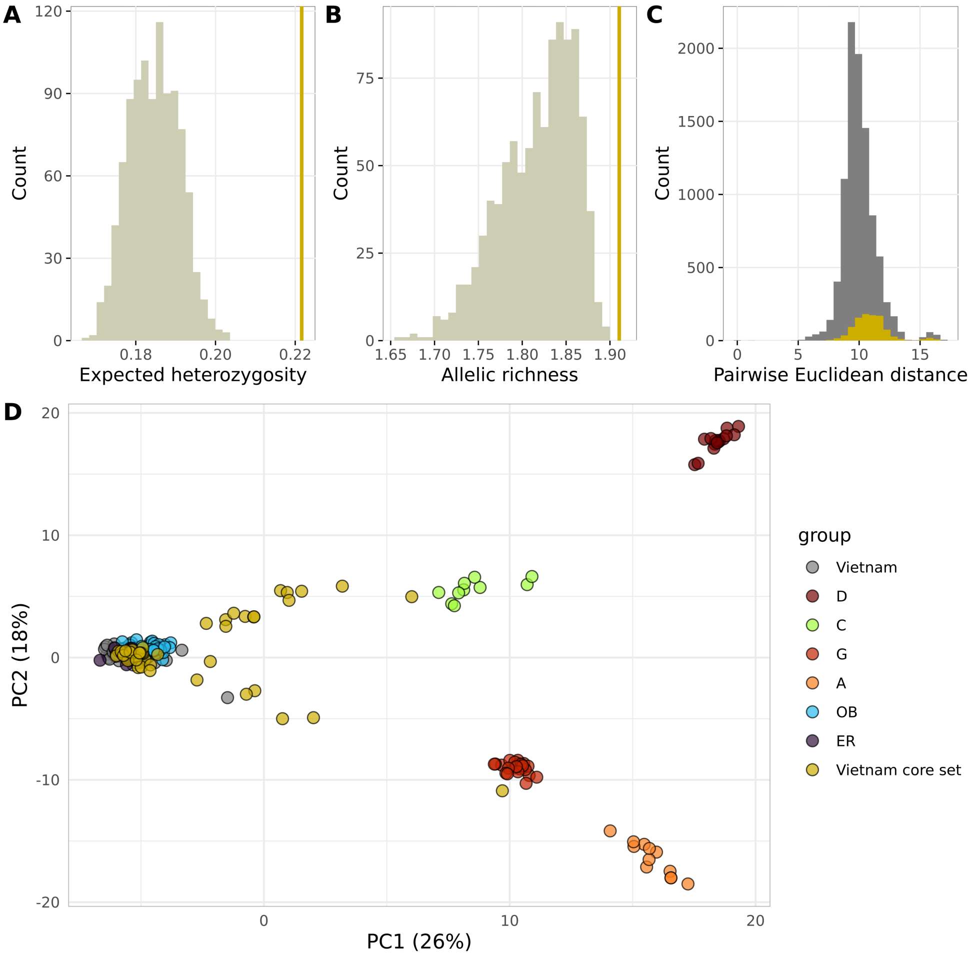


Figure III.4.2: Selection of a core set from the Vietnamese collection. Evaluation of the core set: Histogram of (A) Expected heterozygosity and (B) Allelic richness of 1000 random sets at the same size as the core set (45 individuals); the corresponding values of the core set are marked by yellow vertical lines. (C) Histogram of pairwise Euclidean distances between individuals in the whole collection (gray bars) and in the core set (yellow bars). (D) Individuals selected for the core set are highlighted (yellow) in PCA projection.

III.4.3 Local ancestry analysis of the core set accessions

A total of 13,991,298 whole-genome biallelic SNPs were obtained when combining the WASI core set and the reference set of 55 African wild accession sequences (Tournebize et al., 2022). These SNPs were used for local ancestry inference at the whole-genome level. Using a subset of 5 kb thinned SNPs, the African reference set was differentiated into five genetic groups by sNMF (Supplementary Figure S-III.8.4), similar as previously described in III.4.1 except that the closely related groups A and G were merged due to a smaller sample size. Thus, ancestry at the chromosome level of the Vietnamese core individuals was inferred from the five ancestral groups (D, C, AG, OB, and ER) as source populations.

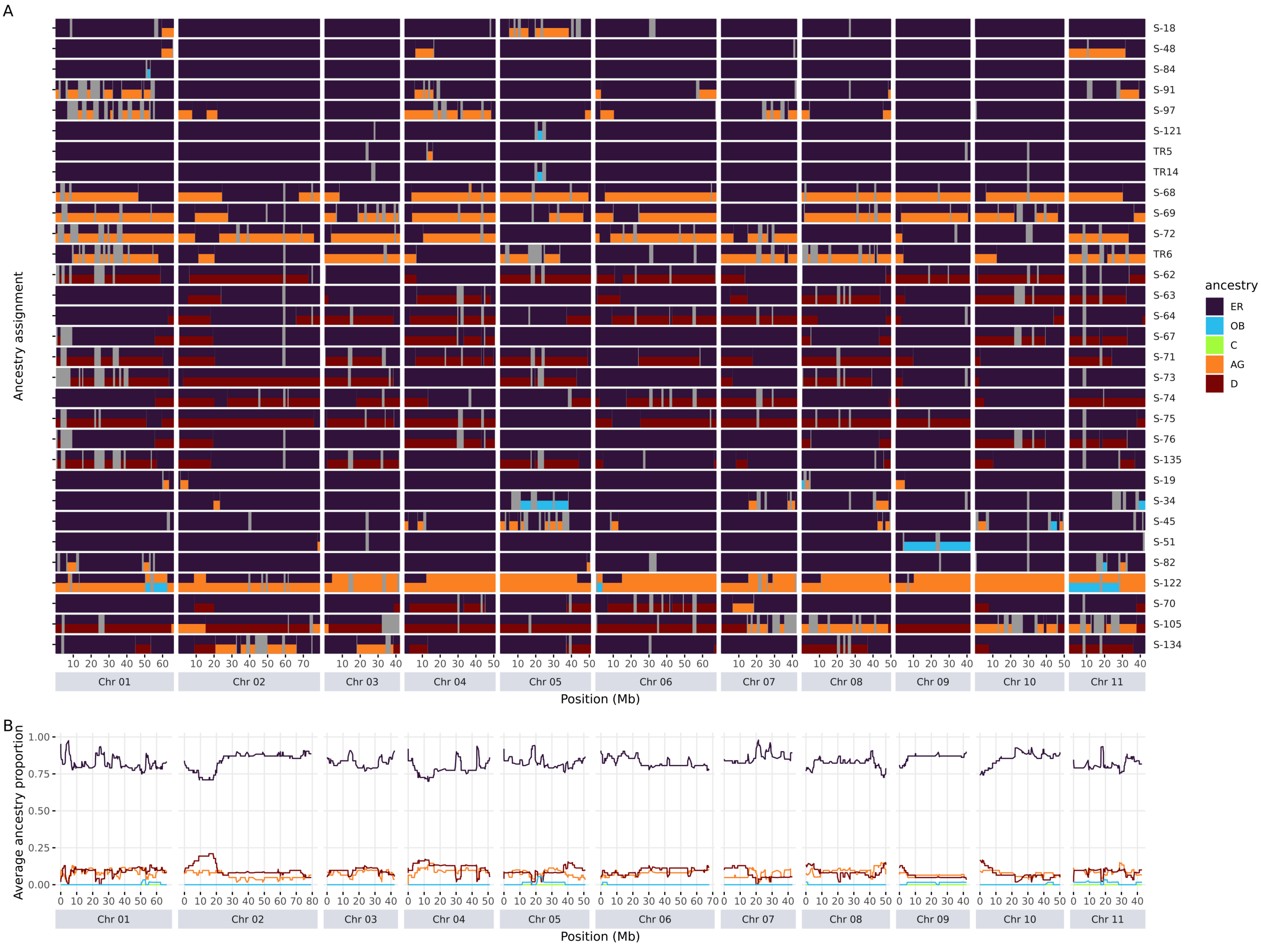


Figure III.4.3: Genome-wide introgression in 31 admixed individuals of the core set. (A) Local ancestry proportion (0%, 50%, 100%) inferred along the chromosomes. The ancestry groups are inferred by colors (AG - orange, OB - blue, D - red, ER - dark purple), and undetermined regions are in gray. (B) Averaged ancestry inference across the admixed individuals at each SNP position. The remaining 14 accessions of the core set were estimated with only ER ancestry.

The total ancestry proportion associated with the local ancestry inference in each individual was congruent with the genetic structure analyzed by sNMF, with correlation coefficient r = 0.97. All of the Vietnamese individuals presented chromosome segments of the group ER (Figure III.4.3 and Supplementary Table S-III.8.1), with at least 47% of the genomic windows assigned to ER. Only 0% to 14% of the genome remained undetermined due to uncertainties in the detection method (Vi et al., 2023). Of the 45 core set accessions, different admixture patterns and tract lengths were detected. Fourteen were of ER origin only, while 31 individuals were detected as admixed. Among the two-way admixed genotypes, eight individuals showed < 10% admixture from group AG or OB, four individuals had AG segments accounted for > 20%, 10 individuals had segments from group D in > 10%. Nine individuals were three-way admixed, which were either ER x AG x OB or ER x D x AG (Figure III.4.3A). The lengths of admixture blocks varied from 1 to 76.7 Mb (Supplementary Figure S-III.8.5), with a median at 5.4 Mb for the AG ancestry, and 9.8 Mb for the D ancestry.

To assess the distribution pattern of the admixed segments along the genome, we estimated the average ancestry proportions across all the 31 hybrids at each SNP (Figure III.4.3B). Despite the varied admixture patterns and segment lengths, in overall, the AG and D ancestries had almost equal contributions at the SNP level and were evenly distributed throughout the genome (Figure III.4.3B). In some chromosomes, the level of admixture was slightly higher at the beginning or end than the other regions, such as in chromosomes 2, 9, and 10. The elevated admixture levels were observed in regions with high recombination rates by alignment to the genetic map along the genome (data from Mérot-L'Anthoëne et al., 2019 available in MoccaDB, https://moccadb.ird.fr/ and in the Supporting Information of (Brazier and Glémin, 2022; Supplementary Figure S-III.8.6).

III.5 Discussion

In this study, we evaluated the genetic diversity of *C. canephora* that has been transmitted from Africa to Vietnam, using genome-wide SNP data. The development of the 261 SNP array has allowed us to rapidly collect data from different genotyping resources, effectively assess genetic diversity (origin), and identify clones or redundant individuals in order to select core accessions. Whole-genome sequencing helped to accurately detect admixture tracts along the genome (Thornton and Bermejo, 2014), which would provide a better understanding of hybridization and association with beneficial traits in Vietnamese Robusta coffee.

The wild Robusta individuals from central Africa were classified differently when using different genotype data (KASPar genotyping and whole-genome re-sequencing), different sets of SNP markers (261 SNPs and whole-genome SNPs), and different analyses (PCA and sNMF). As there was lower genetic differentiation between groups E and R, A and G, and O and B (Mérot-L'Anthoëne et al., 2019), and unequal numbers of samples, they could be merged into the same clusters (ER, AG and OB, respectively). However, they were all congruent with previous studies (Mérot-L'Anthoëne et al., 2019; Tournebize et al., 2022; Vi et al., 2023), and did not affect ancestry assignment of the Vietnamese accessions.

Of the previously classified genetic groups, the groups native to the RDC, E and R, were abundantly represented in the 126 Vietnamese Robusta accessions. However, the Vietnamese individuals were likely not inbred, as suggested by the low inbreeding coefficient. They were also related to wild ER accessions from different sources (such as DRC, Central African Republic, Sankuru and Congo). This finding is consistent with the Vietnamese genotypes studied previously (Garavito et al., 2016; Akpertey et al., 2021) although we provided more precise source origins. They confirm historical records of the early diffusion of Robusta (Congolese) types under the name "Robusta" from the Yangambi region to Southeast Asia via the trials of the Java Coffee Research Station, Indonesia - the first important breeding and worldwide distribution center of Robusta (Coste, 1955; Cramer, 1957; Verleysen et al., 2023). There were also reports mentioning that varieties from Congo were initially introduced to the Central Highlands (ICO, 2019). In particular, our eleven ancient accessions from the Dak Lak Museum collection, which might be the earliest introduced varieties, had an ancestral proportion of 86% to 100% from the ER group. Therefore, Robusta coffee cultivated in Vietnam was mainly derived from genetic resources from the Congo Basin.

However, Congolese populations are not the only wild sources of origin. We found, in at least 31 accessions, traces of all other genetic groups, except group C (which included individuals from northwest Congo, southeast Cameroon, and southwest Central African Republic, Gomez et al., 2009). The two major admixed sources in the hybrid accessions of the WASI core set were from groups D (Guinean group) and AG (Conilon-Angola group), and with a lesser extent group OB (East CAR-Uganda origin). Two of the ancient accessions represented 10% or 12% of the OB group, which aligned with historical records in Java. When Robusta seeds from Congo were initially sent to Java under the name "Robusta" (ex *Coffea robusta* L. Linden), it was quickly accepted by farmers because of its productivity and apparent resistance to coffee leaf rust (Cramer, 1957). These materials were later enriched with other sources such as Sankuru-DRC (ex *Coffea canephora* f. sankuruensis De Wild), "Quillou"/"Kouillou" varieties ("Conilon" variety name is a deformation of Kouillou) with no precise details of their African origin, or Uganda (ex *Coffea ugandae* Cramer) (Cramer, 1957; Ferwerda, 1948).

These different introgressed origins were detected in almost every part of the genome within a Congolese genomic background. As shown by the average ancestry proportion (figure III.4.3B), the probability of admixture seemed to be even along the genome, and slightly higher at the ends of some chromosomes which could be a results of high recombination rates (Brazier and Glémin, 2022; Mérot-L'Anthoëne et al., 2019). The admixture tracts were distributed distinctly in each individual, but all suggested patterns of backcrossing. The hybrids with longer admixture segments might be the result of recent hybridization, and conversely, those with smaller admixture segments might have experienced more cycles of backcrossing, probably before being introduced into Vietnam from Java. To precisely estimate their admixture time, it may require further analysis on local ancestry inference at haplotype level (Duranton et al., 2019).

The different genetic groups presented in the hybrids were considered complementary in terms of agronomic traits, and therefore, might be useful for breeding (Leroy et al., 1993b; Montagnon et al., 1998b; L. N. L. Oliveira et al., 2018). Moreover, our results suggest that introgression of various fragment sizes with potentially favorable traits is possible in Robusta in a few backcross generations, since we observed various sizes of admixture patterns evenly distributed along the genome. The local ancestry inference distribution and genetic recombination rate map could provide useful background information for selection of introgression breeding (Pratap et al., 2021). For example, varieties presenting ancestry segments of desirable genes could be targeted, and with knowledge of recombination rates, possible outcomes could be predicted (Hospital, 2001). In addition, as undesired alleles are often introgressed along with the targeted ones due to genetic linkage (Pratap et al., 2021), varieties with small admixed segments might not present these linkages and could therefore avoid the introgression of undesirable traits. The fact that we already have different combinations of admixture within our core-set accessions will ease and fasten the breeding schemes, when choosing the most suitable clones as parental lines in the crosses (Pratap et al., 2021).

Most breeding programs have relied mainly on heterosis resulting from crossing Congolese and Guinean groups to improve coffee quality, such as bean size or resistance to diseases (Alkimim et al., 2021; Leroy et al., 1993b; Montagnon et al., 1998b; L. N. L. Oliveira et al., 2018). In WASI, there have been attempts to select and produce new varieties with higher performance (Phan, 2017). Eleven elite varieties have been recognized, which have higher yield and bean quality, with tolerance to late irrigation timing. In the present study, ten of them were included in the germplasm collection and the core collection. Despite the expectation that these elite varieties could be hybrids of different genetic groups, seven out of the ten elite varieties were composed of only the group ER, two presented only minor proportions of group AG or OB, and only one variety had about 20% admixture from group AG. This could be explained by the limited genetic resources of the complementary groups. Another possible reason is that the Congolese ER varieties might be better suited to the growing condition in the Central Highlands, compared to the other groups, as they are more resistant to rust leaf disease and have a late maturation time adapted to the dry season pattern (L. F. V. Ferrão et al., 2019;Phan, 2017). The only elite variety with significant admixture from the group Conilion-Angola AG group benefited late ripening time and leaf rust resistance from the Congolese ER group, but had lower yield than the other elite clones (Phan, 2017). The genomic regions associated with yield (M. A. G. Ferrão et al., 2023) possibly might be introgressed by the group AG in this variety. Further studies with more samples and comprehensive phenotyping are needed to confirm this hypothesis.

In the future, Vietnamese Robusta coffee will need to increase its resilience to cope with climate change and associated abiotic and biotic stresses, which are expected to have a significant impact on the Central Highlands (Bunn et al., 2015; Dinh et al., 2022). The accessions with admixture segments from Conilon-Angola or Guinean origin were found to be more drought tolerant and resistant to pests and diseases (Montagnon et al., 1998b; Marraccini et al., 2012; Vieira et al., 2013), and may help improve coffee adaptation in this scenario. They will potentially become more superior varieties, or be used as breeding material to introduce adaptive regions into the elite varieties. A deeper understanding of the genetic-climate relationship in coffee (Kiwuka et al., 2021; Tournebize et al., 2022) might also help to optimize the choice of breeding materials. Moreover, breeding strategies should also focus on conserving or introducing more diversity of different genetic groups from other sources to explore new beneficial genetic markers.

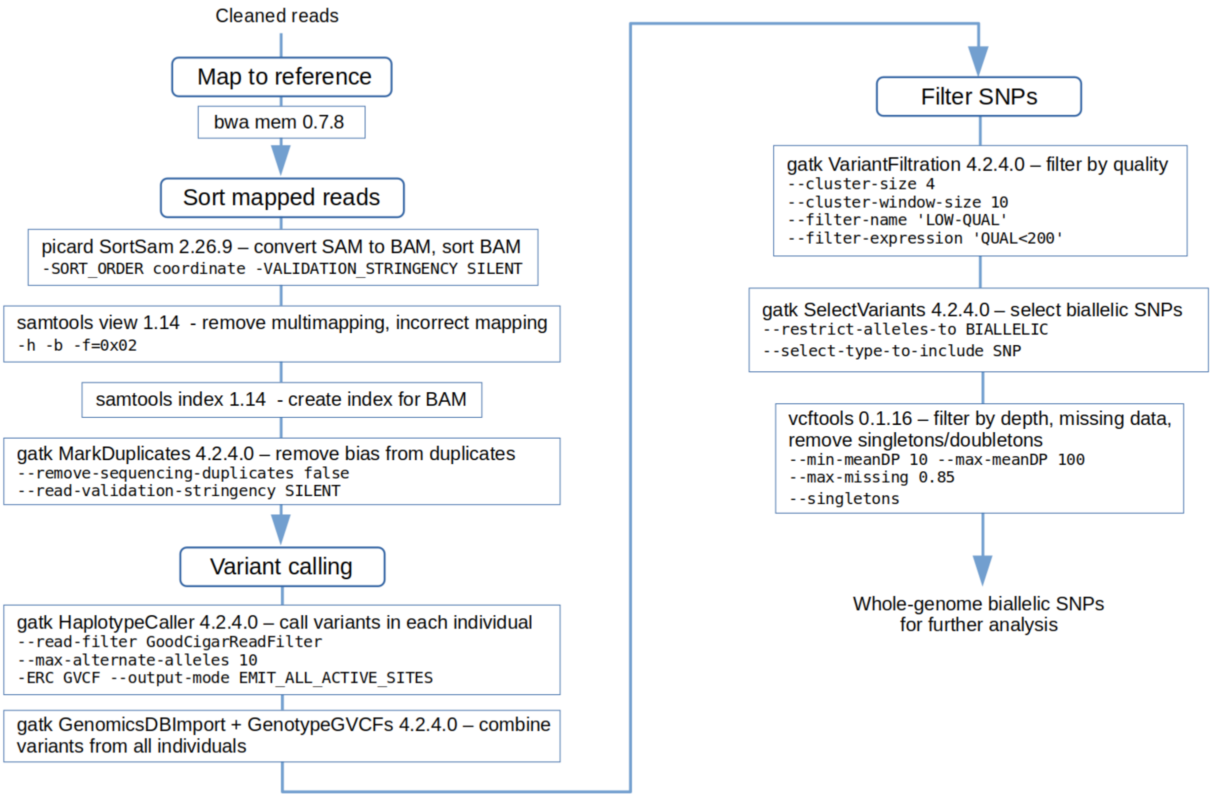
III.6 Acknowledgements

The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/- http://www.southgreen.fr

III.7 Data Availability

The raw sequencing data of the African accessions were obtained from the NCBI SRA database under project accession number PRJNA803612 (Tournebize et al., 2022). The raw sequencing data of the 45 Vietnamese core accessions are available in the NCBI SRA database under project accession number PRJNA950219. Other genotypes at 261 SNPs are available upon request (valerie.poncet@ird.fr).

III.8 Supplementary data


Cleaned reads
Map to reference
bwa mem 0.7.8
Sort mapped reads
picard SortSam 2.26.9 – convert SAM to BAM, sort BAM
samtools view 1.14 - remove multimapping, incorrect mapping
samtools index 1.14 - create index for BAM
gatk MarkDuplicates 4.2.4.0 – remove bias from duplicates
Variant calling
gatk HaplotypeCaller 4.2.4.0 – call variants in each individual
gatk GenomicsDBImport + GenotypeGVCFs 4.2.4.0 – combine variants from all individuals
Filter SNPs
gatk VariantFiltration 4.2.4.0 – filter by quality
gatk SelectVariants 4.2.4.0 – select biallelic SNPs
vcftools 0.1.16 – filter by depth, missing data, remove singletons/doubletons
Whole-genome biallelic SNPs for further analysis

```
-SORT_ORDER coordinate -VALIDATION_STRINGENCY SILENT
```

```
-h -b -f=0x02
```

```
--remove-sequencing-duplicates false
--read-validation-stringency SILENT
```

```
--read-filter GoodCigarReadFilter
--max-alternate-alleles 10
-ERC GVCF --output-mode EMIT_ALL_ACTIVE_SITES
```

```
--cluster-size 4
--cluster-window-size 10
--filter-name 'LOW-QUAL'
--filter-expression 'QUAL<200'
```

```
--restrict-alleles-to BIALLELIC
--select-type-to-include SNP
```

```
--min-meanDP 10 --max-meanDP 100
--max-missing 0.85
--singletons
```

Figure S-III.8.1: Variant calling workflow following GATK best practice and SNP filtering. Software used: BWA mem 0.7.8 (H. Li, 2013), Picard Tools 2.26.9 (https://broadinstitute.github.io/picard/), SAMtools 1.14 (Danecek et al., 2021), GATK 4.2.4.0 (Poplin et al., 2017), vcftools 0.1.16 (Danecek et al., 2021).

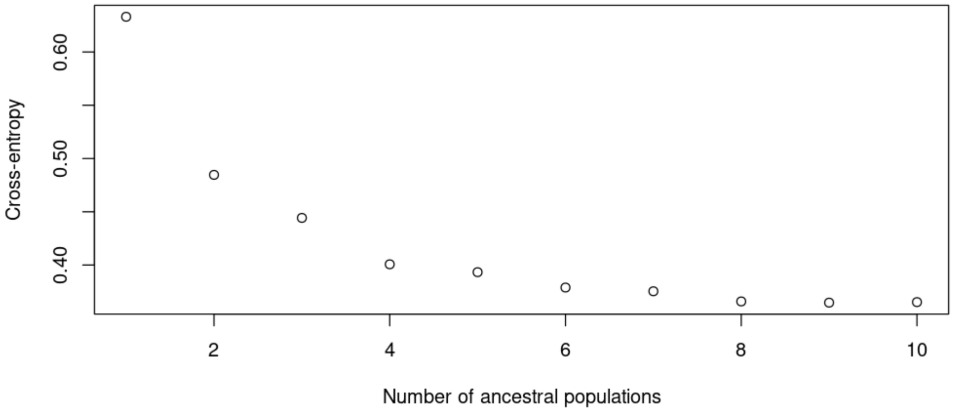


Figure S-III.8.2: Cross-entropy criterion of sNMF runs on all the individuals with varied numbers of K.

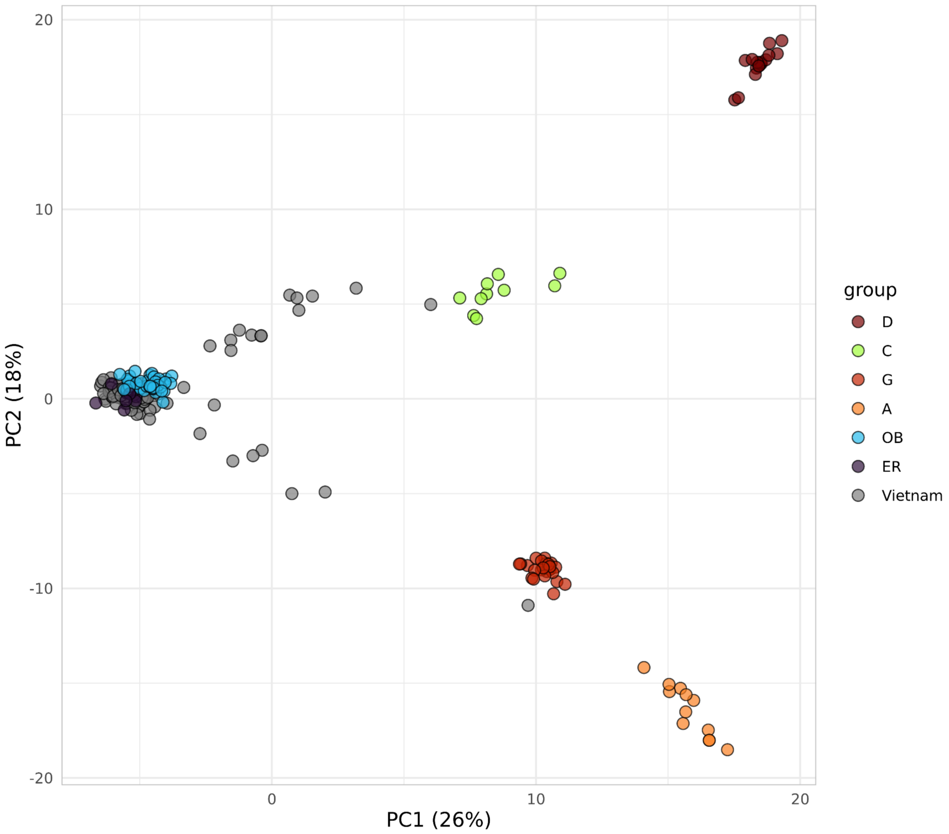


Figure S-III.8.3: PCA results of the 126 WASI Vietnamese germplasm accession and African genetic group refrences, using 261 SNPs.

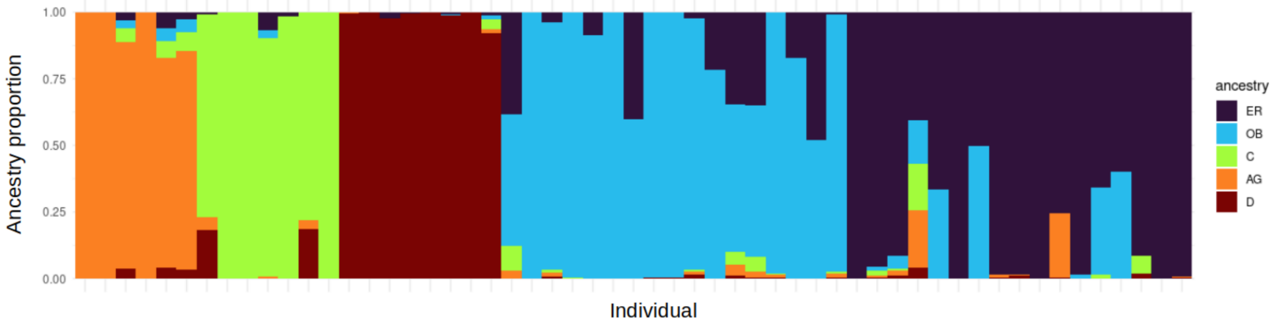


Figure S-III.8.4: Genetic structure of the African reference individuals based on sequencing data. Five ancestral groups were estimated from the best sNMF run using a subset of 5 kb thinned SNPs.

Table S-III.8.1: Global ancestry proportion of the core individuals based on local ancestry results.

ID	AG	C	D	ER	OB	undetermined
S-15	0	0	0	0.813	0	0.187
S-18	0.027	0	0	0.939	0	0.034
S-19	0.011	0	0	0.980	0.001	0.007
S-34	0.015	0	0	0.910	0.022	0.053
S-45	0.030	0	0	0.911	0.003	0.056
S-48	0.040	0	0	0.953	0	0.007
S-51	0.001	0	0	0.955	0.029	0.014
S-62	0	0	0.286	0.656	0	0.059
S-63	0	0	0.164	0.794	0	0.042

S-64	0	0	0.208	0.758	0	0.034
S-67	0	0	0.128	0.835	0	0.037
S-68	0.341	0	0	0.625	0	0.034
S-69	0.284	0	0	0.650	0	0.066
S-70	0.010	0	0.110	0.854	0	0.026
S-71	0	0	0.272	0.676	0	0.052
S-72	0.258	0	0	0.667	0	0.075
S-73	0	0	0.204	0.732	0	0.064
S-74	0	0	0.222	0.741	0	0.037
S-75	0	0	0.337	0.626	0	0.038
S-76	0	0	0.127	0.831	0	0.042
S-77	0	0	0	0.998	0	0.002
S-78	0	0	0	0.995	0	0.005
S-81	0	0	0	0.989	0	0.011
S-82	0.010	0	0	0.949	0.002	0.038
S-84	0	0	0	0.996	0.001	0.002
S-90	0	0	0	0.989	0	0.011
S-91	0.052	0	0	0.892	0	0.056
S-97	0.085	0	0	0.843	0	0.072
S-98	0	0	0	0.992	0	0.008
S-105	0.100	0	0.322	0.442	0	0.136
S-121	0	0	0	0.989	0.002	0.009
S-122	0.758	0	0	0.164	0.033	0.045
S-134	0.041	0	0.103	0.804	0	0.052
S-135	0	0	0.142	0.799	0	0.059
S-142	0	0	0	1.000	0	0.000
TR4	0	0	0	0.998	0	0.002
TR5	0.002	0	0	0.988	0	0.010
TR6	0.194	0	0	0.694	0	0.112
TR9	0	0	0	0.985	0	0.015
TR10	0	0	0	0.989	0	0.011
TR11	0	0	0	0.997	0	0.003
TR12	0	0	0	0.993	0	0.007
TR13	0	0	0	0.991	0	0.009
TR14	0	0	0	0.984	0.002	0.013
TR15	0	0	0	0.991	0	0.009

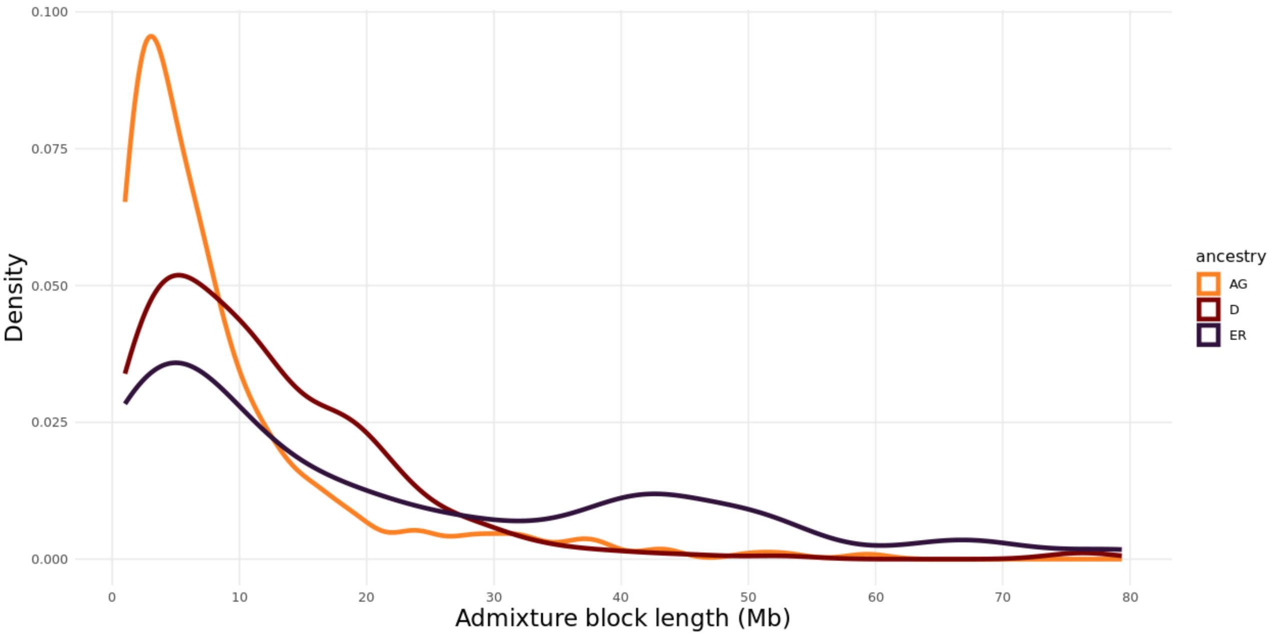


Figure S-III.8.5: Distribution of admixture block lengths. The admixture blocks were constituted of continuous admixture patterns. As the local ancestry inference was not based on phased sequencing data, the exact ancestry segments were unknown, but based on parsimony principle, the probability of having 2 crossover events is higher than 3 or more events. The admixture blocks defined here therefore were at the upper limit and most likely lengths.

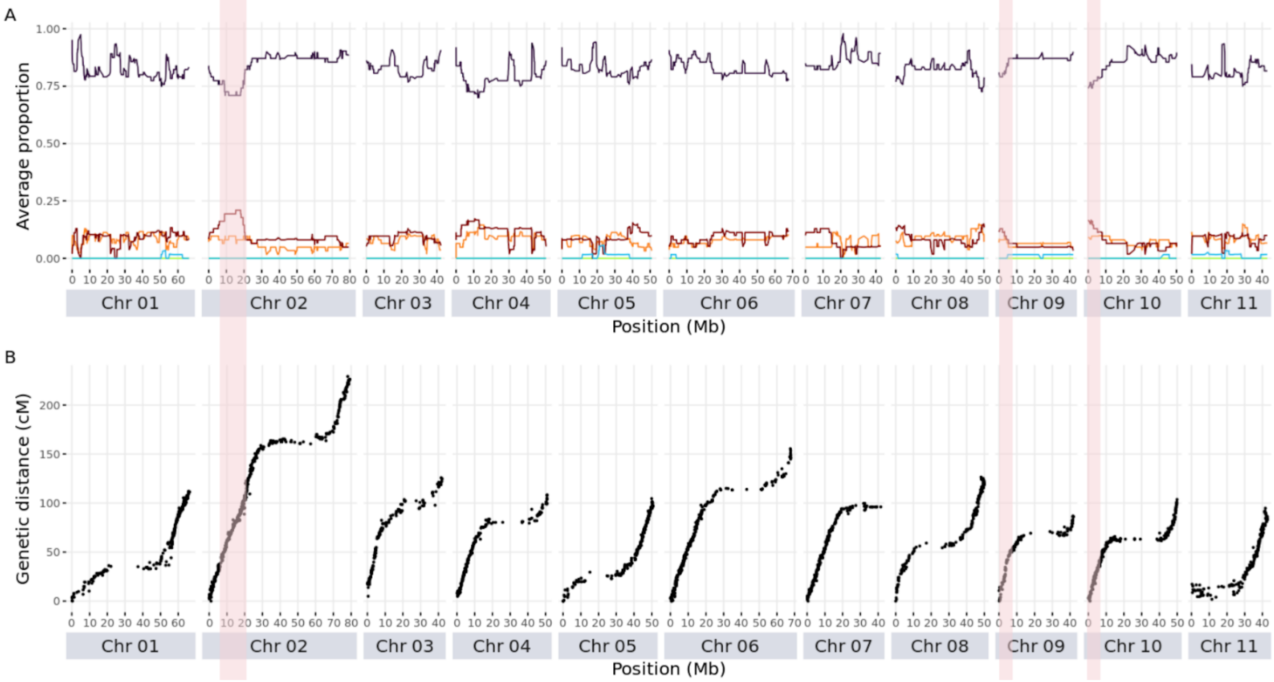


Figure S-III.8.6: Distribution of the genetic groups and recombination rate along the genome. (A) Average proportion of the genetic groups at each SNP position. (B) Genetic distances along the chromosomes obtained from Mérot-L'Anthoëne et al., 2019.

IV Genomic suitability of wild Robusta coffee to local climate in Vietnam

The genetic origin and diversity of Vietnamese Robusta accession were analyzed previously, in chapter II and III, which showed close relationships with a Congolese genetic group (native to the Congo basin regions). This genetic group has been intensively introduced into Vietnam because of its potentials of high yield and resistance to leaf rust. Climatic conditions are also important factors affecting the performance of coffee. However, the suitability of different genetic origins to local climate in Vietnam has not been integrated in previous breeding and conservation approaches. In this chapter, we predicted the suitable materials for the local climate in the future, based on the climatic difference between the wild distributions of Robusta in Africa and the planted regions in Vietnam, and the identification of genetic variants linked to climate.

Vi et al.

IV.1 Abstract

Crops are generally cultivated outside of their native range. Optimal climatic suitability for crop production is raising more concerns, especially in the context of climate change. Robusta coffee (*Coffea canephora*) is indigenous to west and central Africa, but is also widely grown on other continents. Vietnam has been the world's largest Robusta producer since the 2000s, but is currently facing the risk of yield loss due to climate change. Assessing the suitability of wild genetic material to local climate conditions in Vietnam is essential for sustainable improvement of Vietnamese Robusta coffee. We first compared climatic conditions between native environments in Africa and that in Vietnam, using bioclimatic variables. The wild populations were highly differentiated into five main genetic groups, but only one of these (group ER from DRC) was mainly found in cultivated elite varieties in Vietnam, with a minor proportion of group AG from Gabon and Angola. A negative relationship between climate distance to the native range of group ER and coffee yields at the planted areas was found, and used to predict future coffee yields. The climate distance suggested higher suitability of the minor group (AG) in Vietnam. Using a reference-free approach, more than 18M of k-mers (substrings of sequence reads at 31-bp length) were detected in association with bioclimatic variables. Functional annotation of these candidate k-mers identified putative proteins related to gene regulation. They were used to predict the genetic changes that the wild individuals would require to fit the local environment (genetic offset) in the present and future Genetic offsets revealed variation between the different groups, and also suggested that the most suited genotypes came from the group AG. The results would be useful for planning strategies to improve adaptability of Vietnamese Robusta coffee.

IV.2 Introduction

The increase in food demand in proportion with human population growth has led to geographical expansion and migration of crops worldwide (Khoury et al., 2016). The new migrated environment may be different from the native environment to which the crops have initially adapted. Mismatches between global distribution and climatic suitability have been found in many crops, such as potato, rice, or cassava (Mahaut et al., 2022). Furthermore, because of global climate change, the suitable areas for crop planting, including perhaps their centers of origin, may be shifted. For example, the cultivated areas of wheat shifted northward in Europe from 1973 to 2012 to avoid rising temperature (Sloat et al., 2020). In recent years, climate change has increased the yield of some crops (e.g. maize and soybean) in America but also decreased it in Europe, Africa and Asia (Ray et al., 2019). It is therefore important to evaluate the adaptive capacity of crops at both spatial and temporal scales.

Coffea canephora is a species sensitive to climate change (Tournebize et al., 2022). This species produces Robusta coffee, contributing 40% of total coffee production in the world (ICO, 2019), and thus is an important crop in many countries. *C. canephora* is native to large regions from west Africa to Congo basin stretching southwards to Angola (Berthaud, 1986; A. P. Davis et al., 2006). It exhibit different genetic groups (Mérot-L'Anthoëne et al., 2019) distributed in a relatively wide range of environmental conditions, from evergreen to seasonally dry humid tropical forests (A. P. Davis et al., 2006). Since the beginning of the 20th century, Robusta has been widely dispersed to other parts of the world including Vietnam, for breeding and expanding cultivation (Montagnon et al., 1998b). The dispersal of Robusta and climate change may increase mismatch between local suitability and distribution in the future. Environmental factors (e.g. temperature, precipitation) showed high impacts on Robusta productivity and bean quality as well as other agronomic traits (Anim-Kwapong and Boamah, 2010; Kath et al., 2020; Kath et al., 2021). Based on the forecasted climate change, the worldwide suitability of Robusta is predicted to significantly reduce by 2050 (Bunn et al., 2015). Vietnam, the world's largest Robusta producer and exporter since the 2000s, will lose 50% of current production by 2050 (ICO, 2019). Robusta in Vietnam is grown mostly in the Central Highlands with 97% of total planting area (ICO, 2019). The climatic conditions in this region are suitable for Robusta growing, with annual temperature ranging from 18 to 26 °C, and annual rainfall at 1200 to 3000 mm (Ngo-Thanh et al., 2018). However, in recent years, severe climate change, such as increasing temperatures, abnormal rainfall patterns and disasters, has been more frequently observed in the Central Highlands (Baker, 2016; Tan et al., 2013). It is estimated that up to 83% of current suitable planting area will be lost by 2060-2070, following several global climate models (Dinh et al., 2023).

However, most of the previous studies on Robusta habitat suitability (Bunn et al., 2015; Dinh et al., 2023) focused on ecological mismatch but ignored the adaptive genetic variation and the potential of populations to respond to future changes. Intraspecific adaptations have been indeed identified within *C. canephora* native range though the detection of candidate SNPs correlated with environmental variables such as isothermality, seasonal temperature and seasonal precipitation (de Aquino et al., 2022; Tournebize et al., 2022). Genes associated with abiotic stress such as heat, drought, and salt stress were also characterized (de Aquino et al., 2022; de Carvalho et al., 2013; de Carvalho et al., 2014; Marraccini et al., 2012; Torres et al., 2019). Moreover, the different genetic groups were found to have varying degrees of resistance to biotic/abiotic stress, e.g greater sensitivity to drought in a group in Congo basin (Montagnon et al., 1998a). Considering local climatic adaptations to predict genomic vulnerability, Tournebize et al., 2022, found that it was independent of ecological vulnerability in wild African Robusta populations: the most ecologically vulnerable regions were not necessarily those for which the genomic vulnerability of populations was highest, and vice versa. Therefore, prediction of genetic changes in response to climate change needs to be assessed and combined with changes in ecological suitability to estimate the global climate change-driven vulnerability (Chen et al., 2022).

To predict the genetic adaptability of a species to a new environment, it is necessary to understand genomic variations driving local adaptation. Genotype-environment association (GEA) approaches are commonly used to estimate the correlations between environment (e.g. bioclimatic variables) and genetics (e.g. allelic frequencies), under assumption that populations are adapted to their local environment (Rellstab et al., 2021). Most of the methods are based on linear models, which incorporate environmental factors as fixed effects, and neutral genetic structure as random effects (Rellstab et al., 2015). These relationships can be extrapolated in space and time, to forecast the changes in genomic composition required in new or future environments (Bay et al., 2018; Capblancq and Forester, 2021; Fitzpatrick and Keller, 2015; Rellstab et al., 2016), which is referred to as "genomic offset" or "genetic offset" (Rellstab et al., 2021). The level of genomic offset can also be used to predict effects on fitness traits driven by environment (Capblancq et al., 2020; Gain et al., 2023).

Most GEA studies make use of SNP markers, because of their abundance across the genome and efficiency in detection (Rellstab et al., 2015). However, genome scans using SNPs may fail to detect important variants, such as structural variants (SVs), copy number variations (CNVs), or variants in regions that are not present on the reference genome (Voichek and Weigel, 2020). Numerous SVs have been found in or near genes linked to resistance to biotic or abiotic stress in many plants, which potentially contribute to adaptation (Songsomboon et al., 2021). In *Theobroma cacao*, more than 160,000 SVs (insertions, deletions, tandem duplications, inversions, and translocations) were found in 31 genomes, and among them, 864 outlier SVs were linked to genes associated with pathogen defense (Hämälä et al., 2021). In pearl millet, a large number of SVs were found to be related to genes involved in heat tolerance, and some SVs even affected the gene expression (Yan et al., 2023). In wheat, CNVs in two genes, Ppd-B1 and Vrn-A1, presented association with flowering time (Würschum et al., 2015).

A way to detect and identify such associated variants is to employ reference-free approaches using k-mers. By extracting overlapping short sequences at a constant length k (k-mers, with k often at 31) from all the whole-genome sequencing (WGS) reads, such variants could be captured and lead to different k-mer sequences or counts (Rahman et al., 2018; Voichek and Weigel, 2020). Using k-mer presence/absence data, several studies have identified novel genetic variants associated with phenotype, such as k-mers associated with 96 metabolites in tomato leading to identification of a new coding sequence outside of its reference genome (Voichek and Weigel, 2020), or k-mers related to cob color in maize that were not detected by SNP-based method (C. He et al., 2021). This reference-free approach is therefore also promising to scan all potential genomic variants in populations for environmental association and can reduce computational cost (as it does not require mapping).

In this study, we assessed, at the genomic level, the climatic suitability of Robusta coffee of various African origins for the climate of the Central Highlands. Based on distribution-climate models, Dinh et al., 2022, predicted a decrease in climate suitability in most of the planted area yet a slight increase in a small region in the north of the Central Highlands. Using genomic data, the genetic-climate association would allow a more complete assessment of the effects of climate change on Robusta coffee in Vietnam. As about 50% of *C. canephora* genome are transposable elements (Denoeud et al., 2014), a large portion of its variants might not be detected by using reference-based mapping methods. Therefore, we applied the reference-free approach to identify k-mers associated with bioclimatic variables in wild African *C. canephora* populations. This could allow the detection of new putative genomic regions or genes under selection that lie outside the single reference genome. These k-mer-environment relationships enabled estimation of the genetic offset of the wild populations when introduced from Africa to Vietnam, to predict the most suitable genetic materials for the current and future local climates.

IV.3 Materials and Methods

IV.3.1 Sequence and diversity analysis

A total of 60 wild African accessions were collected from different geographical locations covering all the known genetic groups in the native range (Mérot-L'Anthoëne et al., 2019). Whole-genome sequencing (WGS) data of 55 accessions were obtained from BioProject PRJNA803612 (Tournebize et al., 2022), and additional whole genome sequences of 5 accessions were obtained by Novaseq PE150 sequencing (DNA extraction, library preparation, and sequencing were performed by BGI Hong Kong). For the genetic diversity in Vietnam, 10 elite accessions as the most commonly planted varieties in Vietnam (Phan, 2017) were used, and their WGS data were obtained from BioProject PRJNA950219 (Vi et al., 2023).

Prior to analysis, all the sequencing reads were filtered by fastp 0.23.1 (–detect_adapter_for_pe) to remove adapters, and by cutadapt 3.1 (-m 35 -q 20,20) to trim bases < Q20 (Martin, 2011). The sequencing depth of each individuals was > 10X after filtering.

A SNP data set of all the individuals was generated as a control to validate the discrimination capacity of the k-mer data. SNPs calling was performed following GATK Best Practices as applied in chapter III). The reads were mapped on the reference genome (Salojärvi et al., 2023) version 1.8 by bwa mem 0.7.8 (H. Li, 2013) and variants called by GATK HaplotypeCaller 4.2.4.0 (Poplin et al., 2017). Filters on quality (QUAL > 200), depth (meanDP < 100 and > 10), singletons/doubletons, missing data (< 15%), minor allele frequency (MAF > 0.05), and thin (5 kb), were applied to obtain the final data set of 102,754 biallelic SNPs, by using GATK 4.2.4.0, BCFtools 1.9 (Danecek et al., 2021), and VCFtools 0.1.16 (Danecek et al., 2011).

We obtained k-mer sequences (k = 31) and their presence/absence information in all the African individuals using iKiss pipeline version 1.0.0b2 (https://forge.ird.fr/diade/iKISS). Firstly, k-mers were extracted and counted in each individual by KMC3 (Kokot et al., 2017). To avoid bias from sequencing error, k-mers in canonized form occurring less than two times in each individual were removed. The remaining k-mers of all individuals were then combined and filtered using kmersGWAS according to two criteria: (1) present in at least two individuals ("-mac 2"), and (2) in each canonized/non-canonized form in at least 20% of individuals from which it appeared in ("-p 0.2") (Voichek and Weigel, 2020). Finally, the presence/absence table of filtered k-mers was created and randomly divided into subsets, each containing 1M k-mers. These random sets facilitated the downstream analysis in terms of computational efficiency and minimizing of bias in k-mer sampling.

The same pipeline was applied to all the African and Vietnamese individuals, to obtain presence/absence of k-mer in the Vietnamese population.

The genetic structure of the African accessions was assessed by sNMF (Frichot et al., 2014), using both SNP and k-mer (a random subset of 1M k-mers) data. We performed sNMF with the number of ancestral populations K ranging from 1 to 10, and making 10 iterations for each K value. The optimal value for K was assessed using plot of averaged cross-entropy values for the 10 repetitions of each K. The run with the lowest cross-entropy of the most optimal K value was used to assign ancestry proportions.

IV.3.2 Bioclimatic data and climate distance between African and Vietnam

Climatic conditions were estimated using the 19 bioclimatic variables from Worldclim 2.1 (https://www.worldclim.org/). Bioclimatic data in Africa were obtained for Present (near current, years 1970-2000) at 57 geo-references at 30 seconds resolution.

Bioclimatic variables for Vietnam were extracted at 640 occurrences of Robusta cultivated areas in the Central Highlands, covering 38 districts (Bunn et al., 2015). We retrieved data at 30 seconds resolution for both present (years 1970-2000) and future (years 2041-2060) conditions. For the future data, we selected three Global Climate Models (GCMs) that proved to be the most accurate models for Vietnam: CNRM-CM6-1, EC-Earth-Veg, MRI-ESM2-0 (Desmet and Ngo-Duc, 2022; Iqbal et al., 2021) and two scenarios from Shared Socio-economic Pathways (SSPs): the most optimistic (ssp126) and the most pessimistic (ssp585) ones.

All the African current bioclimatic data were analyzed using a Principal Component Analysis (PCA). This PCA space was then used to project the Vietnamese variables. We calculated the pairwise distance between African groups and each district in Vietnam. To do so, we calculated the barycenters of all the individuals of a given genetic group in Africa, and of each district in Vietnam for the first five PC axes, and then calculated the Euclidean distance between these points. The distances were calculated for present and future climate conditions (for each GCM-SSP combination) in Vietnam.

Most recent data on coffee yield (ha/ton) in Vietnam (Statistical Office, Vietnam) were averaged over the period 2010-2019, in 38 districts in the Central Highlands. We built a linear regression between the climate distances and the coffee yields at the district level. Using the climate distance in different climate models (GCMs) and scenario (SSPs), Vietnamese coffee yields in the future were predicted with its 95% confidence intervals based on the linear model.

IV.3.3 Genotype-environment association and genetic offset

Associations between k-mers and the 19 bioclim variables were investigated in the African population using two statistical methods: latent factor mixed models (LFMMs) (Gain and François, 2021) and redundancy analysis (RDA) (Capblancq and Forester, 2021). Association tests were performed for each k-mer subset independently. In LFMM, a regression model was constructed with loci and climatic variables, and in which the genetic structure was modeled with latent factors. The number of latent factors, K, was set according to the sNMF results of random k-mers. Latent factors estimated from the different random 1M k-mers sets were expected to be consistent. Significance of k-mers was based on LFMM fitting with ridge penalty. P-values were calibrated by genomic inflation factor (GIF) then converted to q-values for false discovery rate (FDR) control of 0.01. LFMM and significancy tests were performed using the *lfmm* R package version 1.1 (https://github.com/bcm-uga/lfmm). For RDA, we applied the method developed by Capblancq and Forester, 2021. The k-mer matrix was fitted by the bioclim variables, and loaded on a principal component analysis. The loading of k-mers on the first two axes were then used for significance tests to compute p-values with GIF correction and 0.01 FDR control. Finally, the putative adaptive (candidate) k-mers were defined as the ones detected by both LFMM and RDA.

To identify the genomic position of all the k-mers, they were mapped against the *C. canephora* reference genome version 1.8 (Salojärvi et al., 2023), by bwa-aln (H. Li, 2013) with default parameters. The mapped sequences were then filtered by samtools (Danecek et al., 2021) (with options “-F 4 -q 10”) to remove unmapped k-mers and k-mers with MAPQ < 10.

To functionally annotate the putative adaptative k-mers, we merged them into longer sequences (contigs). They were assembled with a minimum overlap of 15 bp using mergeTags (https://github.com/Transipedia/dekupl-mergeTags). Only contigs with length 400 bp were used for blastx on the NCBI non-redundant protein database of Viridiplantae with filtering by e-value < 1e-6, using Blast 2.12.0 (Altschul et al., 1990). Gene ontology (GO) terms associated with the best hits were identified by Blast2GO 6.0.3 (Conesa et al., 2005).

Finally, we estimated the genetic offset between the present environments in Africa with the present and future environments in Vietnam using the candidate k-mers. Only individuals assigned to a group with > 70% probability were considered. We applied the geometric genetic offset defined by Gain et al., 2023, which is based on the effect sizes of environmental factors on allelic frequencies, corrected for the genetic confounding factors. Specifically, a presence/absence matrix of the candidate k-mers was regressed on the matrices of current African bioclimatic variables and latent factors, using linear models. The matrix of latent factors was obtained from LFMM results of a random set of 1M k-mers and the number of latent factors was determined as the optimal K value in sNMF analysis. Effect sizes of the regression were then used to project the expected k-mer values of each African individual at 640 occurrences in Vietnam (with bioclimatic variable in each of the present and future scenario). The genetic offset was estimated by the mean of squared differences between the expected k-mers and actual k-mers.

IV.4 Results

IV.4.1 Genetic structure of wild African and cultivated Vietnamese Robusta populations

From the sequence data of 60 African individuals, a total of 13,991,298 biallelic filtered SNPs were obtained, and 102,754 SNPs with only a single SNP per 5kb window ("–thin 5000") and MAF > 0.05 were retained for genetic structure analysis. The same sequence dataset led to a matrix of 1,856,279,562 k-mers.

Analysis of the genetic structure of the African populations using data from a random k-mer subset produced results that were very consistent with the analysis of SNP data (Supplementary Figure S-IV.8.1) (correlation coefficient r = 0.99, p < 0.001). The most optimal K value – the number of ancestral groups, was determined at 5 based on the cross-entropy criteria, in both cases. The five groups matched the geographic origin of the samples. We named these groups based on the previous classification (Mérot-L'Anthoëne et al., 2019) with some groupings of genetically close groups: (1) group D corresponded to individuals from West Africa, (2) group C corresponded to individuals from Cameroon region, (3) group AG to individuals from Gabon and Benin region (grouping of group A and G), (4) group OB corresponded to individuals from Uganda and southern Central African Republic region (grouping of group O and B), and (5) group ER corresponded to individuals from the Democratic Republic of Congo (grouping of groups E and R). We assigned a given sample to one of the five groups, if the ancestry proportion to this group was higher than 70%. We classified from 6 to 13 individuals in each of the five groups, for a total of 46 individuals. Fourteen were considered admixed.

We also assessed the genetic composition of the 10 previously sequenced Vietnamese individuals, by including them in the population structure analysis using a k-mer subset (Supplementary Figure S-IV.8.2). All of these accessions exhibited a strong ER genetic background, nine of them presented more than 90% ER ancestry proportions and one had 81% ER and 19% AG ancestry proportions.

IV.4.2 Climatic suitability and coffee yields in Vietnam

We assessed how the five African groups might differ in terms of their local climates using 19 bioclimatic variables. The groups differed in overall but there is also some overlap between their climatic envelopes (Figure IV.4.1A and B). The first two PCA axes were mainly determined by temperature-related variables (bio1 to bio11), which collectively contributed 61.5% and 71.5% of the total variance of PC1 and PC2, respectively (Supplementary Figure S-IV.8.3). The variable contributing the most to PC1 was bio6 (Min Temperature of Coldest Month, 8.3%), and the variable contributing most to PC2 was bio10 (Mean Temperature of Warmest Quarter, 11.2%). Among the precipitation-related variables, bio14 (Precipitation of Driest Month) and bio15 (Precipitation Seasonality) were mainly contributing but with only 7.2 and 6.1%, respectively.

We projected the bioclimatic data for the 640 Vietnamese occurrences, in the near present (1970-2000) and future (2041-2060), onto the PCA space of the 5 African groups (Figure IV.4.1A and B, and Supplementary Figure S-IV.8.4). On the first two PCs, the climatic envelop of Vietnam overlapped with the current climate of all the wild groups. The overlap shifted towards group D or ER climatic envelops in the future climatic conditions. The Euclidean climatic distances between each Vietnamese districts and each of the African genetic groups were calculated using the first five PCs which explained 96% of total variance (Supplementary Figure S-IV.8.3A). On average over all Vietnam districts, the closest African group was AG (Supplementary Figure S-IV.8.5). The climatic distance to this group was significantly lower than to the other groups for both present and future climate, despite a slight increase of the distance in the future. The ER genetic group, the main genetic group at the origin of Vietnam's elite Robusta varieties, had the second highest climatic distance (behind the OB group) in all current and future scenarios. This suggested that Vietnamese clones of ER origin might not be the best suited

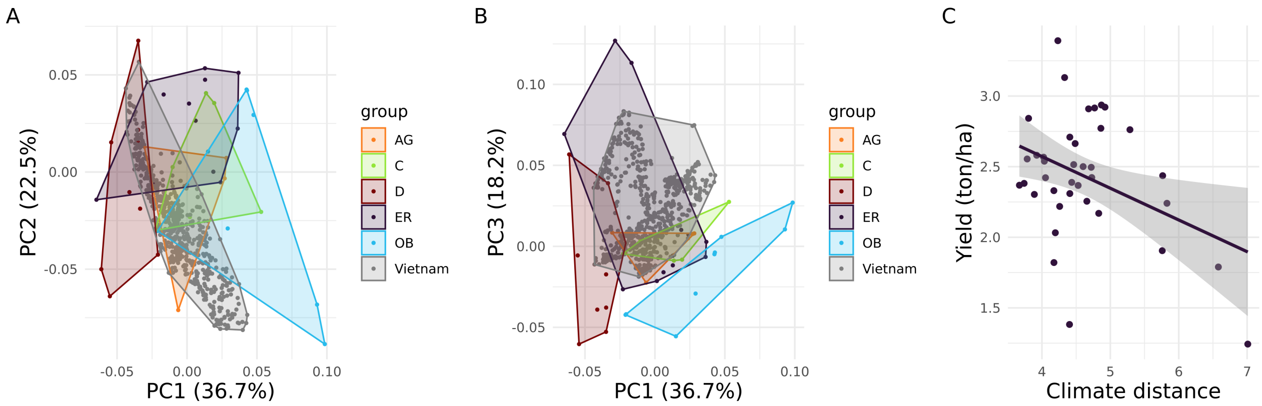


Figure IV.4.1: Climate envelopes and correlation between climatic distances and coffee yield in the present. Plot (A) and (B) Principal component analysis (PCA) of the 19 bioclimatic variables, for the axes PC1 vs PC2, and PC1 vs PC3, respectively. Climatic conditions in native occurences in Africa were treated as active variables while climatic conditions in Vietnam were treated as passive. The lines delimit the environmental envelopes and contain occurrences dots whose color corresponds to their assignment to a genetic group, or the 640 Vietnamese occurrences in grey. (C) Correlation between current climate distances between each Vietnamese district to the African group ER and current coffee yield in Vietnam at district level.

to their current condition in Vietnam and that group AG might be more adapted. Under future predictions, only the climatic distance to group D was predicted to slightly decrease, while the distance will likely remain the same for groups C and ER, or increase for the other groups.

As the current elite varieties in Vietnam originate mainly from the ER group, we compared the climate distances of the African ER group to each of the districts in Vietnam with their average coffee yields, and detected a significant negative correlation (r = -0.392, p-value = 0.015). Based on this correlation, coffee yields were predicted under future climates (Figure IV.4.2 and Supplementary Figure S-IV.8.6). The results were similar regardless of the climate models and scenarios. About a third of the districts, mostly in the west of Gia Lai and Dak Lak province, were expected to show yield reduction, while other districts would experience yield increases. The results based on the CNRM-CM6-1 model and ssp585 scenario, with the most significant changes in yield, predicted for Chu Puh district, Gia Lai province, a decrease up to 1 ton/ha, corresponding to 29.7% of its current yield, and for M'Drak district, Dak Lak province, an increase of 1.2 ton/ha - 86.9% of its current yield.

IV.4.3 Genomic variants associated with climatic factors in African populations

To identify candidate genetic variants underlying climate adaptation, associations between k-mers and the 19 bioclim variables were investigated in the African population using both LFMM and RDA methods. With LFMM, each k-mer was tested with each environmental variable independently. In total, 98,299,004 k-mers were significantly associated with 18 bioclimatic variables (FDR = 0.01). The temperature-related variables (bio1-bio11) were associated with a greater number of k-mers than precipitation-related variables (bio12-bio19). The variables bio2 (mean diurnal range) and bio7 (temperature annual range) were highly correlated (correlation coefficient r = 0.7), and had the highest number of significant k-mers, 71,151,395 and 83,955,391, respectively. No k-mer was found significantly associated with bio16 (precipitation of wettest quarter). In RDA, each k-mer was tested on a PCA space of all environmental variables. The bio7 variable was removed from the tests because of its high correlation with the bio2 variable. This method identified 177,127,153 significantly associated k-mers (FDR = 0.01). A total of 18,403,474 k-mers were commonly detected by both approaches.

Among all the k-mers extracted from the African populations, 1,087,090,346 k-mers (58.6% of the total k-mers) were mapped onto the reference genome with MAPQ > 10, meaning that almost half of the variants were not present on the reference genome. Out of 18M candidate k-mers, only 5,249,029 k-mers (28.5%) were mapped

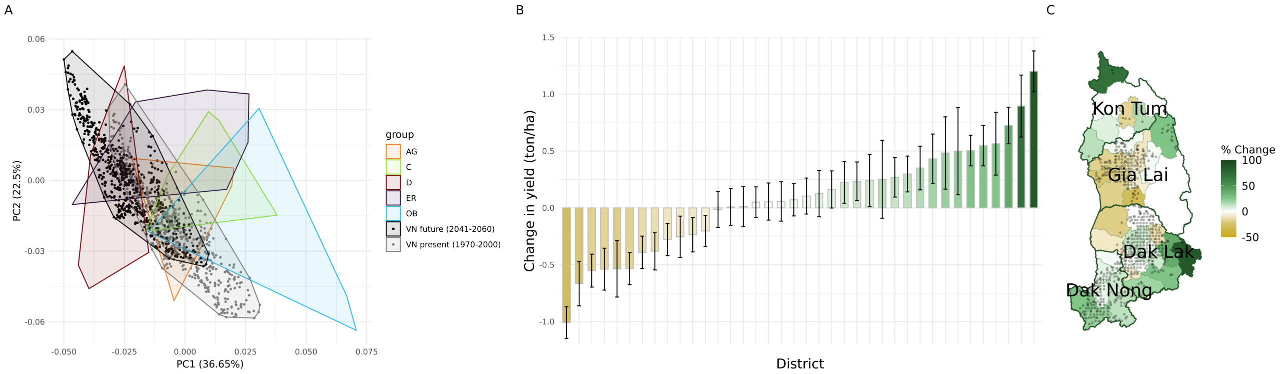


Figure IV.4.2: Climatic change and coffe yield predictions for the future. (A) Projection of Vietnamese bioclimatic variables in the present and future on the PCA of present bioclimatic variables in Africa. The 640 occurrences in Vietnam are shown in grey for the near present (1970-2000) and in black for the future (2041-2060) under the model CNRM-CM6-1 and scenario ssp585. The lines delimit the climatic envelops corresponding to the African and Vietnamese groups. (B) Projected coffee yield changes (ton/ha) in each of the districts in the future (climate model CNRM-CM6-1 and scenario ssp585), based on the ER-districts climate distance-yield correlation. The columns show the average changes, and the error bars the lower and upper limits. (C) Mapping of all the districts with the average change in coffee yields compared with current yields (% Change). Occurrences are represented by black dots. The color and intensity in figure B and C indicates the % change.

Table IV.4.1: Blast results and GO terms associated with contigs

Seq Name	**Seq Length**	**Hit Name**	**Description**	**Sim Mean**	**Align Length**	**GO IDs**	**GO Names**	**SubjectName**
contig_1	468	ref|XP_027064499.1| uncharacterized	uncharacterized protein LOC113690699	92.26	155	P:GO:0090502; F:GO:0003676; F:GO:0004523; F:GO:0046872	P:RNA phosphodiester bond hydrolysis, endonucleolytic; F:nucleic acid binding; F:RNA-DNA hybrid ribonuclease activity; F:metal ion binding	*Coffea arabica*
contig_2	468	ref|XP_027088519.1| uncharacterized	uncharacterized protein LOC113709870	91.8	122	P:GO:0090502; F:GO:0003676; F:GO:0004523	P:RNA phosphodiester bond hydrolysis, endonucleolytic; F:nucleic acid binding; F:RNA-DNA hybrid ribonuclease activity	*Coffea arabica*
contig_63	417	ref|XP_027086722.1| uncharacterized	uncharacterized protein LOC113708466	90.58	138	P:GO:0000723; P:GO:0006281; P:GO:0006310; P:GO:0006974; P:GO:0032508; F:GO:0000166; F:GO:0003678; F:GO:0004386; F:GO:0005524; F:GO:0016787	P:telomere maintenance; P:DNA repair; P:DNA recombination; P:DNA damage response; P:DNA duplex unwinding; F:nucleotide binding; F:DNA helicase activity; F:helicase activity; F:ATP binding; F:hydrolase activity	*Coffea arabica*
contig_115	567	ref|XP_027102884.1| uncharacterized	uncharacterized protein LOC113724156	88.11	185	P:GO:0090502; F:GO:0003676; F:GO:0004523; F:GO:0046872	P:RNA phosphodiester bond hydrolysis, endonucleolytic; F:nucleic acid binding; F:RNA-DNA hybrid ribonuclease activity; F:metal ion binding	*Coffea arabica*
contig_69	414	ref|XP_027124143.1| uncharacterized	uncharacterized protein LOC113740824	87.59	137	F:GO:0003676	F:nucleic acid binding	*Coffea arabica*
contig_13	519	ref|XP_027121946.1| uncharacterized	uncharacterized protein LOC113738868	85.07	134	F:GO:0003676; F:GO:0008270	F:nucleic acid binding; F:zinc ion binding	*Coffea arabica*
contig_80	676	ref|XP_027089582.1| uncharacterized	uncharacterized protein LOC113710661	82.18	174	P:GO:0090502; F:GO:0003676; F:GO:0004523; F:GO:0046872	P:RNA phosphodiester bond hydrolysis, endonucleolytic; F:nucleic acid binding; F:RNA-DNA hybrid ribonuclease activity; F:metal ion binding	*Coffea arabica*
contig_112	401	ref|XP_027096084.1| uncharacterized	uncharacterized protein LOC113715980	78.2	133	P:GO:0090502; F:GO:0003676; F:GO:0004523; F:GO:0046872	P:RNA phosphodiester bond hydrolysis, endonucleolytic; F:nucleic acid binding; F:RNA-DNA hybrid ribonuclease activity; F:metal ion binding	*Coffea arabica*
contig_26	591	ref|XP_027088588.1| uncharacterized	uncharacterized protein LOC113709944	77.12	153	F:GO:0003676; F:GO:0008270	F:nucleic acid binding; F:zinc ion binding	*Coffea arabica*
contig_65	410	ref|XP_027062732.1| uncharacterized	uncharacterized protein LOC113689099	74.36	78	P:GO:0015074; F:GO:0003676; C:GO:0043227	P:DNA integration; F:nucleic acid binding; C:membrane-bounded organelle	*Coffea arabica*
contig_54	479	gb|KYP72617.1| Retrovirus-related	Retrovirus-related Pol polyprotein from transposon TNT 1-94	70.44	159	P:GO:0015074; F:GO:0003676; F:GO:0008270	P:DNA integration; F:nucleic acid binding; F:zinc ion binding	*Cajanus cajan*
contig_86	749	ref|XP_027089582.1| uncharacterized	uncharacterized protein LOC113710661	69.47	131	P:GO:0090502; F:GO:0003676; F:GO:0004523; F:GO:0046872	P:RNA phosphodiester bond hydrolysis, endonucleolytic; F:nucleic acid binding; F:RNA-DNA hybrid ribonuclease activity; F:metal ion binding	*Coffea arabica*
contig_29	441	ref|XP_027103123.1| uncharacterized	uncharacterized protein LOC113724413	63.25	117	P:GO:0090502; F:GO:0003676; F:GO:0004523; F:GO:0046872	P:RNA phosphodiester bond hydrolysis, endonucleolytic; F:nucleic acid binding; F:RNA-DNA hybrid ribonuclease activity; F:metal ion binding	*Coffea arabica*
contig_6	412	ref|XP_027092412.1| serine/threonine-protein	serine/threonine-protein phosphatase 7 long form homolog	61.02	59	P:GO:0010073; P:GO:0048507	P:meristem maintenance; P:meristem development	*Coffea arabica*

on the reference genome. The number of candidate k-mers on each chromosome ranged from 277,568 (on chromosome 3) to 921,025 k-mers (on chromosome 2).

To annotate the biological function of the candidate k-mers, we first assembled them to obtain longer sequences (contigs) with at least 400 bp length. Only 507,029 k-mers were unassembled, and 97% of the candidate k-mers could be assembled into 1,702,000 contigs. The contig lengths ranged from 32 to 796 bp, and the number of assembled k-mers per contig varied from 2 to 576 (Supplementary Figure S-IV.8.7). Among 115 contigs with the length > 400 bp, 45 of them had sequence matches on NCBI non-redundant database of Viridiplantae (50-92% mean similarity). Only 14 contigs were associated to GO terms, which were related to different pathways (molecular function, cellular component, and biological process) (Table IV.4.1). The hit subjects of these 14 contigs were mostly uncharacterized proteins found in C. arabica, but half of them presented the same functions, such as RNA phosphodiester bond hydrolysis, nucleic acid binding, RNA-DNA hybrid ribonuclease activity, and metal ion binding, characteristic of DNA binding proteins such as regulatory ones.

IV.4.4 Genomic suitability of wild African Robusta coffee in Vietnam

To estimate which African populations were best fitted to Vietnamese local climate, we predicted the genetic offset of these individuals under this geographic and climatic change, using the k-mers putatively associated with bioclimatic factors.

Under the current climate, the lowest genetic offsets were found for AG individuals in most regions of the Central Highlands. A few small regions in the south showed a better fit with group ER, and in the east of Dak Lak with groups C and D (Figure IV.4.3A). In contrast, individuals belonging to groups D and OB showed the highest genetic offsets in all districts (Figure IV.4.3C).

Under future climate scenarios, genetic offsets were predicted to increase slightly (Supplementary Figure S-IV.8.8), suggesting that genetic changes required to adapt to local climate in Vietnam would not be significantly affected by climate change. The overall classification of offsets according to the African origin of the genetic material would not change in the future (Figure IV.4.3B and D, and Supplementary Figure S-IV.8.9 and Figure S-IV.8.10). Indeed, group AG would still present the lowest offset in almost all the districts, in five of six climate scenarios. The highest offsets will remain for group D and OB in the future. For a few districts located in the centre of Dak Lak and Gia Lai provinces (Figure IV.4.3D and Supplementary Figure S-IV.8.10), the highest offsets were found for ER individuals under the models MRI-ESM2-0, but for D individuals in the other GCMs.

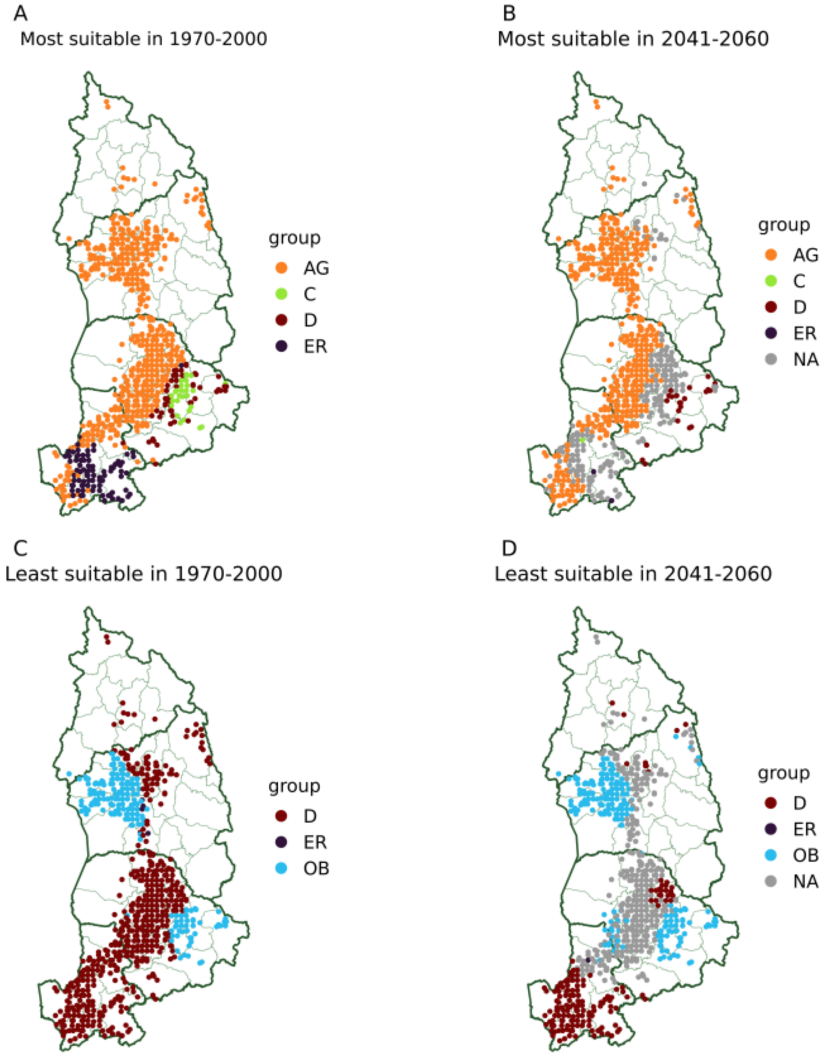

A
Most suitable in 1970-2000
group
AG
C
D
ER
B
Most suitable in 2041-2060
group
AG
C
D
ER
NA
C
Least suitable in 1970-2000
group
D
ER
OB
D
Least suitable in 2041-2060
group
D
ER
OB
NA

Figure IV.4.3: African genotypes best or least suited to Vietnamese local climate. Figure A and C show genetic groups of genotypes with the lowest and highest genetic offsets, respectively, in the current climate for 640 occurrences in Vietnam. Figure B and D show the genetic groups with the lowest and highest genetic offsets, respectively, for future climate. For each occurrence, the genetic group is shown if it has been ranked in its position in at least five out of 6 future predictions, otherwise, it is represented by a grey dot as NA value.

IV.5 Discussion

Predicting the suitability or vulnerability of Robusta coffee to a given environment is important for future planting and breeding strategies in Vietnam. Previous studies forecast a significant decrease in climatic suitability for Robusta coffee production over Vietnam (Bunn et al., 2015; Dinh et al., 2023). These studies were based on environmental niche modeling using current distribution of coffee-growing areas to determine where coffee production could be maintained in the future in a suitable growing environment. However, the estimation of these climatic vulnerabilities were solely based on environmental data and neglected adaptive genomic variation, which plays an important role in the responses of populations under climate change.

Here, we integrated climatic and genomic information when comparing Robusta coffee climate suitability and genetic offset across continents and time. Our analyses are based on the assumption that *C. canephora* species is adapted to its local climate in its areas of origin and that different populations within the species may react in different ways to climate change, depending on their adaptive diversity.

C. canephora native distribution extends over spatially heterogeneous environments and the five main genetic groups occur under different climatic conditions, although there is a degree of overlap in their climatic envelopes. Temperature-related variables, such as minimum temperature of coldest month or mean temperature of warmest quarter, contributed more to climate variability between the groups than precipitation-related variables. Robusta coffee was introduced from Africa to the central highlands in Vietnam. The climatic conditions of this new environment overlap with the African climate envelops. However, the climatic distances show a stronger proximity to the climatic conditions of the AG group, originating from the western coastal regions of Central Africa (Gabon and Angola), whether in the present or in the future.

Most of the genetic diversity introduced to Vietnam and especially the widely cultivated elite varieties, came from the Congo Basin ER group (chapter III). This has also been reported by previous studies (Akpertey et al., 2021; Garavito et al., 2016; Phan, 2017). The introduction of materials from this Congolese group in the past may have been motivated by their robustness and their potential tolerance to leaf rust (Montagnon et al., 1998a), while the climatic suitability of the introduced varieties might have neglected. The mean climatic distance between Vietnamese districts and African ER group is about 1.5 higher than with AG group in the near present (1970-2000). In the future, these climate differences tend to slightly reduce for group ER, but increase for group AG.

Bioclimatic and other environmental factors have been used as predictors of Robusta coffee yield in Vietnam using regression-based models (Dinh et al., 2022). They observed that the sensitivity of yield anomalies to weather varied substantially between provinces and even districts. In our study, since the Vietnamese varieties come from the Congolese ER group, we calculated the climate differences observed between each Vietnamese district and the African ER group. The climate distances to this group were correlated to Robusta yields at the district level. Our linear regression model allowed us to predict yield in the future (2041-2060) with a yield reduction for the central regions of the Central Highlands (west Dak Lak and Gia Lai provinces), and increased yields in other areas such as the northern and south-eastern regions. Therefore, climatic suitability of group ER in Vietnam may not currently be optimal, but could increase in some regions in the future. This geographic distribution of expected yield changes in Vietnam is consistent with changes in suitable planting areas predicted under future climate by the previous studies (Bunn et al., 2015; Dinh et al., 2023). A limitation of our climatic distance approach is that climatic heterogeneity within group was not considered. In particular, the Robusta cultivated in Vietnam could have come from a limited climatic zone within the distribution of the ER group.

To better forecast the potential of populations to respond to climatic changes, we incorporated the climate-adaptive genetic variations in our models. We applied an innovative approach to identify genetic variants associated with the bioclimatic variables. Compared to conventional methods using SNPs, the k-mer approach enables detection of individual-specific variants including structural variants (Gupta, 2021; Voichek and Weigel, 2020). To detect the putatively adaptive (candidate) k-mers we applied both LFMM and RDA methods, which are the most powerful methods currently used. LFMM controls confounding factors such genetic structure but may have high FDR for weak selection, while RDA has lower FDR but does not take into account population structure. Defining the candidate kmers as the ones detected by both methods should leverage the detection power (Forester et al., 2018; François et al., 2016).

A high number of k-mers (more than 18M) were identified as associated with bioclimatic variables, with a majority of them not detected on the reference genome. The assembly of these candidate k-mers led to more than 59,000 contigs with length from 100 to 796 bp, which might include multi-nucleotide variants (Rahman et al., 2018). From the sequences of 115 contigs with length > 400 bp, 14 putative proteins associated with 50 GO terms were identified. Although their functions could not be fully characterized, they present DNA/RNA-, nucleic acid- and ion-binding residues. Zinc ions binding and nucleic acid binding activities have been shown to play a role by regulating plant gene expression in response to stress (Ciftci-Yilmaz and Mittler, 2008). Tournebize et al., 2022, also identified genes, which were associated with bioclimatic factors in *C. canephora*, involved in response to biotic and abiotic stress. These molecular and biological functions may help the plant to gain adaptability in different conditions. Further studies, such as GWAS are needed to validate their role in climate-adaptive traits.

Using the candidate k-mers, we estimated the climate change-related genetic offset between Africa and Central Highlands environments. The genetic offset is based on the assumption that populations have allele frequencies in equilibrium in their African native regions and that this equilibrium will be disrupted when they are introduced into Vietnam. The higher the genetic offset, the greater the vulnerability. Based on all African occurrence – Vietnamese district comparisons, we observed positive correlations between climate distance and genetic offset, but they varied between the different genetic groups. The African groups with the highest or lowest genetic offsets in Vietnam correspond well to those with the climatic conditions are the least or the most suitable, respectively. For example, group AG had the lowest mean climate distance to Vietnam and the lowest genetic offset in most of the tested districts, which will be mostly unchanged in the future. The opposite was found in group OB. Group D had lower climate distances to Vietnam compared to group OB, but may be the least suited in many areas now and in the future.

Integrating climatic suitability and genomic offset will provide better clues for future breeding strategies. The climatic suitability estimated in our study allowed us to understand how the local climate in Vietnam differs from the original conditions of the species, and how this may affect the Robusta coffee yields. Genetic offsets help to determine which genetic origin within the species range can help to cope with the local climate in Vietnam. The main genetic background of Vietnamese Robusta clones comes from ER group, which is probably not be the most suited to Vietnamese climate. The AG group contributes to a lesser extent to the Robusta diversity present in Vietnam through admixtures in ER Congolese genetic backgrounds (chapter III). However, it has been shown here to have both the highest climatic suitability and lowest genetic offset in the future. Therefore, the introduction of more AG material and the generation of additional ERxAG hybrids can be promising for improving the adaptability of coffee to climate change in Vietnam, especially in the central regions where yield loss is expected in the future.

Among other applications, these genetic offsets can guide assisted introduction or breeding programmes, helping to choose complementary non-local but better-adapted genotypes, thereby increasing adaptive diversity and resilience (Aitken and Bemmels, 2016).

IV.6 Acknowledgements

The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/- http://www.southgreen.fr

IV.7 Data Availability

The raw sequencing data of the African accessions were obtained from the NCBI SRA database under project accession number PRJNA803612 (Tournebize et al. 2022). The raw sequencing data of the 10 Vietnamese core accessions are available in the NCBI SRA database under project accession number PRJNA950219. The coffee yield data in Vietnam was obtained from Vietnam's General Statistics Office (GSO) upon request.

IV.8 Supplementary data

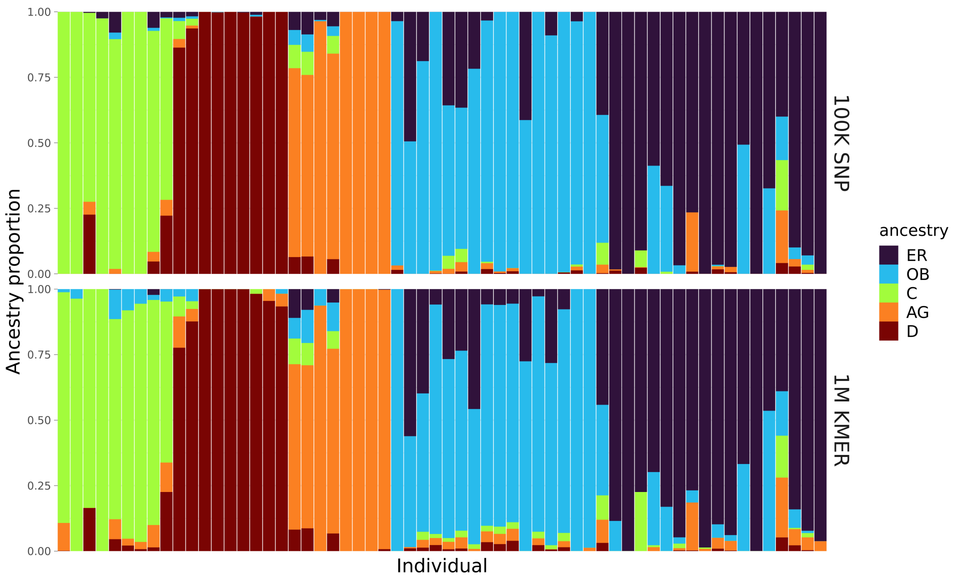


Figure S-IV.8.1: Comparison of sNMF results using 1M random k-mers and 100K genome-wide SNPs. Each column represents ancestry proportions of each African individual.

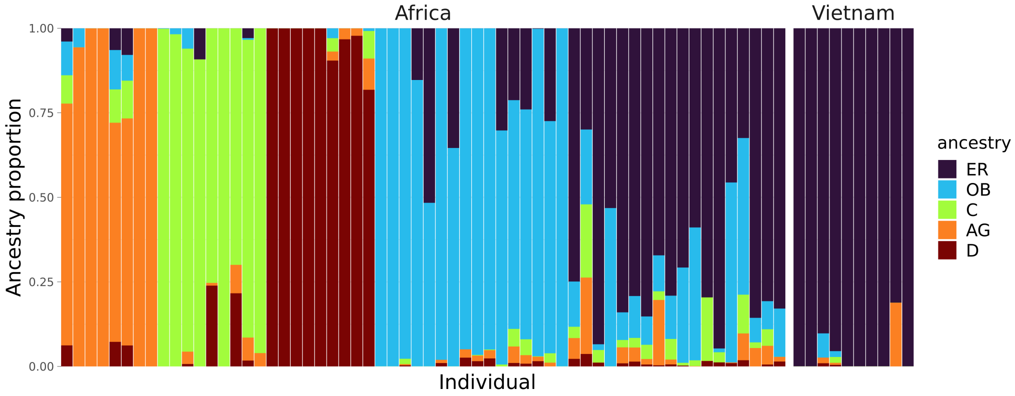


Figure S-IV.8.2: Genetic structure of all the African and Vietnamese elite individuals, using a random k-mer subset. Each column represents ancestry proportions of each individual.

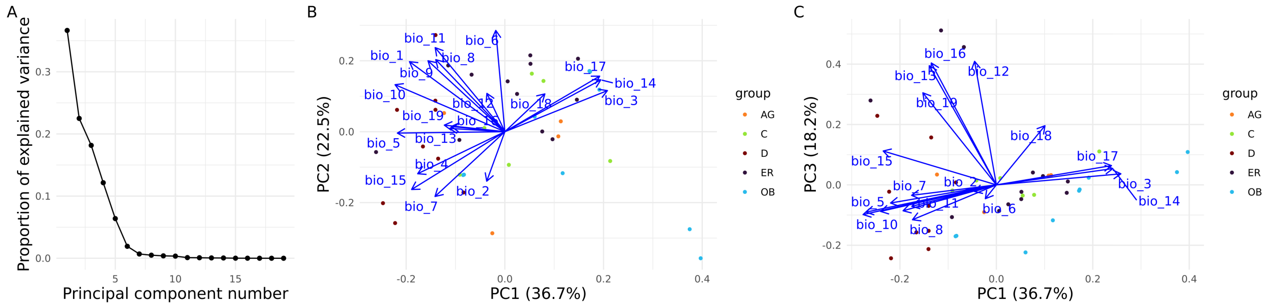


Figure S-IV.8.3: PCA results of African current bioclimatic data. (A) Proportion of variance explained by the principal components. Plot (A) and (B) showed PC1 vs PC2, and PC1 vs PC3, respectively, of bioclimatic data in Africa for the near present (1970-2000). Blue arrows indicate the contribution of the 19 bioclimatic variables.

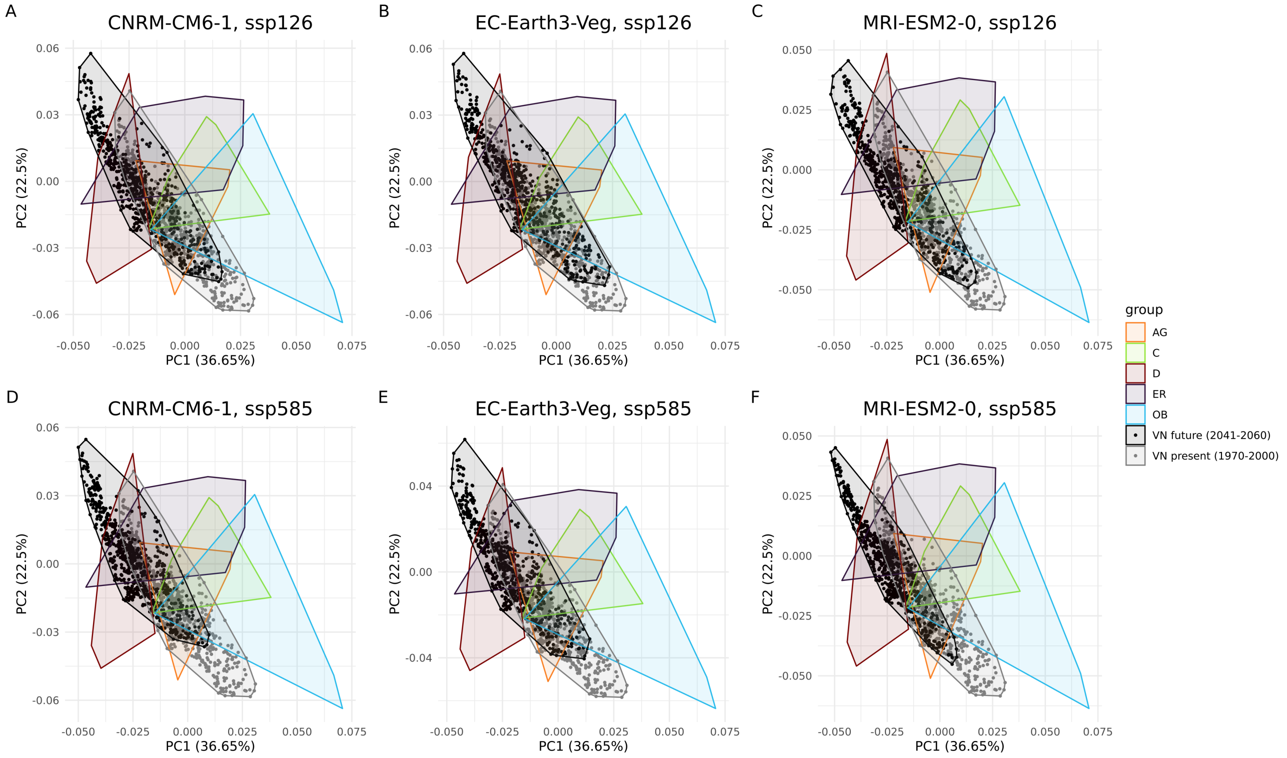


Figure S-IV.8.4: Projection of Vietnamese future bioclimatic data onto the PCA obtained from African current bioclimatic data. All the plots show PCA results from African present bioclimatic data with climatic envelopes delimited by color lines, and the projection of the Vietnamese current bioclimatic data in grey. Future bioclimatic data of Vietnam, under each of the three GCM (CNRM-CM6-1, EC-Earth3-Veg, and MRI-ESM2-0) and the two ssp scenarios (ssp126, and ssp585) are shown in black.

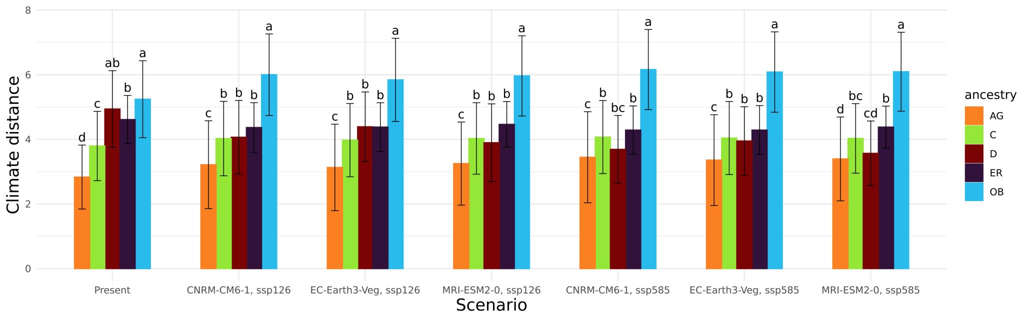


Figure S-IV.8.5: Average climate distance between genetic groups in Africa and Vietnamese districts. Significant levels between groups are presented by the letters (a, b, c, d), based on the Student-Newman-Keuls test with alpha = 0.05.

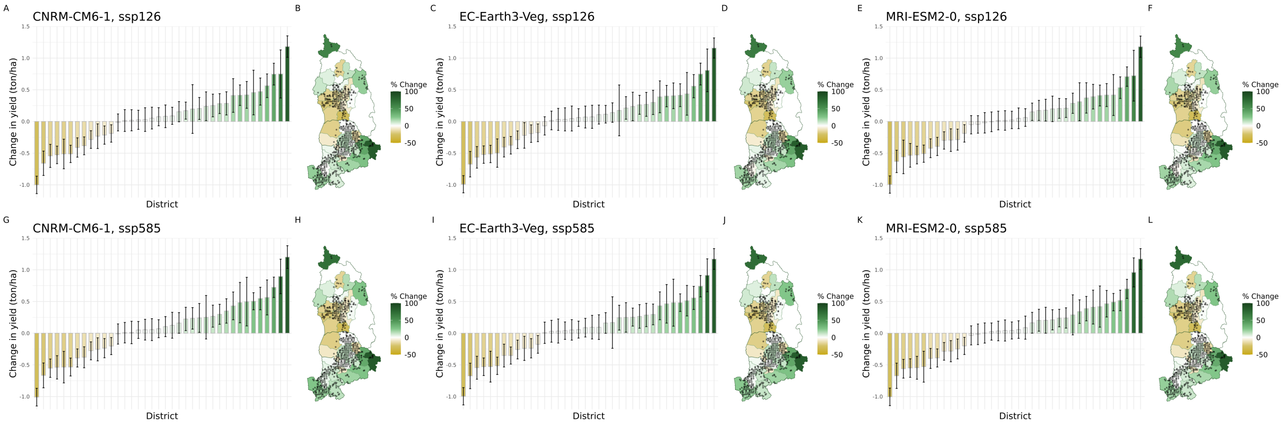


Figure S-IV.8.6: Robusta yields predictions in the Central Highlands (Vietnam) based on climate distance. Plots A and B are corresponding to prediction under CNRM-CM6-1 and ssp126, plots C and D for EC-Earth-Veg and ssp126, plots E and F for MRI-ESM2-0 and ssp126, plots G and H for CNRM-CM6-1 and ssp585, plots I and J for EC-Earth-Veg and ssp585, and plots K and L for MRI-ESM2-0 and ssp585. Projected yield changes (ton/ha) in all districts under the future climate prediction, based on the climate distance-yield correlation. The columns showed the average changes, and the error bars showed the lower and upper limits. The maps show the districts with average yield change compared to the present yields (% Change). The occurrences are represented by black dots. The color and the intensity in figure B and C indicates the % change.

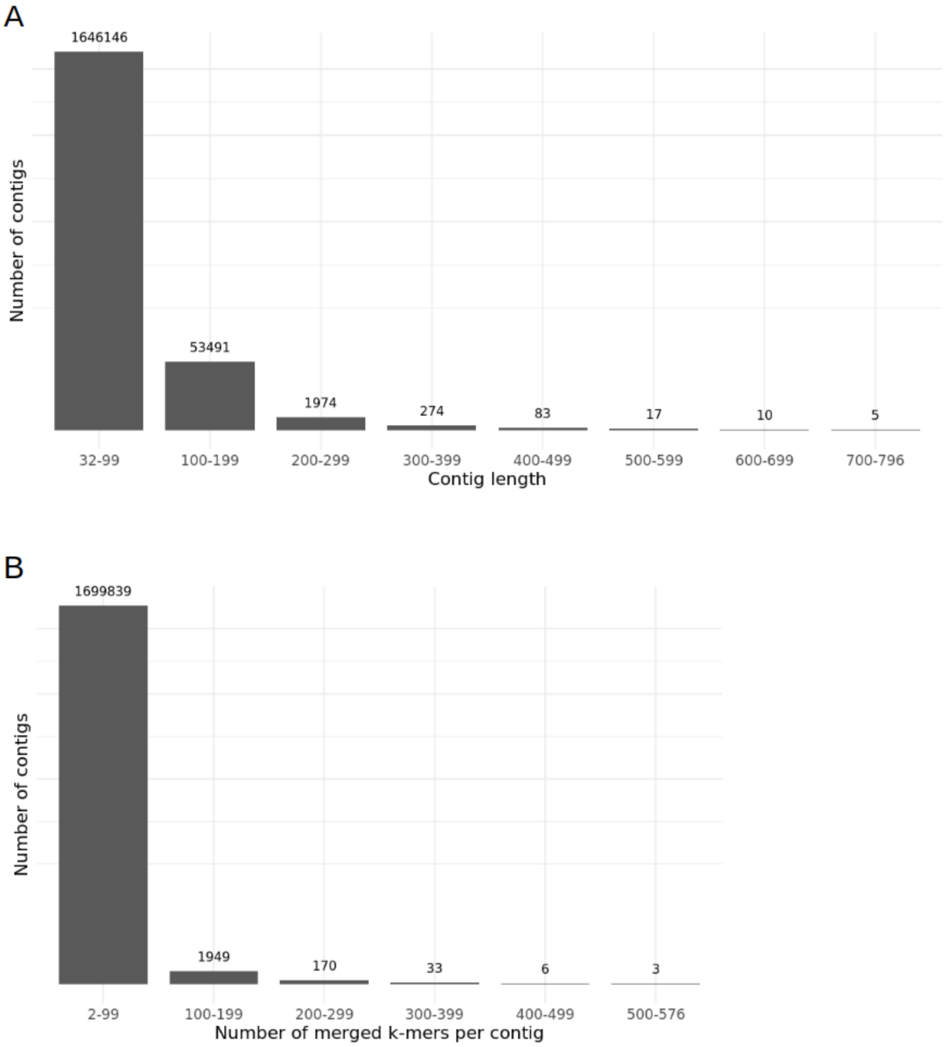


Figure S-IV.8.7: Contigs assembed from 18M candidate k-mers. (A) Distribution of contig lengths (bp). (B)Distribution of number of assembled k-mers per contig.

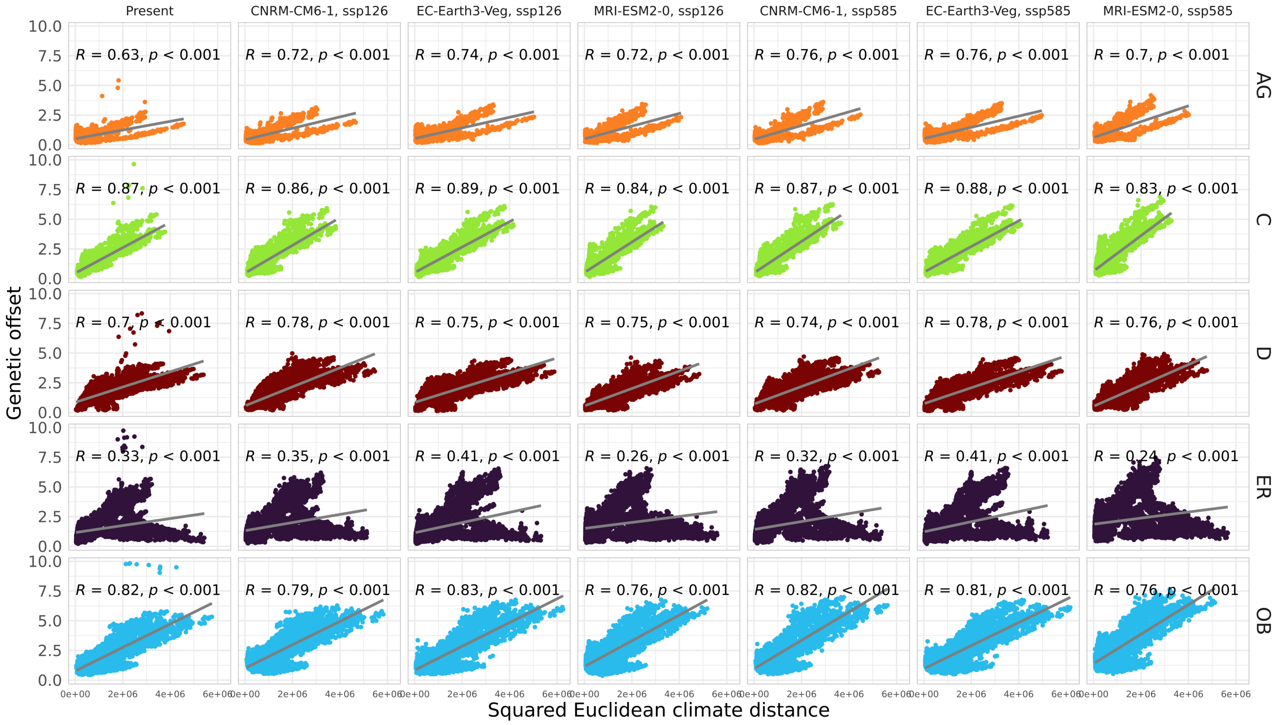


Figure S-IV.8.8: Correlation between genetic offset and climate distance. Each point represents one African individual in one occurrence in Vietnam. The squared Euclidean climate distances were calculated using the 19 bioclimatic variables.

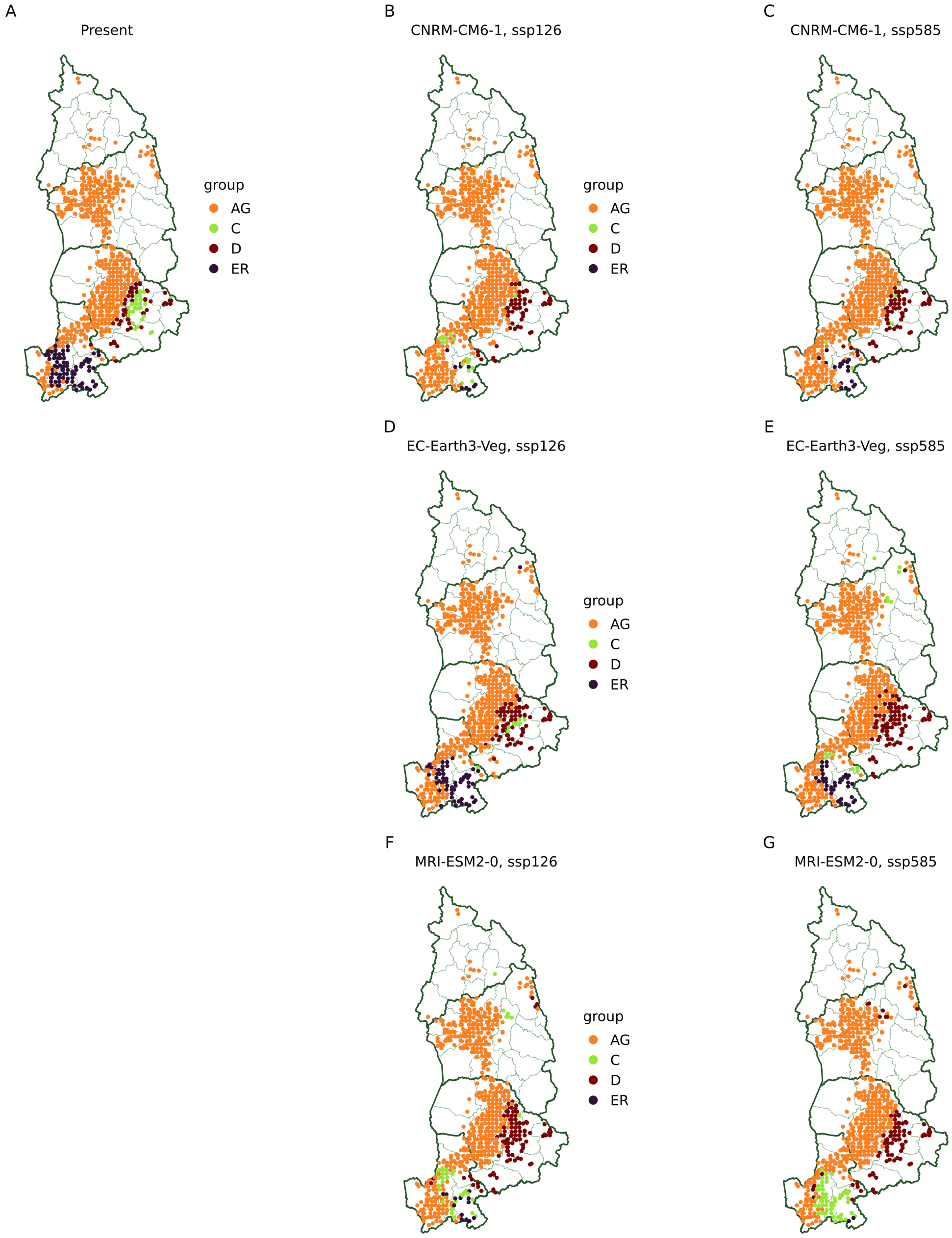


Figure S-IV.8.9: Genetic groups of the genotypes with the lowest genetic offset in 640 occurrences in Vietnam.

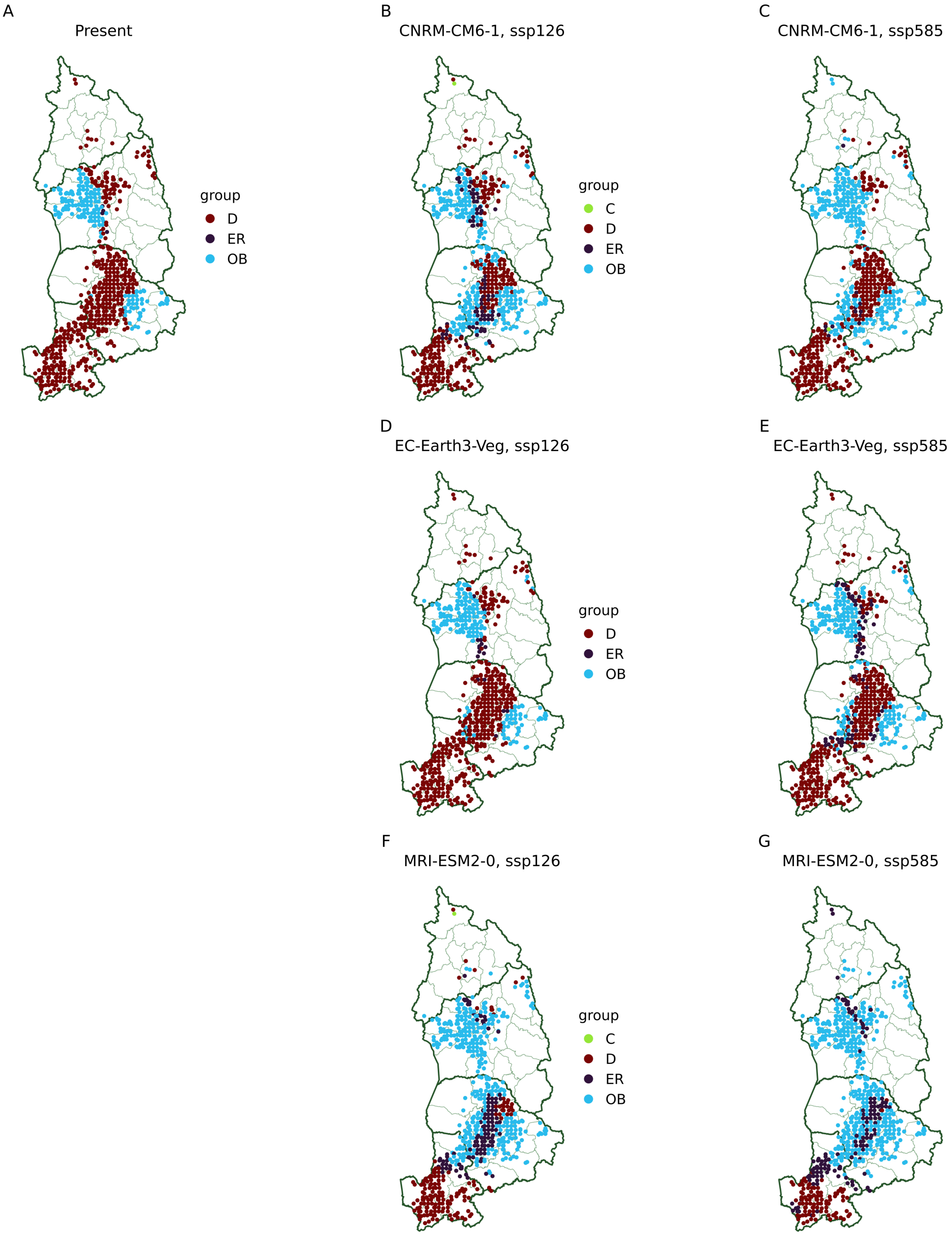


Figure S-IV.8.10: Genetic groups of the genotypes with the highest genetic offset in 640 occurrences in Vietnam.

V General discussion

Under future climate conditions, Vietnamese Robusta coffee trees (*C. canephora*) are at risk of becoming less suitable to the local climate, resulting in a loss of coffee production. This problem will have huge impacts on agriculture, ecology, and the economy of Vietnam, the world's largest Robusta producer. One sustainable measure to mitigate the threats of climate change on Robusta coffee is to establish breeding plans, focusing on the development of new varieties with high adaptability and tolerance to abiotic/biotic stresses. To facilitate future breeding programs, this PhD project has addressed two main research questions: What part of the genetic diversity from the wild source populations has been transmitted to Vietnamese Robusta coffee from the wild source populations, and Which genetic diversity is suitable to the predicted local climate in Vietnam in the future. The different chapters included specific discussions on the narrower topics of each study. Here, we discuss the key findings, limitations, prospects and implications of this research.

V.1 Key findings on genetic diversity and suitability of Robusta coffee in Vietnam

All the Robusta varieties cultivated in the world originated from Africa. The wild populations in our studies have been well classified into five genetic groups corresponding to the well-known geographical groups (Mérot-L'Anthoëne et al., 2019): D in Guinean regions, C in Cameroon, OB in eastern Central African Republic (CAR) and Uganda, AG in Gabon and Angola, and ER in DRC. Robusta coffee trees have not experienced long-time domestication and strong selection like other crops (Meyer et al., 2012), since they have only been widely cultivated since the 1900s (Berthaud, 1986; Cramer, 1957; Montagnon et al., 1998a). Therefore, tracing back the origin of the cultivated accessions and relate them to the wild genetic diversity can be straightforward.

Robusta coffee was introduced to Vietnam in the early 20th century, presumably from Java, Indonesia (Phan, 2017; ICO, 2019). Java was the first breeding center of Robusta coffee with contribution of Congolese accessions, which were previously known as "Robusta" varieties from the Yangambi region in northern Democratic Republic of the Congo (DRC) (Coste, 1955; Cramer, 1957), and unknown varieties from Uganda and Gabon (Montagnon et al., 1998b). The varieties early introduced into Vietnam appear to have been mainly Congolese genotypes (group E and R from DRC), as shown by the main genetic groups contributing to the ancient and current accessions in the Central Highlands. However, about one fourth of the 126 studied materials also showed contribution from one or two other sources, in a Congolese genetic background. Introgressed segments from all the other groups, with the exception of group C, were detected. We have not found any "pure" individuals from these groups (with the only exception of group ER), although their presence in the rest of the germplasm bank could be possible. Although there is evidence of both recent and long-time admixture events, it is unknown whether inter-group crosses occurred before or after diffusion to Vietnam. Since 1976, researchers and breeders at the WASI (Western Highlands Agriculture and Forestry Science Institute located in Dak Lak) have selected and developed new varieties, and identified elite clones with high productivity, good cup quality, and tolerance to leaf rust (Phan, 2017; ICO, 2019). Most of the elite clones originated solely from the Congolese group ER, except for one resulting from a backcross between the ER and AG groups. These elite varieties are the most commonly cultivated in Vietnam (ICO, 2019), and used in varietal mixtures in plantations (due to its self-incompatibility). Therefore, diversity of the Vietnamese local clones is relatively limited and restricted to group ER origin.

The early diffusion of Robusta accessions from the Congo Basin to Southeast Asia (Coste, 1955; Cramer, 1957) might be because of its vigor characteristics (e.g large leaves, high bean weight) and high resistance to leaf rust (Leroy et al., 1993a; Leroy et al., 1997; Montagnon et al., 1998a). However, one aspect that has probably been underestimated is whether the climate in this region is suitable for this genetic group, or whether another genetic group might better suit these local conditions. Many studies have shown the effects of environmental conditions on coffee yield, identified genes or QTLs associated with abiotic stress, and loci associated with climatic variations (Bunn et al., 2015; de Carvalho et al., 2013; de Carvalho et al., 2014; Dinh et al., 2022; Kath et al., 2020; Kath et al., 2021; Kiwuka et al., 2021; Marraccini et al., 2012; Torres et al., 2019; Tournebize et al., 2022). Therefore, more attention needs to be paid to the climatic mal-adaptation of the cultivated accessions and to assessing which genetic source might be better suited to local climate.

Climate distances between the cultivated districts in Vietnam and the native regions of the ER group, which corresponds to the largest contributor to Vietnamese germplasm, were found significantly correlated with coffee yields. By 2060, coffee yields in some districts are expected to decrease by up to 30% of current yields. The genetic group ER might not be the most suitable to the local climate of the Central Highlands, where bioclimatic conditions are closer to the native regions of the AG group. This speculation has been confirmed by investigating the genotype-environment relationship in the wild *C. canephora* populations. Wild populations are assumed to present genetic variants adapted to their local climate (Rellstab et al., 2021). Their equilibrium in the native environment will be disrupted in a new environment, which can be estimated by the genetic offset and used to assess which genotypes have the lowest or highest risks of mal-adaptation. A large number of genetic variants (18M k-mers) were identified as associated with bioclimatic variables, suggesting a polygenic adaptation model consistent with the fact that most adaptive traits are polygenic in forest trees (de Miguel et al., 2022; Sork et al., 2013). Further studies with more samples should be led to understand adaptive traits and genotype-phenotype interaction underlying local adaptation in Robusta coffee. Based on these genotype-environment relationship and correction for confounding factors of population structure (Gain et al., 2023), wild accessions of group AG seemed to be also the most suitable genotypes, i.e. with the lowest genetic offsets, to local climate of most regions of the Central Highlands, in the present and by 2060. The results indicated a mismatch between the widely cultivated genotypes in Vietnam and the predicted genotypes suitable to local environment.

V.2 How does the genetic diversity translate to phenotypic diversity?

Morpho-agronomic traits in coffee are mediated by both genetic heredity and environment, but some might have high heritability (Leroy et al., 1993b; Montagnon et al., 1998a). A set of characterization descriptors including both qualitative and quantitative traits, which are highly heritable and expressed equally in different environments, has been developed (Anthony, 1996). These morphological characters can be used as markers to identify varietal diversity and complement genetic variability, to better understand evolutionary change in *C. canephora* (Loor Solórzano et al., 2017) and local adaptation (Láruson et al., 2020). Morphological data had similar capacity as molecular markers in discriminating two main varietal groups - “Robusta” (group ER) and “Conilon” (group A) (L. F. V. Ferrão et al., 2013), and could detect variability between clonal accessions (Leroy et al., 1993b). Moreover, several morphological traits in coffee are indicators of coffee productivity and quality, and therefore are important criteria for selection in breeding programs. For example, bean’s characteristics, such as the shape, color, and moisture, are ones of attributes for coffee quality standard defined by the International Organization for Standardization (ISO) (Leroy et al., 2006). Spinelli et al., 2018, found hulled coffee yield was strongly influenced by the number of plagiotropic branches and the number of rosettes per productive branch in 130 Robusta clones in Brazil. Akpertey et al., 2022, found that diameter of laterals and number of nodes per lateral were significantly associated with yield in 166 Ghanaian local varieties. Such phenotype evaluations should also be applied in the Vietnamese Robusta coffee, in order to assess the redundancy of genetically similar genotypes, and determine selection criteria for breeding programs. However, phenotyping the WASI germplasm collection is fairly challenging because the plants are at different ages. Many plant characteristics varying at different growth stages, such as tree height, leaf size, or number of flowers, which need to be assessed on materials of the same age.

In *C. canephora*, several traits of agronomic interest were reported as group-specific characteristics, such as drought tolerance in group AG, or resistance to leaf rust in group ER (Montagnon et al., 1998a; P. Musoli et al., 2009). Breeding programs often relies on the exploitation of heterosis or the introgression of these traits of interest through inter-group breeding (Leroy et al., 1993b; Leroy et al., 1997; Montagnon et al., 1998a; Montagnon et al., 1998b). Outputs of inter-breeding are highly varied and it requires multiple trials and selections (Campuzano-Duque and Blair, 2022. Understanding architecture of these traits, such as QTLs and additive effects, can help to accelerate breeding trials. Up to date, there have been only few studies using GWAS in *C. canephora* species, and most of them were based on single-locus model while yield and many important agronomic traits were proven to be polygenic (M. A. G. Ferrão et al., 2023; P. C. Musoli et al., 2013). As Vietnamese Robusta coffee is increasingly threatened by drought and nematode, and as the associated tolerance/resistance traits are also complex traits, studying multi-locus GWAS of these characters would be a promising approach (Segura et al., 2012). Moreover, by using a population with a sufficiently high number of admixed individuals, it is also possible to identify ancestry segments associated with targeted traits, and adaptive introgression (Fetter and Keller, 2023). Understanding quantitative genetics of the desirable traits could help to effectively select suitable parental materials for crossing and producing hybrid vigor, and assist selection stages (M. A. G. Ferrão et al., 2023).

V.3 Can prediction of genomic suitability be improved?

Local adaptation is driven by interactions between environment, phenotype, and genotype (Figure V.3.1). Environmental variations exert selective pressures on fitness-related traits, and sequentially on allele frequencies. This complex system could be tractable by several approaches at different levels (Rellstab et al., 2015). One approach, applied in this project, is directly detecting the relationship between environment and genotype. This approach could predict which genotypes would be the most suitable to a specific environment, but may fail to predict their phenotypic response. The way populations from different climate zones respond to a new environment, and the traits that are responsible for the responses of populations can be studied by means of common garden experiments (Schwinning et al., 2022). In addition, genetic variants controlling adaptive traits can be characterized using GWAS or QTL mapping and, finally, candidate genes can be validated using gene expression profiling.

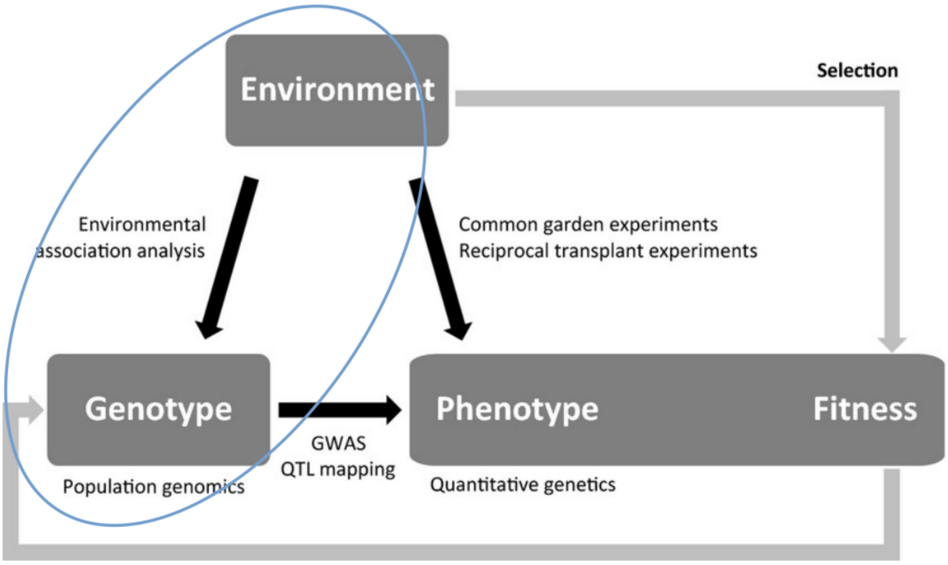


Figure V.3.1: Scheme of detecting signals of local adaptation (figure from Rellstab et al., 2015). Local adaptation is mediated by three components, environment, phenotype, and genotype. Variation in genotype and phenotype can be evaluated by population genomics and quantitative genetics. Gray arrows indicate the selection pressure from environment to phenotype (fitness), and from phenotype to genotype. Black arrows indicate the direction of evolution process, i.e. environment causes evolution changes on genotype and phenotype, and phenotype is affected by both environment and genotype. Interaction between components can be detected, i.e. environment association analysis for identifying genetic markers linked to environmental factors, common garden experiments and reciprocal transplant experiments for detection of fitness traits associated with environment, and GWAS and QTL mapping for identification of genetic markers controlling phenotype. Blue circle indicates the study areas that were investigated in this project.

Genomic offsets of wild *C. canephora* individuals from Africa environments to Vietnam climate have been estimated, to predict the degree of expected mal-adaptation or suitability to the present and future local environment. Genomic offset is an estimate of genetic changes that an individual needs to adapt to a new environment, based on an assumption that it has been adapted to its current environment (Gain et al., 2023). Applying genomic offset on establishing breeding programs is not straightforward, and requires further validation and better understanding of the method's limitations (Lind et al., 2023; Rellstab et al., 2021). The performance of genetic offset relies on many factors, and could be improved by several ways. First, the outcomes of genetic offset in future climates depends on the climate prediction models. Therefore, the choice of appropriate GCMs and improvement in predictive models are important. Second, the selection of climatic predictors can affect the offset estimates, for example, a broader predictor set can sometimes outperform a narrower predictor set (Lachmuth et al., 2023). Third, as using candidate loci in geometric genetic offset improves the prediction power (Gain et al., 2023), even though the selection of loci was proven to have insignificant effect in other genetic statistics (Fitzpatrick et al., 2021; Lachmuth et al., 2023; Láruson et al., 2022), the detection power of selective markers may also affect the offset prediction. Even the most powerful methods, LFMM and RDA, showed a high level of false positives, and require stringent error rate controls (Forester et al., 2018). Forth, measurements of population performance in common gardens (Fitzpatrick et al., 2021), and GWAS to pinpoint genomic regions possibly underlying the adaptive traits (such as fruit weight, or fruit size) can validate the accuracy of genetic offset method, and provide more insights into the mediation of environment on fitness traits of the species. Genetic offset approaches assumes a certain level of mal-adaptation when an adapted population is transplanted from its native environment to a new environment (Capblancq et al., 2020; Rellstab et al., 2021). However, does this assumption always hold? Does increase in genetic offsets always lead to decrease in fitness? Predictive models of genetic offset can be enhanced or revisited by taking into account the fitness quantification of putative loci (Rellstab et al., 2021). Furthermore, adaptive capacity of a species depends on many other genetic factors and population dynamics, such as the reproductive mode, generation time, migration demography, occurrence density, etc., which can be integrated into the prediction of adaptability in the future (Lind et al., 2023; Thurman et al., 2020).

Based on genotype-environment models, not only genomic offset of the wild adapted populations can be estimated, but suitability of new genotypes in existing or new environment is also possibly interpolated. This approach, if applicable, will be useful when wild populations are not accessible in a new environment, or to predict the adaptability of hybrids from breeding programs.

V.4 Suggestion for future breeding plans

Although further studies are needed to improve knowledge of the diversity and adaptive genetics of Robusta coffee in Vietnam, the present results could provide some recommendations for conservation and breeding plans.

As the genotypes of Vietnamese Robusta coffee are essentially of the same geographic origin and closely related to one genetic group, the Congolese ER group, introduction of diversity from the other groups will be useful (H. R. Oliveira et al., 2020; Rius and Darling, 2014) . The AG genetic group originating from Gabon and Angola (including the cultivated "Conilon" type), in particular, should be the focus of more attention. It may have greater adaptive potential in Vietnamese local environment, and has shown heterosis when crossed with group ER ("Robusta" type) materials for varietal improvement. More generally, sources of genetic diversity could be largely sought in crop wild relatives, traditional varieties that already carry beneficial traits (Swarup et al., 2021), or introgression lines which might have new combination of alleles or linkage breakup between beneficial and non-beneficial alleles at linked sites (Flint-Garcia et al., 2023).

Based on these two varietal types, Robusta (ER) and Conilon (AG), three main breeding strategies can be proposed: intra-population recurrent selection within groups (RS), inter-population (reciprocal) recurrent selection between groups (RRS), intra-population recurrent selection within the hybrid (crossed) population (RSH) (Alkimim et al., 2021). RS is commonly used to improve Robusta coffee varieties (Montagnon et al., 1998b), and has been applied effectively in Cote D'Ivoire (Leroy et al., 1997). RRS consists of several cycles of inter-crossing parents from 2 groups, selecting parents that produce the best hybrids, then intra-group crossing of the best parents to improve each group (Montagnon et al., 1998b). The chance of having superior hybrids increases after each cycle. RS, which was proposed in Columbia, is a simpler derivative of RSS, with the improvement of a single group by intra-group crossing (Campuzano-Duque and Blair, 2022). RSH is based on selection and intra-group crossing within the hybrid population resulting from the crossing of two groups. This method has not been widely used in coffee, but was shown as simpler and more effective in terms of genetic gain than RRS for all ranges of traits, while RRS was only suitable to improve vegetative vigor (Alkimim et al., 2021). In the case of Vietnamese Robusta, as two of the most focused selection criteria are drought tolerance and nematode resistance, RSH might be a good option. In the WASI germplasm bank, some elite varieties have already been characterized for their genetic structure, as well as several hybrids (especially ERxAG hybrids) displayed various admixture patterns. These accessions could constitute potential materials for starting breeding populations.

Genomics and genetics are playing an increasingly important role in crop breeding. M. A. G. Ferrão et al., 2023, recently proposed an improved RRS program with genomic-assistance (Figure V.4.1). Thanks to the power of identifying genetic markers associated with traits or environment, phenotypic selection, which is laborious and time-consuming, could be replaced by molecular-assisted selection or genomic prediction. Applying genomic selection to Brazillian Robusta breeding has reduced the duration of the selection cycle by half, with high selective efficiency (Alkimim et al., 2020). This breeding strategy will possibly take over from the traditional breeding program, but still requires extensive knowledge of key-targeted traits. A better understanding of the genetic basis of agronomic traits, beneficial phenotypes, and local adaptation will be the key to coffee improvement.

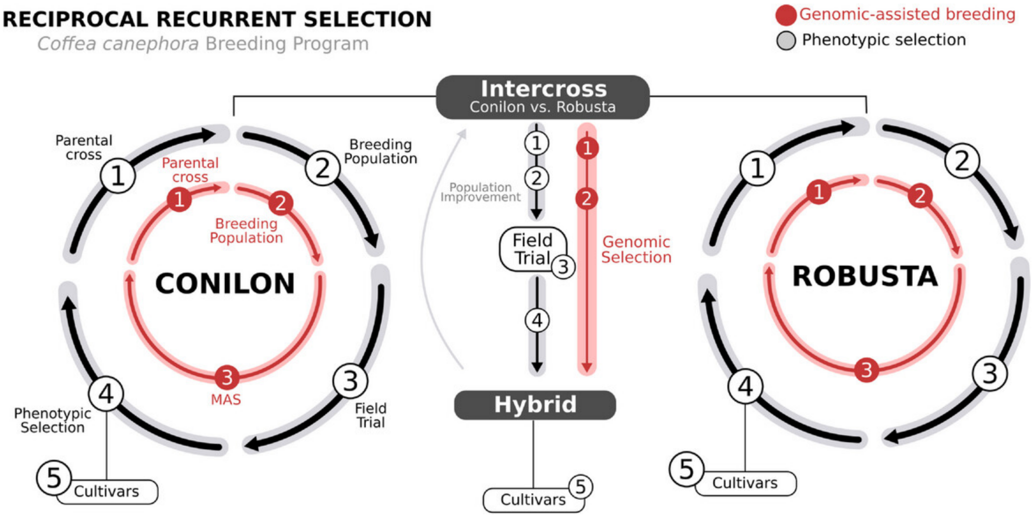


Figure V.4.1: Genomic-assisted RRS and traditional RRS (figure from M. A. G. Ferrão et al., 2023. The first stage is improving two parental populations independently, and the second stage is generating inter-group hybrids. The traditional RRS on both stages was based on selection of only phenotype, and consists of four main steps: (1) defining parental cross, (2) developing breeding population, (3) testing breeding population on field trial, and (4) selecting best genotypes based on phenotypic criteria. With genomic-assisted RRS, the process can be sped up, with three steps on the first stage and two on the second stage. On the first stage, molecular markers associated with desirable traits can support selection of parental cross and predict potential genotypes instead of field trials and phenotypic selection (molecular-assisted selection, MAS). On the second stage, genomic prediction (step 2) can forecast directly the genotypes of inter-group hybrids from the selected parents (step 1).

Bibliography

Aitken, S. N., & Bemmels, J. B. (2016). Time to get moving: assisted gene flow of forest trees [ISBN: 1752-4571 Publisher: Wiley Online Library]. *Evolutionary applications*, *9*(1), 271–290.

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, *1*(1), 95–111. https://doi.org/10.1111/j.1752-4571.2007.00013.x

Akpertey, A., Anim-Kwapong, E., & Adu-Gyamfi, P. K. K. (2022). Broadening the gene pool of Robusta coffee (*C. canephora*): Characterization and character association analysis of local collections based on agro-morphological traits. *Ecological Genetics and Genomics*, *24*, 100126. https://doi.org/10.1016/j.egg.2022.100126

Akpertey, A., Padi, F. K., 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. https://doi.org/10.3389/fpls.2020.612593

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals [Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab]. *Genome Res.*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alkimim, E. R., Caixeta, E. T., Sousa, T. V., Gois, I. B., Silva, F. L. d., Sakiyama, N. S., Zambolim, L., Alves, R. S., & Resende, M. D. V. d. (2021). Designing the best breeding strategy for *Coffea canephora*: Genetic evaluation of pure and hybrid individuals aiming to select for productivity and disease resistance traits [Publisher: Public Library of Science]. *PLoS ONE*, *16*(12), e0260997. https://doi.org/10.1371/journal.pone.0260997

Alkimim, E. R., Caixeta, E. T., Sousa, T. V., Resende, M. D. V., da Silva, F. L., Sakiyama, N. S., & Zambolim, L. (2020). Selective efficiency of genome-wide selection in Coffea canephora breeding. *Tree Genetics & Genomes*, *16*(3), 41. https://doi.org/10.1007/s11295-020-01433-3

Allaby, R. G., Ware, R. L., & Kistler, L. (2019). A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evolutionary Applications*, *12*(1), 29–37. https://doi.org/10.1111/eva.12680

Almazroui, M., Saeed, F., Saeed, S., Nazrul Islam, M., Ismail, M., Klutse, N. A. B., & Siddiqui, M. H. (2020). Projected Change in Temperature and Precipitation Over Africa from CMIP6. *Earth Syst Environ*, *4*(3), 455–475. https://doi.org/10.1007/s41748-020-00161-x

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amalraj, V. A., Balakrishnan, R., Jebadhas, A. W., & Balasundaram, N. (2006). Constituting a core collection of Saccharum spontaneum L. and comparison of three stratified random sampling procedures [ISBN: 0925-9864 Publisher: Springer]. *Genetic Resources and Crop Evolution*, *53*, 1563–1572.

Anagbogu, C. F., Bhattacharjee, R., Ilori, C., Tongyoo, P., Dada, K. E., Muyiwa, A. A., Gepts, P., & Beckles, D. M. (2019). Genetic diversity and re-classification of coffee (Coffea canephora Pierre ex A. Froehner) from South Western Nigeria through genotyping-by-sequencing-single nucleotide polymorphism analysis. *Genet Resour Crop Evol*, *66*(3), 685–696. https://doi.org/10.1007/s10722-019-00744-2

Anim-Kwapong, E., & Boamah, A. (2010). Genetic and environmental correlations between bean yield and agronomic traits in Coffea canephora. *Plant Breed. Crop Sci*, *2*(4), 64–72.

Anthony, F., Combes, M., Astorga, C., Bertrand, B., Graziosi, G., & Lashermes, P. (2002). The origin of cultivated Coffea arabica L. varieties revealed by AFLP and SSR markers. *Theor Appl Genet*, *104*(5), 894–900. https://doi.org/10.1007/s00122-001-0798-8

Anthony, F. (1996). *Descriptors for Coffee (Coffea Spp. and Psilanthus Spp.)* [Google-Books-ID: P1coCx8eeLsC]. Bioversity International.

Arango-López, J., Orozco-Arias, S., Salazar, J. A., & Guyot, R. (2017). Application of Data Mining Algorithms to Classify Biological Data: The Coffea canephora Genome Case. In A. Solano & H. Ordoñez (Eds.), *Advances in Computing* (pp. 156–170). Springer International Publishing. https://doi.org/10.1007/978-3-319-66562-7_12

Babar, A. D., Zaka, A., Naveed, S. A., Ahmad, N., Aslam, K., Asif, M., Maqsood, U., Vera Cruz, C. M., & Arif, M. (2022). Development of Basmati lines by the introgression of three bacterial blight resistant genes through marker-assisted breeding. *Euphytica*, *218*(5), 59. https://doi.org/10.1007/s10681-022-03013-z

Badu-Apraku, B., & Yallou, C. G. (2009). Registration of Striga-Resistant and Drought-Tolerant Tropical Early Maize Populations TZE-W Pop DT STR C4 and TZE-Y Pop DT STR C4. *Journal of Plant Registrations*, *3*(1), 86–90. https://doi.org/10.3198/jpr2008.06.0356crg

Bai, Y., & Lindhout, P. (2007). Domestication and Breeding of Tomatoes: What have We Gained and What Can We Gain in the Future? *Annals of Botany*, *100*(5), 1085–1094. https://doi.org/10.1093/aob/mcm150

Baker, P. S. (2016). *Coffee and climate change in the Central Highlands of Vietnam*. Coffee&Climate, Hanns R. Neumann Stiftung, Germany.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., & Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, *28*(10), 1359–1367. https://doi.org/10.1093/bioinformatics/bts144

Bay, R. A., Harrigan, R. J., Underwood, V. L., Gibbs, H. L., Smith, T. B., & Ruegg, K. (2018). Genomic signals of selection predict climate-driven population declines in a migratory bird [Publisher: American Association for the Advancement of Science]. *Science*, *359*(6371), 83–86. https://doi.org/10.1126/science.aan4380

Bergelson, J., & Purrington, C. B. (1996). Surveying Patterns in the Cost of Resistance in Plants [Publisher: University of Chicago Press]. *The American Naturalist*. https://doi.org/10.1086/285938

Berger, J. D., Buirchell, B. J., Luckett, D. J., & Nelson, M. N. (2012). Domestication bottlenecks limit genetic diversity and constrain adaptation in narrow-leafed lupin (Lupinus angustifolius L.) *Theor Appl Genet*, *124*(4), 637–652. https://doi.org/10.1007/s00122-011-1736-z

Berthaud, J. (1986). *Les ressources genetiques pour l'amelioration des cafeires africains diploides* (Vol. 188). ORSTOM.

Bramel, P., Krishnan, S., Horna, D., Lainoff, B., & Montagnon, C. (2017). *Global Conservation Strategy for Coffee Genetic Resources*.

Bräutigam, K., & Cronk, Q. (2018). DNA Methylation and the Evolution of Developmental Complexity in Plants. *Front Plant Sci*, *9*, 1447. https://doi.org/10.3389/fpls.2018.01447

Brazier, T., & Glémin, S. (2022). Diversity and determinants of recombination landscapes in flowering plants (I. R. Henderson, Ed.). *PLoS Genet*, *18*(8), e1010141. https://doi.org/10.1371/journal.pgen.1010141

Britten, R. J., & Kohne, D. E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science*, *161*(3841), 529–540. https://doi.org/10.1126/science.161.3841.529

Brown, A. H. D. (1989). Core collections: a practical approach to genetic resources management [ISBN: 0831-2796 Publisher: NRC Research Press Ottawa, Canada]. *Genome*, *31*(2), 818–824.

Bunn, C., Läderach, P., Ovalle Rivera, O., & Kirschke, D. (2015). A bitter cup: climate change profile of global production of Arabica and Robusta coffee. *Climatic Change*, *129*(1), 89–101. https://doi.org/10.1007/

s10584-014-1306-x

Cc4vpTimes Cited:6Cited References Count:42

Burgarella, C., Barnaud, A., Kane, N. A., Jankowski, F., Scarcelli, N., Billot, C., Vigouroux, Y., & Berthouly-Salazar, C. (2019). Adaptive introgression: An untapped evolutionary mechanism for crop adaptation. *Frontiers in Plant Science*, *10*. Retrieved March 18, 2022, from https://www.frontiersin.org/article/10.3389/fpls.2019.00004

Campuzano-Duque, L. F., & Blair, M. W. (2022). Strategies for Robusta Coffee (Coffea canephora) Improvement as a New Crop in Colombia [ISBN: 2077-0472 Publisher: MDPI]. *Agriculture*, *12*(10), 1576.

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., & Keller, S. R. (2020). Genomic prediction of (mal) adaptation across current and future climatic landscapes [ISBN: 1543-592X Publisher: Annual Reviews]. *Annual Review of Ecology, Evolution, and Systematics*, *51*, 245–269.

Capblancq, T., & Forester, B. R. (2021). Redundancy analysis: A Swiss Army Knife for landscape genomics [ISBN: 2041-210X Publisher: Wiley Online Library]. *Methods in Ecology and Evolution*, *12*(12), 2298–2309.

Ceglar, A., Zampieri, M., Toreti, A., & Dentener, F. (2019). Observed Northward Migration of Agro-Climate Zones in Europe Will Further Accelerate Under Climate Change. *Earth's Future*, *7*(9), 1088–1101. https://doi.org/10.1029/2019EF001178

Charmetant, P. (1994). Lowlands coffee in Papua New Guinea research programmes.

Charr, J.-C., Garavito, A., Guyeux, C., Crouzillat, D., Descombes, P., Fournier, C., Ly, S. N., Raharimalala, E. N., Rakotomalala, J.-J., Stoffelen, P., Janssens, S., Hamon, P., & Guyot, R. (2020). Complex evolutionary history of coffees revealed by full plastid genomes and 28,800 nuclear SNP analyses, with particular emphasis on Coffea canephora (Robusta coffee). *Molecular Phylogenetics and Evolution*, *151*, 106906. https://doi.org/10.1016/j.ympev.2020.106906

Charrier, A., & Berthaud, J. (1990). Use and value of genetic resources of Coffea for breeding and their long-term conservation. [ISBN: 0344-5615]. *Mitteilungen aus dem Institut für Allgemeine Botanik Hamburg*, 53–64.

Chen, Y., Jiang, Z., Fan, P., Ericson, P. G. P., Song, G., Luo, X., Lei, F., & Qu, Y. (2022). The combination of genomic offset and niche modelling provides insights into climate change-driven vulnerability. *Nat Commun*, *13*(1), 4821. https://doi.org/10.1038/s41467-022-32546-z

Chevalier, A. (1929). Les caféiers du globe. Fasc. 1: Généralités sur les caféiers [Publisher: Paul Lechevalier].

Ciftci-Yilmaz, S., & Mittler, R. (2008). The zinc finger network of plants [ISBN: 1420-682X Publisher: Springer]. *Cellular and Molecular Life Sciences*, *65*, 1150–1160.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Corrêa, R. L., Sanz-Carbonell, A., Kogej, Z., Müller, S. Y., Ambrós, S., López-Gomollón, S., Gómez, G., Baulcombe, D. C., & Elena, S. F. (2020). Viral Fitness Determines the Magnitude of Transcriptomic and Epigenomic Reprograming of Defense Responses in Plants. *Mol Biol Evol*, *37*(7), 1866–1881. https://doi.org/10.1093/molbev/msaa091

Cortés, A. J., & López-Hernández, F. (2021). Harnessing Crop Wild Diversity for Climate Change Adaptation [Number: 5 Publisher: Multidisciplinary Digital Publishing Institute]. *Genes*, *12*(5), 783. https://doi.org/10.3390/genes12050783

Cortés, A. J., López-Hernández, F., & Blair, M. W. (2022). Genome–Environment Associations, an Innovative Tool for Studying Heritable Evolutionary Adaptation in Orphan Crops and Wild Relatives. *Frontiers in Genetics*, *13*. Retrieved September 22, 2023, from https://www.frontiersin.org/articles/10.3389/fgene.2022.910386

Coste, R. (1955). The coffee plants and coffees of the world. Volume I. The coffee plants. [Publisher: Editions Larose, Paris.]. *The coffee plants and coffees of the world. Volume I. The coffee plants.*

Cottin, A., Penaud, B., Glaszmann, J.-C., Yahiaoui, N., & Gautier, M. (2019). Simulation-based evaluation of three methods for local ancestry deconvolution of non-model crop species genomes. *G3 (Bethesda)*, *10*(2), 569–579. https://doi.org/10.1534/g3.119.400873

Coyne, C. J., Kumar, S., von Wettberg, E. J., Marques, E., Berger, J. D., Redden, R. J., Ellis, T. N., Brus, J., Zablatzká, L., & Smýkal, P. (2020). Potential and limits of exploitation of crop wild relatives for pea, lentil, and chickpea improvement. *Legume Science*, *2*(2), e36. https://doi.org/10.1002/leg3.36
e36 LEG3-2019-098.R2

Cramer, P. J. S. (1957). *A review of literature of coffee research in Indonesia* (F. L. Wellman, Ed.; Vol. 262). SIC Editorial, Inter American Institute of Agricultural Sciences.

Crouzillat, D., Rigoreau, M., Mérot-L'Anthoene, V., Tranchant-Dubreuil, C., Poncet, V., & de Kochko, A. (2016). Improvement of Robusta coffee cup quality (Coffea canephora).

Cubry, P., De Bellis, F., Pot, D., Musoli, P., & Leroy, T. (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*(2), 483–501. https://doi.org/10.1007/s10722-012-9851-5

Cui, L., Hanika, K., Visser, R. G. F., & Bai, Y. (2020). Improving Pathogen Resistance by Exploiting Plant Susceptibility Genes in Coffee (Coffea spp.) [Number: 12 Publisher: Multidisciplinary Digital Publishing Institute]. *Agronomy*, *10*(12), 1928. https://doi.org/10.3390/agronomy10121928

Cuppen, E. (2007). Genotyping by Allele-Specific Amplification (KASPar). *CSH Protoc*, *2007*, pdb.prot4841. https://doi.org/10.1101/pdb.prot4841

DaMatta, F. M., & Ramalho, J. D. C. (2006). Impacts of drought and temperature stress on coffee physiology and production: a review [Publisher: Brazilian Journal of Plant Physiology]. *Braz. J. Plant Physiol.*, *18*, 55–81. https://doi.org/10.1590/S1677-04202006000100006

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Brief Funct Genomics*, *9*(5-6), 416–423. https://doi.org/10.1093/bfgp/elq031

Davis, A., TOSH, J., Ruch, N., & Fay, M. (2011). Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. *Botanical Journal of the Linnean Society*, *167*. https://doi.org/10.1111/j.1095-8339.2011.01177.x

Davis, A. P., Chadburn, H., Moat, J., O'Sullivan, R., Hargreaves, S., & Lughadha, E. N. (2019). High extinction risk for wild coffee species and implications for coffee sector sustainability [ISBN: 2375-2548 Publisher: American Association for the Advancement of Science]. *Science advances*, *5*(1), eaav3473.

Davis, A. P., Govaerts, R., Bridson, D. M., & Stoffelen, P. (2006). An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). *Bot. J. Linn. Soc.*, *152*(4), 465–512. https://doi.org/10.1111/j.1095-8339.2006.00584.x

De Beukelaer, H., Davenport, G. F., & Fack, V. (2018). Core Hunter 3: flexible core subset selection. *BMC Bioinformatics*, *19*(1), 203. https://doi.org/10.1186/s12859-018-2209-z

De Kochko, A. (2018). Deciphering the Allotetraploid Genome of *Coffea arabica* L.

De Ollas, C., Morillón, R., Fotopoulos, V., Puértolas, J., Ollitrault, P., Gómez-Cadenas, A., & Arbona, V. (2019). Facing Climate Change: Biotechnology of Iconic Mediterranean Woody Crops. *Frontiers in Plant Science*, *10*. Retrieved October 9, 2023, from https://www.frontiersin.org/articles/10.3389/fpls.2019.00427

de Aquino, S. O., Kiwuka, C., Tournebize, R., Gain, C., Marraccini, P., Mariac, C., Bethune, K., Couderc, M., Cubry, P., Andrade, A. C., Lepelley, M., Darracq, O., Crouzillat, D., Anten, N., Musoli, P., Vigouroux, Y., de Kochko, A., Manel, S., François, O., & Poncet, V. (2022). Adaptive potential of Coffea canephora from Uganda in response to climate change. *Molecular Ecology*, *31*(6), 1800–1819. https://doi.org/10.1111/mec.16360

de Carvalho, K., Bespalhok Filho, J. C., dos Santos, T. B., de Souza, S. G. H., Vieira, L. G. E., Pereira, L. F. P., & Domingues, D. S. (2013). Nitrogen Starvation, Salt and Heat Stress in Coffee (Coffea arabica L.):

Identification and Validation of New Genes for qPCR Normalization. *Mol Biotechnol*, *53*(3), 315–325. https://doi.org/10.1007/s12033-012-9529-4

de Carvalho, K., Petkowicz, C. L. O., Nagashima, G. T., Bespalhok Filho, J. C., Vieira, L. G. E., Pereira, L. F. P., & Domingues, D. S. (2014). Homeologous genes involved in mannitol synthesis reveal unequal contributions in response to abiotic stress in Coffea arabica. *Mol Genet Genomics*, *289*(5), 951–963. https://doi.org/10.1007/s00438-014-0864-y

de Miguel, M., Rodríguez-Quilón, I., Heuertz, M., Hurel, A., Grivet, D., Jaramillo-Correa, J. P., Vendramin, G. G., Plomion, C., Majada, J., Alía, R., Eckert, A. J., & González-Martínez, S. C. (2022). Polygenic adaptation and negative selection across traits, years and environments in a long-lived plant species (Pinus pinaster Ait., Pinaceae). *Molecular Ecology*, *31*(7), 2089–2105. https://doi.org/10.1111/mec.16367

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., & Guarino, L. (2017). Past and Future Use of Wild Relatives in Crop Breeding. *Crop Science*, *57*(3), 1070–1082. https://doi.org/10.2135/cropsci2016.10.0885

Denham, T., Barton, H., Castillo, C., Crowther, A., Dotte-Sarout, E., Florin, S. A., Pritchard, J., Barron, A., Zhang, Y., & Fuller, D. Q. (2020). The domestication syndrome in vegetatively propagated field crops. *Annals of Botany*, *125*(4), 581–597. https://doi.org/10.1093/aob/mcz212

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., Aury, J.-M., Bento, P., Bernard, M., Bocs, S., Campa, C., Cenci, A., Combes, M.-C., Crouzillat, D., Da Silva, C., ... Lashermes, P. (2014). The Coffee Genome Provides Insight into the Convergent Evolution of Caffeine Biosynthesis. *Science*, *345*(6201), 1181–1184. https://doi.org/10.1126/science.1255274

Dereeper, A., Guyot, R., Tranchant-Dubreuil, C., Anthony, F., Argout, X., de Bellis, F., Combes, M.-C., Gavory, F., de Kochko, A., Kudrna, D., Leroy, T., Poulain, J., Rondeau, M., Song, X., Wing, R., & Lashermes, P. (2013). BAC-end sequences analysis provides first insights into coffee (Coffea canephora P.) genome composition and evolution. *Plant Mol Biol*, *83*(3), 177–189. https://doi.org/10.1007/s11103-013-0077-5

de Ruiter, M. C., Couasnon, A., van den Homberg, M. J. C., Daniell, J. E., Gill, J. C., & Ward, P. J. (2020). Why We Can No Longer Ignore Consecutive Disasters. *Earth's Future*, *8*(3), e2019EF001425. https://doi.org/10.1029/2019EF001425
e2019EF001425 2019EF001425

Desmet, Q., & Ngo-Duc, T. (2022). A novel method for ranking CMIP6 global climate models over the southeast Asian region. *International Journal of Climatology*, *42*(1), 97–117. https://doi.org/10.1002/joc.7234

Dias-Alves, T., Mairal, J., & Blum, M. G. B. (2018). Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species. *Mol Biol Evol*, *35*(9), 2318–2326. https://doi.org/10.1093/molbev/msy126

Dinh, T. L. A., Aires, F., & Rahn, E. (2022). Statistical Analysis of the Weather Impact on Robusta Coffee Yield in Vietnam. *Frontiers in Environmental Science*, *10*. Retrieved September 13, 2022, from https://www.frontiersin.org/articles/10.3389/fenvs.2022.820916

Dinh, T. L. A., Aires, F., & Rahn, E. (2023). *Climate change impacts on Robusta coffee production over Vietnam* (preprint). Environmental Sciences. https://doi.org/10.1002/essoar.10512723.1

Dirzo, R., & Raven, P. H. (2003). Global State of Biodiversity and Loss. *Annual Review of Environment and Resources*, *28*(1), 137–167. https://doi.org/10.1146/annurev.energy.28.050302.105532

Donath, A., & Stadler, P. F. (2018). Split-inducing indels in phylogenomic analysis. *Algorithms for Molecular Biology*, *13*(1), 12. https://doi.org/10.1186/s13015-018-0130-7

Durand, T., De Wildeman, É., Micheli, M., Briquet, J., Hallier, H., & Pax, F. (1898). Materiaux pour la Flore du Congo [ISBN: 0037-9557 Publisher: JSTOR]. *Bulletin de la Société Royale de Botanique de Belgique/Bulletin van de Koninklijke Belgische Botanische Vereniging*, 44–128.

Duranton, M., Bonhomme, F., & Gagnaire, P.-A. (2019). The spatial scale of dispersal revealed by admixture tracts. *Evolutionary Applications*, *12*(9), 1743–1756. https://doi.org/10.1111/eva.12829

Dussert, S., Lashermes, P., Anthony, F., Montagnon, C., Trouslot, P., Combes, M.-C., Berthaud, J., Noirot, M., & Hamon, S. (2003). Coffee (Coffea canephora). *Genetic diversity of cultivated tropical plants. Science Publishers, Plymouth*, 239–258.

Dussert, S., Lashermes, P., Anthony, F., Montagon, C., Trouslot, P., Combes, M.-C., Berthaud, J., Noirot, M., & Hamon, S. (1999). Le caféier, *Coffea canephora*. *Diversité génétique des plantes tropicales cultivées*. CIRAD.

Engelmann, F., Dulloo, M., Astorga, C., Dussert, S., & Anthony, F. (2007). *Conserving coffee genetic resources: complementary strategies for ex situ conservation of coffee (Coffea arabica L.) genetic resources a case study in Catie, Costa Rica*. Bioversity international.

Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA*, *6*(1), 24. https://doi.org/10.1186/s13100-015-0055-3

Farah, A. (2009). 15 - Coffee as a speciality and functional beverage. In P. Paquin (Ed.), *Functional and Speciality Beverage Technology* (pp. 370–395). Woodhead Publishing. https://doi.org/10.1533/9781845695569.3.370

Ferrão, L. F. V., Caixeta, E. T., Souza, F. d. F., Zambolim, E. M., Cruz, C. D., Zambolim, L., & Sakiyama, N. S. (2013). Comparative study of different molecular markers for classifying and establishing genetic relationships in Coffea canephora. *Plant Syst Evol*, *299*(1), 225–238. https://doi.org/10.1007/s00606-012-0717-2

Ferrão, L. F. V., Ferrão, R. G., Ferrão, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., & Garcia, A. A. F. (2019). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity (Edinb)*, *122*(3), 261–275. https://doi.org/10.1038/s41437-018-0105-y

Ferrão, M. A. G., da Fonseca, A. F. A., Verdin Filho, A. C., Volpi, P. S., Ferrão, M. A. G., da Fonseca, A. F. A., & Volpi, P. S. (2007). Origem, dispersão geográfica, taxonomia e diversidade genética de Coffea canephora. [Publisher: In: FERRÃO, RG; FONSECA, AFA da.; BRAGANÇA, SM; FERRÃO, MAG; DE MUNER, LH . . .].

Ferrão, M. A. G., da Fonseca, A. F. A., Volpi, P. S., de Souza, L. C., Comério, M., Filho, A. C. V., Riva-Souza, E. M., Munoz, P. R., Ferrão, R. G., & Ferrão, L. F. V. (2023). Genomic-assisted breeding for climate-smart coffee. *The Plant Genome*, *n/a*(n/a), e20321. https://doi.org/10.1002/tpg2.20321

Ferrão, M. A. G., Mendonça, R. F. d., Fonseca, A. F. A., Ferrão, R. G., Senra, J. F. B., Volpi, P. S., Verdin Filho, A. C., & Comério, M. (2021). Characterization and genetic diversity of *Coffea canephora* accessions in a germplasm bank in Espírito Santo, Brazil [Publisher: Crop Breeding and Applied Biotechnology]. *Crop Breed. Appl. Biotechnol.*, *21*. https://doi.org/10.1590/1984-70332021v21n2a32

Ferwerda, F. P. (1948). Coffee Breeding in Java [Publisher: New York Botanical Garden Press]. *Economic Botany*, *2*(3), 258–272. Retrieved August 21, 2023, from https://www.jstor.org/stable/4251903

Fetter, K., & Keller, S. (2023). Admixture mapping and selection scans identify genomic regions associated with stomatal patterning and disease resistance in hybrid poplars.

Fitzpatrick, M. C., Chhatre, V. E., Soolanayakanahally, R. Y., & Keller, S. R. (2021). Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests [ISBN: 1755-098X Publisher: Wiley Online Library]. *Molecular Ecology Resources*, *21*(8), 2749–2765.

Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation [ISBN: 1461-023X Publisher: Wiley Online Library]. *Ecology letters*, *18*(1), 1–16.

Flint-Garcia, S., Feldmann, M. J., Dempewolf, H., Morrell, P. L., & Ross-Ibarra, J. (2023). Diamonds in the not-so-rough: Wild relative diversity hidden in crop genomes [Publisher: Public Library of Science]. *PLOS Biology*, *21*(7), e3002235. https://doi.org/10.1371/journal.pbio.3002235

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, *27*(9), 2215–2233. https://doi.org/10.1111/mec.14584

François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, *25*(2), 454–469. https://doi.org/10.1111/mec.13513

Frankel, O. H., Arber, W., Llimensee, K., Peacock, W. J., & Starlinger, P. (1984). Genetic manipulation: impact on man and society [Publisher: Cambridge University Press Cambridge]. *Genetic perspective of germplasm conservation*, 161–470.

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.*, *6*(8), 925–929. https://doi.org/10.1111/2041-210X.12382

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Gain, C., & François, O. (2021). LEA 3: Factor models in population genetics and ecological genomics with R. *Molecular Ecology Resources*, *21*(8), 2738–2748. https://doi.org/10.1111/1755-0998.13366

Gain, C., Rhoné, B., Cubry, P., Salazar, I., Forbes, F., Vigouroux, Y., Jay, F., & François, O. (2023). A Quantitative Theory for Genomic Offset Statistics. *Molecular Biology and Evolution*, *40*(6), msad140. https://doi.org/10.1093/molbev/msad140

Garavito, A., Montagnon, C., Guyot, R., & Bertrand, B. (2016). Identification by the DArTseq method of the genetic origin of the *Coffea canephora* cultivated in Vietnam and Mexico. *BMC Plant Biol.*, *16*(1), 242. https://doi.org/10.1186/s12870-016-0933-y

Garg, S., Fungtammasan, A., Carroll, A., Chou, M., Schmitt, A., Zhou, X., Mac, S., Peluso, P., Hatas, E., Ghurye, J., Maguire, J., Mahmoud, M., Cheng, H., Heller, D., Zook, J. M., Moemke, T., Marschall, T., Sedlazeck, F. J., Aach, J., … Li, H. (2021). Chromosome-scale, haplotype-resolved assembly of human genomes [Number: 3 Publisher: Nature Publishing Group]. *Nat Biotechnol*, *39*(3), 309–312. https://doi.org/10.1038/s41587-020-0711-0

Gaut, B. S., Seymour, D. K., Liu, Q., & Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication [Number: 8 Publisher: Nature Publishing Group]. *Nature Plants*, *4*(8), 512–520. https://doi.org/10.1038/s41477-018-0210-1

Geza, E., Mugo, J., Mulder, N. J., Wonkam, A., Chimusa, E. R., & Mazandu, G. K. (2019). A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Brief Bioinform*, *20*(5), 1709–1724. https://doi.org/10.1093/bib/bby044

Gimase, J. M., Thagana, W. M., Omondi, C. O., Cheserek, J. J., Gichimu, B. M., Gichuru, E. K., Ziyomo, C., & Sneller, C. H. (2020). Genome-Wide Association Study identify the genetic loci conferring resistance to Coffee Berry Disease (Colletotrichum kahawae) in Coffea arabica var. Rume Sudan. *Euphytica*, *216*(6), 86. https://doi.org/10.1007/s10681-020-02621-x

Goetz, L. H., Uribe-Bruce, L., Quarless, D., Libiger, O., & Schork, N. J. (2014). Admixture and clinical phenotypic variation [Publisher: Karger Publishers]. *Hum. Hered.*, *77*(1-4), 73–86. https://doi.org/10.1159/000362233

Gomez, C., Dussert, S., Hamon, P., Hamon, S., Kochko, A. d., & Poncet, 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. https://doi.org/10.1186/1471-2148-9-167

1471-2148 (Electronic)1471-2148 (Linking)Journal ArticleResearch Support, Non-U.S. Gov't

Gong, M., Chen, S.-N., Song, Y.-Q., & Li, Z.-G. (1997). Effect of Calcium and Calmodulin on Intrinsic Heat Tolerance in Relation to Antioxidant Systems in Maize Seedlings. *Functional Plant Biol.*, *24*(3), 371. https://doi.org/10.1071/PP96118

[TLDR] It is suggested that external Ca2+ can enhance the intrinsic heat tolerance of maize seedlings, which requires the entry of externalCa2+ into cells across plasma membranes and the mediation of intracellular calmodulin, and is associated with the increase of antioxidant system activity.

Goudet, J. (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x

Griffin, J. J., Ranney, T. G., & Pharr, D. M. (2004). Heat and drought influence photosynthesis, water relations, and soluble carbohydrates of two ecotypes of redbud (Cercis canadensis) [ISBN: 2327-9788 Publisher: American Society for Horticultural Science]. *Journal of the American Society for Horticultural Science*, *129*(4), 497–502.

Gross, B. L., Henk, A. D., Richards, C. M., Fazio, G., & Volk, G. M. (2014). Genetic diversity in Malus ×domestica (Rosaceae) through time in response to domestication. *American Journal of Botany*, *101*(10), 1770–1779. https://doi.org/10.3732/ajb.1400297

Gu, R., Fan, S., Wei, S., Li, J., Zheng, S., & Liu, G. (2023). Developments on Core Collections of Plant Genetic Resources: Do We Know Enough? [Number: 5 Publisher: Multidisciplinary Digital Publishing Institute]. *Forests*, *14*(5), 926. https://doi.org/10.3390/f14050926

Guan, K., Sultan, B., Biasutti, M., Baron, C., & Lobell, D. B. (2017). Assessing climate adaptation options and uncertainties for cereal systems in West Africa. *Agricultural and Forest Meteorology*, *232*, 291–305. https://doi.org/10.1016/j.agrformet.2016.07.021

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics*, *196*(3), 625–642. https://doi.org/10.1534/genetics.113.160697

Gupta, P. K. (2021). GWAS for genetics of complex quantitative traits: Genome to pangenome and SNPs to SVs and k-mers. *BioEssays*, *43*(11), 2100109. https://doi.org/10.1002/bies.202100109

Hajjar, R., & Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica*, *156*(1), 1–13. https://doi.org/10.1007/s10681-007-9363-0

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., dePamphilis, C. W., & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *118*(35), e2102914118. https://doi.org/10.1073/pnas.2102914118

Hamon, P., Duroy, P.-O., Dubreuil-Tranchant, C., Mafra D'Almeida Costa, P., Duret, C., Razafinarivo, N. J., Couturon, E., Hamon, S., de Kochko, A., Poncet, V., & Guyot, R. (2011). Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae). *Mol Genet Genomics*, *285*(6), 447–460. https://doi.org/10.1007/s00438-011-0617-0

Hamon, P., Grover, C. E., Davis, A. P., Rakotomalala, J.-J., Raharimalala, N. E., Albert, V. A., Sreenath, H. L., Stoffelen, P., Mitchell, S. E., Couturon, E., Hamon, S., de Kochko, A., Crouzillat, D., Rigoreau, M., Sumirat, U., Akaffou, S., & Guyot, R. (2017). Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. *Molecular Phylogenetics and Evolution*, *109*, 351–361. https://doi.org/10.1016/j.ympev.2017.02.009

Hamon, S., Dussert, S., Deu, M., Hamon, P., Seguin, M., Glaszmann, J.-C., Grivet, L., Chantereau, J., Chevallier, M.-H., Flori, A., Lashermes, P., Legnate, H., & Noirot, M. (1998). Effects of quantitative and qualitative principal component score strategies on the structure of coffee, rubber tree, rice and sorghum core collections. *Genet Sel Evol*, *30*(Suppl 1), S237. https://doi.org/10.1186/1297-9686-30-S1-S237

Harlan, J. R. (1971). Agricultural origins: centers and noncenters [Publisher: American Association for the Advancement of Science]. *Science*, *174*(4008), 468–474. https://doi.org/10.1126/science.174.4008.468

Harlan, J. R. (1992). Crops and Man. American Society of Agronomy. *Crop Science Society of America, Madison, Wisconsin*, *16*(2), 63–262.

Hasegawa, T., Wakatsuki, H., Ju, H., Vyas, S., Nelson, G. C., Farrell, A., Deryng, D., Meza, F., & Makowski, D. (2022). A global dataset for the projected impacts of climate change on four major crops [Number: 1 Publisher: Nature Publishing Group]. *Sci Data*, *9*(1), 58. https://doi.org/10.1038/s41597-022-01150-7

He, C., Washburn, J. D., Hao, Y., Zhang, Z., Yang, J., & Liu, S. (2021). Trait association and prediction through integrative k-mer analysis [Publisher: Cold Spring Harbor Laboratory]. *bioRxiv*, 2021.11. 17.468725.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding [Publisher: Frontiers]. *Front. Plant Sci.*, *5*. https://doi.org/10.3389/fpls.2014.00484

Heslop-Harrison, J. S., & Schwarzacher, T. (2007). Domestication, Genomics and the Future for Banana. *Annals of Botany*, *100*(5), 1073–1084. https://doi.org/10.1093/aob/mcm191

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural Variation in the Sequencing Era: Comprehensive Discovery and Integration. *Nat Rev Genet*, *21*(3), 171–189. https://doi.org/10.1038/s41576-019-0180-9

Horimoto, A. R. V. R., Xue, D., Thornton, T. A., & Blue, E. E. (2021). Admixture mapping reveals the association between native American ancestry at 3q13.11 and reduced risk of Alzheimer's disease in Caribbean Hispanics. *Alzheimers Res. Ther.*, *13*(1), 122. https://doi.org/10.1186/s13195-021-00866-9

Hospital, F. (2001). Size of Donor Chromosome Segments Around Introgressed Loci and Reduction of Linkage Drag in Marker-Assisted Backcross Programs. *Genetics*, *158*(3), 1363–1379. https://doi.org/10.1093/genetics/158.3.1363

Hu, G., Feng, J., Xiang, X., Wang, J., Salojärvi, J., Liu, C., Wu, Z., Zhang, J., Liang, X., Jiang, Z., Liu, W., Ou, L., Li, J., Fan, G., Mai, Y., Chen, C., Zhang, X., Zheng, J., Zhang, Y., ... Li, J. (2022). Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars [Number: 1 Publisher: Nature Publishing Group]. *Nat Genet*, *54*(1), 73–83. https://doi.org/10.1038/s41588-021-00971-3

Huang, L., Wang, X., Dong, Y., Long, Y., Hao, C., Yan, L., & Shi, T. (2020). Resequencing 93 accessions of coffee unveils independent and parallel selection during Coffea species divergence. *Plant Mol Biol*, *103*(1), 51–61. https://doi.org/10.1007/s11103-020-00974-4

Hübner, S., & Kantar, M. B. (2021). Tapping diversity from the wild: from sampling to implementation. *Front. Plant. Sci.*, *12*(626565). https://doi.org/10.3389/fpls.2021.626565

Huttner, E., Wenzl, P., Akbari, M., Caig, V., Carling, J., Cayla, C., Evers, M., Jaccoud, D., Peng, K., Patarapuwadol, S., Uszynski, G., Xia, L., Yang, S., & Kilian, A. (2005). Diversity Arrays Technology: A Novel Tool for Harnessing the Genetic Potential of Orphan Crops.

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., Specht, J. E., Shoemaker, R. C., & Cregan, P. B. (2006). Impacts of genetic bottlenecks on soybean genome diversity [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *103*(45), 16666–16671. https://doi.org/10.1073/pnas.0604379103

ICO. (2019). *International Coffee Organization Statistics* (tech. rep.). International Coffee Organisation. http://www.ico.org/trade_statistics.asp

IPCC. (2019). IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. *IPCC Summary for Policymalers*, 1–472.

Iqbal, Z., Shahid, S., Ahmed, K., Ismail, T., Ziarh, G. F., Chung, E.-S., & Wang, X. (2021). Evaluation of CMIP6 GCM rainfall in mainland Southeast Asia [ISBN: 0169-8095 Publisher: Elsevier]. *Atmospheric Research*, *254*, 105525.

Jarvis, A., Lane, A., & Hijmans, R. J. (2008). The effect of climate change on crop wild relatives. *Agriculture, Ecosystems & Environment*, *126*(1), 13–23. https://doi.org/10.1016/j.agee.2008.01.013

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Joukhadar, R., Daetwyler, H. D., Bansal, U. K., Gendall, A. R., & Hayden, M. J. (2017). Genetic diversity, population structure and ancestral origin of Australian wheat. *Front. Plant. Sci.*, *8*(2115). https://doi.org/10.3389/fpls.2017.02115

Karikari, B., Lemay, M.-A., & Belzile, F. (2023). k-mer-Based Genome-Wide Association Studies in Plants: Advances, Challenges, and Perspectives [Number: 7 Publisher: Multidisciplinary Digital Publishing Institute]. *Genes*, *14*(7), 1439. https://doi.org/10.3390/genes14071439

Kath, J., Byrareddy, V. M., Craparo, A., Nguyen-Huy, T., Mushtaq, S., Cao, L., & Bossolasco, L. (2020). Not so robust: Robusta coffee production is highly sensitive to temperature. *Glob Chang Biol*, *26*(6), 3677–3688. https://doi.org/10.1111/gcb.15097

Kath, J., Mittahalli Byrareddy, V., Mushtaq, S., Craparo, A., & Porcel, M. (2021). Temperature and rainfall impacts on robusta coffee bean characteristics. *Climate Risk Management*, *32*, 100281. https://doi.org/10.1016/j.crm.2021.100281

Khoury, C. K., Achicanoy, H. A., Bjorkman, A. D., Navarro-Racines, C., Guarino, L., Flores-Palacios, X., Engels, J. M. M., Wiersema, J. H., Dempewolf, H., Sotelo, S., Ramírez-Villegas, J., Castañeda-Álvarez, N. P., Fowler, C., Jarvis, A., Rieseberg, L. H., & Struik, P. C. (2016). Origins of food crops connect countries worldwide [Publisher: Royal Society]. *Proc. Royal Soc. B*, *283*(1832). https://doi.org/10.1098/rspb.2016.0792

Kiwuka, C., Goudsmit, E., Tournebize, R., Aquino, S. O., Douma, J. C., Bellanger, L., Crouzillat, D., Stoffelen, P., Sumirat, U., Legnaté, H., Marraccini, P., Kochko, A. d., Andrade, A. C., Mulumba, J. W., Musoli, P., Anten, N. P. R., & Poncet, V. (2021). Genetic diversity of native and cultivated Ugandan Robusta coffee (*Coffea canephora* Pierre ex A. Froehner): Climate influences, breeding potential and diversity conservation [Publisher: Public Library of Science]. *PLOS ONE*, *16*(2), e0245965. https://doi.org/10.1371/journal.pone.0245965

Kiwuka, C., Vos, J., Douma, J. C., Musoli, P., Mulumba, J. W., Poncet, V., & Anten, N. P. R. (2023). Intraspecific variation in growth response to drought stress across geographic locations and genetic groups in *Coffea canephora*. *Ecology and Evolution*, *13*(1), e9715. https://doi.org/10.1002/ece3.9715

Kjaer, E. D., Graudal, L., & Nathan, I. (2001). Ex situ conservation of commercial tropical trees: strategies, options and constraints. *Ex situ*, 127–146.

Kochevenko, A., Jiang, Y., Seiler, C., Surdonja, K., Kollers, S., Reif, J. C., Korzun, V., & Graner, A. (2018). Identification of QTL hot spots for malting quality in two elite breeding lines with distinct tolerance to abiotic stress. *BMC Plant Biology*, *18*(1), 106. https://doi.org/10.1186/s12870-018-1323-4

Kokot, M., Długosz, M., & Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, *33*(17), 2759–2761. https://doi.org/10.1093/bioinformatics/btx304

Kumar, R., Sharma, V., Suresh, S., Ramrao, D. P., Veershetty, A., Kumar, S., Priscilla, K., Hangargi, B., Narasanna, R., Pandey, M. K., Naik, G. R., Thomas, S., & Kumar, A. (2021). Understanding Omics Driven Plant Improvement and de novo Crop Domestication: Some Examples. *Front Genet*, *12*, 637141. https://doi.org/10.3389/fgene.2021.637141

Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP Discovery through Next-Generation Sequencing and Its Applications [Publisher: Hindawi]. *International Journal of Plant Genomics*, *2012*, e831460. https://doi.org/10.1155/2012/831460

Ky, C.-L., Barre, P., Lorieux, M., Trouslot, P., Akaffou, S., Louarn, J., Charrier, A., Hamon, S., & Noirot, M. (2000). Interspecific genetic linkage map, segregation distortion and genetic conversion in coffee (Coffea sp.) [ISBN: 0040-5752 Publisher: Springer]. *Theoretical and Applied Genetics*, *101*, 669–676.

Labouisse, J.-P., Cubry, P., Austerlitz, F., Rivallan, R., & Nguyen, H. A. (2020). New insights on spatial genetic structure and diversity of Coffea canephora (Rubiaceae) in Upper Guinea based on old herbaria. *Plant Ecology and Evolution*, *153*(1), 82–100. https://doi.org/10.5091/plecevo.2020.1584

Lachmuth, S., Capblancq, T., Keller, S. R., & Fitzpatrick, M. C. (2023). Assessing uncertainty in genomic offset forecasts from landscape genomic models (and implications for restoration and assisted migration). *Frontiers in Ecology and Evolution*, *11*. Retrieved September 26, 2023, from https://www.frontiersin.org/articles/10.3389/fevo.2023.1155783

Larson, G., Piperno, D. R., Allaby, R. G., Purugganan, M. D., Andersson, L., Arroyo-Kalin, M., Barton, L., Climer Vigueira, C., Denham, T., Dobney, K., Doust, A. N., Gepts, P., Gilbert, M. T. P., Gremillion, K. J., Lucas, L., Lukens, L., Marshall, F. B., Olsen, K. M., Pires, J. C., … Fuller, D. Q. (2014). Current perspectives and the future of domestication studies [Publisher: Proceedings of the National Academy of Sciences]. *Proc. Natl. Acad. Sci. U.S.A.*, *111*(17), 6139–6146. https://doi.org/10.1073/pnas.1323964111

Láruson, Á. J., Yeaman, S., & Lotterhos, K. E. (2020). The Importance of Genetic Redundancy in Evolution. *Trends in Ecology & Evolution*, *35*(9), 809–822. https://doi.org/10.1016/j.tree.2020.04.009

Láruson, Á. J., Fitzpatrick, M. C., Keller, S. R., Haller, B. C., & Lotterhos, K. E. (2022). Seeing the forest for the trees: Assessing genetic offset predictions from gradient forest [ISBN: 1752-4571 Publisher: Wiley Online Library]. *Evolutionary Applications*, *15*(3), 403–416.

Lashermes, P., Andrzejewski, S., Bertrand, B., Combes, M. C., Dussert, S., Graziosi, G., Trouslot, P., & Anthony, F. (2000). Molecular analysis of introgressive breeding in coffee (Coffea arabica L.) *Theor Appl Genet*, *100*(1), 139–146. https://doi.org/10.1007/s001220050019

Lashermes, P. (2018a). *Achieving sustainable cultivation of coffee: breeding and quality traits.* Burleigh Dodds Science Publishing Limited.

Lashermes, P. (2018b). *Achieving sustainable cultivation of coffee: breeding and quality traits.* Burleigh Dodds Science Publishing Limited.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data [Publisher: Public Library of Science]. *PLoS Genet.*, *8*(1), e1002453. https://doi.org/10.1371/journal.pgen.1002453

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P.-A., & Bernatchez, L. (2020). Using haplotype information for conservation genomics. *Trends in Ecology & Evolution*, *35*(3), 245–258. https://doi.org/10.1016/j.tree.2019.10.012

Leroy, T., Montagnon, C., Charrier, A., & Eskes, A. B. (1993a). Reciprocal recurrent selection applied to Coffea canephora Pierre. *Euphytica*, *74*(1), 121–128. https://doi.org/10.1007/BF00033776

Leroy, T., Montagnon, C., Charrier, A., & Eskes, A. B. (1993b). Reciprocal recurrent selection applied to *Coffea canephora* Pierre. I. Characterization and evaluation of breeding populations and value of intergroup hybrids. *Euphytica*, *67*(1), 113–125. https://doi.org/10.1007/BF00022734

Leroy, T., De Bellis, F., Legnate, H., Kananura, E., Gonzales, G., Pereira, L. F., Andrade, A. C., Charmetant, P., Montagnon, C., Cubry, P., Marraccini, P., Pot, D., & de Kochko, A. (2011). Improving the quality of African robustas: QTLs for yield- and quality-related traits in Coffea canephora. *Tree Genetics & Genomes*, *7*(4), 781–798. https://doi.org/10.1007/s11295-011-0374-6

Leroy, T., De Bellis, F., Legnate, H., Musoli, P., Kalonji, A., Loor Solórzano, R. G., & Cubry, P. (2014). Developing core collections to optimize the management and the exploitation of diversity of the coffee Coffea canephora. *Genetica*, *142*(3), 185–199. https://doi.org/10.1007/s10709-014-9766-5

Leroy, T., Montagnon, C., Cilas, C., Yapo, A., Charmetant, P., & Eskes, A. B. (1997). Reciprocal recurrent selection applied to Coffea canephora Pierre. III. Genetic gains and results of first cycle intergroup crosses [ISBN: 0014-2336 Publisher: Springer]. *Euphytica*, *95*(3), 347–354.

Leroy, T., Ribeyre, F., Bertrand, B., Charmetant, P., Dufour, M., Montagnon, C., Marraccini, P., & Pot, D. (2006). Genetics of coffee quality [Publisher: Brazilian Journal of Plant Physiology]. *Braz. J. Plant Physiol.*, *18*, 229–242. https://doi.org/10.1590/S1677-04202006000100016

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [arXiv:1303.3997 [q-bio]]. https://doi.org/10.48550/arXiv.1303.3997
Comment: 3 pages and 1 color figure

Li, R., Hu, F., Li, B., Zhang, Y., Chen, M., Fan, T., & Wang, T. (2020). Whole genome bisulfite sequencing methylome analysis of mulberry (Morus alba) reveals epigenome modifications in response to drought stress [ISBN: 2045-2322 Publisher: Nature Publishing Group UK London]. *Scientific Reports*, *10*(1), 8013.

Li, Z.-M., Zheng, X.-M., & Ge, S. (2011). Genetic diversity and domestication history of African rice (*Oryza glaberrima*) as inferred from multiple gene sequences. *Theor Appl Genet*, *123*(1), 21–31. https://doi.org/10.1007/s00122-011-1563-2

Liebmann, B., Bladé, I., Kiladis, G. N., Carvalho, L. M. V., Senay, G. B., Allured, D., Leroux, S., & Funk, C. (2012). Seasonality of African Precipitation from 1996 to 2009 [Publisher: American Meteorological Society

Section: Journal of Climate]. *Journal of Climate*, *25*(12), 4304–4322. https://doi.org/10.1175/JCLI-D-11-00157.1

Lind, B. M., Candido-Ribeiro, R., Singh, P., Lu, M., Vidakovic, D. O., Booker, T. R., Whitlock, M. C., Yeaman, S., Isabel, N., & Aitken, S. N. (2023). How useful is genomic data for predicting maladaptation to future climate? [Publisher: Cold Spring Harbor Laboratory]. *bioRxiv*, 2023.02. 10.528022.

Liu, S., An, Y., Tong, W., Qin, X., Samarina, L., Guo, R., Xia, X., & Wei, C. (2019). Characterization of genome-wide genetic variations between two varieties of tea plant (Camellia sinensis) and development of InDel markers for genetic research. *BMC Genomics*, *20*(1), 935. https://doi.org/10.1186/s12864-019-6347-0

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Loor Solórzano, R. G., De Bellis, F., Leroy, T., Plaza, L., Guerrero, H., Subia, C., Calderón, D., Fernández, F., Garzón, I., Lopez, D., & Vera, D. (2017). Revealing the Diversity of Introduced *Coffea canephora* Germplasm in Ecuador: Towards a National Strategy to Improve Robusta. *The Scientific World Journal*, *2017*, 1–12. https://doi.org/10.1155/2017/1248954

Mahadani, A. K., Sanyal, G., Mahadani, P., & Bhattacharjee, P. (2018). Modified evolutionary model with insertion and deletion (Indel) for phylogenetic tree construction [Publisher: Inderscience Publishers (IEL)]. *International Journal of Data Mining and Bioinformatics*. Retrieved May 21, 2021, from https://www.inderscienceonline.com/doi/abs/10.1504/IJDMB.2018.095555

Mahaut, L., Pironon, S., Barnagaud, J.-Y., Bretagnolle, F., Khoury, C. K., Mehrabi, Z., Milla, R., Phillips, C., Rieseberg, L. H., Violle, C., & Renard, D. (2022). Matches and mismatches between the global distribution of major food crops and climate suitability [Publisher: Royal Society]. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1983), 20221542. https://doi.org/10.1098/rspb.2022.1542

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Makałowski, W., Gotea, V., Pande, A., & Makałowska, I. (2019). Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. In M. Anisimova (Ed.), *Evolutionary Genomics: Statistical and Computational Methods* (pp. 177–207). Springer. https://doi.org/10.1007/978-1-4939-9074-0_6

Mani, A. (2017). Local ancestry association, admixture mapping, and ongoing challenges [Publisher: American Heart Association]. *Circ. Cardiovasc. Genet.*, *10*(2), e001747. https://doi.org/10.1161/CIRCGENETICS.117.001747

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet.*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Marraccini, P., Vinecky, F., Alves, G., J O Ramos, H., Sonia, E., Vieira, N., Carneiro, F., Sujii, P., C Alekcevetch, J., Silva, V., DaMatta, F., Ferrão, M., Leroy, T., Pot, D., Vieira, L., da Silva, F., & Andrade, A. (2012). Differentially expressed genes and proteins upon drought acclimation in tolerant and sensitive genotypes of *Coffea canephora*. *J. Exp. Bot.*, *63*, 4191–212. https://doi.org/10.1093/jxb/ers103

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads [Number: 1]. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Matuszewski, S., Hermisson, J., & Kopp, M. (2015). Catch Me if You Can: Adaptation from Standing Genetic Variation to a Moving Phenotypic Optimum. *Genetics*, *200*(4), 1255–1274. https://doi.org/10.1534/genetics.115.178574

MERLO, P. M. d. S., & Capixaba, C. (2012). *100 anos de Desafios, Crescimento e Inovação*. Vitória: Bumerangue Produção de Comunicação.

Mérot-L'Anthoëne, V., Tournebize, R., Darracq, O., Rattina, V., Lepelley, M., Bellanger, L., Tranchant-Dubreuil, C., Coulée, M., Pégard, M., Metairon, S., Fournier, C., Stoffelen, P., Janssens, S. B., Kiwuka, C., Musoli, P., Sumirat, U., Legnaté, H., Kambale, J.-L., Ferreira da Costa Neto, J., … Poncet, V. (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*(7), 1418–1430. https://doi.org/10.1111/pbi.13066

Meyer, R. S., DuVal, A. E., & Jensen, H. R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.*, *196*(1), 29–48. https://doi.org/10.1111/j.1469-8137.2012.04253.x

Meyer, R. S., & Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification [Number: 12 Publisher: Nature Publishing Group]. *Nat Rev Genet*, *14*(12), 840–852. https://doi.org/10.1038/nrg3605

Moat, J., Gole, T. W., & Davis, A. P. (2019). Least concern to endangered: Applying climate change projections profoundly influences the extinction risk assessment for wild Arabica coffee. *Glob. Chang. Biol.*, *25*(2), 390–403. https://doi.org/10.1111/gcb.14341

Molinaro, L., Marnetto, D., Mondal, M., Ongaro, L., Yelmen, B., Lawson, D. J., Montinaro, F., & Pagani, L. (2021). A chromosome-painting-based pipeline to infer local ancestry under limited source availability. *Genome Biol. Evol.*, *13*(4), evab025. https://doi.org/10.1093/gbe/evab025

Monat, C., Tranchant-Dubreuil, C., Kougbeadjo, A., Farcy, C., Ortega-Abboud, E., Amanzougarene, S., Ravel, S., Agbessi, M., Orjuela-Bouniol, J., Summo, M., & Sabot, F. (2015). TOGGLE: toolbox for generic NGS analyses. *BMC Bioinform.*, *16*(1), 374. https://doi.org/10.1186/s12859-015-0795-6

Moncada, M. D. P., Tovar, E., Montoya, J. C., González, A., Spindel, J., & McCouch, S. (2015). A genetic linkage map of coffee (Coffea arabica L.) and QTL for yield, plant height, and bean size. *Tree Genetics & Genomes*, *12*(1), 5. https://doi.org/10.1007/s11295-015-0927-1

Montagnon, C., Leroy, T., Cilas, C., & Eskes, A. B. (1993). Differences among clones of Coffea canephora in resistance to the scolytid coffee-twig borer. *International Journal of Pest Management*, *39*(2), 204–209. https://doi.org/10.1080/09670879309371792

Montagnon, C., Leroy, T., & Yapo, A. (1992a). Genotypic and phenotypic diversity of some coffee groups (*Coffea Canephora* Pierre) in the collections - consequences on their use in breeding. *Cafe Cacao The*, *36*(3), 187–198. ://A1992JR40800002
Jr408Times Cited:2Cited References Count:0

Montagnon, C., & Leroy, T. (1993). Réaction à la sécheresse de jeunes caféiers Coffea canephora de Côte d'Ivoire appartenant à différents groupes génétiques. *Café, Cacao, Thé*. Retrieved August 27, 2019, from http://agritrop.cirad.fr/396722/

Montagnon, C., Leroy, T., & Eskes, A. (1998a). Amélioration variétale de *Coffea canephora*. 1 : critères et méthodes de sélection. *Plantations, Recherche, Développement*, 89–98. Retrieved September 13, 2021, from https://agritrop.cirad.fr/390308/

Montagnon, C., Leroy, T., & Eskes, A. (1998b). Amélioration variétale de *Coffea canephora*. 2 : les programmes de sélection et leurs résultats. *Plantations, Recherche, Développement*. Retrieved June 19, 2019, from http://agritrop.cirad.fr/390311/

Montagnon, C., Leroy, T., & Yapo, A. (1992b). Diversité génotypique et phénotypique de quelques groupes de caféiers (*Coffea canephora* Pierre) en collection. Conséquences sur leur utilisation en sélection. *Café, Cacao, Thé*, *36*(3), 187–198.

Montagnon, C., Leroy, T., & Yapo, A. (1993). Caractérisation et évaluation de caféiers Coffea canephora prospectés dans des plantations de Côte-d'Ivoire. *Café, Cacao, Thé*. Retrieved August 21, 2019, from http://agritrop.cirad.fr/396716/

Moriondo, M., & Bindi, M. (2007). Impact of climate change on the phenology of typical Mediterranean crops. *Ital. J. Agrometeorol*, *3*, 5–12.

Moschetto, D., Montagnon, C., Guyot, B., Perriot, J.-J., Leroy, T., & Eskes, A. (1996). Studies on the effect of genotype on cup quality of Coffea canephora.

Musoli, P., Cubry, P., Aluka, P., Billot, C., Dufour, M., De Bellis, F., Pot, D., Bieysse, D., Charrier, A., & Leroy, T. (2009). Genetic differentiation of wild and cultivated populations: diversity of *Coffea canephora* Pierre in Uganda [Publisher: NRC Research Press]. *Genome*, *52*(7), 634–646. https://doi.org/10.1139/G09-037

Musoli, P. C., Cilas, C., Pot, D., Nabaggala, A., Nakendo, S., Pande, J., Charrier, A., Leroy, T., & Bieysse, D. (2013). Inheritance of resistance to coffee wilt disease (Fusarium xylarioides Steyaert) in Robusta coffee (Coffea canephora Pierre) and breeding perspectives. *Tree Genetics & Genomes*, *9*(2), 351–360. https://doi.org/10.1007/s11295-012-0557-9

Nab, C., & Maslin, M. (2020). Life cycle assessment synthesis of the carbon footprint of Arabica coffee: Case study of Brazil and Vietnam conventional and sustainable coffee production and export to the United Kingdom. *Geo: Geography and Environment*, *7*(2), e00096. https://doi.org/10.1002/geo2.96

Nakagawa, T., Doi, M., Nishi, K., Sugahara, T., Nishimukai, H., & Asano, M. (2019). A simple and versatile authenticity assay of coffee products by single-nucleotide polymorphism genotyping. *Bioscience, Biotechnology, and Biochemistry*, *83*(10), 1829–1836. https://doi.org/10.1080/09168451.2019.1618697

Nave, M., Avni, R., Çakır, E., Portnoy, V., Sela, H., Pourkheirandish, M., Ozkan, H., Hale, I., Komatsuda, T., Dvorak, J., & Distelfeld, A. (2019). Wheat domestication in light of haplotype analyses of the Brittle rachis 1 genes (BTR1-A and BTR1-B). *Plant Sci.*, *285*, 193–199. https://doi.org/10.1016/j.plantsci.2019.05.012

Neves, M. F., Trombin, V. G., Lopes, F. F., Kalaki, R., & Milan, P. (2012). World consumption of beverages. In M. F. Neves, V. G. Trombin, F. F. Lopes, R. Kalaki, & P. Milan (Eds.), *The orange juice business: A Brazilian perspective* (pp. 118–118). Academic Publishers. https://doi.org/10.3920/978-90-8686-739-4_31

Ngo-Thanh, H., Ngo-Duc, T., Nguyen-Hong, H., Baker, P., & Phan-Van, T. (2018). A distinction between summer rainy season and summer monsoon season over the Central Highlands of Vietnam. *Theor Appl Climatol*, *132*(3), 1237–1246. https://doi.org/10.1007/s00704-017-2178-6

Noir, S., Anthony, F., Bertrand, B., Combes, M.-C., & Lashermes, P. (2003). Identification of a major gene (Mex-1) from Coffea canephora conferring resistance to Meloidogyne exigua in Coffea arabica. *Plant Pathology*, *52*(1), 97–103. https://doi.org/10.1046/j.1365-3059.2003.00795.x

Noirot, M., PONCET, V., BARRE, P., HAMON, P., HAMON, S., & DE KOCHKO, A. (2003). Genome size variations in diploid african *Coffea* species. *Ann. Bot.*, *92*(5), 709–714. https://doi.org/10.1093/aob/mcg183

Nonato, J. V. A., Carvalho, H. F., Borges, K. L. R., Padilha, L., Maluf, M. P., Fritsche-Neto, R., & Guerreiro Filho, O. (2021). Association mapping reveals genomic regions associated with bienniality and resistance to biotic stresses in arabica coffee. *Euphytica*, *217*(10), 190. https://doi.org/10.1007/s10681-021-02922-9

Norris, E. T., Rishishwar, L., Chande, A. T., Conley, A. B., Ye, K., Valderrama-Aguirre, A., & Jordan, I. K. (2020). Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biol.*, *21*(1), 29. https://doi.org/10.1186/s13059-020-1946-2

Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. L. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theor Appl Genet*, *126*(2), 289–305. https://doi.org/10.1007/s00122-012-1971-y

Oliveira, H. R., Jacocks, L., Czajkowska, B. I., Kennedy, S. L., & Brown, T. A. (2020). Multiregional origins of the domesticated tetraploid wheats. *PLoS ONE*, *15*(1), e0227148. https://doi.org/10.1371/journal.pone.0227148

Oliveira, L. N. L., Rocha, R., Ferreira, F., Spinelli, V., Ramalho, A. R., & Teixeira, A. L. (2018). Selection of *Coffea canephora* parents from the botanical varieties Conilon and Robusta for the production of intervarietal hybrids. *Cienc. Rural*, *48*(4). https://doi.org/10.1590/0103-8478CR20170444

O'Neill, B. C., Carter, T. R., Ebi, K., Harrison, P. A., Kemp-Benedict, E., Kok, K., Kriegler, E., Preston, B. L., Riahi, K., Sillmann, J., van Ruijven, B. J., van Vuuren, D., Carlisle, D., Conde, C., Fuglestvedt, J., Green, C., Hasegawa, T., Leininger, J., Monteith, S., & Pichs-Madruga, R. (2020). Achievements and needs for the climate change scenario framework [Number: 12 Publisher: Nature Publishing Group]. *Nat. Clim. Chang.*, *10*(12), 1074–1084. https://doi.org/10.1038/s41558-020-00952-0

Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics*, *5*. Retrieved February 18, 2022, from https://www.frontiersin.org/article/10.3389/fgene.2014.00204

Paillard, M., Lashermes, P., & Pétiard, V. (1996). Construction of a molecular linkage map in coffee. *Theoret. Appl. Genetics*, *93*(1), 41–47. https://doi.org/10.1007/BF00225725

Paris, H. S. (2015). Origin and emergence of the sweet dessert watermelon, Citrullus lanatus. *Annals of Botany*, *116*(2), 133–148. https://doi.org/10.1093/aob/mcv077

Parkes, B., Defrance, D., Sultan, B., Ciais, P., & Wang, X. (2018). Projected changes in crop yield mean and variability over West Africa in a world 1.5&thinsp;K warmer than the pre-industrial era [Publisher: Copernicus GmbH]. *Earth System Dynamics*, *9*(1), 119–134. https://doi.org/10.5194/esd-9-119-2018

Pearl, H. M., Nagai, C., Moore, P. H., Steiger, D. L., Osgood, R. V., & Ming, R. (2004). Construction of a genetic map for arabica coffee. *Theor Appl Genet*, *108*(5), 829–835. https://doi.org/10.1007/s00122-003-1498-3

Pembleton, L. W., Cogan, N. O. I., & Forster, J. W. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol Ecol Resour*, *13*(5), 946–952. https://doi.org/10.1111/1755-0998.12129

Phan, V. H. (2017). Research results on robusta coffee breeding in Vietnam. *Vietnam Jounal of Science, Technology and Engineering*, *59*(4), 37–41. Retrieved August 17, 2022, from https://docplayer.net/203190594-Research-results-on-robusta-coffee-breeding-in-vietnam.html

Pickersgill, B. (2007). Domestication of Plants in the Americas: Insights from Mendelian and Molecular Genetics. *Annals of Botany*, *100*(5), 925–940. https://doi.org/10.1093/aob/mcm193

Poplin, R., Ruano-Rubio, V., DePristo, M., Fennell, T., Carneiro, M., Van der Auwera, G., Kling, D., Gauthier, L., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M., Neale, B., MacArthur, D., & Banks, E. (2017). *Scaling accurate genetic variant discovery to tens of thousands of sample* (tech. rep.) [Type: article]. https://doi.org/10.1101/201178

Pratap, A., Das, A., Kumar, S., & Gupta, S. (2021). Current Perspectives on Introgression Breeding in Food Legumes. *Frontiers in Plant Science*, *11*. Retrieved February 12, 2023, from https://www.frontiersin.org/articles/10.3389/fpls.2020.589189

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Purugganan, M. D., & Jackson, S. A. (2021). Advancing crop genomics from lab to field [Number: 5 Publisher: Nature Publishing Group]. *Nature Genetics*, *53*(5), 595–601. https://doi.org/10.1038/s41588-021-00866-3

R Core Team, R. (2022). R: A language and environment for statistical computing [Publisher: Vienna, Austria].

Rahman, A., Hallgrímsdóttir, I., Eisen, M., & Pachter, L. (2018). Association mapping from sequencing reads using k-mers (J. Flint, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *7*, e32920. https://doi.org/10.7554/eLife.32920

Ramadiana, S., Hapsoro, D., Evizal, R., Setiawan, K., Karyanto, A., & Yusnita, Y. (2021). Genetic diversity among 24 clones of robusta coffee in Lampung based on RAPD markers [Number: 6]. *Biodiversitas Journal of Biological Diversity*, *22*(6). https://doi.org/10.13057/biodiv/d220614

Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability [Number: 1 Publisher: Nature Publishing Group]. *Nat Commun*, *6*(1), 5989. https://doi.org/10.1038/ncomms6989

Ray, D. K., West, P. C., Clark, M., Gerber, J. S., Prishchepov, A. V., & Chatterjee, S. (2019). Climate change has likely already affected global food production [Publisher: Public Library of Science]. *PLOS ONE*, *14*(5), e0217148. https://doi.org/10.1371/journal.pone.0217148

Raza, A., Razzaq, A., Mehmood, S. S., Zou, X., Zhang, X., Lv, Y., & Xu, J. (2019). Impact of Climate Change on Crops Adaptation and Strategies to Tackle Its Outcome: A Review [Number: 2 Publisher: Multidisciplinary Digital Publishing Institute]. *Plants*, *8*(2), 34. https://doi.org/10.3390/plants8020034

Razafinarivo, N. J., Guyot, R., Davis, A. P., Couturon, E., Hamon, S., Crouzillat, D., Rigoreau, M., Dubreuil-Tranchant, C., Poncet, V., De Kochko, A., Rakotomalala, J.-J., & Hamon, P. (2013). Genetic structure and diversity of coffee (Coffea) across Africa and the Indian Ocean islands revealed using microsatellites. *Annals of Botany*, *111*(2), 229–248. https://doi.org/10.1093/aob/mcs283

Rellstab, C., Dauphin, B., & Exposito-Alonso, M. (2021). Prospects and limitations of genomic offset in conservation management. *Evolutionary Applications*, *14*(5), 1202–1212. https://doi.org/10.1111/eva.13205

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370. https://doi.org/10.1111/mec.13322

Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A. R., Graf, R., Bodénès, C., Sperisen, C., Kremer, A., & Gugerli, F. (2016). Signatures of local adaptation in candidate genes of oaks (Quercus spp.) with respect to present and future climatic conditions [ISBN: 0962-1083 Publisher: Wiley Online Library]. *Molecular Ecology*, *25*(23), 5907–5924.

Rendón-Anaya, M., Wilson, J., Sveinsson, S., Fedorkov, A., Cottrell, J., Bailey, M. E. S., Ruņģis, D., Lexer, C., Jansson, S., Robinson, K. M., Street, N. R., & Ingvarsson, P. K. (2021). Adaptive introgression facilitates adaptation to high latitudes in European aspen (*Populus tremula* L.) *Mol. Biol. Evol.*, *38*(11), 5034–5050. https://doi.org/10.1093/molbev/msab229

Renzi, J. P., Coyne, C. J., Berger, J., von Wettberg, E., Nelson, M., Ureta, S., Hernández, F., Smýkal, P., & Brus, J. (2022). How Could the Use of Crop Wild Relatives in Breeding Increase the Adaptation of Crops to Marginal Environments? *Frontiers in Plant Science*, *13*. Retrieved July 10, 2023, from https://www.frontiersin.org/articles/10.3389/fpls.2022.886162

Rius, M., & Darling, J. A. (2014). How important is intraspecific genetic admixture to the success of colonising populations? *Trends Ecol. Evol.*, *29*(4), 233–242. https://doi.org/10.1016/j.tree.2014.02.003

Romero, G., Vásquez, L. M., Lashermes, P., & Herrera, J. C. (2014). Identification of a major QTL for adult plant resistance to coffee leaf rust (Hemileia vastatrix) in the natural Timor hybrid (Coffea arabica x C. canephora) [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbr.12127]. *Plant Breeding*, *133*(1), 121–129. https://doi.org/https://doi.org/10.1111/pbr.12127

Roncal, J., Guyot, R., Hamon, P., Crouzillat, D., Rigoreau, M., Konan, O. N., Rakotomalala, J.-J., Nowak, M. D., Davis, A. P., & de Kochko, A. (2016). Active transposable elements recover species boundaries and geographic structure in Madagascan coffee species. *Mol Genet Genomics*, *291*(1), 155–168. https://doi.org/10.1007/s00438-015-1098-3

Salojärvi, J., Rambani, A., Yu, Z., Guyot, R., Strickler, S., Lepelley, M., Wang, C., Rajaraman, S., Rastas, P., Zheng, C., Muñoz, D. S., Meidanis, J., Paschoal, A. R., Bawin, Y., Krabbenhoft, T., Wang, Z. Q., Fleck, S., Aussel, R., Bellanger, L., … Descombes, P. (2023). The genome and population genomics of allopolyploid Coffea arabica reveal the diversification history of modern coffee cultivars [Pages: 2023.09.06.556570 Section: New Results]. https://doi.org/10.1101/2023.09.06.556570

Sankararaman, S., Sridhar, S., Kimmel, G., & Halperin, E. (2008). Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.*, *82*(2), 290–303. https://doi.org/10.1016/j.ajhg.2007.09.022

Sant'Ana, G. C., Pereira, L. F. P., Pot, D., Ivamoto, S. T., Domingues, D. S., Ferreira, R. V., Pagiatto, N. F., da Silva, B. S. R., Nogueira, L. M., Kitzberger, C. S. G., Scholz, M. B. S., de Oliveira, F. F., Sera, G. H., Padilha, L., Labouisse, J.-P., Guyot, R., Charmetant, P., & Leroy, T. (2018). Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *8*(1), 465. https://doi.org/10.1038/s41598-017-18800-1

Sathee, L., Jagadhesan, B., Pandesha, P. H., Barman, D., Adavi B, S., Nagar, S., Krishna, G. K., Tripathi, S., Jha, S. K., & Chinnusamy, V. (2022). Genome Editing Targets for Improving Nutrient Use Efficiency and Nutrient Stress Adaptation. *Frontiers in Genetics*, *13*. Retrieved October 9, 2023, from https://www.frontiersin.org/articles/10.3389/fgene.2022.900897

Scalabrin, S., Toniutti, L., Di Gaspero, G., Scaglione, D., Magris, G., Vidotto, M., Pinosio, S., Cattonaro, F., Magni, F., Jurman, I., Cerutti, M., Suggi Liverani, F., Navarini, L., Del Terra, L., Pellegrino, G., Ruosi,

M. R., Vitulo, N., Valle, G., Pallavicini, A., … Bertrand, B. (2020). A single polyploidization event at the origin of the tetraploid genome of Coffea arabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Sci Rep*, *10*, 4642. https://doi.org/10.1038/s41598-020-61216-7

Schubert, R., Andaleon, A., & Wheeler, H. E. (2020). Comparing local ancestry inference models in populations of two- and three-way admixture. *PeerJ*, *8*, e10090. https://doi.org/10.7717/peerj.10090

Schwinning, S., Lortie, C. J., Esque, T. C., & DeFalco, L. A. (2022). What common-garden experiments tell us about climate responses in plants. *Journal of Ecology*, *110*(5), 986–996. https://doi.org/10.1111/1365-2745.13887

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations [Number: 7 Publisher: Nature Publishing Group]. *Nat Genet*, *44*(7), 825–830. https://doi.org/10.1038/ng.2314

Shriner, D. (2013). Overview of Admixture Mapping. *Curr Protoc Hum Genet*, *CHAPTER*, Unit1.23. https://doi.org/10.1002/0471142905.hg0123s76

Shriner, D. (2017). Overview of admixture mapping. *Current Protocols in Human Genetics* (pp. 1–23). Retrieved February 18, 2022, from http://onlinelibrary.wiley.com/doi/abs/10.1002/cphg.44

Shringarpure, S., & Xing, E. P. (2014). Effects of sample selection bias on the accuracy of population structure and ancestry inference. *G3 (Bethesda)*, *4*(5), 901–911. https://doi.org/10.1534/g3.113.007633

Silva, M. d. C., Várzea, V., Guerra-Guimarães, L., Azinheira, H. G., Fernandez, D., Petitot, A.-S., Bertrand, B., Lashermes, P., & Nicole, M. (2006). Coffee resistance to the main diseases: leaf rust and coffee berry disease [Publisher: Brazilian Journal of Plant Physiology]. *Braz. J. Plant Physiol.*, *18*, 119–147. https://doi.org/10.1590/S1677-04202006000100010

Singh, U. M., Dixit, S., Alam, S., Yadav, S., Prasanth, V. V., Singh, A. K., Venkateshwarlu, C., Abbai, R., Vipparla, A. K., Badri, J., Ram, T., Prasad, M. S., Laha, G. S., Singh, V. K., & Kumar, A. (2022). Marker-assisted forward breeding to develop a drought-, bacterial-leaf-blight-, and blast-resistant rice cultivar. *The Plant Genome*, *15*(1), e20170. https://doi.org/10.1002/tpg2.20170

Sloat, L. L., Davis, S. J., Gerber, J. S., Moore, F. C., Ray, D. K., West, P. C., & Mueller, N. D. (2020). Climate adaptation by crop migration [Number: 1 Publisher: Nature Publishing Group]. *Nat Commun*, *11*(1), 1243. https://doi.org/10.1038/s41467-020-15076-4

Smýkal, P., Nelson, M., Berger, J., & Von Wettberg, E. (2018). The Impact of Genetic Changes during Crop Domestication. *Agronomy*, *8*(7), 119. https://doi.org/10.3390/agronomy8070119

Songsomboon, K., Brenton, Z., Heuser, J., Kresovich, S., Shakoor, N., Mockler, T., & Cooper, E. A. (2021). Genomic patterns of structural variation among diverse genotypes of Sorghum bicolor and a potential role for deletions in local adaptation. *G3 (Bethesda)*, *11*(7), jkab154. https://doi.org/10.1093/g3journal/jkab154

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., & Neale, D. B. (2013). Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes*, *9*(4), 901–911. https://doi.org/10.1007/s11295-013-0596-x

Sousa, I. C. d., Nascimento, M., Silva, G. N., Nascimento, A. C. C., Cruz, C. D., Silva, F. F. e., Almeida, D. P. d., Pestana, K. N., Azevedo, C. F., Zambolim, L., & Caixeta, E. T. (2020). Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms [Publisher: São Paulo - Escola Superior de Agricultura "Luiz de Queiroz"]. *Sci. agric. (Piracicaba, Braz.)*, *78*. https://doi.org/10.1590/1678-992X-2020-0021

Spinelli, V. M., Moraes, M. S., Alves, D. S. B., Rocha, R. B., Ramalho, A. R., & Teixeira, A. L. (2018). Contribution of agronomic traits to the coffee yield of Coffea canephora Pierre ex A. Froehner in the western amazon region [Accepted: 2018-12-20T10:31:06Z Publisher: Editora UFLA]. https://doi.org/10.25186/cs.v13i3.1452

Spinoso-Castillo, J. L., Escamilla-Prado, E., Aguilar-Rincón, V. H., Morales Ramos, V., de los Santos, G. G., Pérez-Rodríguez, P., & Corona-Torres, T. (2020). Genetic diversity of coffee (Coffea spp.) in Mexico evaluated by using DArTseq and SNP markers. *Genet Resour Crop Evol*, *67*(7), 1795–1806. https://doi.org/10.1007/s10722-020-00940-5

Suharyanti, N. A., Mizuno, K., & Sodri, A. (2020). The effect of water deficit on inflorescence period at palm oil productivity on peatland. *E3S Web of Conferences*, *211*, 05005.

Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., & Glenn, K. C. (2021). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Science*, *61*(2), 839–852. https://doi.org/10.1002/csc2.20377

Tan, P. V., Thanh, N. D., & Nguyen, H. V. (2013). Climate Change Assessment for Risk Management of Hydro-meteorological Disasters in Coffee Cultivation and Trading in the Vietnam Central Highlands. *Unpublished report 49pp*.

Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.*, *79*(1), 1–12. https://doi.org/10.1086/504302

Teixeira Alexsandro, L., França Souza, F., Rocha, R., Vieira, J., Torres Josemar, D., Rodrigues Karine, M., Moraes, M., Silva Camila, A., Goncalves de Oliveira Victor, E., & Resende Lourenco Joao, L. (2017). Performance of intraspecific hybrids (Kouillou x Robusta) of Coffea canephora Pierre. *African Journal of Agricultural Research*, *12*, 2675–2680. https://doi.org/10.5897/AJAR2017.12446

Thornton, T. A., & Bermejo, J. L. (2014). Local and global ancestry inference, and applications to genetic association analysis for admixed populations. *Genet. Epidemiol.*, *38*(1), S5–S12. https://doi.org/10.1002/gepi.21819

Thurman, L. L., Stein, B. A., Beever, E. A., Foden, W., Geange, S. R., Green, N., Gross, J. E., Lawrence, D. J., LeDee, O., & Olden, J. D. (2020). Persist in place or shift in space? Evaluating the adaptive capacity of species to climate change [ISBN: 1540-9295 Publisher: Wiley Online Library]. *Frontiers in Ecology and the Environment*, *18*(9), 520–528.

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, *25*(10), 2144–2164. https://doi.org/10.1111/mec.13606

Torres, L. F., Reichel, T., Déchamp, E., de Aquino, S. O., Duarte, K. E., Alves, G. S. C., Silva, A. T., Cotta, M. G., Costa, T. S., Diniz, L. E. C., Breitler, J.-C., Collin, M., Paiva, L. V., Andrade, A. C., Etienne, H., & Marraccini, P. (2019). Expression of DREB-Like Genes in Coffea canephora and C. arabica Subjected to Various Types of Abiotic Stress. *Tropical Plant Biol.*, *12*(2), 98–116. https://doi.org/10.1007/s12042-019-09223-5

Tournebize, R. (2017). *Influence des variations spatio-temporelles de l'environnement sur la distribution actuelle de la diversité génétique des populations* (Doctoral dissertation). University of Montpellier. Montpellier, France.

Tournebize, R., Borner, L., Manel, S., Meynard, C. N., Vigouroux, Y., Crouzillat, D., Fournier, C., Kassam, M., Descombes, P., Tranchant-Dubreuil, C., Parrinello, H., Kiwuka, C., Sumirat, U., Legnate, H., Kambale, J.-L., Sonké, B., Mahinga, J. C., Musoli, P., Janssens, S. B., … Poncet, V. (2022). Ecological and genomic vulnerability to climate change across native populations of Robusta coffee (*Coffea canephora*). *Global Change Biology*, *28*(13), 4124–4142. https://doi.org/10.1111/gcb.16191

Tran, H. T. M., Furtado, A., Vargas, C. A. C., Smyth, H., Slade Lee, L., & Henry, R. (2018). SNP in the Coffea arabica genome associated with coffee quality. *Tree Genetics & Genomes*, *14*(5), 72. https://doi.org/10.1007/s11295-018-1282-9

Tran, H. T., Ramaraj, T., Furtado, A., Lee, L. S., & Henry, R. J. (2018). Use of a draft genome of coffee (Coffea arabica) to identify SNPs associated with caffeine content. *Plant Biotechnol J*, *16*(10), 1756–1766. https://doi.org/10.1111/pbi.12912

Tranchant-Dubreuil, C., Ravel, S., Monat, C., Sarah, G., Diallo, A., Helou, L., Dereeper, A., Tando, N., Orjuela-Bouniol, J., & Sabot, F. (2018). TOGGLe, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses. *bioRxiv*, 245480. https://doi.org/10.1101/245480

Trinh, P. Q., de la Peña, E., Nguyen, C. N., Nguyen, H. X., & Moens, M. (2009). Plant-parasitic nematodes associated with coffee in Vietnam. [ISBN: 0869-6918]. *Russian journal of Nematology*, *17*(1), 73.

Trinh, Q. P., Le, T. M. L., Nguyen, T. D., Nguyen, H. T., Liebanas, G., & Nguyen, T. a. D. (2019). Meloidogyne daklakensis n. sp. (Nematoda: Meloidogynidae), a new root-knot nematode associated with Robusta coffee (Coffea canephora Pierre ex A. Froehner) in the Western Highlands, Vietnam [Publisher: Cambridge University Press]. *Journal of Helminthology*, *93*(2), 242–254. https://doi.org/10.1017/S0022149X18000202

van der Vossen, H., Bertrand, B., & Charrier, A. (2015). Next generation variety development for sustainable production of arabica coffee (Coffea arabica L.): a review. *Euphytica*, *204*(2), 243–256. https://doi.org/10.1007/s10681-015-1398-z

Van Hintum, T. J., Brown, A. H. D., & Spillane, C. (2000). *Core collections of plant genetic resources* [Issue: 3]. Bioversity International.

Vanden Abeele, S., Janssens, S. B., Asimonyio Anio, J., Bawin, Y., Depecker, J., Kambale, B., Mwanga Mwanga, I., Ebele, T., Ntore, S., Stoffelen, P., & Vandelook, F. (2021). Genetic diversity of wild and cultivated *Coffea canephora* in northeastern DR Congo and the implications for conservation. *American Journal of Botany*, *108*(12), 2425–2434. https://doi.org/10.1002/ajb2.1769

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., & Sorrells, M. E. (2021). Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends in Plant Science*, *26*(6), 631–649. https://doi.org/10.1016/j.tplants.2021.03.010

Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J. M., & Castanera, R. (2019). A benchmark of transposon insertion detection tools using real data. *Mobile DNA*, *10*(1), 53. https://doi.org/10.1186/s13100-019-0197-9

Verleysen, L., Bollen, R., Kambale, J.-L., Ebele, T., Katshela, B. N., Depecker, J., Poncet, V., Assumani, D.-M., Vandelook, F., Stoffelen, P., Honnay, O., & Ruttink, T. (2023). Characterization of the genetic composition and establishment of a core collection for the INERA Robusta coffee (Coffea canephora) field genebank from the Democratic Republic of Congo. *Frontiers in Sustainable Food Systems*, *7*. Retrieved August 15, 2023, from https://www.frontiersin.org/articles/10.3389/fsufs.2023.1239442

Vi, T., Marraccini, P., Kochko, A. d., Cubry, P., Khong, N. G., & Poncet, V. (2022). Sequencing-based molecular markers for wild and cultivated coffee diversity exploration and crop improvement [Num Pages: 8]. *Coffee Science*. CRC Press.

Vi, T., Vigouroux, Y., Cubry, P., Marraccini, P., Phan, H. V., Khong, G. N., & Poncet, V. (2023). Genome-wide admixture mapping identifies wild ancestry-of-origin segments in cultivated Robusta coffee. *Genome Biology and Evolution*, *15*(5), evad065. https://doi.org/10.1093/gbe/evad065

Vieira, N. G., Carneiro, F. A., Sujii, P. S., Alekcevetch, J. C., Freire, L. P., Vinecky, F., Elbelt, S., Silva, V. A., DaMatta, F. M., Ferrão, M. A. G., Marraccini, P., & Andrade, A. C. (2013). Different molecular mechanisms account for drought tolerance in *Coffea canephora* var. Conilon. *Trop. Plant Biol.*, *6*(4), 181–190. https://doi.org/10.1007/s12042-013-9126-0

Voichek, Y., & Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes [Number: 5 Publisher: Nature Publishing Group]. *Nat Genet*, *52*(5), 534–540. https://doi.org/10.1038/s41588-020-0612-7

Wrigley, G. (1988). Coffee–tropical agriculture series. *Harlow, UK: Longman Scientific & Technical*.

Wu, J., Liu, Y., & Zhao, Y. (2021). Systematic review on local ancestor inference from a mathematical and algorithmic perspective. *Front. Genet.*, *12*, 639877. Retrieved February 18, 2022, from https://www.frontiersin.org/article/10.3389/fgene.2021.639877

Würschum, T., Boeven, P. H. G., Langer, S. M., Longin, C. F. H., & Leiser, W. L. (2015). Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genet*, *16*(1), 96. https://doi.org/10.1186/s12863-015-0258-0

Xu, F., Wang, W., Wang, P., Jun Li, M., Chung Sham, P., & Wang, J. (2012). A fast and accurate SNP detection algorithm for next-generation sequencing data [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *3*(1), 1258. https://doi.org/10.1038/ncomms2256

Xu, Z. Z., & Zhou, G. S. (2006). Combined effects of water stress and high temperature on photosynthesis, nitrogen metabolism and lipid peroxidation of a perennial grass Leymus chinensis. *Planta*, *224*(5), 1080–1090. https://doi.org/10.1007/s00425-006-0281-5

Yadav, B., Jogawat, A., Rahman, M. S., & Narayan, O. P. (2021). Secondary metabolites in the drought stress tolerance of crop plants: A review. *Gene Reports*, *23*, 101040. https://doi.org/10.1016/j.genrep.2021.101040

Yali, W., & Mitiku, T. (2022). Mutation Breeding and Its Importance in Modern Plant Breeding. *JPS*, *10*(2), 64. https://doi.org/10.11648/j.jps.20221002.13

Yamada, S., Kurokawa, Y., Nagai, K., Angeles-Shim, R. B., Yasui, H., Furuya, N., Yoshimura, A., Doi, K., Ashikari, M., & Sunohara, H. (2020). Evaluation of backcrossed pyramiding lines of the yield-related gene and the bacterial leaf Blight resistant genes. *J. Intl. Cooper Agric. Dev*, *18*, 18–28.

Yan, H., Sun, M., Zhang, Z., Jin, Y., Zhang, A., Lin, C., Wu, B., He, M., Xu, B., Wang, J., Qin, P., Mendieta, J. P., Nie, G., Wang, J., Jones, C. S., Feng, G., Srivastava, R. K., Zhang, X., Bombarely, A., . . . Huang, L. (2023). Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet [Number: 3 Publisher: Nature Publishing Group]. *Nat Genet*, *55*(3), 507–518. https://doi.org/10.1038/s41588-023-01302-4

Yang, J. J., Li, J., Buu, A., & Williams, L. K. (2013). Efficient inference of local ancestry. *Bioinform.*, *29*(21), 2750–2756. https://doi.org/10.1093/bioinformatics/btt488

Zhang, D., Vega, F. E., Solano, W., Su, F., Infante, F., & Meinhardt, L. W. (2021). Selecting a core set of nuclear SNP markers for molecular characterization of Arabica coffee (Coffea arabica L.) genetic resources. *Conservation Genet Resour*. https://doi.org/10.1007/s12686-021-01201-y

Zhao, L., Wang, K., Wang, K., Zhu, J., & Hu, Z. (2020). Nutrient components, health benefits, and safety of litchi (*Litchi chinensis* Sonn.): A review. *Compr. Rev. Food. Sci. Food Saf.*, *19*(4), 2139–2163. https://doi.org/10.1111/1541-4337.12590

Zheng, Y., Crawford, G. W., Jiang, L., & Chen, X. (2016). Rice Domestication Revealed by Reduced Shattering of Archaeological rice from the Lower Yangtze valley [Number: 1 Publisher: Nature Publishing Group]. *Sci Rep*, *6*(1), 28136. https://doi.org/10.1038/srep28136

Zhou, Q., Zhao, L., & Guan, Y. (2016). Strong Selection at MHC in Mexicans since Admixture. *PLOS Genetics*, *12*(2), e1005847. https://doi.org/10.1371/journal.pgen.1005847

Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., Xu, D., Chen, H., Wang, Y., Wang, Y.-g., Liu, S., Jiao, C., Lu, H., Wang, J., Yin, C., Jiao, Y., & Lu, F. (2020). *Triticum* population sequencing provides insights into wheat adaptation [Number: 12 Publisher: Nature Publishing Group]. *Nat. Genet.*, *52*(12), 1412–1422. https://doi.org/10.1038/s41588-020-00722-w

Zhu, F., Ahchige, M. W., Brotman, Y., Alseekh, S., Zsögön, A., & Fernie, A. R. (2022). Bringing more players into play: Leveraging stress in genome wide association studies. *Journal of Plant Physiology*, *271*, 153657. https://doi.org/10.1016/j.jplph.2022.153657

Żmieńko, A., Samelak, A., Kozłowski, P., & Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theor Appl Genet*, *127*(1), 1–18. https://doi.org/10.1007/s00122-013-2177-7

Zohary, D., & Hopf, M. (2000). Domestication of plants in the Old World: the origin and spread of cultivated plants in West Asia, Europe and the Nile Valley. [Publisher: Oxford University Press]. *Domestication of plants in the Old World: the origin and spread of cultivated plants in West Asia, Europe and the Nile Valley.*, (Ed.3). Retrieved October 9, 2023, from https://www.cabdirect.org/cabdirect/abstract/20013014838

Robusta coffee is produced from the *Coffea canephora* species of the Rubiaceae flowering-plant family. The genetic diversity of wild *C. canephora* is structured across its natural distribution in intertropical Africa. These different genetic groups differ in their adaptation to various environments. Only a limited portion of its wild genetic diversity has been selected and diffused for cultivation in other parts of the world since the 1900s. Vietnam has been the largest global Robusta producer since the 2000s, but is now facing a risk of coffee yield loss due to climate change. To propose strategies of adaptation to climate change for Vietnamese *C. canephora*, it is crucial to understand the part of genetic diversity contributing to the cultivated germplasm, and assess the suitability of all the wild genetic sources to the current and future local environment. Using population genetic analyses with a reference set of native African accessions, we identified that a genetic group originating from the Congo basin was the main contributor to the Vietnamese accessions. We developed an approach for chromosome painting to detect the distribution of wild ancestry-of-origin segments in cultivated Robusta genomes. We then revealed multiple hybrids of two or three groups with complex genome-wide admixture patterns. Finally, we estimated genomic suitability of wild African accessions to the current and future climate in Vietnam, based on relationships between genetic variants and climatic variables. Our predictions highlighted that the best suitable material comes from a genetic group in Gabon and Angola, which is currently only present in few hybrids in the Vietnamese varieties. The mismatch between the genetic origins of cultivated accessions and the predicted suitable genetic sources would offer useful information for establishing *C. canephora* coffee improvement and conservation plans in Vietnam.

Le café robusta est produit à partir de l'espèce Coffea canephora, de la famille des plantes à fleurs Rubiaceae. La diversité génétique de l'espèce sauvage *C. canephora* est très structurée dans son aire de répartition naturelle en Afrique intertropicale. Ces groupes génétiques présentent des adaptations différentes a leur environnements. Une fraction limitée de cette diversité génétique sauvage a été sélectionnée et diffusée pour la culture dans d'autres parties du monde depuis les années 1900. Le Vietnam est le plus grand producteur mondial de Robusta depuis les années 2000s, mais il risque aujourd'hui des pertes de rendements en café en raison du changement climatique. Pour proposer des strategies d'adaptation au changement climatique des *C. canephora* vietnamiens, il est essentiel de comprendre la part de diversité génétique contribuant aux accessions actuelles cultivées au Vietnam, et d'évaluer l'adéquation des differentes sources génétiques africaines à l'environnement local actuel et futur. L'analyse de la structure génétique incorporant un set de référence d'accessions africaines natives, nous a permis d'identifier un groupe génétique originaire du bassin du Congo comme le principal contributeur à la diversité variétale des *C. canephora* vietnamiens. Nous avons développé une approche pour peindre les chromosomes en fonction des groupes génétiques africains. Il a été ainsi possible d'étudier l'origine d'individus hybrides entre deux ou trois groupes. Enfin, nous avons estimé l'adéquation génomique des accessions sauvages africaines à l'environnement actuel et futur du Vietnam, sur la base des relations statistiques entre les variants génétiques et les variables climatiques. Nos prédictions ont mis en évidence que les accessions natives les mieux adaptées se trouvent au sein d'un groupe génétique distribué au Gabon et en Angola, qui n'est représenté, au sein des variétés Vietnamienne que dans quelques hybrides. Le décalage génétique entre les accessions actuellement cultivées et les sources génétiques prédites comme étant les plus adaptées aux climats actuels et futurs offre des informations utiles pour l'établissement de plans d'amélioration et de conservation du café *C. canephora* au Vietnam.