THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE MONTPELLIER SUPAGRO

En Génétique et amélioration des plantes

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Portée par

Unité de recherche AGAP

Caractérisation de génomes mosaïques de plantes cultivées : évaluation méthodologique et application aux bananiers

Présentée par Aurélien COTTIN

Le 19 juin 2020

Sous la direction de Nabila YAHIAOUI

Devant le jury composé de

Myriam HEUERTZ, Directrice de recherches, HDR, INRAE	Rapporteur
Thierry ROBERT, Maître de conférence, HDR, Sorbonne Université	Rapporteur
Jacques DAVID, Professeur, HDR, Montpellier SupAgro	Président du jury
Gilles CHARMET, Directeur de recherches, HDR, INRAE	Examinateur
Nabila YAHIAOUI, Chercheur, HDR, CIRAD	Directrice de thèse
Mathieu GAUTIER, Directeur de recherches, HDR, INRAE	Invité, co-encadrant

UNIVERSITÉ DE MONTPELLIER

Montpellier SupAgro

Remerciements

Ces remerciements ont été amendés post-soutenance pour les étoffer un peu.

Je remercie en premier lieu Myriam Heuertz et Thierry Robert, les rapporteurs de cette thèse, Gilles Charmet, examinateur, et Jacques David le président du jury, pour leurs retours très positifs, et leurs remarques, questions et la discussion qui a suivi ma soutenance.

Merci à Tom Druet, Xavier Perrier, Nicolas Bierne et Yves Vigouroux, membres de mes comités de thèse, pour leurs conseils et commentaires.

Merci à Christophe Joubert pour la technique de la visioconférence qui malgré le coronavirus a permis une soutenance dans de très bonnes conditions.

Je remercie chaudement ma directrice de thèse Nabila Yahiaoui et mon co-encadrant Mathieu Gautier pour la confiance et le support qu'ils m'ont apporté tout au long de ces trois ans et demi de thèse, au gré de mes pérégrinations scientifiques chaotiques entre *Musa* et Markov. Je vais de plus insister sur le support mêlant patience et bienveillance (la vraie, pas au sens « Startup Nation » du terme) tout au long de la thèse dans les coups de mou du moral, et surtout sur la période de la fin de thèse qui a été pour le moins turbulente. Merci a Jean-Christophe Glaszmann, mon premier directeur de thèse, pour ses conseils et les discussions « riz » pendant la première partie de la thèse.

Je remercie les personnes ayant aidées en plus de mes encadrants sur la préparation du manuscrit, Guillaume et Françoise, ainsi que sur l'oral, re-Guillaume, Franc-Christophe, Frank Curk, David Pot, Xavier Perrier et Jean-Christophe Glaszmann.

Je remercie l'équipe SEG pour son accueil et les moments partagés ensemble ces trois ans, en particulier mes colocataires de bureaux, successivement le microscope et Franc-Christophe. Merci à Olivier pour ses histoires rocambolesques, Catherine pour sa bonne humeur, Guillaume pour son soutien et ses conseils, Franc-Christophe pour les randonnées (en particulier le pèlerinage pour apprendre la projection astrale auprès du gardien de la porte des energies telluriques de Bugarach). Merci aussi à Jean-Yves, Françoise et Angélique.

Je remercie les camarades de l'équipe ID de la pause café (Jean-François, Gautier, Marilyne, Fred) pour les moments café/canapé et les discussions de hautes volées, ainsi que du support moral sur la thèse.

Je remercie successivement pour le soutien et l'aide apportée aux moments de (gros) doutes Nabila et Mathieu, Nicolas Bierne, Hélène Joly, Stéphanie Bocs et Émilie Benoit-Rivier.

Merci également aux camarades doctorants, stagiaires et CDD pour le support multiple, les animations et discussions scientifiques mais aussi et surtout pour la vie sociale à l'écart du boulot, à savoir Marion, Lauriane, Clément, Lisa, Antoine, João, Ian, Thibault, Jeanne, Stella, Aurélie, Abdoulaye, Kelly, Benjamin, Léo, Nicolas, Cédric, Céline, et tout les malheureux que j'ai oublié (pardon !).

Merci à Florence Chazot, Corinne Poitout et Elisabeth Bozsonyik pour la gestion des parties administratives de la thèse. Merci spécial à Bertrand Pitollat pour ses pouvoirs magiques d'administrateur système et l'environnement de travail fort agréable fourni pendant ces 3 ans.

Merci à toutes celles et ceux qui m'ont guidé sur le parcours scientifique, de Fred, Jamy et Sabine sur la télévision publique, à mes profs qui m'ont marqué au lycée et à l'université comme Marcel Eberhart, Matthieu Reichstadt, ou encore Roland Barriot, Gwennaële Fichant, à l'hurluberlu rencontré

au CIRAD en 1ère qui ventait les mérites de la vulgarisation avec une bédé sur l'autruche.

Merci aux personnes comme Henri Broch ou son disciple Richard Monvoisin, liés aux mouvements d'esprit critiques de m'avoir enseigné la philosophie des sciences là où l'Université ne l'a pas fait directement, ainsi qu'à Guillemette Reviron qui m'a permis la mise en pratique de ces concepts.

Merci à l'association Montpellibre et la maison pour tous Albert Camus de m'avoir accueilli et permit d'intervenir en tant que bénévole pour faire de l'éducation populaire à l'informatique et Internet auprès d'un public éloigné de ces concepts mais dont ils ne peuvent plus vraiment se passer.

J'en profite pour glisser un message plus général ici de remerciement envers la communauté du logiciel libre. Cette thèse a majoritairement été réalisée quand c'était possible avec des logiciels libres, et je remercie les communautés du noyau Linux, du système GNU, d'Ubuntu et d'Archlinux, de GNOME, de Libreoffice, de LaTeX, d'Inkscape, de la fondation Mozilla, de Thunderbird, et plein d'autres, mais aussi des langages `R` (en particulier le `tidyverse`, même si c'est plus open source que libre), `perl` (5, 6, 7 ?), `python` (définitivement 3 ou plus) ou d'outils comme `pandoc`. Merci aux communautés de Bioinformatique et de Génétique en général de favoriser au mieux le libre ou à défaut l'open source.

Merci à mes encadrants de stage Sébastien Terrat, Jérôme Bouligand, Roland Barriot, Matthias Zytnicki et Christine Gaspin, qui ont chacun et chacune guidé mes premiers pas dans l'environnement de la recherche scientifique publique française.

Merci à mes amis à distance, du `mumble` ou d'ailleurs, Jordan, Gaucher, Etienne, Quentin, Romain, Louise, Pierrick, mais aussi Régis, Maeva, Bastieng, Audric, Cédric, Simon, Pierre, Aurèle, Valentin, Anne-Sophie, Arsène et tout les autres malheureux que j'ai oublié aussi (encore pardon !).

Merci à ma famille qui m'a soutenue malgré la distance physique et la distance mentale que j'ai établi, mes parents Noël et Chantal, mes frères et sœurs Stéphanie, François, Émilie, Julien ainsi que Patrice, Thomas, François 2, Virginie, Madison et Noëlie.

Enfin, merci à Marion et Mirri d'avoir traversé tout ça avec moi, ces presque trois années ensemble ont été fantastiques.

Je cite la fin des remerciements de la thèse de Marion, parce que c'est bien écrit et que ca s'applique tout autant à moi :

« *Le mérite n'est qu'un mirage, cachant une réalité faite de contexte et de chance. J'ai eu la chance d'être accompagnée au long de ma vie par toutes ces personnes qui ont fait de moi qui je suis, compensant un contexte qui ne permettait pas de prédire ce parcours. Je souhaite à toutes celles et ceux à qui on a un jour dit que ça ne serait pas possible d'avoir le même « mérite » que moi.* »

Résumé

De nombreuses plantes cultivées sont issues d'évènements d'hybridations intersubspécifiques qui sont associés à leur processus de domestication et leur diversification. C'est, par exemple, le cas des bananiers cultivés qui sont des hybrides diploïdes ou triploïdes, propagés par multiplication végétative et issus d'hybridations entre des sous-espèces et des espèces du genre *Musa*, réparties dans différentes régions et îles du Sud-Est asiatique. Les génomes résultant de ces hybridations ont une structure en mosaïque de séquences d'origines ancestrales différentes qui peut être caractérisée par des méthodologies d'inférence locale des états ancestraux (ILEA). Ces méthodologies ont été développées principalement dans le cadre de la génétique humaine, pour des contextes qui ne sont pas forcément compatibles avec ceux de certaines plantes cultivées. L'objectif des travaux de thèse était d'évaluer et d'appliquer des méthodologies pour élucider les structures mosaïques de génomes de plantes cultivées avec le bananier en tant que cas d'étude.

Un programme permettant de simuler des données de génotypage et de comparer des résultats d'inférence locale a été mis en place pour évaluer l'impact de différentes caractéristiques possibles de jeux de données de plantes non modèles sur les performances de méthodes d'ILEA. Trois méthodes d'ILEA publiées ont été comparées via les simulations. Cela a montré que des niveaux élevés de différenciation entre les populations ancestrales ainsi qu'un nombre limité de générations après les évènements d'hybridation permettent une inférence ancestrale plus exacte par les méthodes d'ILEA. Par ailleurs, l'exactitude de l'inférence par ces méthodes a été modérément affectée par un faible nombre de représentants de populations ancestrales, la variation du nombre de populations ancestrales, par l'autofécondation chez une population ancestrale et la multiplication végétative chez les individus hybrides. Ces méthodes peuvent donc être utilisées pour l'inférence ancestrale chez des plantes cultivées si toutes les populations ancestrales sont représentées dans les jeux de données.

Dans un second temps, des données SNP obtenues à partir du reséquençage de 115 accessions de bananiers diploïdes ont été analysées. Ces accessions comprennent des bananiers sauvages de diverses espèces et sous-espèces du genre *Musa* et des bananiers cultivés. Une approche basée sur la détermination de ratios de couverture allélique a été utilisée pour sélectionner des individus non ou peu introgressés, représentants de groupes génétiques ancestraux connus des bananiers cultivés. Cette approche permet de visualiser une origine ancestrale locale à partir de groupes génétiques représentés et a permis aussi la détection de zones du génome des bananiers qu'il n'a pas été possible d'assigner à une origine et qui pourraient être issues de groupes génétiques inconnus. Cela confirme des travaux récemment publiés sur l'existence d'un ou deux contributeurs ancestraux des bananiers cultivés pour lesquels des représentants n'ont pas encore été identifiés. Un sous-jeu de données de 14 accessions cultivées, sans introgressions visibles de contributeurs inconnus, a été analysé par les trois méthodes d'ILEA. Deux d'entre elles ont montré des résultats d'inférence fortement corrélés et des profils de mosaïques similaires à ceux obtenus avec les ratios alléliques. Cela tend à montrer que parmi les méthodes d'ILEA testées, les méthodes basées sur les HMM sont utilisables sur des jeux de données de plantes non modèles si les groupes ancestraux sont caractérisés et ce même avec peu de représentants disponibles. Les 14 accessions étudiées sont majoritairement issues de Nouvelle-Guinée, zone d'origine des pôles *M. a.* ssp. *banksii* et *M. schizocarpa*. Les mosaïques obtenues illustrent une contribution plus répandue que prévu de *M. schizocarpa* aux génomes de bananiers cultivés.

Abstract

Many cultivated plants went through intersubspecific hybridization events that are associated to their domestication and diversification processes. This is for example the case of cultivated bananas that are diploid or triploid hybrids, multiplied through vegetative propagation and deriving from different hybridization events between subspecies and species of the Musa genus, that are spread through different regions and islands in South-East Asia. The genomes resulting from these hybridizations have a mosaic structure of sequences from different origins. This mosaic can be characterized by local ancestry inference (LAI) methods. These methods have been, for the most of them, developed in the framework of human genetic studies, for situations with implicit assumptions that may not always fit plant models. The objective of this thesis was to evaluate and apply LAI methods to elucidate crop plant mosaic genome structures, with a specific focus on banana.

A program allowing simulation of genotyping data and comparison of local ancestry inference results was set up to evaluate the impact of different characteristics that can be found in non model crop plant datasets, on LAI method performances. Three published LAI methods were compared through simulations. The results have shown that elevated differentiation levels between ancestral populations and small numbers of generations after hybridization events allow a more accurate inference. Moreover, inference accuracy was moderately affected by a relatively small number of representatives of ancestral populations, by the variation of the number of ancestral populations, by selfing in one ancestral population or vegetative propagation for admixed individuals. When one ancestral population was not represented in the dataset, the genome regions contributed by this missing population were here variably assigned by the methods to one or the other represented ancestral population. These methods may thus be used for local ancestry inference in cultivated plants but only if all ancestral populations are sampled.

In a second part, SNP data obtained from resequencing of 115 diploid banana accessions were analyzed. These accessions included wild individuals from diverse *Musa* species and subspecies and diploid cultivars. An approach based on the determination of ratios of allele sequence coverage was used to select representatives of banana known genetic groups with no or low levels of introgression. This approach allowed the visualization of local ancestry from ancestral groups represented in the dataset and also allowed the detection of banana genome regions that could not be assigned to a known origin and that may derive from unknown ancestors. This supports recently published work on the existence of one or two ancestral groups contributing to banana cultivars and for which no wild representatives are yet identified. A dataset from 14 cultivated banana accessions without unknown ancestry was analyzed by the three evaluated LAI methods. Two of these methods have shown highly correlated inference results and mosaic profiles very similar to those obtained with the allelic ratios approach for the 14 accessions. This tends to show that among the LAI methods tested, HMM-based methods can be used on non-model plant datasets as long as ancestral groups are characterized, even with few available representatives. The 14 accessions studied mainly originate from New Guinea, the native area of *M. a.* ssp. *banksii* and *M. schizocarpa*. The inferred mosaics illustrate a more widespread contribution than previously shown of *M. schizocarpa* to cultivated banana genomes.

Table des matières

Table des matières **iv**

Table des figures **viii**

Liste des tableaux **xi**

1 Introduction **2**

1.1 Les hybridations interspécifiques et les génomes mosaïques 2

1.2 Les hybridations interspécifiques et intersubspécifiques chez les bananiers 3

1.2.1 Systématique des bananiers 3

1.2.2 Domestication et diversité des bananiers cultivés 7

1.2.3 Etudes moléculaires de la diversité et origines ancestrales des bananiers cultivés 8

1.2.3.1 Etudes à l'échelle globale du génome 8

1.2.3.2 Travaux récents sur les mosaïques des génomes de bananiers . . 12

1.3 Intérêt de la caractérisation des mosaïques 14

1.4 Méthodologies pour caractériser les génomes mosaïques 15

1.4.1 Inférence globale 16

1.4.2 Inférence locale 18

1.5 Le projet de thèse 22

2 Le simulateur de données de génomes mosaïques 'plmgg' **26**

2.1 Contexte 26

2.2 Processus de simulation 28

2.2.1 Génération des populations sources 28

2.2.2 Production des mosaïques 30

2.2.3 Échantillonnage, formatage et statistiques 32
2.3 Validation du simulateur 33
2.3.1 L'hétérozygotie 33
2.3.2 Le déséquilibre de liaison 36
2.4 Illustration du fonctionnement du simulateur 38
2.5 Discussion 43

3 Évaluation de méthodes d'ILEA par simulation 44
3.1 Résumé en français 44
3.2 Abstract 46
3.3 Introduction 47
3.4 Material & methods 49
3.4.1 Simulation tool 49
3.4.2 Simulated scenarios 52
3.4.3 LAI methods 53
3.4.4 Evaluation of the performance of LAI methods 54
3.4.5 Data availability 55
3.5 Results 55
3.5.1 Source differentiation and number of generations since admixture 55
3.5.2 Number of individuals from the source representative populations 56
3.5.3 Number of source populations and absence of source representative individuals 57
3.5.4 Selfing and vegetative propagation 59
3.5.5 Computational performances of LAI methods 61
3.6 Discussion 62
3.6.1 Acknowledgments 65

4 Application de l'inférence ancestrale locale 66

4.1 Contexte 66
4.2 Matériel et méthodes 68
4.2.1 Description du jeu de données 68
4.2.2 Traitement des données de séquençage et appel de variants 68
4.2.3 Exploration préliminaire du jeu de données 69
4.2.3.1 Analyse de la structure par ACP 69
4.2.3.2 Analyse de la structure par inférence globale des états ancestraux 70
4.2.3.3 Calcul de l'hétérozygotie 70
4.2.4 Méthode de « chromosome painting » par ratio de couverture allélique . . 70
4.2.5 Détection d'individus représentatifs des sources et d'individus avec des contributions inconnues par l'approche ARP 72
4.2.6 Inférence locale des états ancestraux 72
4.3 Résultats 74
4.3.1 Structure globale de la diversité 74
4.3.1.1 L'approche par ACP 74
4.3.1.2 Inférence ancestrale globale avec `ADMIXTURE` 76
4.3.1.3 L'hétérozygotie des cultivars et des sauvages 78
4.3.2 Identification des représentants des groupes génétiques sauvages par approche itérative et analyse exploratoire des cultivars par ARP 79
4.3.3 Mosaïques et cohérences des méthodes d'ILEA 82
4.3.3.1 Cohérence entre méthodes d'ILEA 83
4.3.3.2 Mosaïque du sous jeu de données par `ELAI` 86
4.4 Discussion 88
4.4.1 Diversité des bananiers sauvages 88
4.4.2 Diversité et structure des cultivars diploïdes 89
4.4.3 Origine ancestrale locale selon le ratio de couverture allélique 91
4.4.4 ILEA sur le modèle bananier 92

5 Discussion et Perspectives **94**

5.1 Usage des méthodes d'ILEA 95

5.1.1 Différences et points communs entre simulations et données réelles 95

5.1.2 Limitations et perspectives des méthodes d'ILEA 96

5.1.3 Extension du cadre de travail informatique 97

5.2 Perspectives pour l'étude des mosaïques des génomes des bananiers 99

5.2.1 Etude des groupes d'origines inconnues 99

5.2.2 Le cas de la polyploïdie 100

Bibliographie **102**

Figures et tables supplémentaire **121**

Table des figures

1.1 Origines et mosaïques des génomes de quelques espèces d'agrumes hybrides . . . 4

1.2 Arbre phylogénétique de la famille des *Musaceae* 6

1.3 Distribution géographique de *M. balbisiana* et des sous-espèces de *M. acuminata* en Asie du Sud-Est 7

1.4 Caractéristiques principales de la première version de l'assemblage du génome de référence du bananier 11

1.5 Visualisation de la mosaïque AB du chromosome 9 de 5 accessions par ratio de couverture allélique par la méthode de Baurens et al. (2019) 13

1.6 Visualisation de la mosaïque de l'accession AAcv 'Manang' par la méthode de Martin et al. (2020) 14

2.1 Évolution de l'hétérozygotie pendant la phase « forward » du simulateur. 35

2.2 Évolution du déséquilibre de liaison (mesuré par le r^2 moyen par classe de distance physique) en fonction du taux de recombinaison et du nombre de générations 38

2.3 Statistiques générées lors d'une simulation 40

2.4 Représentation des mosaïques d'origine ancestrale de la simulation de démonstration 42

3.1 Overview of the admixture simulation process with `plmgg` 50

3.2 Accuracy of LAI methods with varying levels of differentiation and number of generations (**DiffGenSam** simulation) 55

3.3 Accuracy of LAI methods with varying levels of differentiation and source-representative sample size (**DiffGenSam** simulation) 56

3.4 LAI results for a simulated admixed individual of the **SrcMiss** simulation . . . 58

3.5 Accuracy of LAI methods with varying number of generations of vegetative propagation (**AdmxVegProp** simulation) 60

3.6 Memory usage and computation time of LAI methods with varying number of sources (**SrcNum** simulation) 61

4.1 ACP sur l'ensemble des 115 accessions de bananiers (33 783 marqueurs SNP) . . 74

4.2 ACP sur un groupe de 97 individus comprenant *M. laterita*, *M. rosea*, *M. schizocarpa*, *M. acuminata* et des individus hybrides (33 783 marqueurs SNP) 75

4.3 Représentation des valeurs de validation croisée d'`ADMIXTURE` sur 5 sous-jeux de données, avec 10 répétitions et une valeur de K allant de 2 à 20. 76

4.4 Partitionnement de 112 individus du jeu de données analysés par `ADMIXTURE` pour des valeurs de K de 7, 9, 11 et 13 77

4.5 Exemples de painting allélique d'un individu représentatif d'un groupe génétique sauvage, d'un cultivar dont les origines sont identifiées et d'un cultivar avec origine manquante à la troisième itération. 81

4.6 Résultat de l'ARP pour les accessions 'DYN114_Heva' et 'DYN112_Guyod' à l'étape trois de l'analyse itérative 82

4.7 Comparaison des inférences par paires de méthodes (`ELAI`, `SABER` et `WINPOP`) en fonction du paramètre de temps depuis l'hybridation et des sous-jeux de données utilisés 84

4.8 Résultats de l'ILEA et « painting » allélique de l'accession 'Guyod' 85

4.9 Résultats de l'inférence locale d'`ELAI` sur les 14 individus hybrides 87

S.1 Accuracy of LAI methods with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation) 122

S.2 Accuracy of LAI methods with varying sample size of the third source-representatives (**SamBal** simulation) 123

S.3 Memory usage of LAI methods with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation) . . 124

S.4 Computation time of LAI methods with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation) . 125

S.5 Time since admixture estimation ($\widehat{t_{adm}}$) of `SABER` with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation) 126

S.6 Résultat de l'ARP pour les accessions 'DYN319-Pisang_Segun' et 'DYN-ITC0897-Banksii_ITCO897' à l'étape trois de l'analyse itérative 127

S.7 Résultat de l'ARP pour les accessions 'TC1701-Musa_acuminata_ssp_sumatrana' et 'DYN398-Truncata' à l'étape deux de l'analyse itérative 128

S.8 Résultat de l'ARP pour les accessions 'DYN-ITC1028-Agutay' et 'DYN040-Borneo' à l'étape deux de l'analyse itérative 129

S.9 Résultats de l'inférence locale de `SABER` sur les 14 individus hybrides 130

S.10 Résultats de l'inférence locale de `WINPOP` sur les 14 individus hybrides 131

S.11 Stabilité des ILEA entre les différents sous-jeux de données dans des comparaisons par paires. 132

S.12 Stabilité des ILEA par rapport au paramètre du temps depuis l'hybridation dans des comparaisons par paires 133

S.13 Résultats de l'ILEA et de l'ARP de l'accession 'Gulum' 134

S.14 Résultats de l'ILEA et de l'ARP de l'accession 'Tomolo' 135

S.15 Résultats de l'ILEA et de l'ARP de l'accession 'Papat' 136

S.16 Résultats de l'ILEA et de l'ARP de l'accession 'Sena' 137

S.17 Résultats de l'ILEA et de l'ARP de l'accession 'Gorop' 138

S.18 Résultats de l'ILEA et de l'ARP de l'accession 'Manameg Red' 139

S.19 Résultats de l'ILEA et de l'ARP de l'accession 'Spiral' 140

S.20 Résultats de l'ILEA et de l'ARP de l'accession 'Sihir' 141

S.21 Résultats de l'ILEA et de l'ARP de l'accession 'Aivip' 142

S.22 Résultats de l'ILEA et de l'ARP de l'accession 'Katual n2' 143

S.23 Résultats de l'ILEA et de l'ARP de l'accession 'Vudo Beo' 144

S.24 Résultats de l'ILEA et de l'ARP de l'accession 'Wompa' 145

S.25 Résultats de l'ILEA et de l'ARP de l'accession 'Tonton Kepa' 146

Liste des tableaux

1.1 Synthèse des principales études par marqueurs moléculaires publiées sur la diversité génétique des bananiers 9

1.2 Caractéristiques des principales méthodes ILEA citées 21

2.1 Notations utilisées pour les paramètres du simulateur `plmgg` 29

2.2 Exemple de matrice de contribution avec $S = 3$ sources et $P = 5$ populations en « forward » 31

3.1 Summary of simulation parameters. 52

3.2 Accuracy of LAI methods with varying number of source populations (SrcNum scenario). 57

3.3 Accuracy of LAI methods with varying proportions of an unknown source population in admixed populations (SrcMiss scenario). 59

3.4 Accuracy of LAI methods with varying proportions of selfing in a source-representative population (SrcSelf simulation). 59

4.1 Valeurs de F_{ST} mesurées entre les groupes du sous-jeu de données. 83

S.1 Origines, status et identifiants des 115 accessions de bananiers utilisées 147

S.2 Pourcentage de site hétérozygote par accession. 150

S.3 Nombre d'allèles associés aux groupes pour les 3 itérations de l'approche « ARP » 151

S.4 Résumé de la séléction d'individu représentatif des pôles par ACP, `Admixture` et « ARP ». 152

Chapitre 1

Introduction

1.1 Les hybridations interspécifiques et les génomes mosaïques

Des processus d'hybridations entre espèces et sous espèces sont retrouvés tout au long de l'histoire de la domestication des plantes cultivées. Ils ont contribué à l'émergence de nouveaux phénotypes, à la diversification et à l'expansion des plantes cultivées (Arnold, 2004; Miller and Gross, 2011; Purugganan, 2019). Ces hybridations peuvent avoir lieu de plusieurs façons : entre populations sauvages, entre populations cultivées et sauvages, ainsi qu'entre populations domestiquées transportées (migrations humaines) ou échangées (échanges commerciaux) et populations locales. Enfin, les hybridations sont aussi utilisées dans l'agriculture moderne pour produire de nouvelles variétés.

La représentation des origines ancestrales des hybrides le long de leur génome peut former une figure ressemblant à une mosaïque. Cette mosaïque des origines ancestrales est produite par les recombinaisons génétiques [1] entre des chromosomes de différentes origines et montre des blocs de séquences d'origines différentes le long des chromosomes.

De nombreux exemples d'hybridations intersubspécifiques chez les plantes cultivées sont décrits dans la littérature. On en trouve par exemple chez le blé (*Triticum sp.*) et le riz (*Oryza sativa* L.), qui sont parmi les principales céréales cultivées dans le monde. Le blé dur (*T. durum*, 2n=4x=28) et le blé tendre (*T. aestivum*, 2n=6x=42) sont issus d'hybridations entre des blés sauvages accompagnées de polyploidisation et l'un des géniteurs sauvages diploïdes serait aussi un hybride (El Baidouri et al., 2017; Glémin et al., 2019).

Le riz cultivé est une plante autogame diploïde. Sa diversité est principalement composée de 4 groupes génétiques (certains notés sous-espèces) : Indica, Japonica, Aus et Basmati (Zhao et al., 2010; Civáň et al., 2019; Santos et al., 2019). La plupart des riz cultivés portent des traces

1. Échange de matériel génétique par enjambement (« crossing-over ») entre chromosomes homologues pendant la méiose

d'introgressions des différents groupes. Les riz aromatiques (Basmati) seraient issus d'hybridation intersubspécifique entre Japonica et Aus il y a 4000 à 2400 ans, et probablement aussi avec Indica (Civáň et al., 2019; Santos et al., 2019).

Les agrumes (*Citrus sp.*) sont un autre exemple frappant de l'importance de l'hybridation chez les plantes cultivées. Les agrumes cultivés (majoritairement des espèces diploïdes, mais aussi quelques espèces triploïdes et tétraploïdes) sont issues d'hybridations entre quatre principaux taxons ancestraux : *C. reticulata*, *C. maxima*, *C. medica* et *C. micrantha* (Curk et al., 2015, 2016; Wu et al., 2018; Ahmed et al., 2019). On retrouve chez les agrumes cultivés des espèces issues de croisements simples (une génération) comme le bigaradier *C. aurantium* et des espèces issues de plusieurs croisements comme le citronnier *C. limon* et le bergamotier *C. bergamia* (Figure 1.1).

Les hybridations interspécifiques et intersubspécifiques présentées recouvrent différents cas de structures de mosaïques ancestrales. Chez le blé, il existe des gènes permettant de contrôler l'appariement des chromosomes lors de la méiose (locus *Ph1* par exemple), limitant les recombinaisons entre sous-génomes (Dubcovsky et al., 1995). Dans le cas des agrumes cultivées, des niveaux de morcellement des mosaïques relativement faibles à nul sont identifiés (Curk et al., 2016). Certaines espèces hybrides ayant sans doute été stabilisées car propagées végétativement après un seul croisement, les recombinaisons génétiques n'ont pas pu faire évoluer la structure hybride initialement constituée. Quelques espèces d'agrumes sont néanmoins issues de plusieurs évènements d'hybridations, comme le bergamotier *C. bergamia* et le citronnier *C. limon* montrés en figure 1.1 et montrent une mosaïque plus morcelée. Dans le cas du riz, des hybridations ont lieu à un niveau intersubspécifique, dans un contexte diploïde et de multiplication sexuée (majoritairement par autofécondation). Les mosaïques s'étant formées après les hybridations peuvent être très fragmentées (Santos et al., 2019).

Les bananiers cultivés sont également issus d'hybridations interspécifiques et intersubspécifiques (Carreel et al., 1994; Perrier et al., 2011). Ils sont le cas d'étude dans le cadre de ces travaux de thèse et les connaissances sur leurs origines sont détaillées dans la partie suivante.

1.2 Les hybridations interspécifiques et intersubspécifiques chez les bananiers

1.2.1 Systématique des bananiers

Les bananiers sont des herbes géantes, monocotylédones de l'ordre des Zingibérales, de la famille des *Musaceae*. La famille des *Musaceae* est composée de trois genres ; *Musa*, *Musella* et

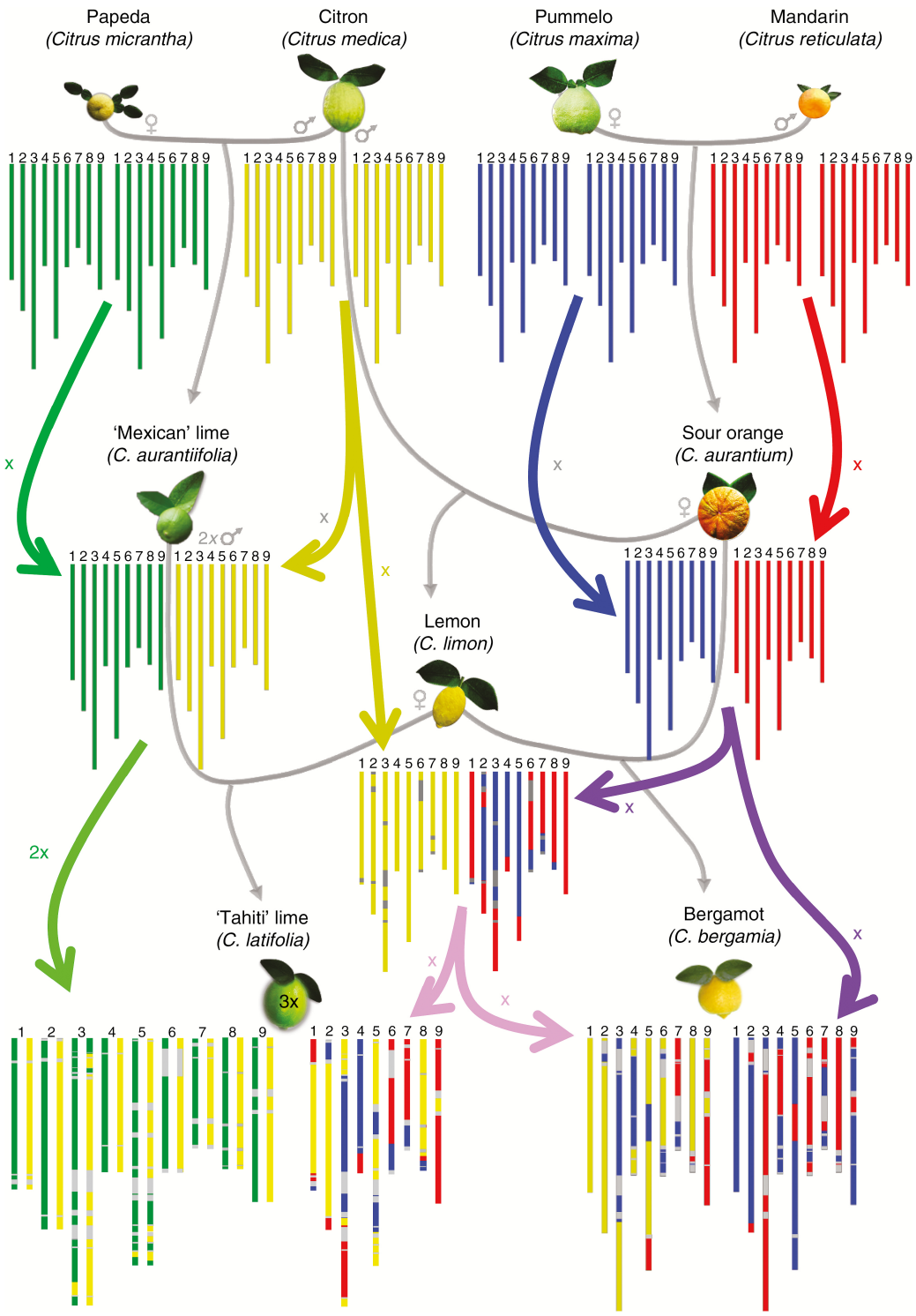

FIGURE 1.1 – Origines et mosaïques des génomes de quelques espèces d'agrumes hybrides. Les quatre groupes ancestraux sont représentés sur la partie supérieure de la figure. Les hybrides sont représentés sous les groupes ancestraux. Les flèches grises indiquent les croisements, et les flèches colorées indiquent le niveau de ploïdie du gamète (x ou 2x). Source : Figure 8 de Ahmed et al. (2019), Annals of Botany.

Ensete originaires de l'Indo-Burma, correspondant à la partie continentale de l'Asie du Sud-Est (Janssens et al., 2016) (Figure 1.2). Le genre *Ensete* s'étend de l'Afrique tropicale (notamment en Éthiopie) à l'Asie (Väre and Häkkinen, 2011) et comprend entre autre l'espèce *Ensete ventricosum* qui est cultivée pour son bulbe comestible. Le genre *Musella* comprend une seule

espèce cultivée à des fins ornementales (Ma et al., 2011). C'est le genre *Musa* qui comprend la majorité des espèces de la famille des *Musaceae* (environ 70 espèces, Janssens et al., 2016) ainsi que les bananiers cultivés pour la banane dessert, mais aussi pour la banane à cuire ou à bière et la production de fibres végétales. Ce genre a été subdivisé en quatre à cinq sections sur base de critères morphologiques et de nombres de base chromosomiques différents : *Australimusa* (x=10), *Callimusa* (x=10/9), *Rhodochlamys* (x=11) et *Eumusa* (x=11) et quelquefois *Ingentimusa* (une seule espèce *M. ingens*, x=7). Suite à différents travaux d'analyses moléculaires, cette subdivision a été révisée en une structuration en deux sections : *Musa* (x=11) et *Callimusa* (x=10/9/7) (Häkkinen, 2013). Les travaux de Janssens et al. (2016) proposent un scénario de dispersion des *Musaceae* de l'Asie du Sud-Est jusqu'à la Nouvelle-Guinée et l'Australie du Nord-Ouest, en lien avec les évènements géologiques et climatiques de la région (du Miocène au Pliocène). Ils soutiennent également une subdivision du genre *Musa* en deux clades (notés I et II sur la figure 1.2). Le clade I qui regroupe les anciennes sections *Australimusa*, *Callimusa* et *Ingentimusa* se serait diversifié depuis (environ) 26,3 millions d'années. Le clade II regroupe *Eumusa* et *Rhodochlamys* avec une radiation qui aurait débutée il y a environ 20,9 millions d'années (Janssens et al., 2016). La classification en quatre sections principales est cependant encore utilisée pour des raisons pratiques (Christelová et al., 2017).

Ce sont les bananiers de l'ex-section *Eumusa* qui sont à l'origine de la très grande majorité des bananiers cultivés, à l'exception des bananiers Fe'i (*Australimusa*). *Eumusa* contient les espèces *M. acuminata* (génome A, 2n=2x=22) et *M. balbisiana* (génome B, 2n=2x= 22) qui sont à l'origine de la majorité des cultivars ainsi que *M. schizocarpa* (génome S, 2n=2x= 22) qui serait présente dans quelques cultivars. L'espèce *M. balbisiana* est retrouvée sur la partie continentale de l'Asie du Sud-Est, en Inde, en Chine, et aux Philippines. L'espèce *M. schizocarpa* est, elle, retrouvée sur l'île de Nouvelle-Guinée. Ces deux espèces seraient relativement peu diversifiées, surtout en comparaison de *M. acuminata*, la principale composante des bananiers cultivés.

L'espèce *M. acuminata* est découpée selon les auteurs en 8 à 10 sous-espèces plus ou moins inter-fertiles sur base de critères géographiques et morphologiques (Simmonds, 1962; Perrier et al., 2009). Ces sous-espèces sont : *M. a.* ssp. *banksii*, *M. a.* ssp. *burmannica*, *M. a.* ssp. *burmannicoides*, *M. a.* ssp. *errans*, *M. a.* ssp. *malaccensis*, *M. a.* ssp. *microcarpa*, *M. a.* ssp. *siamea*, *M. a.* ssp. *truncata*, *M. a.* ssp. *zebrina*, et selon les sources également *M.a.var.* ou *ssp. sumatrana*. Elles sont réparties de la partie continentale de l'Asie du Sud-Est à l'île de Nouvelle-Guinée (figure 1.3). La plupart des espèces de bananiers sont monoïques, c'est-à-dire qu'elles possèdent des fleurs femelles et des fleurs mâles. Les fleurs femelles apparaissant avant les fleurs mâles, des pollinisations entre plantes différentes sont possibles même si des (auto)-pollinisations par des rejets d'une même souche ne sont pas exclues (Simmonds, 1962). La sous-espèce *M. a.* ssp. *banksii*, ainsi que *M. schizocarpa* se distinguent par la présence de fleurs basales hermaphrodites (Simmonds, 1962). Elles sont plutôt autogames, ce qui provoque

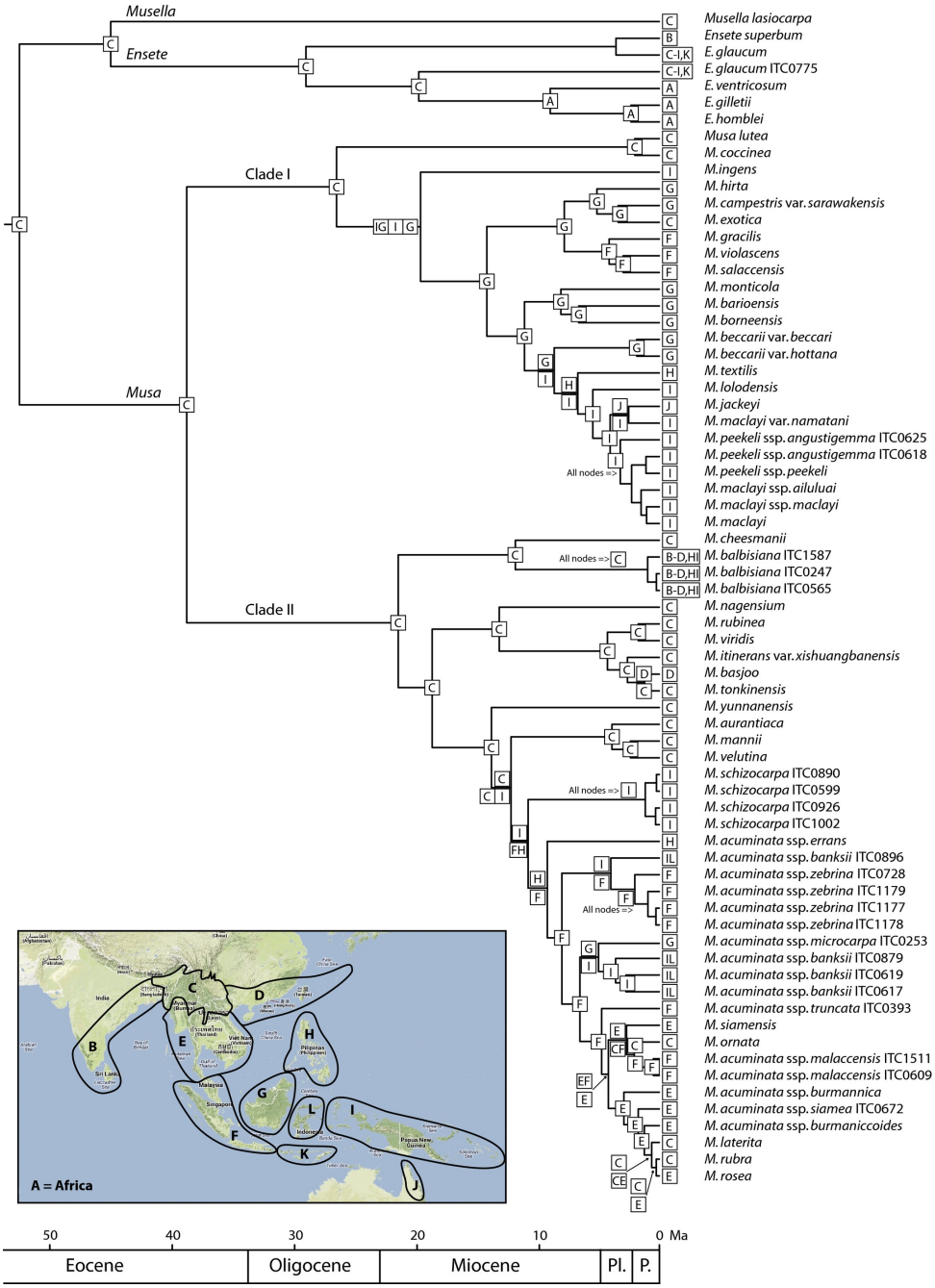

FIGURE 1.2 – Arbre phylogénétique de la famille des *Musaceae*. Les zones géographiques correspondant aux régions ancestrales avec le plus fort support et indiquées sur la carte et l'arbre sont : (A) Afrique, (B) Sud-ouest de l'Inde et Sri Lanka, (C) Nord de l'Indo-Burma, (D) Sud de la Chine, (E) Sud de l'Indo-Burma, (F) Sumatra et péninsule malaisienne, (G) Bornéo, (H) Philippines, (I) Nouvelle-Guinée, (J) Nord-ouest de l'Australie, (K) petites îles de la Sonde, (L) Célèbes. Sur la frise chronologique, le Pliocène et le Pleistocène sont indiqués respectivement par « Pl. » et « P. ». Source : Figure 1 de Janssens et al. (2016), New Phytologist

une structuration génétique particulière en comparaison des autres espèces ou sous-espèces (niveau d'hétérozygotie plus faible que chez les sous-espèces allogames). De plus, des variations structurales chromosomiques de grande taille, en particulier des translocations réciproques ont été détectées entre les génomes de certaines des sous-espèces de *M. acuminata* (Shepherd, 1999; Martin et al., 2017; Dupouy et al., 2019).

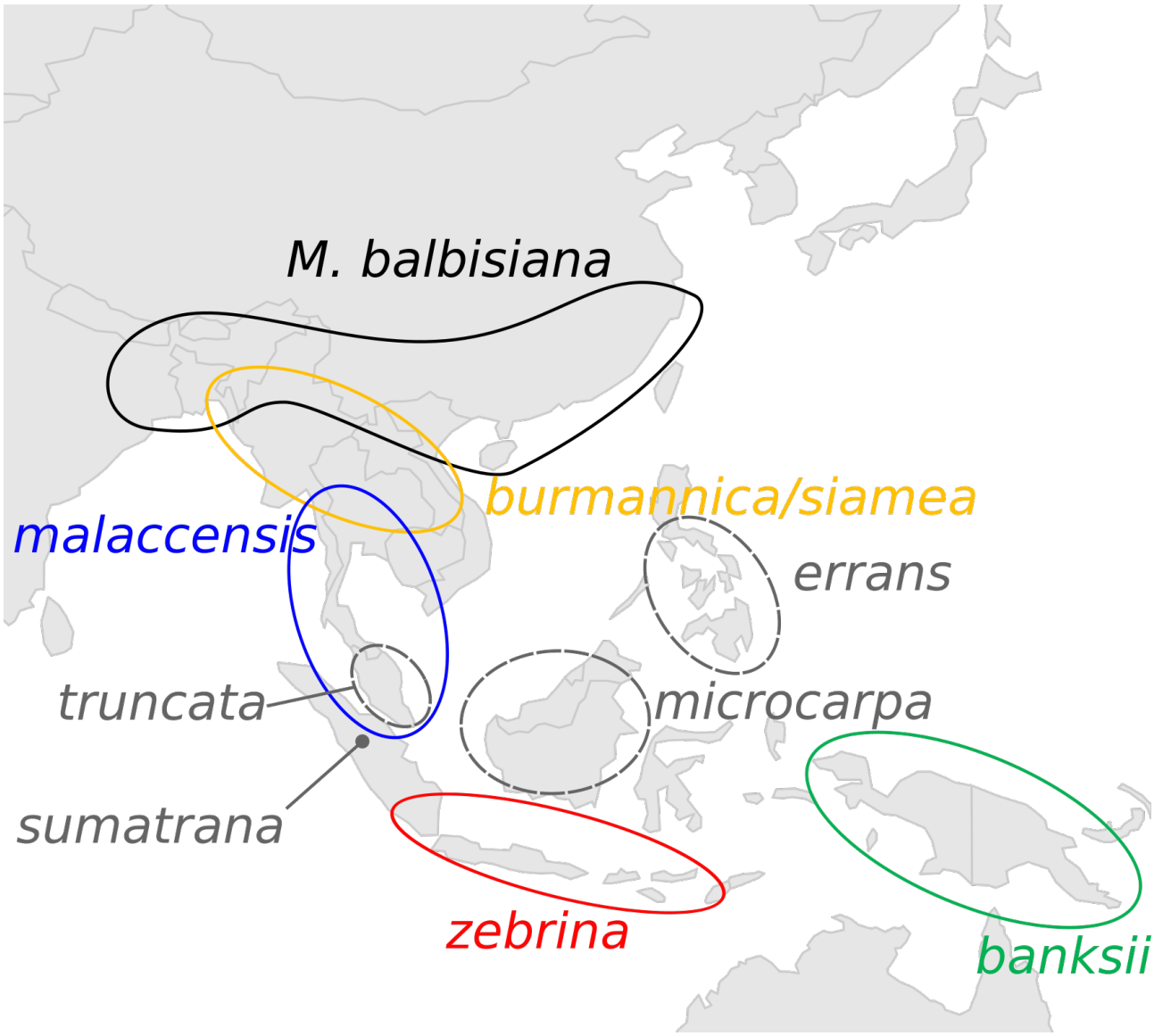

Figure 1.3 – Distribution géographique de *M. balbisiana* et des sous-espèces de *M. acuminata* en Asie du Sud-Est. Les espèces et sous-espèces représentées en couleurs sont des contributrices identifiées chez les bananiers cultivés. Source : adapté de la Figure 1 de Perrier et al. (2011), PNAS.

1.2.2 Domestication et diversité des bananiers cultivés

L'origine et le processus de domestication des bananiers cultivés ont fait l'objet d'études croisant botanique, biologie moléculaire, archéologie et linguistique (Simmonds and Shepherd, 1955; Simmonds, 1962; De Langhe, 2009; De Langhe et al., 2009; Perrier et al., 2009). Ces travaux ont été synthétisés et un scénario de domestication a été proposé (Perrier et al., 2011). Dans ce scénario, des hybridations se seraient produites entre des plantes fertiles de différentes sous-espèces de *M. acuminata* grâce aux migrations des populations humaines de l'Asie du Sud-Est durant l'Holocène ($\leq$ 11 700 ans). Celles-ci auraient transporté avec elles des bananiers (protocultivars) hors de leur milieu d'origine vers les lieux de migration. Ces mouvements ont donc remis en contact des sous-espèces séparées géographiquement (et par des variations

structurales chromosomiques), ce qui a permis la formation d'hybrides diploïdes. Ces individus hybrides ont pu ensuite se croiser de nouveau entre eux ou avec les populations ou espèces sauvages locales. Des anomalies lors de la production de gamètes ont permis la formation de gamètes diploïdes et l'apparition d'individus triploïdes. Des hybrides diploïdes et triploïdes stériles ou à fertilité réduite ont été sélectionnés par les populations humaines pour leurs fruits sans graines, à développement parthénocarpique (développement du fruit sans fécondation). Différents hybrides ont donc été sélectionnés et ont été propagés par multiplication végétative à l'aide de rejets, sur de grandes échelles de temps. Certains hybrides, principalement ceux issus des génomes d'origine *M. acuminata* (noté A) et *M. balbisiana* (noté B), ont été propagés de l'Asie à l'Afrique et au Pacifique puis aux Amériques. Ils ont été classifiés en différents groupes génomiques : AA, AAA, AB, AAB, ABB (Simmonds and Shepherd, 1955). Plus marginalement, des hybrides AT, AAT (T dénotant l'origine *M. textilis*) et AS (S dénotant l'origine *M. schizocarpa*) sont retrouvés dans la diversité des bananiers cultivés. Au sein des groupes génomiques, la multiplication végétative des hybrides a produit des variants phénotypiques dits « somaclonaux », les variants issus d'un même évènement sexué définissant un sous-groupe. Actuellement, les bananiers cultivés à plus grande échelle sont triploïdes comme les bananiers dessert du sous-groupe Cavendish (groupe AAA) qui produisent près de la moitié de la production mondiale de bananes (Lescot, 2018), ou les bananiers à cuire du sous-groupe Plantain (groupe AAB). Les cultivars (cv) diploïdes sont moins répandus. On trouve une grande diversité de cultivars au sein du groupe AA en Papouasie-Nouvelle-Guinée et en Asie du Sud-Est. Au sein du groupe AA, seuls les sous-groupes clonaux 'Mchare' et 'Figue' sont trouvés en dehors des zones d'origines (Perrier et al., 2009, 2019).

1.2.3 Etudes moléculaires de la diversité et origines ancestrales des bananiers cultivés

1.2.3.1 Etudes à l'échelle globale du génome

La structuration de la diversité génétique des bananiers sauvages et cultivés a été l'objet de plusieurs études moléculaires (Table 1.1), portant à la fois sur le matériel génétique nucléaire et cytoplasmique (Carreel et al., 1994, 2002; Raboin et al., 2005; Boonruangrod et al., 2009; Perrier et al., 2009, 2011; Hippolyte et al., 2012; Sardos et al., 2016a,b; Christelová et al., 2017; Baurens et al., 2019; Martin et al., 2020). Une étude du génome nucléaire avec des marqueurs RFLP [2] (Carreel et al., 1994) , couvrait 70 individus séminifères (sauvages) de *M. acuminata*,

2. Polymorphismes de longueurs de fragments de restrictions (« Restriction Fragment Length Polymorphism, RFLP »). Ils sont détectés à partir de la digestion par des enzymes de restrictions de l'ADN des individus analysés, puis de leur séparation par électrophorèse et enfin de la visualisation de fragments spécifiques par l'hybridation de sondes. Des fragments d'ADN de différentes tailles peuvent alors être observés entre les différents individus.

M. balbisiana et *M. schizocarpa* ainsi que 90 individus parthénocarpiques diploïdes (cultivars). Des rapprochements ont été proposés entre les sous-espèces *M. a.* ssp. *burmannica* et *M. a.* ssp. *burmannicoides* et également entre *M. a.* ssp. *banksii* et *M. a.* ssp. *errans*, *M. a.* ssp. *malaccensis* et *M. a.* ssp. *truncata*. Une structuration importante permettait de différencier clairement *M. a.* ssp. *banksii*, *M. a.* ssp. *malaccensis*, *M. a.* ssp. *zebrina* et *M. a.* ssp. *burmannica*. Les hybrides AB ont été validés et la validation moléculaire d'hybrides AS a aussi permis d'étayer la participation de *M. schizocarpa* à la diversité des bananiers cultivés. Enfin, il a été montré que la majorité des cultivars de Papouasie-Nouvelle-Guinée étaient proches de *M. a.* ssp. *banksii*, les autres cultivars étant sans doute hybrides de différentes sous-espèces.

Table 1.1 – Synthèse des principales études par marqueurs moléculaires publiées sur la diversité génétique des bananiers

Nombre d'accessions	Type de marqueurs	Nombre de marqueurs	Analyse génétique	Étude
160	sonde RFLP	30	Analyse factorielle sur tableau de distance (AFTD)	Carreel et al. (1994)
305	sonde RFLP (mito. & chloro.)	28 + 14	AFTD	Carreel et al. (2002)
176	sonde RFLP	36	Comptage de marqueurs	Raboin et al. (2005)
172	SSR	22	AFC , arbre phylogénétique (NJ)	Perrier et al. (2009)
52	SNP – ciblé ARNr 18S	29	Analyse en composantes principales (ACP)	Boonruangrod et al. (2009)
561	SSR	22	Phylogénie (NJ)	Hippolyte et al. (2012)
575	DArT	498	Clustering	Sardos et al. (2016a)
105	DArt, SNP	498, 5 544	Clustering, Phylogénie (NJ)	Sardos et al. (2016b)
695	SSR	19	Phylogénie (UPGMA)	Christelová et al. (2017)
21	SNP (RADSeq)	19 585	Ratio de couverture des marqueurs diagnostics	Baurens et al. (2019)
24	SNP (RNAseq)	197 876	AFC + Clustering	Martin et al. (2020)

Des marqueurs RFLP chloroplastiques et mitochondriaux (Carreel et al., 2002) ont aussi été utilisés pour élucider les relations d'apparentements entre bananiers sauvages et cultivés et pour caractériser les lignées maternelles et paternelles de bananiers cultivés ; les chloroplastes étant à transmission maternelle et les mitochondries à transmission paternelle chez le bananier (Fauré et al., 1994). Cela a permis de mettre en évidence les ascendants diploïdes de certaines accessions triploïdes, ainsi qu'une prévalence de l'origine ancestrale *M. a.* ssp. *banksii* chez les cultivars et la détection d'un profil cytoplasmique de *M. a* .ssp. *errans* dans de nombreux cultivars. Les études par Raboin et al. (2005) et Hippolyte et al. (2012) ont ensuite permis de préciser les prédictions sur les ascendants diploïdes des principaux sous-groupes triploïdes ('Cavendish' notamment).

Une synthèse des études moléculaires a été proposée par Perrier et al. (2009, 2011), qui incluent de plus une analyse phylogénétique (« neighbor joining ») de la diversité à l'aide de marqueurs SSR[3]. Les sous-espèces *M. a.* ssp. *burmannica*, *M. a.* ssp. *burmannicoides* et *M. a.* ssp. *siamea* formeraient un seul complexe génétique *M. a.* ssp. *burmannica/siamea*, qui

3. Les microsatellites (« Single Sequence Repeats », SSR) sont des séquences de petites tailles répétées plusieurs fois d'affilée, retrouvées en nombre dans les génomes eucaryotes. Le nombre de copies des SSR varie, et ce polymorphisme permet de les utiliser comme marqueurs génétiques.

n'aurait pas eu une contribution importante aux bananiers cultivés. Les sous-espèces *M. a.* ssp. *banksii*, *M. a.* ssp. *malaccensis* et *M. a.* ssp. *zebrina*, seraient les principales sous-espèces de *M. acuminata* retrouvées chez les cultivars. Les sous-espèces *M. a.* ssp. *errans*, *M. a.* ssp. *microcarpa* et *M. a.* ssp. *truncata* ne présentent pas une structuration aussi claire que les 4 autres sous-espèces mais peu de représentants sont en collection. Les cultivars diploïdes AAcv forment des clusters répartis sur un continuum entre *M. a.* ssp. *banksii* et les trois autres sous-espèces, *M. a.* ssp. *malaccensis*, *M. a.* ssp. *burmannica/siamea* et *M. a.* ssp. *zebrina*. Les cultivars les plus proches de *M. a.* ssp. *banksii* sont originaires de Nouvelle-Guinée et sont les moins hétérozygotes. Les clusters plus éloignés ont une hétérozygotie plus forte à mesure que les contributions de *M. a.* ssp. *malaccensis* et *M. a.* ssp. *zebrina* augmentent. Des scénarios expliquant l'origine et la diffusion des bananiers triploïdes dans la partie insulaire de l'Asie du Sud-Est, vers l'Inde, l'Afrique et le Pacifique ont aussi été proposés par ces auteurs. La structuration en 4 groupes génétiques des sous-espèces de *M. acuminata* a été renforcée, indépendamment, dans une étude portant sur la séquence 5'ETS [4] des ARN ribosomiques (Boonruangrod et al., 2009).

Plus récemment Sardos et al. (2016a) ont analysé un jeu de données de 498 marqueurs DArT [5] pour 575 accessions (dont 94 sauvages diploïdes et 208 cultivars diploïdes). La caractérisation de la structuration génétique de ces accessions a été réalisée avec le logiciel d'inférence globale des états ancestraux (IGEA) `STRUCTURE` (Pritchard et al., 2000), détaillé dans la partie 1.4 et également avec une analyse en composantes principales (ACP) sur une matrice de distances. Les résultats ont montré deux « pools » génétiques pour les cultivars diploïdes AAcv, l'un de Nouvelle-Guinée dérivé de *M. a.* ssp. *banksii* et un second d'Asie du Sud-Est avec d'autres contributions, suggérant deux foyers et deux processus différents de domestication. Deux autres études (Sardos et al., 2016b; Christelová et al., 2017) identifient également un cluster de cultivars AAcv de Papouasie-Nouvelle-Guinée qui serait directement dérivé de *M. a.* ssp. *banksii*. Par ailleurs, Sardos et al. (2016a) suggèrent également sur la base de leur analyse ACP, que la diversité sauvage *M. acuminata* contribuant aux cultivars n'est pas totalement explorée.

4. Séquence de l'ARNr appelée espaceur externe transcrit (External Transcribed Spacer), utilisée comme marqueur phylogénétique

5. Diversity Arrays Technology : Technologie de génotypage basée sur les puces à ADN, permettant de scanner des marqueurs pré-déterminés.

Encadré 1 : Ressources génomiques et spécificités pour l'étude des mosaïques du bananier

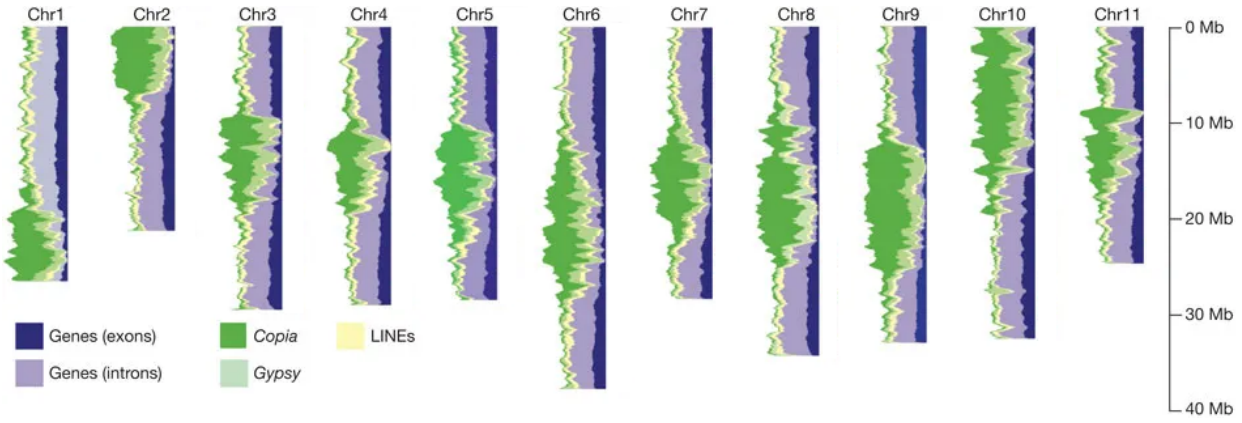

Figure 1.4 – Caractéristiques principales de la première version de l'assemblage du génome de référence du bananier. Distributions des gènes (introns, exons) et éléments transposables (Copia, Gypsy, LINEs) le long des 11 chromosomes du génome de référence. Les centromères sont appauvris en gènes et enrichis en éléments transposables. Les chromosomes 1, 2 et 10 sont potentiellement acrocentriques (centromère sur une extrémité). Source : adapté de la Figure 1 de D'Hont et al. (2012), Nature.

L'assemblage du génome de référence du bananier (Figure 1.4) a été publié en 2012 (D'Hont et al., 2012). L'accession séquencée ('DH-Pahang') est un haploïde doublé, dérivé de l'accession sauvage 'Pahang' de la sous-espèce *M. a.* ssp. *malaccensis*. L'assemblage était constitué de 24 425 « contigs » assemblés en 7 513 « scaffolds » avec une valeur de N50 de 1.3Mb (la moitié de l'assemblage est contenus dans des « scaffolds » de taille supérieure à 1.3 Mb). La taille de l'assemblage était de 472Mb (contre 523Mb de taille de génome estimée par cytométrie en flux), et 36 542 gènes ont été prédits. Une carte génétique a permis l'ancrage de 70 % de cet assemblage aux 11 chromosomes de bananier. Ce génome de référence de *M. acuminata* a été ensuite amélioré (Martin et al., 2016) via une carte optique BioNano et du séquençage en lectures pairées (Illumina) de grands inserts. Le N50 a été plus que doublé (de 1,3Mb à 3Mb), avec un nombre de « scaffolds » réduit à 1 532. La proportion d'assemblage ancrée sur des chromosomes était de 89,5 % pour une taille totale de 451Mb.

D'autres génomes de référence ont été depuis assemblés, correspondant aux espèces *M. balbisiana* et *M. schizocarpa* et aux sous-espèces *M. a.* ssp. *burmannica/siamea*, *M. a.* ssp. *zebrina* et *M. a.* ssp. *banksii* (Belser et al., 2018; Rouard et al., 2018; Wang et al., 2019).

Les ressources génétiques issues des prospections des bananiers sont réparties dans différentes collections dans le monde. La base de données MGIS (Musa germplasm information system, Ruas et al., 2017) répertorie 6 619 accessions de 30 collections de bananiers. La collection de référence est celle de l'International Transit Centre (ITC) (Christelová et al., 2017) avec plus de 1500 accessions qui sont maintenues *in vitro*. La collection CIRAD (365 bananiers en plein champs) est hébergée au Centre de Ressources Génétiques (CRB) plantes tropicales Antilles (`http://crb-tropicaux.com/Portail`). Pourtant, le nombre d'accessions sauvages correspondant aux sous-espèces de bananiers est faible. À titre d'exemple, il y a dans la collection de l'ITC 61 accessions (sur 412 AA) de bananiers sauvages identifiées comme membres de sous-espèces de *M. acuminata*, dont 30 *M. a.* ssp. *banksii*, 16 *M. a.* ssp. *malaccensis*, 7 *M. a.* ssp. *burmannica/siamea* et 6 *M. a.* ssp. *zebrina*.

1.2.3.2 Travaux récents sur les mosaïques des génomes de bananiers

Au-delà des études à l'échelle globale du génome, la disponibilité d'un génome de référence pour le bananier (Encadré 1) et les possibilités de génotypage par séquençage ouvrent de nouvelles perspectives pour caractériser les origines ancestrales le long des chromosomes des bananiers. Des ressources génétiques sont aussi disponibles même si la diversité accessible des bananiers sauvages reste relativement limitée (Encadré 1). Les cultivars de bananiers ont été sélectionnés puis reproduits par multiplication végétative. Le nombre de recombinaisons est sans doute limité et on s'attend à des structures mosaïques assez peu fragmentées comparativement à des plantes à multiplication sexuée.

Des travaux très récents, réalisés au sein de l'équipe d'accueil, ont visé à caractériser ces mosaïques avec des données de polymorphisme nucléotidique SNP [6] issues de données de séquençage et de génotypage par séquençage. L'étude de Baurens et al. (2019) porte sur des hybrides interspécifiques diploïdes et triploïdes entre *M. acuminata* et *M. balbisiana* (AB, AAB, ABB) et vise à comprendre l'impact des structures génétiques A/B sur la recombinaison et la ségrégation des chromosomes. Une approche par marqueurs diagnostics a été utilisée pour caractériser la mosaïque d'origine ancestrale. Ici, 9 accessions *M. acuminata* et 3 accessions *M. balbisiana* ont été utilisées pour générer la liste des marqueurs diagnostics, sur base de leur présence chez tous les individus d'un groupe et leur absence du second groupe (avec un jeu de données de 148 329 SNPs). Un ratio de couverture allélique a été ensuite calculé, correspondant pour chaque allèle au nombre de lectures de séquençage le supportant sur le nombre total de lectures couvrant la position. Pour un diploïde, le ratio de couverture allélique égal à 1 est attendu pour une origine AA ou BB, et égal à 0.5 pour une origine A/B. Dans le cas de triploïdes, les ratios attendus sont soit égaux à 1 si une seule origine est présente soit égaux à 0.33 ou 0.66 selon qu'il y ait une ou deux doses de chaque origine (par exemple 1/3 de A et 2/3 de B pour une région ABB). La figure 1.5 illustre la visualisation obtenue par cette méthode. Des recombinaisons entre les génomes A et B ont été observées chez les hybrides résultant en des déviations locales par rapport à la classification génomique globale des cultivars. Le profil des recombinaisons a suggéré la possibilité que des populations *M. balbisiana* introgréssées soient impliquées dans la formation de cultivars A/B.

Pour caractériser la mosaïque intersubspécifique d'accessions issues de *M. acuminata*, Martin et al. (2020) ont utilisé des marqueurs SNP identifiés avec des données de RNA-Seq ou de séquençage génomique (197 876 marqueurs) sur 24 accessions diploïdes (14 *M. acuminata* sauvages, 9 cultivars AAcv et 1 *M. balbisiana* utilisé comme 'outgroup') et une accession triploïde AAA ('Grande Naine'). Une analyse IGEA avec `ADMIXTURE` (Alexander et al., 2009) des 24 individus a inféré six groupes : quatre groupes ancestraux de *M. acuminata* (*banksii/micro-*

6. « Single Nucleotide Polymorphism », polymorphisme d'un seul nucléotide

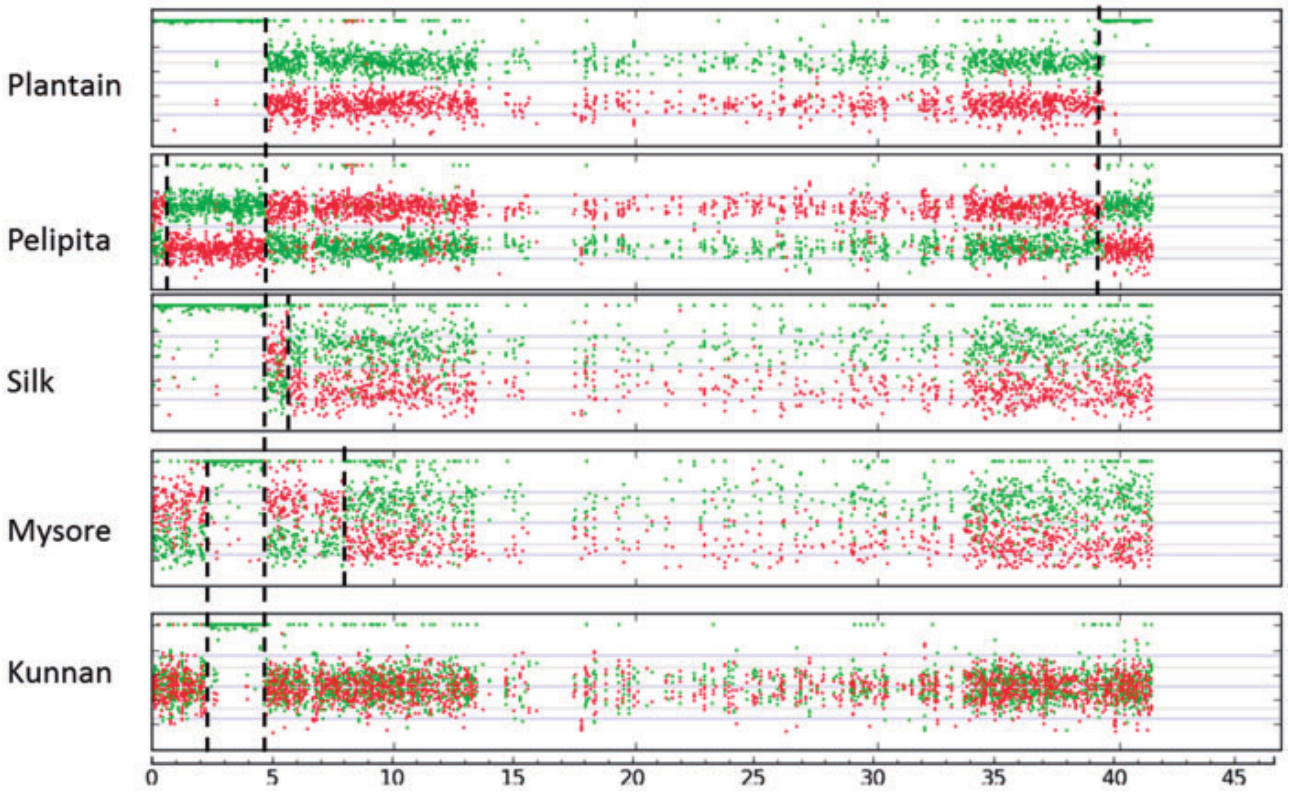

Figure 1.5 – Visualisation de la mosaïque AB du chromosome 9 de 5 accessions par ratio de couverture allélique par la méthode de Baurens et al. (2019). L'axe des ordonnées représente le ratio de couverture de chaque marqueur, et l'axe des abscisses leurs coordonnées génomiques (en mégabase). Les points verts correspondent aux allèles spécifiques de *M. acuminata* (A) et les points rouges aux allèles spécifiques de *M. balbisiana* (B). Les 4 premières accessions ('Plantain', 'Pelipita', 'Silk' et 'Mysore') sont triploïdes, et l'accession 'Kunnan' est diploïde. Les lignes verticales pointillées indiquent des évènements de recombinaison entre les origines A et B. Source : Figure 6 de Baurens et al. (2019), Molecular Biology and Evolution.

carpa type 'Bornéo', *malaccensis*, *zebrina*, *burmannica/siamea*), et un groupe ancestral présent seulement chez des individus hybrides ainsi qu'un groupe pour « l'outgroup » *M. balbisiana*. Une analyse factorielle des correspondances (AFC) a été réalisée sur les allèles des accessions identifiées comme représentatives des six groupes génétiques avec `ADMIXTURE`. Une étape de clustering sur les coordonnées de l'AFC a permis de sélectionner des marqueurs diagnostics pour chacun des six groupes. Ensuite, une probabilité d'appartenance à ces groupes est calculée par fenêtre glissante le long des génomes de chacun des individus. Une probabilité associant une portion de génome à l'un des six groupes ancestraux est transcrite sous forme de blocs de différentes couleurs (« chromosome painting »). Les blocs successifs de même origine sont assignés à un même chromosome homologue. Les mosaïques obtenues chez les cultivars ont montré des structures impliquant de trois à cinq groupes génétiques selon les individus (Martin et al., 2020). La figure 1.6 montre un exemple de mosaïque obtenue, ici sur l'accession 'Manang' (mosaïque la plus complexe avec au moins cinq origines).

Les mosaïques obtenues ont aussi mis en évidence des introgressions chez des accessions sauvages. Le groupe ancestral théorique (g5) a été proposé à partir d'un cultivar hautement

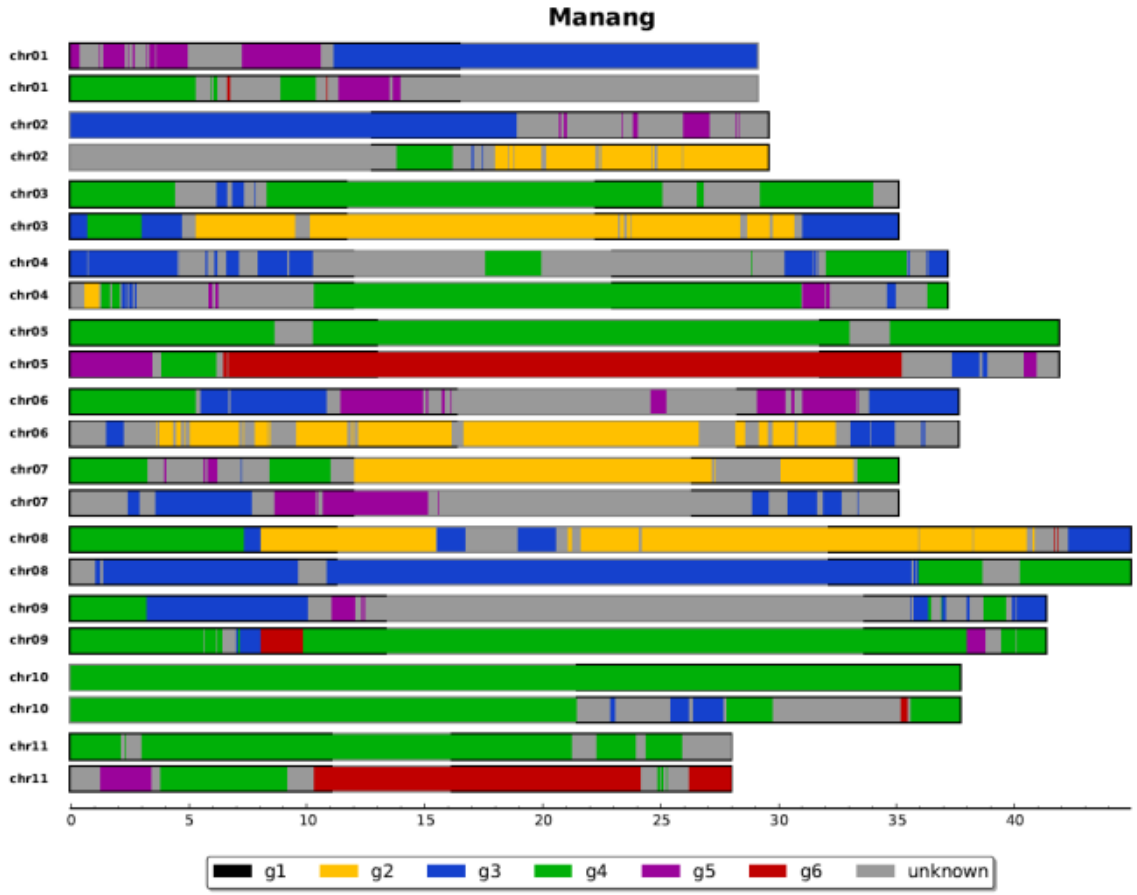

FIGURE 1.6 – Visualisation de la mosaïque de l'accession AAcv 'Manang' par la méthode de Martin et al. (2020). L'axe des abscisses représente les coordonnées génomiques (en mégabase). Les origines ancestrales sont représentées par des blocs de couleurs : 'burmannica/siamea' (g2) en jaune, 'malaccensis' (g3) en bleu, 'banksii/Borneo' (g4) en vert, 'Pisang Madu' (g5) en violet, 'zebrina' (g6) en rouge. Le groupe (g1) en noir correspond à « l'out group » *M. balbisiana*. La couleur grise représente les zones où l'assignation à une origine ancestrale n'a pas été possible. Source : Figure 6 C de Martin et al. (2020), The Plant Journal.

hétérozygote 'Pisang Madu', qui correspondrait potentiellement à une voire deux origines non déterminées.

1.3 Intérêt de la caractérisation des mosaïques

Au-delà des études de l'origine hybride de certains génomes par des approches globales, la caractérisation des mosaïques d'origines ancestrales chez les plantes peut permettre de caractériser de façon plus précise les différentes composantes ancestrales et de mieux comprendre les processus ayant mené à la formation des cultivars, à leur domestication et à leur diversification. Ces informations peuvent aussi contribuer à guider le choix de géniteurs et de stratégies de croisements dans le cadre de l'amélioration variétale. Enfin, cela peut permettre de relier la variation de certaines caractéristiques phénotypiques à des origines ancestrales.

Les mosaïques caractérisées chez le manioc *Manihot esculenta* (Bredeson et al., 2016) ou l'ananas *Ananas comosus* (Chen et al., 2019), ont apporté des informations sur le nombre de groupes ancestraux, le niveau de contribution chez les individus cultivés ou encore le nombre de croisements.

Les mosaïques de bananiers récemment publiées donnent une idée plus précise de leur composition génomique (Baurens et al., 2019; Martin et al., 2020) et ont mis en évidence des recombinaisons interspécifiques ou des mosaïques relativement complexes ce qui suggère des nombres de croisements plus importants que prédit initialement. Dans le cas des agrumes, Ahmed et al. (2019) ont fait ressortir à partir des mosaïques qu'ils ont inférées des schémas de croisements expliquant l'origine d'agrumes diploïdes, triploïdes et tétraploïdes (figure 1.1). Cela permet d'identifier les géniteurs ancestraux à améliorer puis de reproduire les schémas de croisements identifiés afin de créer de nouvelles variétés améliorées.

Des liens entre des caractères d'intérêts et des groupes ancestraux ont été montrés dans le cas d'introgressions adaptatives. Une revue de Burgarella et al. (2019) cite deux exemples s'appuyant sur le logiciel d'inférence locale des états ancestraux (ILEA) `HAPMIX` (Price et al., 2009) chez les plantes : le maïs et le peuplier. Une étude sur le maïs (Hufford et al., 2013) évalue les introgressions entre les populations de maïs cultivé et de téosinte *Zea mays.* ssp. *mexicana*, résistantes aux hautes altitudes. Neuf régions chromosomiques d'introgressions de *Zea mays.* ssp. *mexicana* ont été détectées chez les populations cultivées ce qui aurait favorisé leur dissémination dans des milieux où ces populations n'étaient pas originellement adaptées. De la même manière, la caractérisation des mosaïques a été utilisée pour une étude sur le peuplier en Amérique du Nord, entre les espèces *Populus trichocarpa* et *Populus balsamifera* (Suarez-Gonzalez et al., 2016). L'espèce *P. balsamifera*, adaptée au froid est retrouvée jusque dans les régions du Nord du Canada. Il a été montré que les hybrides portant des introgressions spécifiques de *P. balsamifera* avaient une meilleure adaptation au froid et se retrouvaient plus au Nord que les populations *P. trichocarpa*.

1.4 Méthodologies pour caractériser les génomes mosaïques

La définition des groupes ancestraux est une étape incontournable pour la caractérisation des génomes mosaïques. Celle-ci est classiquement réalisée avec des méthodes d'inférence globale des états ancestraux (IGEA). Les méthodes d'inférence locale des états ancestraux (ILEA) vont ensuite pouvoir être appliquées en utilisant les groupes définis. Une grande diversité de méthodes d'ILEA a été développée dans le contexte de la génétique humaine.

Chez les plantes, les méthodes IGEA sont souvent utilisées. Des exemples ont déjà été cités chez le bananier (Sardos et al., 2016a,b; Martin et al., 2020), le manioc (Bredeson et al., 2016), le riz (Wang et al., 2017), le cacao (Cornejo et al., 2018) ou encore l'ananas (Chen et al., 2019). Par contre, pour la caractérisation locale des contributions ancestrales on trouve souvent des développements spécifiques au cas d'étude plutôt que l'usage de méthodes d'ILEA publiées. L'usage de marqueurs diagnostics est plus souvent retrouvé dans ces cas-là, par exemple pour les agrumes (Curk, 2014; Wu et al., 2018), le manioc (Bredeson et al., 2016) et le bananier (Baurens et al., 2019; Martin et al., 2020).

Les deux parties suivantes présentent les méthodes d'IGEA et d'ILEA développées dans le contexte de la génétique humaine.

1.4.1 Inférence globale

L'inférence globale d'états ancestraux (IGEA) a pour but de détecter la structure génétique d'une population et la contribution des groupes détectés pour chaque individu. Il existe deux grands types de méthodes (Liu et al., 2013; Alhusain and Hafez, 2018), celles basées sur des modèles (appelées aussi paramétriques), et les méthodes non-paramétriques basées sur la réduction de dimension ou sur les distances.

`STRUCTURE` (Pritchard et al., 2000), une des méthodes d'IGEA paramétrique les plus utilisées par la communauté scientifique (Porras-Hurtado et al., 2013), repose sur un modèle de classification hiérarchique non supervisée. Ce modèle modélise la probabilité des génotypes observés (les données en entrée) à chaque individu sur chacun des marqueurs génétiques en fonction i) des proportions globales d'origine des individus pour un nombre de groupes (clusters) prédéfinis, les groupes pouvant être interprétés comme des populations ancestrales ; et ii) des fréquences des allèles de chaque marqueur dans ces groupes. `STRUCTURE` s'appuie sur un cadre bayésien et utilise un algorithme de Markov Chain Monte Carlo (MCMC) pour l'estimation des paramètres du modèle. Dans sa version initiale, le modèle supposait que les allèles reçus par les individus étaient indépendants conditionnellement aux origines ancestrales des individus et aux fréquences alléliques dans les clusters (ce qui s'apparente à l'hypothèse d'Hardy-Weinberg) et que les marqueurs n'étaient pas en déséquilibre de liaison (association non aléatoire entre allèles de différents marqueurs dans une population, noté DL).

Les développements qui ont suivi ont principalement eu pour objectif i) de lever cette dernière hypothèse d'absence de DL entre marqueurs, ii) d'introduire des modèles de relations entre les clusters (Falush et al., 2003) ; et iii) de réduire les temps de calculs. Des algorithmes de maximisation de la vraisemblance de type EM (Expectation-Maximization) ont été mis en place (appliqués à des données SNPs massives) dans le cas des programmes `FRAPPE` (Tang

et al., 2005) et ADMIXTURE (Alexander et al., 2009; Alexander and Lange, 2011) pour estimer les paramètres du modèle initial considéré par STRUCTURE. ADMIXTURE est bien plus rapide que STRUCTURE, tout en étant plus précis que FRAPPE. Une évolution de la méthode STRUCTURE, fastSTRUCTURE a été proposée par (Raj et al., 2014). Elle remplace l'estimation des paramètres réalisée via MCMC par une approche dite bayésienne variationnelle, qui permet d'exprimer cette estimation comme un problème d'optimisation. Cela permet à fastSTRUCTURE de produire une inférence plus rapide de deux ordres de grandeur par rapport à STRUCTURE tout en ayant des performances similaires voire meilleures qu'ADMIXTURE.

Les méthodes non-paramétriques se basent sur des approches de classification précédées par de la réduction de dimension ou par des calculs de distance. La réduction de dimension est classiquement réalisée par ACP SMARTPCA (Patterson et al., 2006) ; adegenet (Jombart, 2008) ; ipPCA Intarapanich et al. (2009), ou encore avec des factorisations matricielles (Liu and Zhao 2006 ; sNMF Frichot et al. 2014). Une étape de clustering est ensuite réalisée pour former des groupes parmi les individus analysés. Les méthodes basées sur les distances utilisent des matrices de distances comme résumé de l'information génétique partagée entre les individus. La distance principalement utilisée est « l'allele sharing distance » ou distance en allèles partagés (AWclust Gao and Starmer 2007 ; iNJclust Limpiti et al. 2014). Dans le cas d'AWclust, le clustering utilisé est hiérarchique, basé sur la méthode de Ward. La méthode iNJclust utilise le « neighbor joining » pour construire un arbre des individus, puis affine itérativement les sous-arbres jusqu'à obtenir des clusters homogènes.

Le paramètre fondamental des méthodes d'IGEA est le nombre de clusters supposés dans chaque jeu de données, noté classiquement K. À l'usage, si K est connu d'avance, les méthodes d'IGEA permettent simplement d'obtenir pour chaque individu les proportions des populations ancestrales le composant. Si K n'est pas connu, il est recommandé d'explorer plusieurs valeurs de K et d'utiliser les critères fournis par les méthodes pour définir le choix de K. Les méthodes paramétriques fournissent des mesures pour évaluer les valeurs optimales de K selon l'inférence obtenue, par exemple des approches de 'cross-validation' (Alexander et al., 2009). Les méthodes non paramétriques utilisent des évaluations de la qualité du clustering pour déterminer la meilleure valeur de K, avec par exemple la statistique « gap » pour AWclust, qui évalue la dispersion des éléments de chaque cluster. Il faut toutefois noter que la valeur de K optimale dépend autant (voire plus) des caractéristiques du jeu de données (nombre d'individus, nombre de marqueurs) que de l'histoire des populations.

L'interprétation des clusters obtenus n'est pas une étape triviale et fait l'objet de publications détaillant les protocoles à suivre (Porras-Hurtado et al., 2013; Lawson et al., 2018). Des études comparatives entre les méthodes permettent de mieux guider le choix des utilisateurs, en plus des comparaisons proposées dans la publication liée à chaque méthode (par exemple Stift et al., 2019).

1.4.2 Inférence locale

Les méthodes d'inférence locale d'état ancestral (ILEA) sont apparues rapidement après les méthodes d'IGEA. Leur objectif est d'inférer les probabilités d'origine ancestrale à une échelle locale, c'est-à-dire en chaque position du génome d'un individu. Les origines locales vont correspondre en espérance aux origines globales considérées dans les méthodes d'IGEA. Ces méthodes ont été principalement développées dans le contexte de la génétique humaine, à la fois pour mieux comprendre les dynamiques des populations et pour permettre d'affiner les études d'associations entre maladies et groupes génétiques (Seldin et al., 2011; Liu et al., 2013; Padhukasahasram, 2014; Geza et al., 2018).

Les revues bibliographiques des méthodes d'inférence proposent généralement une division en méthodes paramétriques contre non-paramétriques, quand elles inspectent à la fois les approches d'IGEA et d'ILEA (Seldin et al., 2011; Liu et al., 2013; Padhukasahasram, 2014). La dernière revue se concentrant sur les méthodes d'ILEA propose, elle, une séparation entre méthodes qui modélisent le déséquilibre de liaison contre méthodes qui l'ignorent voire le retirent de l'analyse (Geza et al., 2018).

Les méthodes d'ILEA les plus populaires reposent sur des modèles de Markov à états cachés (hidden markov models, HMM) (Rabiner, 1989). La première méthode d'ILEA a été introduite dans une extension du logiciel `STRUCTURE` évoqué plus haut par Falush et al. (2003) en se basant sur un HMM.

Un HMM modélise un enchaînement d'états non visibles ou cachés (ici, les origines ancestrales) portés par une séquence (chaîne) observée. Le modèle spécifie i) la probabilité de chaque état possible à la première position de la chaîne ; ii) les probabilités de transition en chaque position de la chaîne en fonction de la position précédente ; et iii) les probabilités, appelées probabilités d'émission, des observations (ici les génotypes) conditionnellement à chaque état caché. Un HMM est dit d'ordre 1 si la probabilité de transition vers un état en une position donnée de la chaîne ne dépend que de l'état à la position immédiatement voisine (et par extension d'ordre n si les probabilités de transition vers un état dépendent des états aux n positions voisines précédentes).

Dans le cas de la détection d'états ancestraux, la séquence visible est une séquence de marqueurs le long d'un chromosome, dont chaque élément est codé en 0,1 ou 2 (correspondant au nombre d'allèles alternatifs porté en chaque position pour un individu diploïde). La séquence d'état caché est l'état ancestral de chacun des marqueurs, que l'on peut appeler par exemple A et B pour deux origines. Les probabilités de transitions vont donc piloter le changement d'état ancestral, et seront classiquement dépendantes de paramètres biologiques comme le taux de recombinaison et le nombre de générations depuis l'hybridation. Les probabilités d'émissions

vont généralement être liées aux fréquences alléliques dans les différents groupes ancestraux.

Par exemple, `ANCESTRYMAP` (Patterson et al., 2004) prend en compte une carte physique et le taux de recombinaison pour affiner l'inférence des origines ancestrales dans le cadre d'études d'associations de maladies génétiques. Ces méthodes reposent sur la capture du signal du déséquilibre de liaison (DL) entre marqueurs d'une même origine, appelé « admixture LD ».

Les méthodes suivantes ont étendu les HMM utilisés pour récupérer un autre niveau de DL, appelé « background LD » qui correspond au DL entre marqueurs dans les populations ancestrales (Tang et al., 2006; Sundquist et al., 2008; Price et al., 2009). La première de ces méthodes à HMM étendu est SABER (Tang et al., 2006), utilisant un HMM d'ordre 1 (que les auteurs définissent comme HMM markovien, MHMM) mais avec des probabilités d'émission définies sur deux positions successives. L'information de phasage (ou information haplotypique, c'est-à-dire la co-localisation des allèles sur un même chromosome homologue) est exploitée par ces modèles afin d'affiner les inférences. Les méthodes basées sur les HMM hiérarchiques (HHMM) qui étendent les HMM, dont `HAPAA` (Sundquist et al., 2008) et `HAPMIX` (Price et al., 2009), vont utiliser l'information des phases tout en modélisant ou contrôlant les erreurs possibles à l'estimation des haplotypes (phasage). Des limitations techniques vont apparaître en même temps que les modèles se complexifient, rendant par exemple l'exploration de scénarios avec un grand nombre de groupes ancestraux trop coûteuse d'un point de vue computationnel.

D'autres approches ont donc été proposées par la suite pour segmenter le problème d'inférence, afin d'améliorer encore une fois la précision tout en contournant les problèmes de complexité des modèles (Baran et al., 2012; Churchhouse and Marchini, 2013; Guan, 2014). Par exemple, `LAMPLD` (Baran et al., 2012) utilise un modèle similaire à ceux d'`HAPAA` et `HAPMIX`, en l'appliquant sur des fenêtres le long des chromosomes, permettant de découper l'inférence en sous-parties et de réduire le temps de calcul. Dans une approche similaire, `MULTIMIX` (Churchhouse and Marchini, 2013) fonctionne avec un HMM sur des fenêtres de marqueurs qui sont résumés à leur origine via un modèle de distribution normale multivariée (MVN, « multivariate normal distribution »). La méthode `ELAI` (Guan, 2014) décompose le problème d'ILEA d'une manière différente. La méthode modélise explicitement via un premier HMM des haplotypes ancestraux (dont le nombre est supérieur au nombre de groupes ancestraux) à partir des données observées qui vont être détectées le long des individus analysés. Un second niveau de HMM associe ces haplotypes ancestraux générés par le premier HMM à des groupes ancestraux. Enfin, une méthode très récente, `MOSAIC` (Salter-Townshend and Myers, 2019) propose de la même façon un HMM à deux couches, permettant de modéliser les haplotypes ancestraux et les groupes d'origines ancestrales. De plus, il infère les relations entre les groupes ancestraux depuis les données, ce qui lui permet de gérer les cas où les populations représentatives des groupes ancestraux sont éloignées des « vrais » groupes ancestraux, et même de détecter la participation de groupes non référencés.

Les méthodes d'ILEA ne modélisant pas le déséquilibre de liaison sont moins nombreuses que les méthodes de la catégorie passée en revue précédemment. `LAMP` et son évolution `WINPOP` (Sankararaman et al., 2008b; Paşaniuc et al., 2009) fonctionnent sur le principe d'une analyse par fenêtres glissantes de SNP. Une approche de clustering est utilisée pour déterminer l'appartenance aux groupes ancestraux dans chaque fenêtre à partir des fréquences alléliques connues des groupes ancestraux. Chaque SNP est ensuite assigné à un des groupes en synthétisant les regroupements effectués sur chacune des fenêtres le recouvrant par vote à la majorité. Là où `LAMP` utilise une taille de fenêtre fixe et suppose une absence de recombinaison au sein des fenêtres, son extension `WINPOP` utilise une taille de fenêtre variable et considère la possibilité d'un évènement de recombinaison par fenêtre, ce qui lui permet d'être plus précis que `LAMP` quand les populations sont génétiquement proches. `RFMIX` (Maples et al., 2013) utilise une approche discriminante basée sur un classificateur (« conditional random fields »). La méthode utilise des fenêtres au sein desquelles les groupes ancestraux sont déterminés avec un classificateur paramétré par des forêts aléatoires entraînées sur des représentants des groupes. `LOTER` (Dias-Alves et al., 2018) reformule le problème d'ILEA en un problème d'optimisation, où chaque individu hybride est localement une copie d'un individu représentatif des groupes. Il permet alors d'inférer la mosaïque de l'hybride en parcourant un chemin dans les individus représentatifs des sources (chaque individu étant une piste) où le seul paramètre du modèle est une pénalité sur le changement d'individu.

TABLE 1.2 – Caractéristiques des principales méthodes ILEA citées

Logiciel	Nb. populations	Modélisation DL	Paramètres biologiques	Paramètres statistiques	Phases références	Phases hybrides	Étude
Structure v2	-	HMM	Prop. ancestrales		Non	Non	Falush et al. (2003)
ANCESTRYMAP	2	HMM	Carte physique, taux recombinaison, prop. ancestrales		Non	Non	Patterson et al. (2004)
SABER	-	MHMM	Carte physique, nb. générations, prop. ancestrales		Non	Non	Tang et al. (2006)
HAPAA	-	HHMM	Nb. générations, divergence génétique		Oui	Oui	Sundquist et al. (2008)
LAMP	-	-	Carte physique, taux recombinaison	Seuil DL	Non	Non	Sankararaman et al. (2008b)
HAPMIX	2	HHMM	HMM x2		Oui	Non	Price et al. (2009)
WINPOP	-	-	Carte physique, nb. générations, taux recombinaison	Seuil DL	Non	Non	Paşaniuc et al. (2009)
LAMP-LD	2,3,5	HHMM	Carte physique	Taille fenêtre	Oui	Non	Baran et al. (2012)
MULTIMIX	-	MVN	Carte génétique		Non	Non	Churchhouse and Marchini (2013)
RFMIX	-	-	Carte génétique, nb. générations	Taille fenêtre	Oui	Oui	Maples et al. (2013)
ELAI	-	HMM x2	Carte physique, nb. générations	nb. clusters inférieurs et supérieurs	Non	Non	Guan (2014)
LOTER	-	-		Pénalité switch	Oui	Oui	Dias-Alves et al. (2018)
MOSAIC	-	HMM x2	Carte physique, taux recombinaison		Oui	Oui	Salter-Townshend and Myers (2019)

La colonne 1 indique le nom de la méthode. La colonne 2 indique le nombre théorique de populations pris en charge par la méthode, « - » indiquant qu'il n'y a pas de limite théorique (dans les faits, les méthodes sont rarement testées sur des scénarios avec plus de 3 populations ancestrales). La colonne 3 indique comment est modélisé le déséquilibre de liaison (HMM x2 indique un modèle avec deux HMM imbriqués). Les colonnes 4 et 5 indiquent les principaux paramètres nécessaires aux méthodes. Les colonnes 6 et 7 indiquent si les données génétiques des références et des hybrides doivent être phasées. Table adaptée de Geza et al. (2018), Briefings in Bioinformatics.

Les principaux paramètres (Table 1.2) des méthodes sont le temps depuis l'évènement d'hybridation (qui peut éventuellement être estimé), les cartes génétiques ou le cas échéant une carte physique et un taux de recombinaison (centiMorgan/Mégabase), le taux de mutation et le nombre de populations ancestrales. La taille des fenêtres est aussi un paramètre important pour les méthodes basées sur les fenêtres.

Les paramètres biologiques et démographiques des populations humaines sont majoritairement connus mais c'est beaucoup moins le cas chez d'autres organismes, par exemple le bananier. Par exemple, il est attendu chez les bananiers cultivés un nombre réduit de générations après les évènements d'hybridation (du fait de la multiplication végétative), mais il n'y a pour le moment pas de données précises sur ce paramètre.

Les informations de phasage sont aussi nécessaires pour une grande partie des méthodes soit pour les populations de références (ancestrales) soit pour les populations hybrides (par ex. Table 1.2), alors que l'obtention de données phasées n'est pas toujours possible chez les organismes non modèles. Du phasage statistique (Browning and Browning 2011) peut être réalisé mais cela nécessite de grands échantillons des populations. Certaines méthodes (`HAPAA`, `HAPMIX`, `RFMIX`, `LOTER` et `MOSAIC`) modélisent l'erreur de phasage et permettraient de mitiger l'impact de ces erreurs sur l'inférence (Guan, 2014).

Les méthodes retenues comme potentiellement adaptées à l'inférence sur des données de plantes cultivées non modèles sont `SABER`, `WINPOP` et `ELAI`. Elles permettent à la fois d'inférer les origines sur des scénarios avec plus de deux origines ancestrales et fonctionnent avec des données non phasées.

1.5 Le projet de thèse

Des évènements d'hybridations interspécifiques et intersubspécifiques ont eu un impact important dans la structure en origine ancestrale des génomes de nombreuses plantes cultivées (Zhao et al., 2010; Perrier et al., 2011; El Baidouri et al., 2017; Wu et al., 2018; Cornejo et al., 2018; Civáň et al., 2019; Chen et al., 2019). Élucider à une échelle fine les structures mosaïques des génomes de plantes cultivées permet de mieux comprendre l'histoire de leur domestication et de leur diversification, et peut apporter des informations utiles pour les programmes d'améliorations génétiques par l'identification de l'origine phylogénomique de caractères d'intérêt agronomiques et une meilleure connaissance des processus/des croisements à l'origine des cultivars.

Des études de la mosaïque ancestrale chez les plantes sont souvent basées sur des approches par allèles diagnostics Curk et al. (2015); Wu et al. (2018). Les avantages de ces méthodes

sont principalement la simplicité d'utilisation et l'absence de paramètres biologiques à utiliser. Celles-ci fonctionnent correctement quand la différenciation entre les groupes ancestraux est forte et que la mosaïque attendue est peu complexe.

Nous avons fait le choix ici de tester l'apport potentiel des méthodes d'ILEA pour des cas de plantes cultivées non modèles. L'attendu principal est l'amélioration de l'inférence des mosaïques avec la prise en compte du déséquilibre de liaison dans les modèles. Cependant, les méthodes d'ILEA sont principalement développées dans le cadre de la génétique animale, et humaine particulièrement. Elles sont donc justifiées par des pré-requis pas toujours compatibles avec les modèles d'évolution de certaines plantes qui ont des modes de reproductions (autofécondation, multiplication végétative) qui s'écartent des modèles humains. Enfin, chez certaines plantes cultivées, les populations ancestrales peuvent être multiples, de 4 chez les agrumes (Curk et al., 2015) jusqu'à 10 chez le cacao (Cornejo et al., 2018) et être représentées par un nombre peu élevé d'individus disponibles, comme chez le bananier (Christelová et al., 2017).

L'objectif de cette thèse est d'évaluer et d'appliquer des méthodes d'ILEA pour caractériser les structures mosaïques de plantes cultivées. Il faut pour cela tester le comportement des méthodes d'ILEA dans des cas limites de leurs utilisations, afin d'utiliser les méthodes les plus appropriées pour caractériser les mosaïques des génomes de plantes en fonction de leurs spécificités et contraintes biologiques. Les méthodes seront appliquées à la caractérisation des mosaïques des bananiers.

Les méthodes d'ILEA pouvant fonctionner avec plus de deux populations et avec des données non phasées ont été évaluées sur des simulations. En effet, le phasage des données est une étape pouvant être problématique dans le contexte d'un faible nombre d'individus représentant les groupes ancestraux. Le chapitre 2 décrit le simulateur de données utilisé et sa validation. Il fonctionne en deux étapes, une première produisant des populations ancestrales, et une seconde les hybridant pour obtenir des mosaïques.

Dans un deuxième temps, trois méthodes d'ILEA ont été sélectionnées et évaluées sur des simulations couvrant l'impact du nombre de populations sources, du nombre d'individus représentatifs, de la propagation végétative et de l'autofécondation ou encore l'impact d'une population ancestrale manquante. Ces travaux ont fait l'objet d'une publication dans la revue G3 : Genes | Genomes | Genetics, présentée dans le chapitre 3.

Enfin, le chapitre 4 présente l'analyse d'un jeu de données SNP de 115 bananiers diploïdes sauvages et cultivés. Il a d'abord fait l'objet d'une approche exploratoire pendant laquelle nous avons proposé une méthode pour visualiser les mosaïques à partir des données de couvertures en lectures de séquençage des allèles. Nous avons ensuite appliqué et comparé entre elles les méthodes d'ILEA sur un sous-jeu de données de 25 individus.

Les simulations de données, la comparaison des méthodes d'ILEA ainsi que l'exploration et l'analyse du jeu de données bananier sont conduits sur la plateforme SouthGreen hébergée par le CIRAD. Elle met à disposition de sa communauté d'utilisateurs 1104 cœurs répartis sur 23 nœuds avec 192 GB de RAM chacun, avec un espace très large de stockage de données. Ces travaux de thèse, financés par Agropolis Fondation et le CIRAD, se sont déroulés dans le cadre du projet « GenomeHarvest : Mobilizing biomathematics/bioinformatics and genomics/genetics to decipher genome organization and dynamics as pathways to crop improvement » financé par Agropolis Fondation. Ils ont également bénéficié des ressources en données de séquençage de bananiers du projet France Génomique DYNAMO en collaboration avec le GENOSCOPE.

Chapitre 2

Le simulateur de données de génomes mosaïques 'plmgg'

2.1 Contexte

Les méthodes d'ILEA présentées dans le chapitre précédent ont toutes été testées avec des données simulées dans le but de mesurer leur précision, illustrer leurs comportements dans des cas particuliers (par exemple, des populations ancestrales très proches) et pour les comparer aux autres méthodes d'ILEA déjà publiées (Falush et al., 2003; Patterson et al., 2004; Tang et al., 2006; Sundquist et al., 2008; Sankararaman et al., 2008b; Price et al., 2009; Paşaniuc et al., 2009; Baran et al., 2012; Churchhouse and Marchini, 2013; Maples et al., 2013; Guan, 2014; Dias-Alves et al., 2018; Salter-Townshend and Myers, 2019). Il existe trois grands types d'approches pour la simulation en génétique des populations, i) les simulations remontant dans le temps (approche par coalescence ou « backward »), ii) les simulations avançant dans le temps (approche « forward »), iii) les simulations « sideways » qui utilisent des jeux de données déjà existants pour produire des données simulées (Liu et al., 2008; Yuan et al., 2012; Hoban et al., 2012).

Brièvement, la simulation par coalescence consiste à reconstruire la généalogie d'un échantillon de gènes (sous la forme d'un arbre bifurqué) en simulant les évènements (par exemple, date des évènements de coalescence entre deux lignées) jusqu'à leur ancêtre commun (Kingman, 1982). Les mutations sont ensuite distribuées sur l'arbre en fonction de paramètres de mutations. Ce type de simulations présente comme avantage principal la rapidité d'exécution. Les modèles sont simples et l'algorithmique des graphes permet l'usage d'optimisation dans les programmes. Parce que les individus ne sont pas modélisés explicitement, il est difficile de modéliser via la coalescence des scénarios complexes, comme ceux que l'on trouve chez les plantes cultivées. Un des premiers simulateurs disponibles utilisant la coalescence est `ms`, proposé par Hudson (2002). Dans les méthodes d'ILEA présentées, seul Falush et al. (2003) utilisent la simulation en coalescence pour l'évaluation du modèle.

La simulation en « forward » est une approche centrée sur les individus. Ils sont modélisés explicitement et sont porteurs de l'information génétique. À chaque génération, les individus suivent un cycle de vie de type naissance, reproduction et mort. L'avantage principal de cette approche est le très grand contrôle sur les évènements rythmant la génération des populations, et le suivi possible de chaque individu simulé. Son désavantage principal est son coût computationnel qui est beaucoup plus important que ceux des simulateurs en coalescence dans la mesure où l'ensemble de la population est simulée. Les débuts de la simulation en « forward » sont un peu plus anciens que la simulation en coalescence, avec `VORTEX` par Lacy (1993).

Les méthodes « sideways » ont été permises avec l'avènement des gros jeux de données génétiques comme les jeux de données HGDP (Cann, 2002) ou HAPMAP (The International HapMap Consortium, 2005). Elles consistent généralement en un ré-échantillonnage et des étapes de méioses artificielles sur les chromosomes individuels. Les méthodes d'ILEA `HAPAA`, `LAMP`, `WINPOP`, `LAMP-LD` et `RFMIX` (Sundquist et al., 2008; Sankararaman et al., 2008b; Paşaniuc et al., 2009; Baran et al., 2012; Maples et al., 2013) ont été évaluées sur des simulations issues d'approches « sideways » simples en réalisant des recombinaisons artificielles d'haplotypes ségrégeant dans deux à quatre populations humaines représentées dans les jeux de données de génotypage HAPMAP ou HGDP (Cann, 2002; The International HapMap Consortium, 2005; The International HapMap 3 Consortium, 2010). Les simulations concernant les méthodes `ANCESTRYMAP`, `SABER` et `MULTIMIX` (Patterson et al., 2004; Tang et al., 2006; Churchhouse and Marchini, 2013) se basent sur des HMM pour générer pour chaque individu une séquence d'origine ancestrales. Cette séquence va servir de patron pour construire un génotype complet à partir des portions correspondantes des données réelles utilisées. Les génotypes sont générés différemment selon les trois méthodes : i) les génotypes sont générés à partir des fréquences alléliques du jeu de données utilisé (`ANCESTRYMAP`), ii) les génotypes sont directement échantillonnés sur le jeu de donnée utilisé (`SABER`), iii) les génotypes sont générés à partir du modèle de DL de Li and Stephens (2003) (`MULTIMIX`). Enfin, `HAPMIX`,`ELAI` et `LOTER` (Price et al., 2009; Guan, 2014; Dias-Alves et al., 2018) ont été évaluées sur des simulations utilisant une distribution exponentielle pour générer la structure d'origine ancestrale (en générant le long du chromosome simulé les points de recombinaisons) puis copie pour chaque segment entre deux recombinaisons les portions correspondantes à un individu choisi au hasard dans les populations de référence.

Dans le cadre de cette thèse, le choix a été fait de d'implémenter dans un simulateur simple, une approche combinant simulations par coalescence (pour les haplotypes fondateurs) et « forward ». Ce dernier choix visait à satisfaire les trois objectifs principaux de simulations réalistes, à savoir i) avoir des modes de reproductions de type plantes, ii) générer des scénarios complexes avec des changements de ces modes de reproduction au cours du temps et iii) suivre les origines de chaque marqueur pour chaque individu. Le recours à une simulation par coalescence pour générer les haplotypes fondateurs utilisés pour démarrer la simulation en « forward » (plutôt qu'issus de jeu de données préexistants) permettait de ne pas trop se rapprocher d'un cas par-

ticulier de plante cultivée, et aussi de contrôler directement le nombre de groupes ancestraux et leurs liens de parentés (structure de l'arbre des divergences et temps depuis les divergences entre les sources).

Le simulateur de données utilisé au cours de la thèse a été écrit dans sa première version par Benjamin Penaud au cours de son stage de Master 2 en 2016. Le simulateur est alors un script `R` (R Core Team, 2020) monolithique (code et paramètres inclus). Il a été décidé de faire évoluer ce simulateur, pour le rendre plus flexible. D'un point de vue théorique, les deux changements notables sont donc la modification de la création des populations sources mais aussi le traitement de la reproduction. D'un point de vue technique, l'atomisation des fonctions du simulateur a fluidifié fortement son utilisation à large échelle avec l'usage de fichiers de configurations externes. La partie suivante décrit en détail le fonctionnement du simulateur que nous avons nommé « `plmgg` » pour « plant like mosaic genome generator ». Parce que sa partie en « forward » a entièrement été implémentée (pendant le stage puis la thèse), elle nécessite des tests pour s'assurer que le comportement des populations simulées est correct par rapport à des attendus simples de génétiques des populations. Parce que nous allons parler de populations ancestrales théoriques, elles seront appelées « sources » dans les chapitres 2 et 3. Dans le chapitre 4, nous utiliserons les termes de « groupes ancestraux » pour les groupes génétiques cohérents dans la diversité des bananiers.

2.2 Processus de simulation

Le processus de simulation se déroule en deux grandes étapes. La première étape produit des populations sources. La seconde étape utilise les populations sources et un schéma de croisement pour produire des populations hybridées et des populations représentatives des sources. Les outils dans le paquet permettent ensuite d'échantillonner et d'exporter ces populations ainsi que des statistiques sur la simulation. Les données produites sont des données de variants SNP bialléliques et non phasées, dans un contexte diploïde. Chaque individu est composé d'un seul chromosome. Le nombre de marqueurs produits est fonction de différents paramètres de la simulation par coalescence (taille efficace de la population N_e , temps de différenciation τ et taux de mutation θ, voir table 2.1) et n'est pas contrôlé directement dans la simulation.

2.2.1 Génération des populations sources

L'objectif de cette étape est de produire des populations sources en contrôlant leur nombre et leur niveau de différenciation génétique. Le simulateur dans sa première version utilisait un algorithme naïf où i) les individus étaient générés avec un nombre de marqueurs fixe, ii) les

TABLE 2.1 – Notations utilisées pour les paramètres du simulateur `plmgg`

Symbole	Détail
S ; s	Nombre et indice des populations ancestrales
P ; p	Nombre et indice des populations dérivées des populations ancestrales
N_e	Taille efficace haploïde de la population de départ (coalescence)
τ	Temps de divergence (coalescence)
θ ; μ	Taux de mutation à l'échelle (coalescence) et taux par site (« forward »)
ρ ; r	Taux de recombinaison à l'échelle (coalescence) et taux par site (« forward »)
G ; g	Nombre de générations et indice (« forward »)
n_P ; n_p	Nombre d'individus diploïdes pour l'ensemble des populations et pour la population p
o_P ; o_p	Nombre d'individus diploïdes échantillonnés à la fin de la phase « forward » pour l'ensemble des populations et pour la population p
h	Indice des haplotypes
b	Position du point de cassure pour la recombinaison (« forward »)

variants étaient tirés via une loi binomiale et iii) les individus étaient séparés en sous populations et subissaient un nombre de générations de panmixie dépendant de la différenciation voulue.

Cette méthode avait pour avantage sa facilité de développement et d'utilisation. En revanche, elle ne permettait pas de contrôler précisément la différenciation entre les populations et était coûteuse en temps de calcul.

Pour palier à ce problème et renforcer la solidité de la simulation, nous nous sommes tournés vers les simulateurs classiquement utilisés en génétique des populations. Le choix s'est porté sur les logiciels de simulation par coalescence et plus précisément le simulateur `scrm` (Staab et al., 2015). Le simulateur `scrm` permet de reproduire les résultats très proches de ceux de `ms`(Hudson, 2002), tout en étant en moyenne 50 fois plus rapide grâce à des approximations sur le processus de recombinaison. De plus, il a été interfacé avec `R` via le paquet `coala` (Staab and Metzler, 2016), ce qui permet de l'utiliser directement dans le simulateur.

Dans le cas de la production des sources, la coalescence est utilisée de façon simple avec une population source de taille N_e divergeant en S populations sources différenciés. Un seul évènement de divergence est programmé avec un temps de différenciation τ sur une échelle de diffusion (c'est-à-dire, $\tau = t/4N_e$ où N_e est la taille efficace de chaque population supposée identique) en supposant une phylogénie en étoile pour l'histoire des populations sources, c'est-à-dire qu'elles se différencient toutes de manière indépendante depuis leur population ancestrale

commune. Le simulateur prend en entrée un taux de mutation par base μ qui est mis à l'échelle ($\theta = 4N_e\mu$) pour l'étape de coalescence. Le taux de recombinaison à l'échelle $\rho = 4N_e r$ est calculé à partir de d'un taux de recombinaison par base r (fixé arbitrairement, voir chapitre 3 pour cas pratique).

2.2.2 Production des mosaïques

L'étape de production des individus hybrides est effectuée par une simulation en « forward », basée sur une description des contributions ancestrales à chaque population et une liste de modes de reproduction par générations pour produire des structures en mosaïque d'origine ancestrale. Les populations non mélangées qui dérivent des sources seront dites « représentatives des sources ». Le simulateur permet quatre « modes » de reproductions qui sont :

- La propagation végétative c'est-à-dire : un clonage d'individu sans évènement de méiose (les chromosomes du descendant sont la copie conforme de ses parents, à la mutation somatique près).
- L'autofécondation : les deux gamètes d'un descendant proviennent du même parent.
- Le croisement intra-population : les deux gamètes d'un descendant proviennent de deux individus différents issus de la même population.
- Le croisement inter-population : les deux gamètes d'un descendant proviennent de deux individus différents issus de deux populations différentes.

La matrice de contribution $P \times S$ (avec P le nombre de populations simulées en « forward », c'est-à-dire populations hybrides et populations représentatives des sources) fixe le nombre d'haplotypes de chaque source qui contribue à chaque population simulée (voir Table 2.2). Chaque élément de la matrice est un nombre pair strictement positif. La somme des éléments en ligne est égale à la taille finale de la population simulée. Pour produire une population représentative d'une source, il suffit de ne mettre que des individus de cette source. Pour faire une population hybride, il faut échantillonner des individus de différentes sources dans la population (la structure de la mosaïque sera déterminée par les croisements).

TABLE 2.2 – Exemple de matrice de contribution avec $S = 3$ sources et $P = 5$ populations en « forward »

	s_1	s_2	s_3
p_1	300	0	0
p_2	0	300	0
p_3	0	0	300
p_4	150	150	0
p_5	100	100	100

L'illustration correspond à un échantillonnage de 300 individus par source en coalescence. Les populations p_1, p_2, p_3 sont ici représentatives de sources. La population p_4 est hybride des sources s_1 et s_2, et la population p_5 est hybride des trois sources.

La matrice de reproduction $P \times G$ (avec G le nombre de générations de la partie « forward ») fixe la proportion de chaque mode de reproduction par génération pour chaque population. Chaque élément de la matrice est un vecteur de taille 4, avec des valeurs comprises entre 0 et 1, et dont la somme fait 1. Ses valeurs correspondent dans l'ordre aux proportions d'événement de propagation végétative, d'autofécondation, de croisement intra- et de croisement inter-populationnel. Pour les croisements inter-populationnels, la proportion est complétée par le numéro de ligne correspondant à la seconde population.

Concrètement, le vecteur s'écrit avec les valeurs séparées par des barres obliques « `/` ». La population extérieure est précédée par un caractère « `p` » dans la notation (qui s'élude si ce mode de reproduction n'est pas utilisé). Par exemple, une génération avec 1/4 de chaque mode de reproduction se note « `0.25/0.25/0.25/0.25p2` » pour un croisement inter-population avec la population 2. Il est possible d'y accoler un caractère « `:` » pour indiquer une répétition pendant un nombre de générations. Par exemple, une population qui se reproduit à 99 % par autofécondation et à 1 % par croisement pendant $G = 50$ générations se note « `0/0.99/0.01/0:50` ». Enfin, il est possible de distribuer différentes modulations des proportions au hasard via un caractère « `,` ». Par exemple, une population qui se reproduit par autofécondation très majoritairement (ce qui se note par exemple « `0/0.9/0.1/0:45` ») et qui reçoit aléatoirement du matériel génétique d'une autre population au hasard (noté par exemple « `0/0/0/1p2:5` ») se notera « `0/0.9/0.1/0:45,0/0/0/1p2:5` ». Ces raccourcis permettent d'éviter une écriture lourde pour des schémas de reproduction dépassant plus d'une dizaine de générations.

Un individu diploïde est représenté par une paire de vecteurs de taille M (le nombre de SNPs) contenant l'état allélique observé sur chaque chromosome. Chaque allèle est codé 0 (état ancestral) ou 1 (état dérivé). Cela se traduit dans le programme par des variables booléennes. Une population p est représentée par une matrice de dimension $2n_p \times M$ (avec n_p le nombre diploïde d'individus de la population p). Les haplotypes des individus sont reliés par leur position dans la matrice. Concrètement, le premier individu est composé des haplotypes (h_1,h_2), le second des haplotypes (h_3,h_4) et le dernier des haplotypes (h_{2n_p-1}, h_{2n_p}).

De la même manière, l'état ancestral d'un individu est représenté par une paire de vecteurs de taille M d'indices indiquant l'état ancestral. L'état ancestral d'une population est donc de la même façon une matrice $2n_p \times M$. La valeur prise par chaque élément du vecteur ou de la matrice est l'indice s de la population source.

Tout changement génétique est donc une transformation de la matrice des génotypes. Par mesure de simplification, les évènements de mutation ne créent pas de nouveaux allèles et correspondent donc simplement à une inversion du booléen dans la matrice des haplotypes($0 \rightarrow 1$ et $1 \rightarrow 0$).

Un croisement entre deux individus s'effectue par la production de deux « gamètes » par méiose, un gamète étant ici un haplotype issu d'une recombinaison de deux haplotypes parentaux. Une méiose est effectuée en trois étapes. Deux haplotypes (notés h_1 et h_2) sont récupérés aléatoirement. Un point de cassure b est déterminé par un choix aléatoire d'une position comprise entre 2 et $M - 1$. Le nouvel haplotype est alors construit en concaténant simplement h_1 des positions 1 à b et h_2 des positions $b + 1$ à M.

Concrètement, le simulateur manipule les populations en utilisant des listes de matrices. Chaque population p est représentée par une matrice d'haplotypes ($2n_p \times M$) et une matrice d'origine de même dimension. Pour simuler la génération $g+1$ à partir de la g, la fonction créée la même structure (listes de doubles matrices) vide. La fonction va « remplir » les matrices $g + 1$ d'haplotypes et d'origines en sélectionnant au hasard des individus des matrices g et en les croisant. Un individu peut donc être « parent » plusieurs fois. Le nombre d'individus par mode de reproduction est calculé à partir de la matrice de croisement. Dans le cas où les divisions ne tombent pas juste, la somme des restes est utilisée pour ajuster le nombre de chaque mode de reproduction. Par exemple, dans le cas de l'autofécondation fortement majoritaire (99 %) pour 150 individus (donc 300 haplotypes), la correction fournira aléatoirement les combinaisons 149 ;1 ou 148 ;2. Une fois que tous les haplotypes d'une génération sont générés, la mutation est appliquée via une loi binomiale $B(M, \mu)$ avec le nombre de marqueurs M comme nombre d'expériences et le taux de mutation par base μ comme probabilité de succès. L'opération est répétée autant de fois que le nombre de générations fixé par la matrice de croisement.

2.2.3 Échantillonnage, formatage et statistiques

Une fois que toutes les populations ont subi les étapes de reproduction, il reste à échantillonner et exporter les données. Le paramètre fourni au simulateur est un vecteur de taille P contenant le nombre d'individus o_p à échantillonner pour chaque population p. Le format de base d'export est le variant calling format (vcf) (Danecek et al., 2011). C'est un format tabulé pour décrire les variants. Il est composé de 10 colonnes obligatoires contenant des infor-

mations diverses sur les marqueurs (comme leur position ou leur encodage) et d'une colonne par individu qui contient les comptages du marqueur (nombre de lectures supportant chaque allèle). Ici, toutes les informations autres que la présence du marqueur et la position sont artificielles et ajoutées à la volée à l'écriture du jeu de données. Les positions chromosomiques sont fixées à partir des positions produites lors de la simulation initiale par coalescence. Le choix du vcf est motivé par son usage courant pour le stockage de données génétiques. Des fonctions supplémentaires permettent d'obtenir des statistiques, à savoir une estimation de l'indice de fixation F_{ST} (Weir and Cockerham, 1984), la taille des segments issus des différentes sources et l'hétérozygotie.

2.3 Validation du simulateur

L'objectif de la simulation est la production de mosaïques en mélangeant des individus de populations différenciées. Il est important que l'information génétique des individus pendant la phase en « forward » soit conforme aux attendus de la génétique des populations au moins dans les cas simples où ceux-ci peuvent être dérivés ou approchés analytiquement. Nous avons donc évalué l'évolution de l'hétérozygotie et de l'étendue du déséquilibre de liaison (DL) au cours des générations en « forward » dans des scénarios simplistes correspondant aux différents modes de reproduction testés séparément sur une seule population. La partie en coalescence des simulations a été réalisée avec $N_e = 500$, une seule population ($S = 1$), et un taux de mutation θ calculé en supposant $\mu = 1.10^{-8}$.

2.3.1 L'hétérozygotie

L'hétérozygotie est une mesure du niveau de diversité génétique d'une population. Pour un site donné (SNP) elle correspond à la probabilité que deux gamètes tirés au hasard portent un allèle différent (ce qui correspond à la proportion d'individus hétérozygotes dans une population à l'équilibre Hardy-Weinberg). L'hétérozygotie est influencée principalement (pour les marqueurs neutres) par le taux de mutation et l'histoire démographique de la population (évolution de la taille de la population) et par conséquent par les évènements de reproduction qui s'y produisent. Typiquement, les évènements de reproduction entre individus apparentés conduisent à une réduction de l'hétérozygotie, la forme la plus extrême étant l'autofécondation. Plus généralement, une réduction de la taille de la population conduit aussi à une perte de l'hétérozygotie (phénomène de dérive génétique). L'introduction de nouveaux variants (par mutation ou migration d'individus issus d'autres populations) va au contraire augmenter l'hétérozygotie. Dans le cadre de nos simulations, du fait du faible taux de mutation et du faible nombre de générations que l'on considère, les principaux facteurs pouvant influencer l'hétéro-

zygotie de nos populations, lorsqu'elles sont simulées en isolement sont la dérive génétique et le mode de reproduction.

D'un point de vue théorique (Hedrick, 2011), l'hétérozygotie H_t au temps t dans une population panmictique de taille N évolue en fonction de l'hétérozygotie au temps H_0 suivant l'équation suivante :

$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0 \tag{2.1}$$

En cas d'autofécondation, l'hétérozygotie est réduite par deux à chaque génération, comme décrit par l'équation suivante :

$$H_t = \left(\frac{1}{2}\right)^t H_0 \tag{2.2}$$

Il faut noter que ces formules restent valables si l'on considère l'hétérozygotie moyenne (sur l'ensemble des marqueurs).

Dans le cas de la propagation végétative, il n'y a pas d'attendu théorique proposé. Supposant que la population est assez grande, il ne devrait pas y avoir d'impact sur l'hétérozygotie. Dans le cas d'une petite population, il est possible qu'un petit nombre d'individus envahissent la population, conduisant à une chute de l'hétérozygotie.

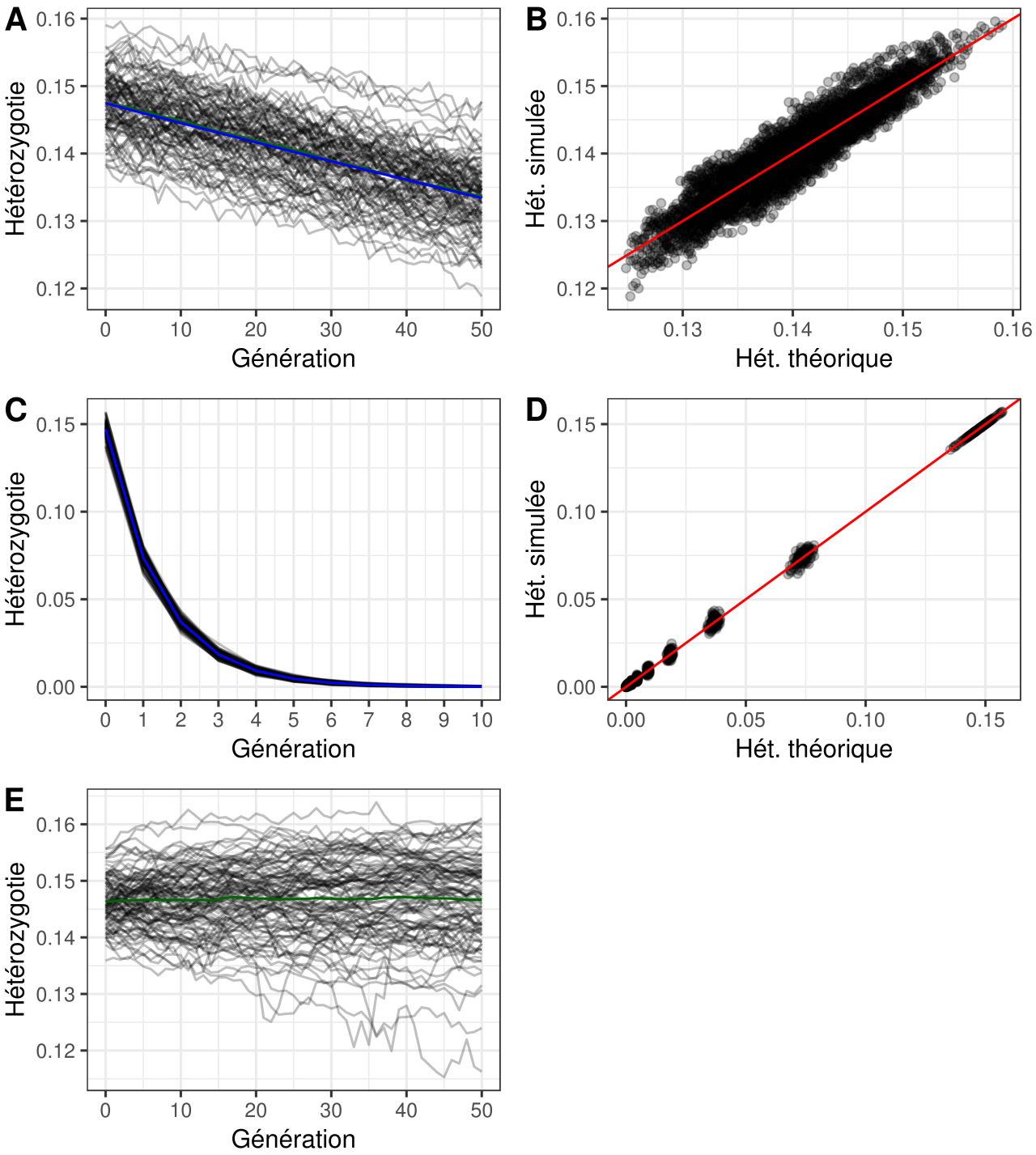

FIGURE 2.1 – Évolution de l'hétérozygotie pendant la phase « forward » du simulateur. (A) Évolution de l'hétérozygotie moyenne (en ordonnée) en fonction des générations (en abscisse) pour la simulation en reproduction sexuée classique. (B) Représentation pour chaque point de données de sa valeur d'hétérozygotie théorique (abscisses) et simulée (ordonnées) dans le cas de la reproduction sexuée classique. (C) Évolution de l'hétérozygotie (en ordonnée) en fonction des générations (en abscisse) pour la simulation en autofécondation. (D) Représentation pour chaque point de données de sa valeur d'hétérozygotie théorique (abscisse) et simulée (ordonnée) dans le cas de l'autofécondation. (E) Évolution de l'hétérozygotie en fonction des générations pour la simulation en multiplication végétative. Pour les figures (A, C, E), les courbes noires transparentes correspondent chacune à une simulation, les courbes bleues correspondent à l'attendu théorique calculé d'après la moyenne des simulations au temps 0 et les courbes vertes correspondent à la moyenne de chaque simulation par génération. Les courbes vertes et bleues sont confondues sur les figures (A) et (C). Pour les figures B et D, les points sont représentés avec une valeur de transparence fixe afin de faire ressortir la densité et la droite $y = x$ est en rouge.

Les simulations ont été conduites avec une seule population source produite par coalescence. Pour chaque simulation, la population « forward » est de taille $n_P = 250$ et le mode de reproduction est maintenu $G = 50$ générations. Chacune des simulations a été répétée 100 fois. Pour chaque simulation à chaque génération, l'hétérozygotie moyenne de la population est calculée. Les valeurs théoriques sont calculés pour chaque simulation à partir de la valeur d'hétérozygotie à la fin de la production de la population par coalescence. L'écart entre les valeurs théoriques et observées est mesuré par la valeur absolue de la différence pour chaque paire de valeur.

En ce qui concerne la reproduction sexuée classique, une chute lente de l'hétérozygotie est attendue. C'est bien la tendance qui est observée (Figure 2.1 A). L'écart moyen entre l'hétérozygotie théorique et simulée est de $1,76.10^{-3} \pm 1,40.10^{-3}$ (moyenne et écart-type). Cet écart-type du même ordre de grandeur que la moyenne elle-même est illustré par la Figure 2.1 B.

Pour l'autofécondation (Figure 2.1 C), la différence entre l'attendu et la simulation est d'un ordre de grandeur plus faible, avec un écart moyen de $1,77.10^{-4} \pm 5,97.10^{-4}$. Les nuages de points de la Figure 2.1 D mettent en avant la réduction par un facteur 2 de l'hétérozygotie à chaque génération, avec des valeurs proches de 0 dès 10 générations.

Enfin, la simulation de la propagation végétative semble en moyenne ne pas avoir d'impact sur l'hétérozygotie, à 50 générations tout du moins (Figure 2.1 E). La valeur absolue de l'écart entre la moyenne de l'hétérozygotie à 0 et 50 générations est de $3,17.10^{-4}$. Toutefois, certaines simulations montrent une chute un peu plus importante (de 0,136 à 0,116) de l'hétérozygotie. C'est explicable par la possibilité d'une « invasion » de la population par un petit nombre de clones.

2.3.2 Le déséquilibre de liaison

Le déséquilibre de liaison (DL) correspond à l'association non aléatoire entre les allèles de différents marqueurs et il est généralement d'autant plus élevé que les marqueurs sont proches. En effet, la recombinaison génétique entre les marqueurs tend à le réduire. Cependant, d'autres forces évolutives, comme la sélection peuvent maintenir un niveau de DL élevé sur de longues distances autour de variants favorables. De même, le niveau de base du DL entre marqueurs non liés génétiquement (sur des chromosomes différents par exemple) est également inversement corrélé à la taille de la population. Dans le cas de la simulation, l'évolution du DL permet de savoir si le simulateur reproduit correctement les évènements de recombinaison et leur impact génétique sur la structure des haplotypes. La population a suivi un régime de reproduction sexuée panmictique (croisement intra-population) pendant $G = 100$ générations. Nous utiliserons ici le r^2, mesure du DL proposé par (Hill and Robertson, 1968) correspondant au coefficient

de corrélation entre les états alléliques (notés 0 pour l'allèle de référence et 1 pour l'allèle alternatif) à deux marqueurs bialléliques dans l'ensemble des chromosomes de la population. Cette mesure est comprise entre 0 (les marqueurs sont indépendants) et 1 (les marqueurs portent la même information génétique). Une approximation proposée dans les travaux de McVean (2002) permet de prédire l'espérance du r^2 entre une paire de marqueurs distants de c (où c représente le taux de recombinaison par méiose entre les deux marqueurs) à l'équilibre mutation-dérive. Avec $\rho = 4N_e c$, l'attendu du DL mesuré par le r^2 est égal en espérance à :

$$E(r^2) = \frac{10 + \rho}{22 + 13\rho + \rho^2} \tag{2.3}$$

Le r^2 sur les données simulées a été calculé avec le logiciel `Haploview` (Barrett et al., 2005), avec les paramètres de base (filtre sur la fréquence allélique mineure inférieure à 0.05, distance maximale considérée entre marqueurs de 500kb). Les calculs ont été produits pour les générations g_1, g_{25}, g_{50}, g_{75}, g_{100}. Les valeurs moyennes de r^2 entre paires de marqueurs sont regroupées par classe de distance entre marqueurs afin de faciliter la représentation et le calcul de l'attendu théorique.

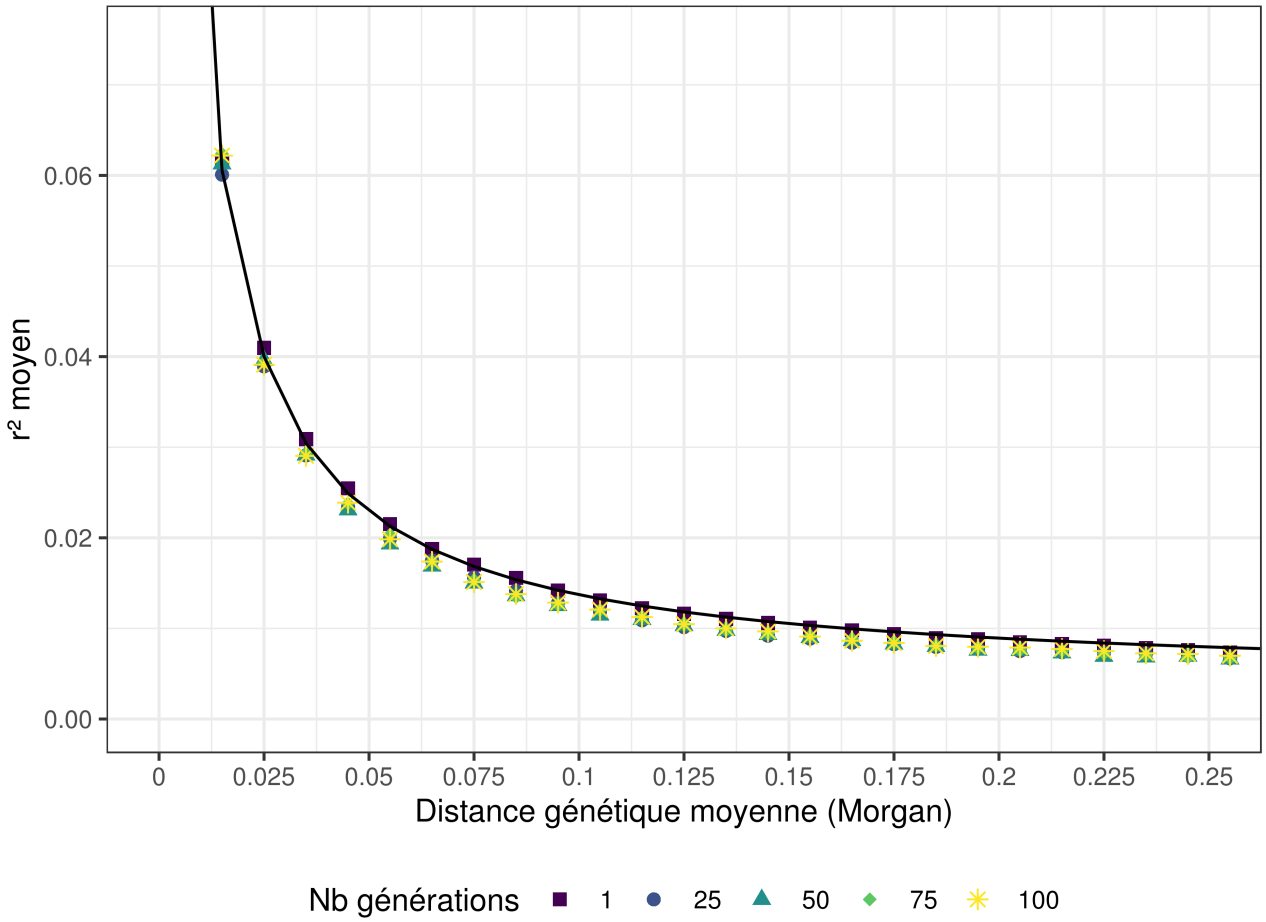

FIGURE 2.2 – Évolution du déséquilibre de liaison (mesuré par le r^2 moyen par classe de distance physique) en fonction du taux de recombinaison et du nombre de générations. L'axe des abscisses représente le taux de recombinaison moyen (distance génétique) entre paires de marqueurs. L'axe des ordonnées représente le r^2 entre marqueurs calculé par `Haploview`. La courbe noire indique l'attendu théorique à partir du taux de recombinaison et de la taille de la population. Les points indiquent les valeurs de r^2 par classe de distance. Ils sont codés en forme et en couleur pour le nombre de générations au moment de l'analyse du DL.

Les courbes de DL observées (Figure 2.2) suivent la forme de l'attendu théorique, avec un DL fort sur les courtes distance et qui se réduit quand la distance augmente. Le DL observé dans la simulation est légèrement plus faible que les valeurs théoriques, avec un écart entre DL attendu moyen et DL observé moyen à g_1 de $8,5.10^{-4} \pm 2,3.10^{-3}$ et de $2,2.10^{-3} \pm 6,1.10^{-3}$ entre DL attendu moyen et DL observé moyen à g_{100}. L'écart moyen entre le r^2 des populations à la génération g_1 et les populations à la génération g_{100} est de $1,43.10^{-3} \pm 3,86.10^{-3}$.

2.4 Illustration du fonctionnement du simulateur

Cette partie présente un exemple de simulation avec le simulateur `plmgg` qui illustre le fonctionnement des différents modes de reproduction et l'impact sur les mosaïques d'origines

ancestrales. L'objectif est de simuler des populations hybrides en montrant les différentes possibilités offertes par le simulateur. Les quatre modes de reproduction vont donc être utilisés, sur une simulation de $G = 50$ générations. La taille des populations sources est fixée à $N_e = 500$ haplotypes et le taux de mutation par base est fixé particulièrement bas ($\mu = 1.10^{-11}$) pour accélérer la simulation (en réduisant le nombre de sites). Trois populations sources ($S = 3$), avec une différenciation de $\tau = 0,4$ vont être générées pour produire $P = 6$ populations, trois populations représentatives de sources (p_1, p_2, p_3) et trois populations hybrides (p_4, p_5, p_6). La population p_3 représentante de la source s_3 se reproduira majoritairement par autofécondation (95 %). La population p_4 sera une hybride des deux sources s_1 et s_2 tandis que les deux autres p_5, p_6 seront des hybrides des trois sources. La population p_5 recevra une migration constante de la part de la population p_4 (10 % de croisement inter-population pour 90 % en intra). La population p_6 passera en mode reproduction par propagation végétative à 99 % (et 1 % de croisement intra-population) après 10 générations de croisements intra-population.

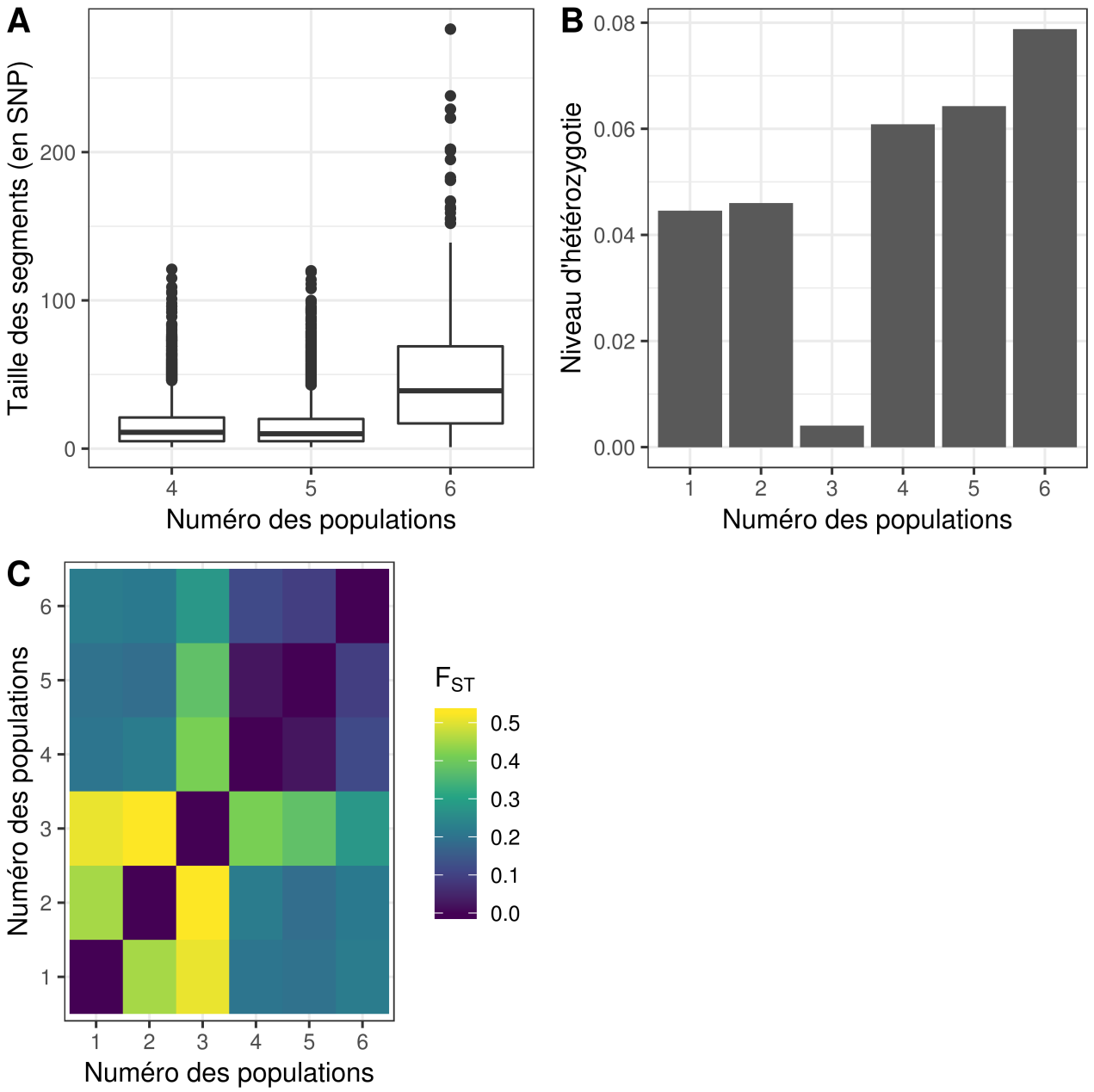

FIGURE 2.3 – Statistiques générées lors d'une simulation. Les diagrammes en boites (A) indiquent la taille des segments ancestraux en SNP (ordonnées) chez les populations hybrides (abscisses). Le diagramme en barre (B) indique le niveau d'hétérozygotie (ordonnées) pour chacune des populations (abscisses). La carte thermique (C) indique le niveau de différenciation mesurée par la statistique F_{ST} entre chaque population (abscisses et ordonnées).

Le simulateur produit des statistiques sur la différenciation entre populations (F_{ST}), la mosaïque (taille des blocs) et l'hétérozygotie (Figure 2.3). Le premier panneau (Figure 2.3 A) montre bien que la population p_6 se reproduisant par propagation végétative présente des tailles plus grandes des blocs de mosaïque. La conséquence sur l'hétérozygotie du niveau élevé d'autofécondation de la population p_3 est fortement apparente sur la Figure 2.3 B, avec un niveau extrêmement faible. L'autofécondation a aussi renforcé la différenciation de la source, avec un F_{ST} plus important entre la population p_3 et les autres populations représentatives et

hybrides. De la même manière, le flux de la population p_4 vers la p_5 est visible sur la Figure 2.3 C avec une différenciation très faible entre ces populations. La mosaïque produite par la simulation est visible via un échantillon représenté par la figure 2.4. La population p_5 ayant reçu un flux continu de la p_4, il n'y a quasiment plus de blocs de la source s_3 (jaune) dans les 5 individus représentés, par rapport à la proportion de la source s_3 dans la population p_6. Les tailles des segments de la population p_6 sont bien plus importantes que celle des deux autres populations. De plus, les deux individus 33-34 et 35-36 montrent la même mosaïque, effet de la propagation végétative.

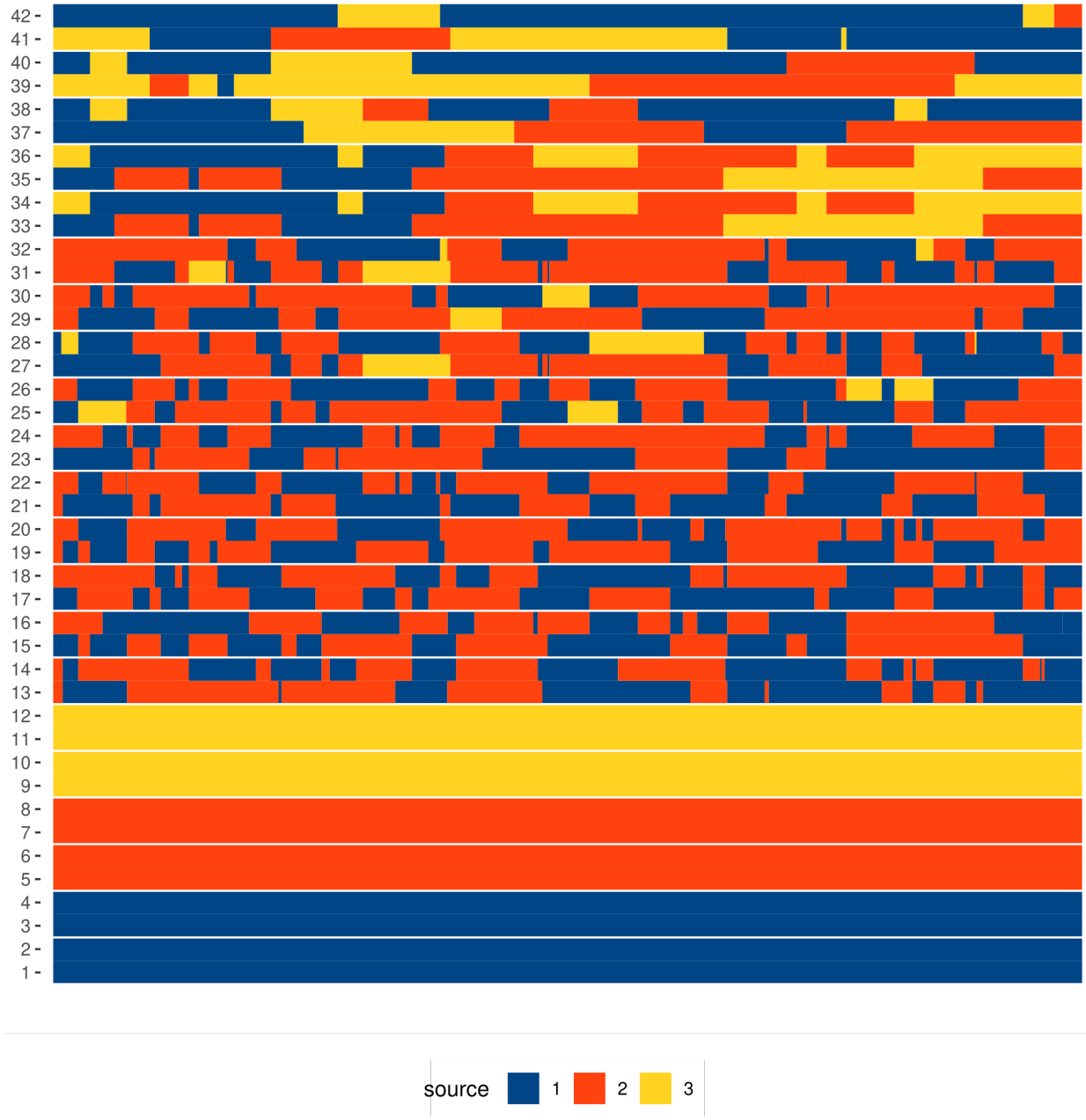

FIGURE 2.4 – Représentation des mosaïques d'origine ancestrale de la simulation de démonstration. Les sources s_1, s_2, s_3 sont représentées respectivement en bleu, rouge et jaune. Les numéros en ordonnées indiquent les haplotypes. Ici, les six premiers individus (haplotypes 1 à 12) appartiennent aux populations représentatives des sources p_1, p_2, p_3. Les 15 individus suivants (haplotypes 13 à 42) appartiennent respectivement aux populations p_4, p_5, p_6. La population p_4 est hybride des sources p_1 et p_2, et se reproduit par croisement intra-population. La population p_5 est hybride des trois sources, et se reproduit par croisement intra-population avec en plus des croisements inter-population avec la population p_4. La population p_6 est hybride des trois sources, et se reproduit par propagation végétative, précédée par 10 générations de croisements intra-population

2.5 Discussion

Globalement, le processus de simulation en « forward » du simulateur se comporte comme attendu sur nos tests simples. Le maintien de l'information génétique globale de la population semble être validé par les résultats sur l'hétérozygotie, montrant que l'information génétique globale est correctement transmise sur 50 générations sur une population de taille relativement petite (250 individus diploïdes). La structure locale des haplotypes semble être maintenues au fur et à mesure des générations, comme le montrent les mesures de DL par le r^2. Le DL à la génération g_{100} est répartie de manière similaire à l'attendu théorique et au DL de la génération g_1.

Le simulateur de données spécifiquement développé ici permet de produire de manière efficace des individus hybrides. L'outil est mis à disposition sous forme d'un paquet `R` (R Core Team, 2020) qui contient des fonctions sous forme de boîte à outils pour définir de manière très flexible les scénarios à simuler. Avec un petit nombre d'opérations, il est possible de simuler des populations diverses et d'obtenir des statistiques générales sur le jeu de données. Le simulateur a toutefois des limites.

D'un point de vue purement technique, le simulateur est codé en `R` pur et repose pour la partie « forward » sur des matrices qui sont découpées et rassemblées pour imiter la recombinaison. Cette logique simple a permis un développement rapide mais à un impact probablement fort sur la vitesse de calcul, et consomme beaucoup de mémoire. Il est difficile de simuler des populations de plus de 150 individus avec des nombres de marqueurs supérieurs à 30 000. Les optimisations possibles sont nombreuses (tel que l'utilisation d'un autre langage pour les parties intensives du code ou l'usage de structures de données et d'algorithmes différents).

Du point de vue génétique, le simulateur apporte une approche simpliste des mécanismes à l'œuvre dans les populations et les individus. Le simulateur a été pensé pour contraindre à une recombinaison par génération dans la partie en coalescence comme en « forward ». De plus, la probabilité de recombinaison est étendue de manière uniforme sur le « génome ». Il y a aussi des effets importants sur la diversité génétique du point de vue des modèles de plantes cultivées, comme par exemple les générations chevauchantes ou la sélection humaine sur des traits d'intérêt qui ne sont pas pris en compte ici. Cependant, il est tout à fait possible de lui fournir des données génétiques existantes pour travailler à partir de populations sources plus réalistes.

Chapitre 3

Évaluation par simulation de trois méthodes pour l'inférence ancestrale locale de génomes de plantes cultivées non-modèles

Ces travaux ont fait l'objet d'une publication le 6 février 2020 dans la revue G3 : Genes | Genomes | Genetics sous le titre 'Simulation-Based Evaluation of Three Methods for Local Ancestry Deconvolution of Non-model Crop Species Genomes' (10.1534/g3.119.400873), par Aurélien Cottin [†] [‡], Benjamin Penaud[†‡], Jean-Christophe Glaszmann[†‡], Nabila Yahiaoui[†‡] et Mathieu Gautier [§].

3.1 Résumé en français

Les méthodes d'ILEA ont majoritairement été construites, testées et utilisées dans le cadre de la génétique humaine (Hui et al., 2017; Geza et al., 2018). Elles permettent d'exploiter des jeux de données de taille importante (par exemple The International HapMap Consortium, 2005) et se sont focalisées sur des scénarios génétiques avec peu de populations impliquées (majoritairement 2 et 3). S'il existe bien des plantes cultivées dont les ressources génétiques sont importantes, comme le riz (The 3,000 rice genomes project, 2014), de nombreuses plantes cultivées non modèles n'ont pas de ressources génétiques accessibles comparables ni des données dont les phases sont identifiées. D'autres difficultés s'ajoutent à cela, par exemple le nombre de groupes ancestraux élevés ou encore des modes de reproduction non retrouvés dans la génétique humaine (autofécondation et propagation végétative).

Nous avons donc évalué la qualité des inférences produites par les méthodes d'ILEA sur des scénarios simulées incorporant des problématiques pouvant être retrouvées chez des plantes cultivées, en utilisant le simulateur présenté dans le chapitre 2. Les méthodes d'inférences

†. CIRAD, UMR AGAP, F-34398 Montpellier, France
‡. AGAP, Univ. Montpellier, CIRAD, INRAE, Montpellier SupAgro,Montpellier, France
§. INRAE, UMR CBGP, F-34988 Montferrier-sur-Lez Cedex, France

utilisées dans l'étude ont été choisies sur deux critères : la capacité de travailler avec plus de deux populations sources et avec des données non phasées. Les méthodes choisies sont `SABER` (Tang et al., 2006), `WINPOP` (Paşaniuc et al., 2009) et `ELAI` (Guan, 2014).

Les méthodes ont été paramétrées avec les paramètres exacts correspondant aux simulations (nombre de génération après l'hybridation, nombre de populations). L'écart entre les mosaïques simulées et inférées a été évalué avec une mesure de la précision α comprise entre 0 et 1, 1 correspondant à une mosaïque inférée identique à la mosaïque simulée. La consommation de ressources informatiques a été mesurée à travers l'usage de mémoire et du temps de calcul.

Six scénarios ont été mis en place pour générer des données simulées, qui ont été analysées par chaque méthode d'ILEA. Nous avons dans un premier temps testé dans un scénario à 3 populations sources, des paramètres classiques comme l'impact de la différenciation entre les populations sources, le nombre de générations après l'hybridation et la taille des échantillons des populations utilisées. Nous avons ensuite évalué l'impact d'un déséquilibre dans la taille des échantillons des populations représentatives, du nombre de populations ancestrales (de 3 à 6), et d'une population ancestrale manquante au moment de l'inférence. Enfin, nous avons testé l'impact sur les inférences, de la propagation végétative chez les hybrides et de l'autofécondation chez les représentants d'une population ancestrale.

Le temps depuis l'évènement d'hybridation et la différenciation entre les populations sont les deux paramètres qui ont montré le plus d'effet sur l'inférence des méthodes d'ILEA. Plus la différenciation est forte, plus les inférences sont précises (cela facilite la discrimination entre les populations) et plus le nombre de générations après l'hybridation est élevé, moins les inférences sont précises (les mosaïques sont plus fragmentées). La méthode `WINPOP` particulièrement adaptées aux populations sources peu différenciées s'est bien montrée comme étant la plus précise dans ces cas. C'est la méthode `ELAI` qui était la plus précise pour des nombres de générations plus élevés du moment que les populations ancestrales étaient bien différenciées. La taille d'échantillon des populations représentatives (avec un minimum de 5 individus par population source) a eu un impact limité du moment que les populations ancestrales étaient suffisamment différenciées. Le déséquilibre dans ces tailles d'échantillons a eu un impact plus important sur la précision des inférences pour `ELAI`. Un nombre plus élevé de populations ancestrales a eu lui un impact non négligeable sur la précision des méthodes `SABER` et `WINPOP`, et moindre sur la méthode `ELAI` mais les performances computationelles d'`ELAI` ont été affectées. L'autofécondation a un impact limité sur la précision des inférences, quand elle a lieu dans une population représentative. La propagation végétative rend plus difficile l'estimation du paramètre du temps depuis l'hybridation puisque qu'elle « fige » la structure génétique le long des générations. Les simulations avec propagation végétative ont été analysées avec un paramètre de temps depuis l'hybridation surestimé, et nous avons montré que cela avait un impact réduit sur la précision des inférences, sauf dans le cas de `SABER`. L'absence de représentants d'une population source

qui a contribué aux hybrides analysés est par contre un cas critique pour les méthodes d'ILEA testés. Les régions dérivant de cette population manquante ont été ici attribuées aux sources disponibles apparemment sans biais particulier d'assignation aux différentes populations présentes. Cette absence de biais était toutefois probablement due à une différenciation similaire entre toutes les populations simulées.

Globalement, il semble que les méthodes d'ILEA évaluées ici sont fiables pour des contextes de plantes cultivées, même avec plus de deux populations ancestrales à la condition que toutes les populations sources ancestrales soient représentées. La méthode `ELAI` parait la plus adaptée dans le cas de populations différenciées et bien représentées, et est aussi la plus précise dans les cas avec plus de trois populations ancestrales. La méthode `WINPOP`, prévue pour des populations proches et des évènements d'hybridations récents parait effectivement être la plus adapté dans ce type de cas.

Les figures supplémentaires sont présentes à la fin du manuscrit de thèse, correspondant aux figures S.1 à S.5. Les matériels supplémentaires (Fichiers S.1 à S.3) sont téléchargeables sur Figshare (`https://doi.org/10.25387/g3.10266149`).

3.2 Abstract

Hybridizations between species and subspecies represented major steps in the history of many crop species. Such events generally lead to genomes with mosaic patterns of chromosomal segments of various origins that may be assessed by local ancestry inference methods. However, these methods have mainly been developed in the context of human population genetics with implicit assumptions that may not always fit plant models. The purpose of this study was to evaluate the suitability of three state-of-the-art inference methods (`SABER`, `ELAI` and `WINPOP`) for local ancestry inference under scenarios that can be encountered in plant species. For this, we developed an R package to simulate genotyping data under such scenarios. The tested inference methods performed similarly well as far as representatives of source populations were available. As expected, the higher the level of differentiation between ancestral source populations and the lower the number of generations since admixture, the more accurate were the results. Interestingly, the accuracy of the methods was only marginally affected by i) the number of ancestries (up to six tested) ; ii) the sample design (i.e., unbalanced representation of source populations) ; and iii) the reproduction mode (e.g., selfing, vegetative propagation). If a source population was not represented in the data set, no bias was observed in inference accuracy for regions originating from represented sources and regions from the missing source were assigned differently depending on the methods. Overall, the selected ancestry inference methods may be used for crop plant analysis if all ancestral sources are known.

3.3 Introduction

Inter-(sub)-specific hybridizations have shaped the genomes of many crop species as for example in wheat (El Baidouri et al., 2017), rice (Zhao et al., 2010), citrus (Wu et al., 2014, 2018), banana (Perrier et al., 2011) or apple (Cornille et al., 2012). They can be a consequence of germplasm transport by humans bringing together plants from related but differentiated species, subspecies or populations, or of gene flow between cultivated plants and neighboring wild relatives (e.g. Semon et al., 2005; Perrier et al., 2011). At the genome level, such admixture events can result in a mosaic pattern of chromosomal segments of various origins. The complexity of the mosaic will depend on the demographic characteristics of the populations (e.g., number of source ancestries and the timing of admixture events). As already demonstrated in other species such as humans, characterizing the genome mosaic may in turn provide valuable insights into the genetic history of populations (e.g. Moreno-Estrada et al. (2013); Hellenthal et al. (2014), and in the context of domesticated plant species, it may lead to a better understanding of crop domestication and diversification history. It might also help identifying the origin of introgressed variants underlying agricultural traits of interest (Burgarella et al., 2019) and possibly support breeding strategies to produce improved hybrids.

Over the past twenty years, the development of high-density genotyping and sequencing technologies has promoted the development of accurate approaches to infer genetic ancestry of individuals based on genotyping data. Historically, the first proposed methods aimed at characterizing individual ancestries on a genome-wide scale by estimating the relative contributions of a given number of underlying ancestries. The most popular of these methods are based on unsupervised clustering approaches as introduced by Pritchard et al. (2000) in the Structure software, where clusters were interpreted as proxies for ancestries. An extension of the Pritchard's work by Falush et al. (2003), further allowed to perform local ancestry inference (LAI), i.e. to infer the ancestral origin at a local chromosome scale in individual genomes. Since then and as reviewed in Geza et al. (2018), more than 20 LAI methods have been published extending, in particular, this pioneering work by scaling up to high throughput genotyping data or by leveraging phased data for more accurate inferences.

Most LAI approaches have been developed in the context of human genetics studies for which their properties have been extensively characterized (Liu et al., 2013; Padhukasahasram, 2014; Hui et al., 2017; Geza et al., 2018). Human studies relying on LAI approaches usually aim at assessing admixture between two or three populations and may benefit from a rich amount of genetics resources (The International HapMap Consortium, 2005; The Wellcome Trust Case Control Consortium, 2007) with, in particular, dense haplotype data for reference populations and/or admixed samples. For plant species, large-scale sequencing or genotyping resources are also increasingly available for some crops such as rice (The 3,000 rice genomes project, 2014) or

barley (Milner et al., 2019). However, for many other species of interest such resources remain scarce which implies that haplotype data may not yet be fully accessible, particularly for non-autogamous species. In addition, ancestries may be multiple as exemplified by the cacao tree (*Theobroma cacao*) germplasm which is composed of 10 major genetically differentiated groups with up to six-way admixed individuals (Cornejo et al., 2018) or pineapple (*Ananas comosus*) with up to four-way admixed individuals between cultivar groups and varieties (Chen et al., 2019). Moreover, populations representative of contributing ancestries may be unavailable or represented by only a few individuals. To that respect, the case for banana is particularly illustrative since hybridization events involving well-differentiated *Musa acuminata* subspecies are predicted to be involved in the formation of some major cultivars (Perrier et al., 2009, 2011). Yet, some of these subspecies are represented only by a few individuals (Christelová et al., 2017) or some contributors may not be represented in available germplasm (Sardos et al., 2016a). Moreover, it remains unclear how some features, regarding reproduction modes (e.g., selfing or vegetative reproduction) that may be encountered in plant models, may affect the performance of LAI methods. Hence, in fruit crops such as citruses or banana, individuals resulting from inter(sub)specific hybridizations were further multiplied by vegetative propagation (also termed clonal propagation). Thus, they do not form a population but rather a collection of individuals sometimes of different origins with ancestry mosaics of relatively large blocs depending on the number of sexual generations they may have undergone. Datasets of vegetatively propagated individuals may thus be heterogeneous in terms of ancestry structure and in terms of time in generations since admixture events. The number of sexual generations since admixture is a parameter that is often required by LAI programs (Geza et al., 2018) and it can be difficult to correctly estimate it for plants that have been vegetatively propagated sometimes since hundreds of years. On the other hand, high selfing rates result in increased levels of homozygosity and generally in reduced diversity levels compared to outcrossing species (Brandvain et al., 2013; Barrett et al., 2014) while introducing additional levels of structuring of haplotype diversity when selfing and outcrossing populations are analyzed together. Finally, polyploidy that is a feature of many crop plants (e.g. wheat, sugarcane, potato, and major banana cultivars) is still a complex case to handle for LAI as genotypes are difficult to infer.

The purpose of this study was to evaluate the accuracy of LAI approaches to perform ancestry deconvolution, based on genotyping data simulated under scenarios that can be representative of plant species models. Given the lack of available methods dealing with polyploids, we only considered diploid individuals. Among the 22 LAI approaches recently reviewed by (Geza et al., 2018), we chose to evaluate three methods – `SABER` (Tang et al., 2006) ; `ELAI` (Guan, 2014) and `WINPOP` (Paşaniuc et al., 2009) because they do not require prior phasing of the data and they could cope with more than two ancestries. We developed an R package to perform simulations and we focused our evaluation on the influence of i) the level of divergence between the source populations ; ii) the number of generations since admixture for the

admixed populations ; iii) the number of contributing ancestries and their representation in the analyzed data sets ; and iv) the mode of reproduction such as selfing (in a source representative population) or vegetative propagation (in the admixed population).

3.4 Material & methods

3.4.1 Simulation tool

We developed an R package (named `plmgg` for plant-like mosaic genome generator) to simulate individual chromosome-wide genotyping data from an arbitrary number of populations P deriving from S differentiated source populations under scenarios that may include hybridization events and modes of reproduction representative of plant model evolution (i.e., selfing or vegetative propagation). The simulation approach is depicted in Figure 3.1 and consisted of the three following successive steps i) coalescent simulation of a sample of founder chromosomes from S differentiated source populations ; ii) forward in-time simulation of P populations deriving from the S source populations with complex demographic scenarios involving various modes of reproduction and admixture ; and iii) sampling of individuals from the P populations to generate the genotyping data sets. For coalescent simulations, we relied on the `scrm` algorithm (Staab et al., 2015) implemented in the R package `coala` (Staab and Metzler, 2016) to simulate $n_{\mathbf{S}}^{(h)} = \sum_{s=1}^{\mathbf{S}} n_s^{(h)}$ founder chromosomes (i.e., haploid (h) individuals) from S predefined source populations (where $n_s^{(h)}$ is the number of chromosomes from source population s). Sources were assumed to derive from a single ancestral population under a pure-drift model of divergence with a star-shaped history. The divergence scenario of the source populations was specified with three parameters : i) the divergence time τ measured in units of $4N_e$ (i.e., $\tau = \frac{t}{4N_e}$ where t is the number of generations since the ancestral population and N_e is the haploid effective source population size assumed to be the same for the S source populations) ; ii) the scaled mutation rate $\theta = 4N_e\mu$ (where μ is the mutation rate per site and per generation) ; and iii) the scaled recombination rate $\rho = 4N_e r$ (where r is the recombination rate per site and per generation). For the purpose of this study, τ was varied to control the level of differentiation among the source populations (see below) while both θ and ρ were set equal to 10^{-4} as obtained for instance if one assumes $\mu = 2.5 \times 10^{-8}$ mutation and $r = 2.5 \times 10^{-8}$ recombination per site and per generation in a population of haploid size $N_e = 10^3$).

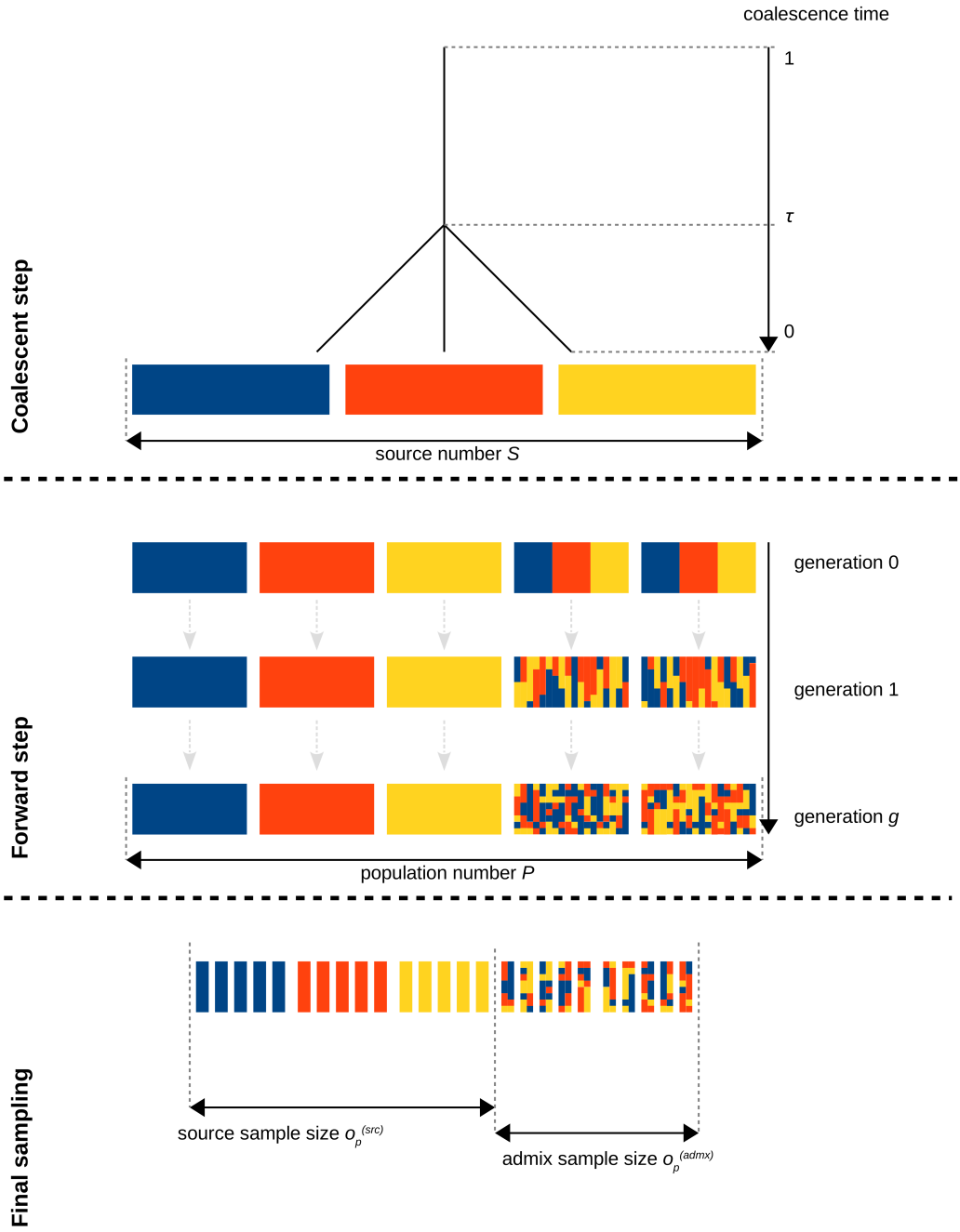

FIGURE 3.1 – Overview of the admixture simulation process with `plmgg`. The coalescent step produces S source populations (here, three sources are represented in blue, red and yellow) that differentiated at τ. In the forward step, source-representatives populations and admixed populations are generated from sampling of the source populations. Then, each population follows for a number of generations g, a user-defined reproduction process that allows to select and combine reproduction modes (within population random mating, across population random mating, selfing, vegetative reproduction). In the last step, a sampling is performed on each population of the forward step to generate a data set for analysis.

In the second simulation step, $n_P = \sum_{p=1}^{P} n_p$ diploid individuals belonging to P populations (where n_p is the number of diploid individuals from population p) were first generated by randomly sampling two chromosomes (without recombination and with replacement) among the $n_S^{(h)}$ founder ones according to the S pre-defined source contributions (in a $P \times S$ ances-

try proportion matrix). The n_P individuals were further reproduced over G generations, in a forward-in-time process, by specifying for each generation g (in a $P \times G$ matrix) the proportions of the four following possible population-specific modes of reproduction : i) within population random mating ; ii) across population random mating ; iii) selfing ; and iv) vegetative reproduction (consisting of randomly reproducing an individual's chromosome pair from one generation to the next). For sexual reproduction events (i.e., random mating and selfing), parental gametes were generated by randomly distributing one crossing-over between the two parental chromosomes which amounts to assume a 1 Morgan length chromosome map. In addition, mutations that only affected existing variant positions (switch to the alternate SNP allele) were introduced at each generation at the rate μ defined above (whatever the reproduction mode). In other words, no new segregating sites appeared after the initial coalescent phase of the simulation (first simulation step described above).

In the third and last step of the simulation, $o_P = \sum_{p=1}^{P} o_p$ diploid individuals belonging to the P populations were sampled to generate the data set to be analyzed (where o_p represents the number of genotyped diploid individuals from population p that are randomly sampled with replacement from the corresponding n_p individuals available at generation G). After filtering out monomorphic SNPs, the simulation output consists of i) the genotyping data set in vcf (Danecek et al., 2011) and `plink` ped (Purcell et al., 2007) ; ii) the true local ancestry for each individual at each SNP position which may be displayed with plotting functions ; and iii) summary statistics including pairwise population F_{ST} (Weir and Cockerham, 1984), population heterozygosities and ancestry block sizes.

3.4.2 Simulated scenarios

TABLE 3.1 – Summary of simulation parameters.

SIMULATION	τ	S	G	$P^{(src)}$	$P^{(admx)}$	$o_p^{(src)}$	$o_p^{(admx)}$	$Un(\%)$	$Slf(\%)$
DiffGenSam	**0.05 - 0.40**	3	**05-50**	3	2	**05 - 40**	40	0	0
SamBal	0.20	3	50	3	2	**2-20**[a]	40	0	0
SrcNum	0.20	**3-6**	50	**3-6**	**2**	**05**	40	0	0
SrcMiss	0.20	**4**	50	3[b]	**2**	**20**	40	0-15	0
SrcSelf	0.20	3	50	3	2	20	40	0	0-99
AdmxVegProp	0.20	3	50	3	10	20	**10**	**0**	**0**

τ, differentiation time ; S,number of sources ; G, number of generations after the admixture event ; $P^{(src)}$, numbers of source-representative populations ; $P^{(admx)}$,number of admixed populations ; $o_p^{(src)}$, sample size of source-representative populations ; $o_p^{(admx)}$,sample size of admixed populations ; *Un*, percentage of unknown source ; *Slf*, percentage of selfing in the third source-representative population.

[a] 2-20 indicate the variation of $o_p^{(src)}$ on the third source-representative population, while the two others are fixed at $o_p^{(src)} = 20$.

[b] $P^{(src)}$ here equals to $S - 1$ to simulate a missing source-representative population.

Six scenarios detailed in Table 3.1, each replicated 50 times, were considered for this study. The number of founder chromosomes was set to $n_s^{(h)} = 300$ for each source population (thereby mimicking bottlenecks involved by the domestication process from a small number of wild relatives) and the number of diploid individuals was set to $n_p = 150$ for all the populations (i.e., the source and admixed populations). Forward simulations were run for $G = 50$ generations maintaining S non-admixed populations as source population proxies and two populations originating from an admixture event between three or more ancestries that occurred from $t_{adm} =$ 5 to 50 generations ago. Unless otherwise stated, the sampled data set consisted of $o_p = 20$ diploid individuals for each ancestry representative population and $o_p = 40$ individuals for each admixed population. The scenarios were split into three groups to investigate the effect of i) the ancestry representative sample size ; ii) the number of sources ; and iii) the reproduction modes. First, the **DiffGenSam** scenarios (Table 1) aimed at evaluating the impact of the amount of differentiation between $S = 3$ source populations (with τ varying from 0.05 to 0.40) ; the number of generations since admixture for the two admixed populations (from $t_{adm} = 5$ to 50) ; and the sample size (from $o_p = 5$ to 40) of each of the three ancestry-representative populations. Five other scenarios were subsequently considered to address specific points while setting $\tau =$ 0.20 and $t_{adm} = 50$ (Table 3.1). The **SamBal** scenario aimed at evaluating the impact of unbalanced sample sizes among three ancestry representative populations (i.e., two with 20 sampled individuals and the remaining with 2 to 20 sampled individuals). We also considered two scenarios to address the impact of the number of source populations (**SrcNum** with $S = 3$

to 6 source populations equally contributing to the admixed populations) or the presence of a non-sampled source population contributing to the admixed individuals (**SrcMiss**). In the latter case, $S = 4$ source populations were simulated, but only three of them had representatives in the final data set. The contribution of the "missing" source population to the admixed populations varied from 0.05 to 0.15, the three other sources having equal contributions. Finally, the **SrcSelf** and **AdmxVegProp** scenarios aimed at investigating the impact of alternative modes of reproduction. In the **SrcSelf** scenario, we assumed that one of the three source populations was reproducing with a selfing rate varying from 0 (i.e., no selfing) to 0.99. The **AdmxVegProp** scenario modeled 10 admixed populations (with $n_p = 100$ and $o_p = 10$ for each admixed population) that switched from exclusive within population random mating to exclusive vegetative propagation t_{veg} generations ago, with t_{veg} varying from 0 (i.e. no vegetative propagation) to 45 for the 10 populations. Note that the realized number of SNPs (after filtering steps) in the different simulated data sets ranged from 10^4 to 2.7×10^4 (File S.1).

3.4.3 LAI methods

As mentioned in the introduction, we retained the three LAI methods respectively implemented in the programs `SABER`, `WINPOP` and `ELAI`, that do not require prior phasing of the data and that could cope with more than two ancestries. The two methods `SABER` (Tang et al., 2006) and `ELAI` (Guan, 2014) rely on Hidden Markov Models (HMM) with an explicit modeling of LD, while `WINPOP` (Paşaniuc et al., 2009) relies on a model-based LD-free approach.

More precisely, `SABER` (Tang et al., 2006) extended the HMM by Falush et al. (2003), to account for background LD existing in ancestral populations by modeling the joint distribution of alleles from consecutive markers within each ancestral population. In addition, `SABER` allows modeling an arbitrary number of ancestral groups that may admix at different times estimated by a Likelihood Maximization algorithm (`saberML` function, here initialized with the simulated values). Each individual SNP-specific ancestry estimates were calculated as the posterior probability obtained with the forward-backward algorithm implemented in the `pipeline` function.

`ELAI` (Guan, 2014) implements a two layers HMM to model two different scales of LD : the admixture LD (between alleles from different source populations) and a shorter ranged LD existing between alleles within each source population. This is achieved by introducing a local structuring of haplotypes into i) upper-layer clusters that represent different groups (interpreted as source populations) ; and ii) lower-layer clusters that represent group-specific haplotypes. We here set the number of upper clusters to the number S of simulated sources ; the number of lower clusters to $5S$ as recommended ; and the time since admixture (also required by `ELAI`) to the corresponding simulated one. Model fitting was carried out with the default

Expectation-Maximization (EM) algorithm.

`WINPOP`, included in `LAMP` >=2.3 (Sankararaman et al., 2008a), is a model-based LD-free method that focuses on ancestry informative markers (AIM) to assign local haplotype blocks to their originating source populations (Paşaniuc et al., 2009). `WINPOP` works with variable-size overlapping windows along the chromosomes, and uses a clustering method to assign ancestries in each window, based on estimates of global ancestry proportions. `WINPOP` was used with the simulated recombination rate and default parameters for the configuration files including a LD pruning cutoff of $r^2 = 0.1$ and a fraction of sliding window overlap of 20%.

Both `WINPOP` and `SABER` required estimates of global ancestry proportion. These were obtained by running the default unsupervised hierarchical clustering algorithm implemented in the `ADMIXTURE` software (Alexander et al., 2009) setting the number of clusters to S, the number of simulated source populations.

3.4.4 Evaluation of the performance of LAI methods

To evaluate the performance of the LAI methods, we defined an accuracy metric α to quantify the overall differences between simulated and inferred local ancestries. Let $z_{s,m}^{(i)}$ represent the simulated proportion of ancestry s ($s = 1, \ldots, S$) at SNP position m ($m = 1, \ldots, M$); M being the number of SNPs) for individual i ($z_{s,m}^{(i)}$ =0, 0.5 or 1 since individuals are diploid). Similarly, let $x_{s,m}^{(i)}$ represent the inferred proportion of ancestry. The SNP-specific accuracy for individual i over the S ancestries was then defined as $\alpha_m^{(i)} = 1 - \sum_{s=1}^{\mathbf{S}} \frac{\left| z_{s,m}^{(i)} - x_{s,m}^{(i)} \right|}{2}$. Note that $\sum_{s=1}^{\mathbf{S}} \left| z_{s,m}^{(i)} - x_{s,m}^{(i)} \right|$ is bound between 0 (when $z_{s,m}^{(i)} = x_{s,m}^{(i)}$ for all s) and 2 (when $z_{s,m}^{(i)} = 0$ or $x_{s,m}^{(i)} = 0$ for all s since $\sum_{s=1}^{\mathbf{S}} z_{s,m}^{(i)} = 1$ and $\sum_{s=1}^{\mathbf{S}} x_{s,m}^{(i)} = 1$), hence the division of this sum by 2 in the definition of the accuracy $\alpha_m^{(i)}$ to keep $\alpha_m^{(i)}$ between 0 and 1. The overall accuracy metric α was defined as the average of $\alpha_m^{(i)}$ over the M markers and the individuals belonging to admixed populations (i.e., excluding individuals belonging to source representative populations). For the particular case of **SrcMiss** simulations (Table 3.1) in which one source representative population was missing, SNP positions with the corresponding missing ancestry were excluded from the computation of α. According to our definition, α always lies between 0 and 1 (the higher the α value, the more accurate the inference). For calibration purposes, we also computed a minimal value of α as would be obtained by randomly inferred local ancestries under the assumptions of equal contribution of the sources (i.e., setting $x_{s,m}^{(i)} = 1/S$ for all s). Alternative metrics, such as the coefficient of determination (i.e., sample correlation coefficient between the inferred and true local ancestries) or mean square errors were also evaluated but were not presented since they lead to the same conclusions regarding the ranking of LAI methods.

We finally evaluated computational efficiency of the different LAI programs by recording for

each run of analysis on our computer grid, both the memory usage and the system computing time (`max_vmem` and `ru_wallclock`, respectively) available from the Sun Grid Engine user notification.

3.4.5 Data availability

The R package `plmgg` is available at `https://gitlab.southgreen.fr/acottin/plmgg`. Scripts used to run simulations, LAI programs and inference comparison are available at `https://gitlab.southgreen.fr/acottin/lai-comparison`. Supplemental material is available at `https://gsajournals.figshare.com/s/a9a44d6fb3317f41cb11`.

3.5 Results

3.5.1 Source differentiation and number of generations since admixture

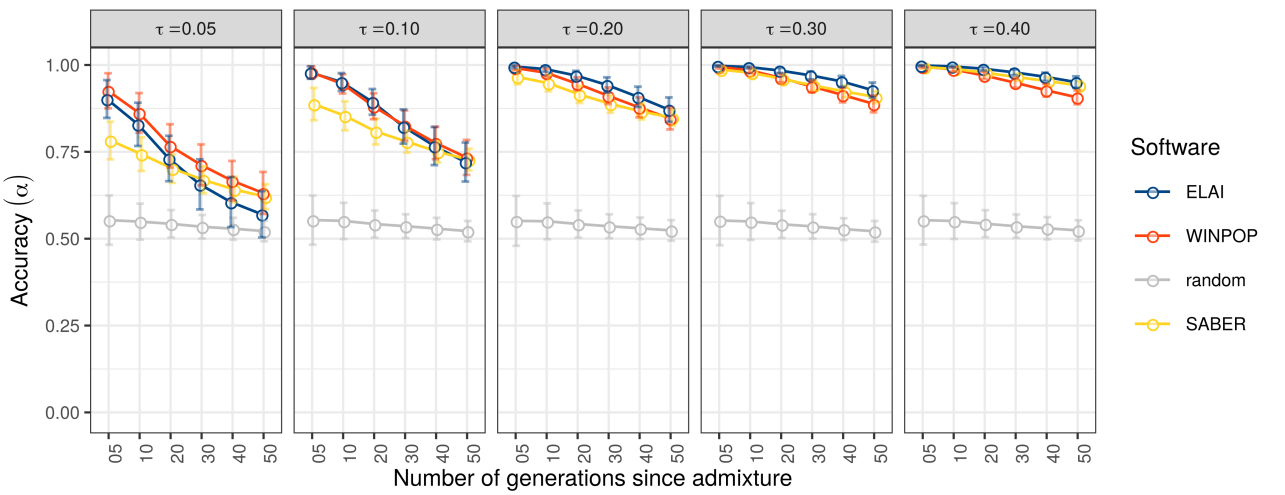

FIGURE 3.2 – Accuracy of LAI methods with varying levels of differentiation and number of generations (**DiffGenSam** simulation). The accuracy (α) of the LAI methods (y-axis) is plotted for different levels of differentiation that vary from 0.05 to 0.4 (vertical tiles) and a number of generations after admixture that varies from 5 to 50 (x-axis). The sample size is set to 20 for the sources and the admixed populations. Each dot is the mean value of 50 repetitions of each simulation. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` scores are plotted in blue, red and yellow, respectively. Accuracy of random inference (proportion of ancestry fixed at 1/3) is plotted in gray.

The impact of the level of differentiation among the sources and the number of generations since admixture on the performance of the three LAI methods was assessed with the **DiffGenSam** scenarios (Table 3.1). The analysis of the generated data sets showed that both the level

of differentiation among sources and the number of generations since admixture had a strong impact on the performance of LAI methods (Figure 3.2, File S.2). Indeed, the accuracy α decreased with an increasing number of generations after admixture (i.e., when ancestry block sizes became smaller) and with decreasing levels of differentiation between source populations (Figure 3.2). Although, the three evaluated LAI approaches performed overall similarly, at the lowest levels of differentiation ($\tau \leq 0.10$), `ELAI` and `WINPOP` were more accurate than `SABER` for more recent admixture events ($t_{adm} \leq 20$) (Figure 3.2, File S.2). In the most favorable situations of high differentiation among the source populations (i.e., $\tau \geq 0.3$), the accuracy α tended towards 1 (i.e., no error) with decreasing time since admixture for all the three LAI methods.

3.5.2 Number of individuals from the source representative populations

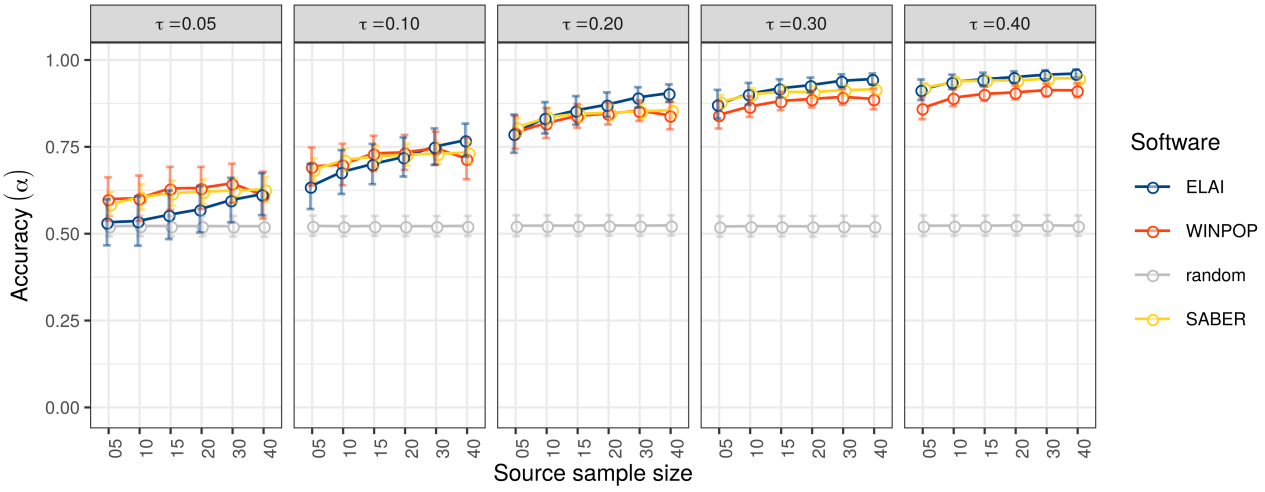

FIGURE 3.3 – Accuracy of LAI methods with varying levels of differentiation and source-representative sample size (**DiffGenSam** simulation). The accuracy (α) of the LAI methods (y-axis) is plotted for different levels of differentiation that vary from 0.05 to 0.4 (vertical tiles) and the size of source-representative sample that varies from 5 to 40 individuals (x-axis). The source sample size is set to 20. Each dot is the mean value of 50 repetitions of each simulation. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` scores are plotted in blue, red and yellow, respectively. Accuracy of random inference (proportion of ancestry fixed at 1/3) is plotted in gray.

The impact of the number of sampled representative individuals for each of the three source populations was also evaluated within the **DiffGenSam** scenarios (Table 3.1). As shown in Figure 3.3, for a given time since admixture (here $t_{adm} = 50$, see Figure S.1 and File S.2 for alternative t_{adm} values) decreasing the number of individuals representative of the source populations (e.g., from $o_p^{(s)} = 20$ as in Figure 3.2 to $o_p^{(s)} = 5$) had a higher impact on accuracy for `ELAI` compared to `WINPOP` and `SABER`. Conversely, except for the highest level of differentiation

among source populations, increasing the number of source representative individuals improved ELAI performances.

We evaluated the robustness of the three LAI methods to unbalanced sample sizes of source-representative populations by analyzing data sets simulated under the **SamBal** scenarios where the number of samples was reduced for one of the three sources (Figure S.2, File S.3). The accuracy of ELAI was lower than both the WINPOP and SABER when sampling was reduced for the third source (e.g. for 2 representatives instead of 20, accuracy of 0.720 for ELAI vs. 0.815 for WINPOP and 0.806 for SABER, File S.3) but it increased when sampling was more balanced reaching accuracy of 0.870 for a completely balanced setting. For WINPOP and SABER the accuracy was only marginally improved, reaching up to 0.850.

According to the results above, to allow better discrimination of the LAI methods in relatively challenging conditions, we chose to perform the remaining evaluations with a number of generations after admixture set to 50, a level of differentiation among sources of $\tau = 0.2$ and 20 individuals per source representative population.

3.5.3 Number of source populations and absence of source representative individuals

TABLE 3.2 – Accuracy of LAI methods with varying number of source populations (SrcNum scenario).

S	ELAI	WINPOP	SABER	RANDOM
3	$0.788 \pm 1.7\ 10^{-3}$ (0.055)	$0.792 \pm 1.5\ 10^{-3}$ (0.048)	$0.806 \pm 0.8\ 10^{-3}$ (0.026)	$0.523 \pm 0.9\ 10^{-3}$ (0.030)
4	$0.765 \pm 1.6\ 10^{-3}$ (0.053)	$0.758 \pm 1.4\ 10^{-3}$ (0.044)	$0.768 \pm 0.9\ 10^{-3}$ (0.030)	$0.411 \pm 0.7\ 10^{-3}$ (0.022)
5	$0.758 \pm 1.6\ 10^{-3}$ (0.053)	$0.730 \pm 1.5\ 10^{-3}$ (0.047)	$0.737 \pm 1.0\ 10^{-3}$ (0.032)	$0.336 \pm 0.6\ 10^{-3}$ (0.018)
6	$0.767 \pm 1.5\ 10^{-3}$ (0.050)	$0.723 \pm 1.5\ 10^{-3}$ (0.048)	$0.717 \pm 1.0\ 10^{-3}$ (0.032)	$0.285 \pm 0.5\ 10^{-3}$ (0.015)

Mean accuracy α, accuracy confidence interval (0.95) and accuracy standard deviation (sd) of ELAI, WINPOP and SABER on simulated data with variation on the number of source (S) from 3 to 6 are indicated. Simulations were conducted with 50 repetitions, $\tau = 0.2$, 50 generations after admixture and 20 individuals sampled from each population. Random inference ($1/S$ for each ancestry) was evaluated like LAI methods.

With the **SrcNum** scenarios (Table 3.1), data sets were simulated for admixture events involving up to six source populations. The analysis of LAI results showed that the accuracy decreased with increasing numbers of sources for all three evaluated LAI approaches (Table 3.2). However, the magnitude of decrease in accuracy from $S = 3$ to $S = 6$ source populations remained moderate with rates equal to 2.7%, 8.7% and 11% for ELAI, WINPOP and SABER respectively (to be compared with the 45% decrease observed with the random inference) (Table 3.2). We further assessed the impact of the absence of individuals from one out of four source

representative populations using data sets simulated under the **SrcMiss** scenario (Table 3.1). Different proportions of this unrepresented source to admixed populations were tested (5, 10 and 15%) and accuracy was measured by excluding regions contributed by the missing source population. As shown in Table 3.3, the accuracy for all methods was stable in regions without the unknown ancestry, whatever the global proportions (at a data set level) of unknown ancestry. This suggested that the absence of individuals from a source representative population in the analyzed data sets did not introduce biases in inferring local ancestries of the represented source populations. Visual inspection of local ancestries inferred in regions containing the missing ancestry did not reveal any particular pattern (e.g., like a higher switching rate among the other represented ancestries). As an example, Figure 3.4 shows the inferred local ancestry mosaic of one individual from a simulated data set with a 10% contribution of the unrepresented source population. In general, the chromosomal regions originating from the missing source population tended to be assigned to different represented ancestries, the assignation also varying according to the LAI method used.

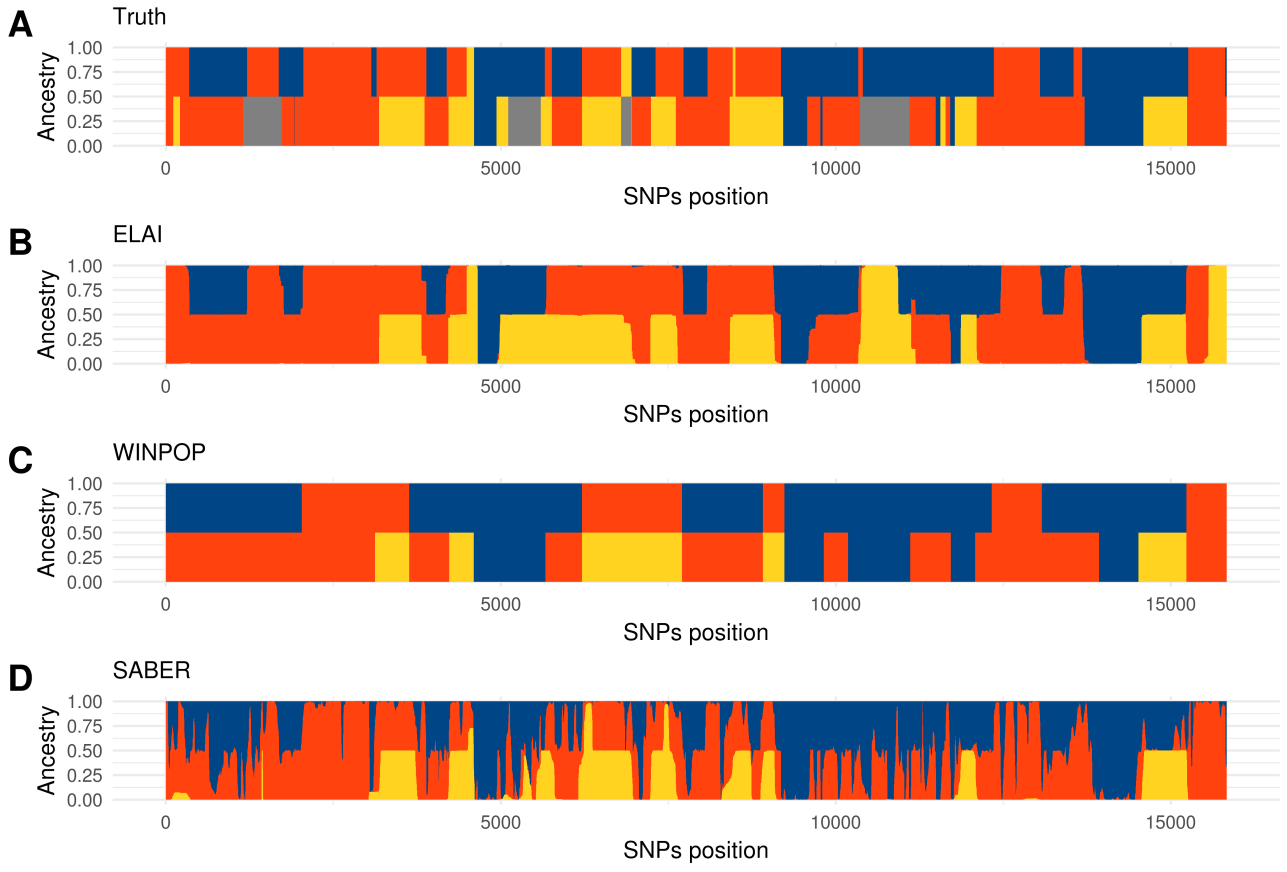

FIGURE 3.4 – LAI results for a simulated admixed individual of the **SrcMiss** simulation. The x-axis indicates simulated SNPs positions and the y-axis represents the stacked ancestry proportions (between 0 and 1). (**A**) represents the true ancestries. (**B**), (**C**) and (**D**) represent the inference from `ELAI`, `WINPOP` and `SABER`, respectively. The three known sources are shown in blue, red and yellow and the unknown source in gray.

TABLE 3.3 – Accuracy of LAI methods with varying proportions of an unknown source population in admixed populations (SrcMiss scenario).

Un(%)	ELAI	WINPOP	SABER	RANDOM
0	0.869 ± 1.2 10^{-3} (0.038)	0.843 ± 1.0 10^{-3} (0.033)	0.846 ± 0.7 10^{-3} (0.023)	0.523 ± 0.9 10^{-3} (0.030)
5	0.863 ± 1.3 10^{-3} (0.042)	0.844 ± 1.1 10^{-3} (0.035)	0.846 ± 0.8 10^{-3} (0.024)	0.522 ± 1.0 10^{-3} (0.031)
10	0.860 ± 1.4 10^{-3} (0.044)	0.842 ± 1.2 10^{-3} (0.038)	0.844 ± 0.8 10^{-3} (0.026)	0.520 ± 1.0 10^{-3} (0.033)
15	0.855 ± 1.5 10^{-3} (0.047)	0.844 ± 1.2 10^{-3} (0.040)	0.844 ± 0.8 10^{-3} (0.026)	0.518 ± 1.0 10^{-3} (0.035)

Mean accuracy α, accuracy confidence interval (0.95) and accuracy standard deviation (sd) of `ELAI`, `WINPOP` and `SABER` on simulated data with different percentage of a fourth source population participating to the admixture event are indicated. Accuracy was computed after removal of the unknown population segments in the admixed individuals, to measure the impact on well represented segments. Simulations were conducted with 50 repetitions, $\tau = 0.2$, 50 generations after admixture and 20 individuals sampled from each population. Random inference ($1/S$ for each ancestry) was evaluated like LAI methods.

3.5.4 Selfing and vegetative propagation

TABLE 3.4 – Accuracy of LAI methods with varying proportions of selfing in a source-representative population (SrcSelf simulation).

Slf(%)	ELAI	WINPOP	SABER	RANDOM
0	0.871 ± 1.1 10^{-3} (0.035)	0.846 ± 1.0 10^{-3} (0.031)	0.848 ± 0.7 10^{-3} (0.023)	0.524 ± 0.9 10^{-3} (0.030)
25	0.873 ± 1.1 10^{-3} (0.035)	0.836 ± 1.1 10^{-3} (0.037)	0.844 ± 0.7 10^{-3} (0.024)	0.523 ± 0.9 10^{-3} (0.029)
50	0.870 ± 1.2 10^{-3} (0.038)	0.831 ± 1.2 10^{-3} (0.038)	0.838 ± 0.8 10^{-3} (0.024)	0.524 ± 0.9 10^{-3} (0.030)
75	0.864 ± 1.2 10^{-3} (0.040)	0.810 ± 1.3 10^{-3} (0.042)	0.828 ± 0.8 10^{-3} (0.026)	0.523 ± 0.9 10^{-3} (0.029)
99	0.863 ± 1.2 10^{-3} (0.039)	0.790 ± 1.8 10^{-3} (0.058)	0.818 ± 1.0 10^{-3} (0.032)	0.522 ± 0.9 10^{-3} (0.029)

Mean accuracy α, accuracy confidence interval (0.95) and accuracy standard deviation (sd) of `ELAI`, `WINPOP` and `SABER` on simulated data with variation on selfing proportion in the third source-representative population are indicated. Simulations were conducted with 50 repetitions, $\tau = 0.2$, 50 generations after admixture and 20 individuals sampled from each population. Random inference ($1/S$ for each ancestry) was evaluated like LAI methods.

Table 3.4 gives the accuracy of the different LAI approaches on data sets simulated under the **SrcSelf** simulation (Table 3.1) in which the third source representative population reproduced with a varying extent of selfing. For the three LAI approaches, increased proportions of selfing in the third source representative population resulted in a decrease of accuracy, to a small extent. Indeed the decrease in accuracy between rates of selfing of 0 and 99% was equal to 0.92%, 6.6% and 3.5% for `ELAI`, `WINPOP` and `SABER`, respectively.

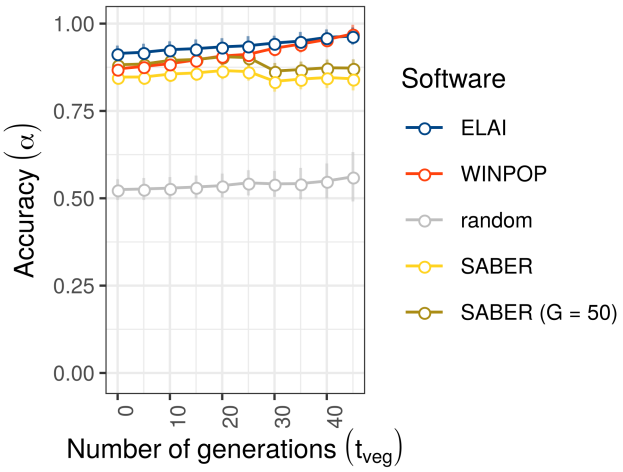

FIGURE 3.5 – Accuracy of LAI methods with varying number of generations of vegetative propagation (**AdmxVegProp** simulation). The accuracy of the LAI methods (y-axis) is plotted for different numbers of generations of vegetative propagation after admixture (t_{veg}) that vary from 0 to 45 (x-axis). The source sample size is set to 20, the differentiation set to 0.2 and the total number of generations after the admixture event set to 50. Each dot is the mean value of 50 repetitions of each simulation. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` scores are plotted in blue, red and yellow, respectively. `SABER` score with fixed number of generations after admixture is plotted in darker yellow. Accuracy of random inference (proportion of ancestry fixed at 1/3) is plotted in gray.

Figure 3.5 plots the accuracies of LAI approaches estimated on data sets simulated under the **AdmxVegProp** scenarios (Table 3.1) consisting of individuals from three source representative populations and 10 admixed populations that switched to an exclusive vegetative propagation mode t_{veg} generations ago (t_{veg} varying from 0 to 45 for the different populations). Note that, the larger t_{veg}, the larger the ancestry block sizes (since the smaller the number of post-admixture recombinations). For both `WINPOP` and `ELAI` based inference, the accuracy increased for increasing values of t_{veg} as expected given larger ancestry block sizes. However, the accuracy of `SABER`, being very similar for individuals with $t_{veg} = 45$ and $t_{veg} = 0$, was mostly not influenced by t_{veg}, although a slight decrease was observed at $t_{veg} = 30$. As this decrease appeared for higher numbers of generations of vegetative propagation, it may be linked to the fact that `SABER` performs its own estimation of time since admixture. To investigate this, a second run of `SABER` was performed without using the time since admixture estimation method (`saberML` function), but with a time since admixture fixed at $t_{adm} = 50$ as for `WINPOP` and `ELAI` (Figure 3.5). `SABER` accuracy was found higher with this fixed number of generations but a decrease at $t_{veg} = 30$ was still observed.

3.5.5 Computational performances of LAI methods

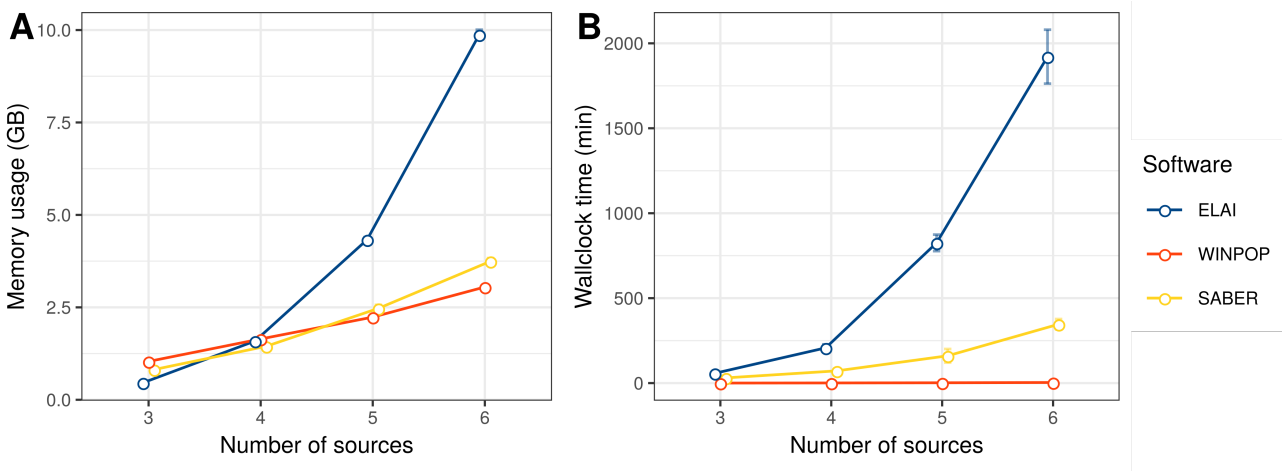

FIGURE 3.6 – Memory usage and computation time of LAI methods with varying number of sources (**SrcNum** simulation). (**A**) Memory usage of LAI methods in giga bytes (y-axis). (**B**) Wallclock time of the LAI methods in minutes (y-axis). The x-axis represents the number of simulated source populations in the **SrcNum** scenario. Each dot is the mean value of 50 repetitions of each simulations. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` performance in memory (**A**) and time (**B**) are plotted in blue, red and yellow, respectively.

Computational performance was measured for all the analyses performed on the simulated data sets. For the **DiffGenSam** scenario ($S = 3$ sources), memory consumption for the different methods ranged from 0.5Gb to 2Gb of RAM and was not highly variable across scenario variations (Figure S.3). `WINPOP` was the fastest of the three LAI methods with a mean running time ranging from 20 s to 60 s in the **DiffGenSam** data sets (with three source representative populations) while `SABER` runs lasted between 30 min and 60 min and `ELAI` runs between 50 min and 4h (Figure S.4). The analysis of data sets simulated under the **SrcNum** scenarios showed that the number of sources had the most significant impact on resource consumption (Figure 3.6), particularly for `ELAI` that used up to 10GB and 30h with $S = 6$ sources. This corresponded to a 20-fold memory and a 38-fold computing time increases as compared with $S = 3$ sources, (Figure 3.6) whereas the overall number of individuals (70 versus 55) and the number of SNPs remained similar. Although memory usage increased steadily for `WINPOP` and `SABER` (from 0.7GB to 3.75GB), the computing time remained low for `WINPOP` (20s to 3min20s) and intermediary (up to 5h) for `SABER`.

3.6 Discussion

The approaches evaluated in this study (implemented in the `SABER`, `WINPOP` and `ELAI` programs) were mostly developed for applications in human populations. The purpose of our study was to carry out a detailed evaluation of the accuracy of these three LAI approaches on data simulated under scenarios with features that may be encountered in studies of plant domestication or diversification involving admixture. For instance, the three methods we considered here were originally tested on data simulated by resampling haplotypes from two to three human populations in scenarios consisting of two-way or three-way admixture with up to a few tens generations post-admixture and including from 100 to 200 genotyped individuals per source representative populations in the analysis (Tang et al., 2006; Paşaniuc et al., 2009; Guan, 2014).

We developed an R package (`plmgg`) to simulate genotyping data under a wider range of scenarios and sample designs that include plant-like features. Even if this simulator has some limitations (it does not simulate recombination hotspots, multiple recombination per chromosomes nor selection), it allowed us to assess the influence on LAI accuracy of the level of differentiation, of multiway admixture with up to six ancestries and of limited sampling of source populations. In addition, the impact of two plant reproduction modes was also evaluated : selfing (in a source representative population) and vegetative propagation (in the admixed population).

Overall, the two main factors that contributed to improve accuracy of all the three tested LAI approaches were the level of divergence between source populations (the higher, the better) and the number of generations since admixture (the smaller, the better) which was not surprising given their expected influence on the complexity of genome mosaics. Indeed, due to both mutations and recombination, divergence between source populations leads to increased differences among their originating haplotypes that facilitates their discrimination. Similarly, increasing the number of generations since admixture, results in shorter ancestral chromosome segment tracks, which are then more difficult to identify. However, it should be noticed that in scenarios with the most extreme level of differentiation among the source populations we considered here ($\tau = 0.4$ which corresponds to a $F_{ST} \simeq 1 - e^{-\tau} \simeq 0.33$ in the pure-drift model of divergence we simulated), LAI accuracy remained acceptable even for the oldest admixture events (50 generations since admixture). In Citrus, average F_{ST} values of 0.44 up to 0.85 were found between the four ancestral taxa depending on studies or marker types (Curk et al., 2015, 2016). In the cacao tree or in pineapple, pairwise F_{ST} ranges between genetic groups were of 0.16 to 0.65 (Cornejo et al., 2018) and 0.28 to 0.94 (Chen et al., 2019), respectively. The lowest part of these ranges are covered in our simulations and higher values of F_{ST} will actually facilitate LAI even with older admixture events. For closely related source populations, LAI approaches only performed well if admixture events were very recent (i.e. below 10 generations). The three methods tested behaved roughly similarly, although `WINPOP` tended to be superior

when source populations were more closely related whereas for more differentiated sources and between 20 and 50 generations after admixture, ELAI tended to be more accurate. This result was consistent with the WINPOP paper (Paşaniuc et al., 2009) that showed that WINPOP performed well with closely related populations, with its improved modeling of recombination and adaptive window length that takes into account local genetic distances between ancestral populations. As for ELAI, its two-layer HMM model helps resolving short ancestry segments that can result from increasing generation numbers after admixture (Guan, 2014). In practice, differentiation among the source populations may be estimated with genotyping data available in the source representative individuals even when few individuals are available (Willing et al., 2012).

The timing of admixture events, required by both ELAI and WINPOP, may also represent in practice a parameter difficult to provide, especially for populations reproducing with vegetative propagation. Also, as we fixed this parameter to its true simulated value when running ELAI and WINPOP programs, our evaluation of these two methods may be overly optimistic. Yet, results obtained on the **AdmxVegProp** scenarios that include several generations of vegetative propagations suggests that both ELAI and WINPOP remain robust to (at least) upwardly biased estimates of the timing of admixture. In practice however, it may be valuable to check the sensitivity of the results obtained with these methods to a biologically sound range of (exponentially) varying values for this parameter. On the other hand, the timing of admixture events may also be estimated as proposed in the SABER framework. We nevertheless observed that in our settings the SABER estimations were inaccurate (see Figure S.5) which suggests in turn that LAI relying on SABER is also robust to biased estimates of the timing of admixture events. Other approaches may thus be preferable to that end, for example those modeling LD decay on a whole genome basis providing sampling allows it (e.g. Loh et al., 2013). Recently, Chen et al. (2019) estimated an average of 37 generations since the onset of admixture events for 22 (primarily) vegetatively propagated pineapple (var. *comosus*) hybrids, with a range of 21-55 generations.

Interestingly, we found that selfing (in a source representative population) or vegetative propagation (in the admixed population) had only a small impact on the inference accuracy. Selfing in a source population is of particular interest for banana as one of the *M. acuminata* subspecies contributing to banana hybrids is predicted to be frequently self-pollinated (Simmonds, 1962). Reproduction by vegetative propagation is favored for many fruit tree crops (Miller and Gross, 2011). Depending on the number of generations of sexual reproduction after admixture, vegetative propagation of admixed individuals can result in different levels of fragmentation of the mosaic structures. As mentioned above, this type of setting, with an overestimation of the generation number parameter had a minor impact on both ELAI and WINPOP, but a more notable impact on SABER inference for individuals where the overestimation was the highest.

Increasing the number of source populations (up to six tested) only marginally affected the accuracy of the tested LAI methods, particularly for `ELAI`. Nevertheless, this also increased the computational burden that became substantial for the `ELAI` program, presumably due to the higher number of model parameters. Hui et al. (2017) developed a tool (LAIT) to run four LAI methods including `WINPOP` and `ELAI` on a data set. They used LAIT to compare LAI methods on two-way and three-way admixture, and showed that `ELAI` performed better than `WINPOP` at the cost of increased resources consumption, which is consistent with our results.

Our results also showed that LAI methods perform similarly well for moderate to high levels of differentiation among source populations, even when the number of source representative individuals is small, which may have favorable practical consequences as it is not always possible to have access to large numbers of source representatives. Yet the three different methods behaved differently given an unbalanced data set, with a minor impact on `SABER` and `WINPOP` compared to `ELAI`. This may be explained by the two layers models of `ELAI` that ties haplotypes structure to ancestries, so that clustering will be hindered by low haplotypic variability. More generally, and in practice, assessing the number of source populations and assigning individuals to them might not be an easy task. Unsupervised clustering approaches (Pritchard et al., 2000; Alexander et al., 2009; Frichot et al., 2014) might be viewed as a reference choice (Stift et al., 2019) provided the source populations are differentiated enough and evenly represented in the data set (Puechmaille, 2016). The `Chromopainter` method (Lawson et al., 2012) allows to determine ancestry sources without individuals assigned as source-representatives, provided that phased data are available.

A most critical issue regarding LAI performances was the absence of representative individuals for a given source. The results obtained on the **SrcMiss** simulations showed no particular bias in attributing the missing population to known ancestries. This result may come from the fact that in our simulation the population tree between the four sources is star shaped. In practice, a star shaped tree is uncommon, one known population may be closely related to the missing population and bias cannot be excluded in this case. Some empirical and specific sampling procedures have been proposed to circumvent the absence of source representatives, in the case of large proportions of unrepresented ancestry in admixed populations (Zhou et al., 2016). Recently, a promising and more generic alternative has been developed in the `MOSAIC` model of Salter-Townshend and Myers (2019) for haplotype data, which allows for extracting information on source populations from related (and possibly admixed) individuals. Yet, phased data that we purposely kept out of consideration may not be accessible for many crop species. Moreover, it has been shown that switch errors that can occur with statistical phasing (Scheet and Stephens, 2006; Browning and Browning, 2011) reduce LAI accuracy (e.g. Guan, 2014). However, haplotype-based LAI approaches such as `RFMix` (Maples et al., 2013), `LOTER` (Dias-Alves et al., 2018) and `MOSAIC` (Salter-Townshend and Myers, 2019) that included switch error modeling demonstrated that, if properly modeled, inaccurate phasing is becoming less of

a threat for LAI accuracy.

LAI on phased data may also be particularly well suited to deal with polyploidy, ploidy being highly variable in crop species (e.g. pineapple 2x, cacao tree 2x, banana 2x and 3x, citrus up to 4x, sugarcane up to 12x) although statistical phasing might be challenging. Alternatively, HMM-based methods such as those proposed by (Corbett-Detig and Nielsen, 2017) for Pool-Seq data may also be of value.

The evaluation of LAI methods accuracy and performance with the `plmgg` R package, showed that LAI methods are usable in the scope of crops genetics, with caution particularly in case of a missing source population. The software `WINPOP` seems suited when source populations are close and admixture events recent. `ELAI` could be particularly adapted for well differentiated and relatively well represented sources, in case of selfing in source populations, for vegetative propagation settings, and multiway admixture although for the latter, computational performance might be a limiting factor. Other parameters more specific to different plant/crop models might be evaluated using the `plmgg` package.

3.6.1 Acknowledgments

This work was supported by funding from the Agropolis Fondation 'GenomeHarvest' project (ID 1504-006) through the French 'Investissements d'avenir' program (Labex Agro :ANR-10-LABX-0001-01) and by CIRAD. We thank Franck Curk, Gilles Costantino, Benjamin Heuclin, Joao D Santos, Guillaume Martin for fruitful discussions and the PIs of the 'GenomeHarvest' project, Manuel Ruiz and Angélique D'Hont. We also thank the two anonymous reviewers for their constructive comments. This work was also supported by the CIRAD—UMR AGAP HPC Data Center of the South Green Bioinformatics platform (`http://www.southgreen.fr/`).

Author contributions : A.C. and B.P. developed the simulation program ; A.C. performed simulation experiments ; J.C.G. contributed to project conception ; A.C., N.Y. and M.G. conceived the project, analyzed results and wrote the paper.

Chapitre 4

Application de l'inférence ancestrale locale à l'analyse de la mosaïque de génomes de bananiers diploïdes

4.1 Contexte

L'utilisation de méthodes d'ILEA a été montrée fiable sur des données simulées avec des scénarios qui pouvaient être rencontrés chez des plantes cultivées non modèles. Elles sont donc a priori adaptées au traitement de données de type plante. L'objet de ce chapitre est l'exploration de la structure de la diversité de bananiers diploïdes avec des données de reséquençage et l'application des méthodes ILEA à un jeu de données réel.

Rappelons ici que la majorité des cultivars de bananiers est issue d'hybridations entre des sous-espèces de *M. acuminata* (génome A) ainsi qu'avec *M. balbisiana* (génome B) (Simmonds and Shepherd, 1955; Perrier et al., 2011). L'espèce *M. acuminata* a été subdivisée en 8 à 10 sous-espèces interfertiles. Des proximités génétiques fortes entre les sous-espèces *burmannica/siamea/burmannicoides* ou entre *banksii/errans/microcarpa* (Carreel et al., 1994; Perrier et al., 2009; Dupouy et al., 2019; Martin et al., 2020) ont été identifiées et trois principaux groupes ayant contribué aux génomes A des bananiers cultivés ont été proposés : *M. a.* ssp. *banksii* (Papouasie-Nouvelle-Guinée), *M. a.* ssp. *malaccensis* (Péninsule malaise) et *M. a.* ssp. *zebrina* (Java, Indonésie) (Carreel et al., 1994, 2002; Boonruangrod et al., 2009; Perrier et al., 2009, 2011). Des hybrides parthénocarpiques entre *M. schizocarpa* (génome S) et *M. acuminata* ont été identifiés (hybrides AS, Carreel et al., 1994). Les trois espèces *M. acuminata*, *M. balbisiana* et *M. schizocarpa* font partie d'une section du genre *Musa* appelée *Eumusa* et il existe une diversité d'espèces regroupées dans d'autres sections, *Australimusa*, *Callimusa* et *Rhodochlamys*. Bien que la majorité des bananiers cultivés découlent d'hybridation entre espèces et sous espèces de la section *Eumusa*, certains dérivent d'autres sections comme les bananiers cultivés appelés F'éi qui dérivent des *Australimusa*, et des hybrides entre *M. textilis* (*Australimusa*) et *M. acuminata* ont été décrits (Dodds, 1946; D'Hont et al., 2000).

Il ressort des différentes analyses de la diversité des bananiers une structure complexe avec pour certains groupes génétiques des difficultés à les définir. En effet, certains bananiers sauvages sont principalement autogames (*M. a.* ssp. *banksii*, *M. schizocarpa*), d'autres sont plus allogames. De plus, peu de représentants des espèces et sous-espèces sauvages sont disponibles dans les collections. Les analyses récentes de mosaïques de bananiers dérivés de *M. acuminata* ont montré que des représentants sauvages peuvent être introgréssés (Martin et al., 2020). Par ailleurs, cette étude a révélé qu'un à deux groupes génétiques pour lesquels aucun représentant sauvage n'a été identifié, sont prédits chez des cultivars dérivés de *M. acuminata*. Pour rappel, les logiciels testés dans le chapitre 2 (`SABER`, `WINPOP` et `ELAI`) ne permettent pas de gérer des groupes inconnus.

Un jeu de données de reséquençage est disponible dans le cadre du projet DYNAMO (France Génomique, en collaboration avec le Génoscope). Ce projet a permis le reséquençage de 159 accessions de bananiers sauvages et cultivés, diploïdes et triploïdes, et l'un de ses objectifs est la caractérisation des génomes mosaïques des bananiers. Nous nous focaliserons ici sur des accessions diploïdes. L'analyse de la diversité diploïde passe par quatre objectifs : i) décrire la structuration génétique globale du jeu de données, ii) identifier des représentants de groupes génétiques sauvages, iii) tester le comportement des méthodes d'ILEA sélectionnées dans le chapitre précédent et iv) évaluer l'apport des mosaïques prédites par ces méthodes. Nous allons dans un premier temps analyser la structure génétique du jeu de données par analyse multivariée (ACP) et par le programme d'IGEA `ADMIXTURE` (Alexander et al., 2009). Dans un second temps, une première analyse locale sera effectuée via une méthode inspirée de l'approche utilisée pour déterminer les mosaïques A/B (*M. acuminata*/*M. balbisiana*) chez des hybrides interspécifiques AB, AAB et ABB de bananiers (Baurens et al., 2019). Cela va permettre d'identifier les meilleurs représentants des groupes génétiques, et de repérer les cultivars ayant des contributions d'origine inconnue. Un sous-jeu de données formé à partir des individus sans contribution ancestrale inconnue, sera analysé avec les méthodes d'ILEA évaluées dans le chapitre précédent. Cela permettra de tester les attendus du fonctionnement des méthodes d'ILEA sur des données de type plantes cultivées du chapitre précédent. La précision des méthodes d'ILEA ne pouvant pas être déterminée sur un jeu de données réel, c'est ici leur stabilité (reproductibilité entre sous-échantillons) et robustesse (variation de l'inférence à différentes valeurs du paramètre temps depuis l'hybridation) qui est analysée. Enfin, les inférences seront comparées aux informations disponibles sur les individus du sous-jeu pour évaluer leur cohérence et l'apport des méthodes d'ILEA.

4.2 Matériel et méthodes

4.2.1 Description du jeu de données

Des données de séquençage de 115 accessions diploïdes par la technologie Illumina (Table S.1) ont été obtenues dans le cadre du projet Dynamo (France Génomique), réalisé en collaboration avec le Génoscope. Les individus sont issus des collections du CRB (Centre de Ressources Biologiques) Plantes Tropicales Antilles CIRAD-INRA en Guadeloupe (France), et de l'ITC (International Transit Centre, Bioversity International/Katholieke Universitat Leuven, Belgique). Le jeu de données diploïde analysé ici est composé de 59 accessions cultivées (c-à-d parthénocarpiques, 53 cultivars AAcv, 2 ABcv, 2 AScv, 2 F'éi *Australimusa*), de 33 accessions sauvages (séminifères) de *M. acuminata* (6 accessions de *M. a.* ssp. *banksii*, 5 accessions des *M. a.* ssp. *burmannica/siamea*, 8 accessions de *M. a.* ssp. *malaccensis*, 5 accessions de *M. a.* ssp. *zebrina*, 4 accessions de *M. a.* ssp. *microcarpa*, 1 accession de *M. a.* ssp. *sumatrana*, 1 accession de *M. a.* ssp. *truncata*, 1 accession de *M. a.* ssp. *errans* et 2 accessions indéterminées), 4 accessions hybrides F1 de *M. acuminata*, 2 accessions *M. schizocarpa*, 5 accessions *M. balbisiana*, 11 autres espèces *Musa* ainsi qu'une accession représentant le genre *Ensete* (*E. ventricosum*). Les accessions sont désignées dans ce document par leur code dans le fichier vcf, la correspondance avec les noms des accessions et les codes de collections est donnée dans le Table S.1.

4.2.2 Traitement des données de séquençage et appel de variants

Le séquençage a été réalisé par le GENOSCOPE sur un séquenceur Illumina HiSeq4500 en lectures pairées de 150 paires de bases. La profondeur de lecture minimale était d'environ 30x pour les accessions diploïdes.

L'appel de variant a été réalisé par Guillaume Martin, à l'aide des outils `process_reseq` et `VcfPreFilter` (Garsmeur et al., 2018). Dans un premier temps, les lectures sont alignées séparément contre le génome de référence ('Pahang HD' version 2) (Martin et al., 2016), avec `bwa mem` (Li and Durbin, 2009; Li, 2013). Les lectures qui ne s'alignent pas et les lectures s'alignant à plusieurs endroits sont retirées, avec `samtools` (Li et al., 2009). Les lectures pairées sont regroupées et les duplicats de PCR sont éliminés, avec `picardtools` (Broad Institute, 2019). Un réalignement local autour des insertions-délétions (indels) est effectué avec `GATK` (McKenna et al., 2010). À partir des fichiers bam obtenus après nettoyage et réalignement, chaque site sur l'ensemble des accessions est analysé, via l'outil `bam_readcount` (`https://github.com/genome/bam-readcount`). Chacun des sites est considéré comme un variant potentiel si les lectures portant la variation ont une qualité d'alignement moyenne supérieure à 10. Le génotype du variant est estimé à partir d'un calcul de probabilité (basé sur

une loi binomiale) d'être hétérozygote ou homozygote de chaque allèle. La probabilité la plus forte détermine le génotype du variant, et le génotype de l'individu au site donné. Le génotype, la profondeur de lecture pour chaque allèle et la profondeur totale sont rapportés dans le fichier VCF. Cette étape (`process_reseq`) va sélectionner un très grand nombre de sites qui vont ensuite être filtrés en plusieurs étapes.

Les sites variants sont inspectés par individu, en tant que point de donnée. Les points de données couverts par au moins 10 lectures et moins de 10 000 lectures sont inspectés (`VcfPreFilter`). Une première étape de filtrage est effectuée en gardant par point de données, tous les allèles couverts par plus de 3 lectures dans une accession et avec une fréquence sur l'ensemble des lectures de plus de 5 %. L'appel de variants est répété pour ces points de données en ne prenant en compte que les allèles retenus. Une seconde étape de filtrage est ensuite réalisée avec `vcfFilter`. Les points de données dont la couverture est inférieure à 10 ou supérieure à 10 000 sont convertis « données manquantes ». De la même manière, les point de données dont l'allèle mineur est soutenu par moins de 3 lectures et moins de 10 % des lectures sont convertis en « données manquantes ». Enfin, les indels (insertions/délétions), les sites monomorphes, les sites multi-alléliques, les sites avec données manquantes et les sites issus de séquences répétées sont retirés. Le jeu de données final pour les 115 accessions diploïdes est composé de 16 891 756 variants stockés dans un fichier vcf.

4.2.3 Exploration préliminaire du jeu de données

Le but est de caractériser la structure globale du jeu de données, de vérifier les a priori sur les individus (par exemple, assignation à une sous-espèce, degré d'hétérozygotie attendu pour les autogames vs. allogames, identification des hybrides) et de détecter les individus susceptibles d'être utilisés comme représentants de groupes ancestraux lors de l'inférence locale.

4.2.3.1 Analyse de la structure par ACP

L'analyse en composante principale (ACP) a été réalisée avec la fonction `prcomp` de `R` (R Core Team, 2020) (ainsi que les figures associées). Un échantillonnage des marqueurs (un marqueur sur 500) est effectué avant l'ACP pour une taille de 33 783 marqueurs répartis sur l'ensemble des chromosomes. Comme le jeu de données comprend des individus particulièrement distants, l'ACP a été réalisée en deux étapes pour en faciliter l'exploration. Une première fois avec tous les individus et une seconde fois avec 95 individus ne s'étant pas détachés sur les premiers axes de la première ACP.

4.2.3.2 Analyse de la structure par inférence globale des états ancestraux

L'analyse globale des états ancestraux a été réalisée avec le logiciel `ADMIXTURE` (Alexander et al., 2009). Les clones naturels (accessions différentes, génétiquement identiques, issues de la propagation végétative) ont été retirés du jeu de données pour ne pas augmenter artificiellement leur impact sur le clustering. Trois paires de clones sont présentes, ce qui amène le jeu de données utilisable à 112 individus. Pour la même raison que pour l'ACP, le jeu de données a été subdivisé en 32 de manière à obtenir environ 500 000 marqueurs par sous-jeu. L'analyse avec `ADMIXTURE` (version 1.23) à été conduite en mode non supervisé, avec une valeur de K (nombre de clusters) allant de 2 à 20, et 10 répétitions par valeurs de K. Pour réduire le temps de calcul, seuls les 5 premiers sous-jeux de données ont été analysés par `ADMIXTURE`. Les logs de sorties d'`ADMIXTURE` concernant les informations de validation croisée ont été mis en forme et synthétisés sous forme de figures avec `R` (R Core Team, 2020). Les résultats de clustering présentés ont été manuellement réarrangés entre les différentes valeurs de K sélectionnées.

4.2.3.3 Calcul de l'hétérozygotie

L'hétérozygotie de chaque individu (nombre de marqueurs hétérozygotes divisé par le nombre de marqueurs total) a été calculée à l'aide de `vcftools` (Danecek et al., 2011) avec l'option `-het`, à partir du fichier vcf des 115 individus.

4.2.4 Méthode de « chromosome painting » par ratio de couverture allélique

Dans le but de permettre une exploration plus approfondie du jeu de données tout en restant sur un principe simple, une méthode de « painting » allélique a été proposée par Guillaume Martin et implémentée dans le cadre de ce travail de thèse. L'idée est d'obtenir le long des chromosomes un signal d'appartenance à des groupes ancestraux désignés. Cette méthode est une variante de l'approche utilisée dans le cadre de la caractérisation des mosaïques AB de Baurens et al. (2019). L'avantage de cette approche est sa simplicité, mais elle ne tient pas compte des liens entre les marqueurs (contrairement à la majorité des méthodes d'ILEA). L'approche repose sur la détection d'allèles représentatifs de groupes génétiques, la détermination d'un ratio de couverture allélique pour ces marqueurs et sa représentation graphique le long des chromosomes.

Dans un premier temps, les allèles spécifiques de groupes ancestraux sont identifiés. La sélection des groupes et des individus composant chaque groupe est manuelle. L'outil utilisé

dans cette étape est `vcf2allPropAndCov` (Baurens et al., 2019). Le programme détecte pour chaque groupe tous les allèles présents dans au moins un individu du groupe et absents de tous les autres individus des autres groupes. Une option permet au programme d'être plus strict sur la sélection d'allèles, en n'acceptant que les allèles présents chez chaque membre du groupe plutôt qu'au moins un membre. Le ratio de couverture est ensuite calculé en divisant le nombre de lectures supportant l'allèle par le nombre de lectures supportant le marqueur. L'information finale obtenue est composée de la position de l'allèle (chromosome + coordonnées), l'allèle lui-même et le ratio de lectures le supportant, pour chaque individu de chaque groupe.

Dans un second temps, une synthèse des listes d'allèles spécifiques est produite à partir de la liste d'allèles par individu représentatif de chaque groupe. Les ratios de chaque allèle sont moyennés sur tous les individus de chaque groupe. Si l'allèle est absent dans un individu, son ratio a une valeur fixée à 0. Ce dernier point est autorisé car, du fait des filtres appliqués, le fichier vcf utilisé pour réaliser cette approche ne contient pas de sites pour lesquels le génotype est inconnu ou pour lesquels il n'y a pas de données de séquençage. À la fin de cette étape, nous obtenons une liste de ratio de couverture allélique par groupe.

Il est maintenant possible de comparer ces listes d'allèles et leurs ratios aux individus cibles de l'approche. Pour chaque individu d'intérêt, chaque allèle retrouvé dans un des groupes est listé, le ratio de couverture allélique calculé sur l'individu et stocké de la même manière que le ratio attendu par groupe.

Enfin, une fenêtre glissante est appliquée (ici 300 000 paires de bases) le long des chromosomes de chaque individu. Un point de donnée est généré par fenêtre en divisant la somme des ratios observés d'allèles par la somme des ratios attendus. Cette opération est réalisée pour chaque groupe génétique, et permet de produire un graphique représentant le ratio d'allèles spécifiques de chaque groupe sur l'individu. Dans le cas d'individus diploïdes, on attend un ratio autour de 0,5 pour deux origines dans le cas d'une région hybride entre deux groupes génétiques et un ratio autour de 1 pour une seule origine. Cette approche permet de corriger (en partie) à la fois les biais de « mapping » des lectures sur la référence en fonction des variants qu'elles portent grâce au ratio calculé sur les groupes ancestraux. La méthode a été imaginée dans le cadre de l'analyse de la diversité des bananiers, là où certains groupes peuvent avoir un nombre potentiel d'allèles diagnostics le long des chromosomes très différent. En prenant en compte le ratio de couverture, en permettant sélectionner les allèles diagnostics même sur les sites hétérozygotes chez les représentants des groupes ancestraux et également en calculant une moyenne des ratios de couvertures par fenêtre, elle permet de réduire l'impact de la différence de fixation des allèles dans les groupes ancestraux.

Cette approche sera désignée dans ce manuscrit par l'abréviation ARP, pour « allele ratio painting ».

4.2.5 Détection d'individus représentatifs des sources et d'individus avec des contributions inconnues par l'approche ARP

À partir des individus détectés comme représentants potentiels des sources par les analyses globales, la méthode ARP a été utilisée de manière itérative. Les individus les plus homozygotes de chaque groupe potentiel (un individu par groupe) ont été utilisés pour réaliser un premier « painting ».

À la fin du premier « painting », chaque mosaïque est vérifiée visuellement. Un second individu par groupe est choisi sur base de i) sa présélection par les analyses globales ii) un profil de « painting » sans régions inconnues ni introgression. Un nouveau « painting » est effectué avec deux individus par groupe : le nouvel individu sélectionné et l'individu de départ.

Trois étapes successives ont permis de sélectionner une liste d'individus représentatifs. Par ailleurs, un jeu de données (individus représentatifs et hybrides) est constitué en retirant les individus portant des origines inconnues.

4.2.6 Inférence locale des états ancestraux

Les méthodes d'ILEA sélectionnées sont utilisées sur un jeu de données dépourvu de groupes ancestraux manquants et constitué de 11 individus représentatifs de quatre groupes ancestraux et 14 individus hybrides. Ce sous-jeu est composé de 2 712 295 marqueurs variant parmi ces 25 individus.

Les méthodes d'ILEA ont été testées précédemment sur des jeux de données simulées composés de 20k à 30k marqueurs, et pouvaient malgré tout avoir une consommation de ressources (temps et mémoire) importante, spécifiquement la méthode `ELAI`. Pour obtenir des inférences dans un temps court (< 24 h) et surtout avec une empreinte mémoire raisonnable (< 32Go), le jeu de données a été là aussi divisé en sous-échantillon. Les sous-échantillons permettront aussi de vérifier la stabilité des méthodes sur les données. Parce que les analyses locales se déroulent indépendamment par chromosome, il n'a pas été nécessaire de faire plus de 3 sous-échantillons. Précisément, le nombre de marqueurs par sous-échantillon est compris entre 62k (chromosome 11) et 106k (chromosome 4).

Il a été montré dans le chapitre 2 que les méthodes d'ILEA sont robustes pour des valeurs de temps depuis l'hybridation surestimées. Ce paramètre sera donc testé sur trois modalités, 5, 20 et 50 générations, pour couvrir le champ exploré lors du chapitre II (5-50) et se rapprocher du nombre de générations supposé dans le cas du bananier cultivé. Il s'agit ici de suppositions et pas d'estimations formelles, qui n'ont pas été faites sur les individus hybrides du bananier.

SABER et WINPOP utilisent les proportions globales d'origines ancestrales en paramètre d'entrée, qui sont obtenues ici avec une analyse IGEA effectuée via ADMIXTURE en mode supervisé, en fixant les représentants des groupes ancestraux. Pour WINPOP, le paramètre du taux de recombinaison est fixé à 0,045/Mb, correspondant à la moyenne sur tout les chromosomes de l'estimation du taux de recombinaison produite par Martin et al. (2017). La valeur du temps depuis l'hybridation pour SABER est fixée sur la valeur du paramètre en entrée et non pas estimée comme dans le chapitre précédent. Le paramètre du nombre de « clusters » inférieurs d'ELAI est recommandé à 5 fois le nombre de groupes par les auteurs (Guan, 2014), ce qui équivaut ici à une valeur de 20. Cependant, il y a en tout seulement 11 individus représentatifs et deux groupes ancestraux sont représentés uniquement par deux individus, le paramètre est donc fixé à 8 (2x le nombre de groupes) pour garder un équilibre entre les populations ancestrales. Comme dans le chapitre précédent, les sorties des logiciels utilisés sont pour SABER et ELAI, la moyenne des estimations d'origines ancestrales a posteriori, et pour WINPOP l'estimation d'origine ancestrale discrétisée à la suite de l'estimation pour chaque marqueur.

Les inférences sont comparées deux à deux pour chaque combinaison de paramètres (entre méthodes, entre sous-échantillons, entre paramètres de temps depuis l'hybridation) pour les 14 individus du sous-jeu de données. La différence entre les deux matrices d'origines ancestrales comparées est mesurée par la moyenne des corrélations entre chaque paires de colonnes d'une même origine. Pour la représentation finale (les figures de mosaïques), les répétitions sur les sous-échantillons de chaque méthode sont combinées et moyennées sur les trois répétitions. Pour combiner ces inférences sur les sous-échantillons portant des jeux de marqueurs différents, le résultat de chaque inférence est étendu par sous-échantillon sur la liste de l'ensemble de marqueurs, puis la moyenne est calculée sur les trois matrices d'origines ancestrales qui sont alors de même dimension.

4.3 Résultats

4.3.1 Structure globale de la diversité

4.3.1.1 L'approche par ACP

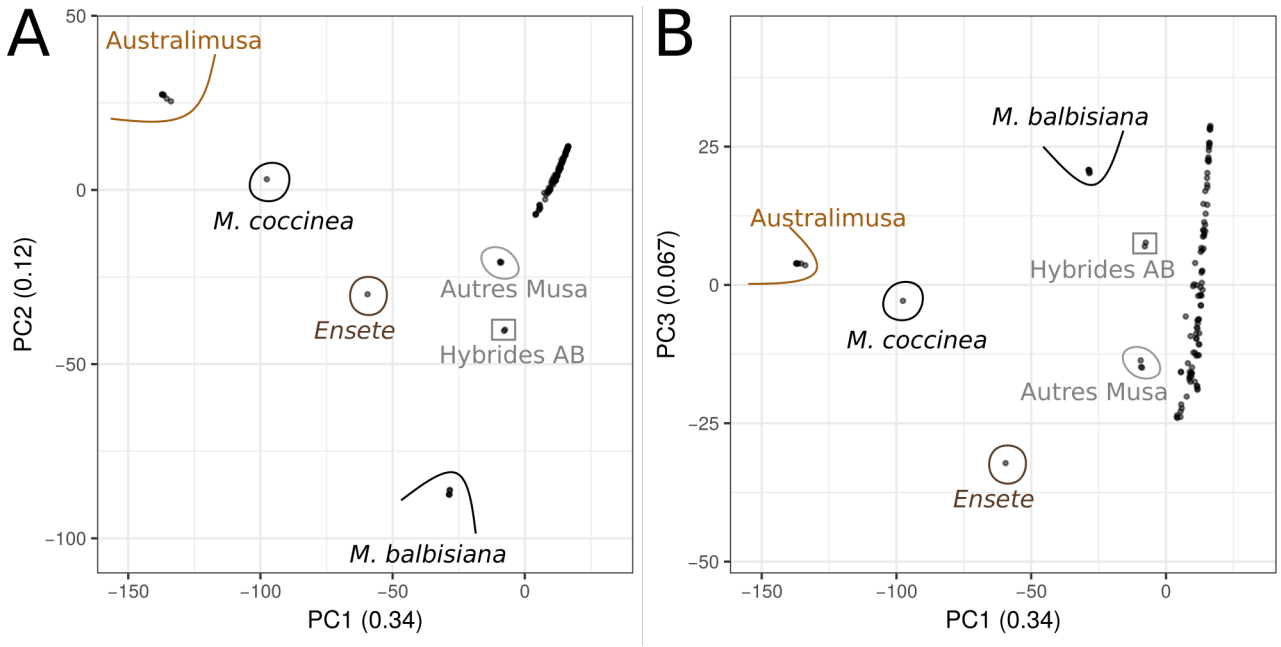

FIGURE 4.1 – ACP sur l'ensemble des 115 accessions de bananiers (33 783 marqueurs SNP). (A) Projection des accessions sur les axes 1 et 2. (B) Projection des accessions sur les axes 1 et 3. Les groupes représentés sont *Australimusa* (marron clair), *Ensete* (marron foncé), *M. coccinea* et *M. balbisiana* (noir), les hybrides AB (encadré gris) et un groupe « autres *Musa* » *Rhodochlamys* composé de *M. velutina*, *M. ornata* et *M. sanguinea* (gris). La variance capturée par les axes est notée à côté de leur légende.

La première ACP (Figure 4.1) montre l'ensemble des accessions réparties sur les trois premiers axes (52 % de l'information). Le premier axe sépare principalement les *Australimusa* et le deuxième axe sépare principalement un groupe *M. balbisiana* des autres accessions et met en évidence les hybrides AB. Les deux premiers axes découpent le jeu de données en 6 groupes, *Ensete*, *Australimusa*, *M. coccinea*, *M. balbisiana*, 3 espèces de l'ancienne section *Rhodochlamys* (*M. velutina*, *M. ornata* et *M. sanguinea*) notés « autres *Musa* » et un groupe de 97 accessions, principalement *M. acuminata* et des hybrides AA, mais aussi *M. schizocarpa*, *M. rosea* et *M. laterita*. Le troisième axe étale la structure interne du groupe des 97 accessions, et discrimine *Ensete*.

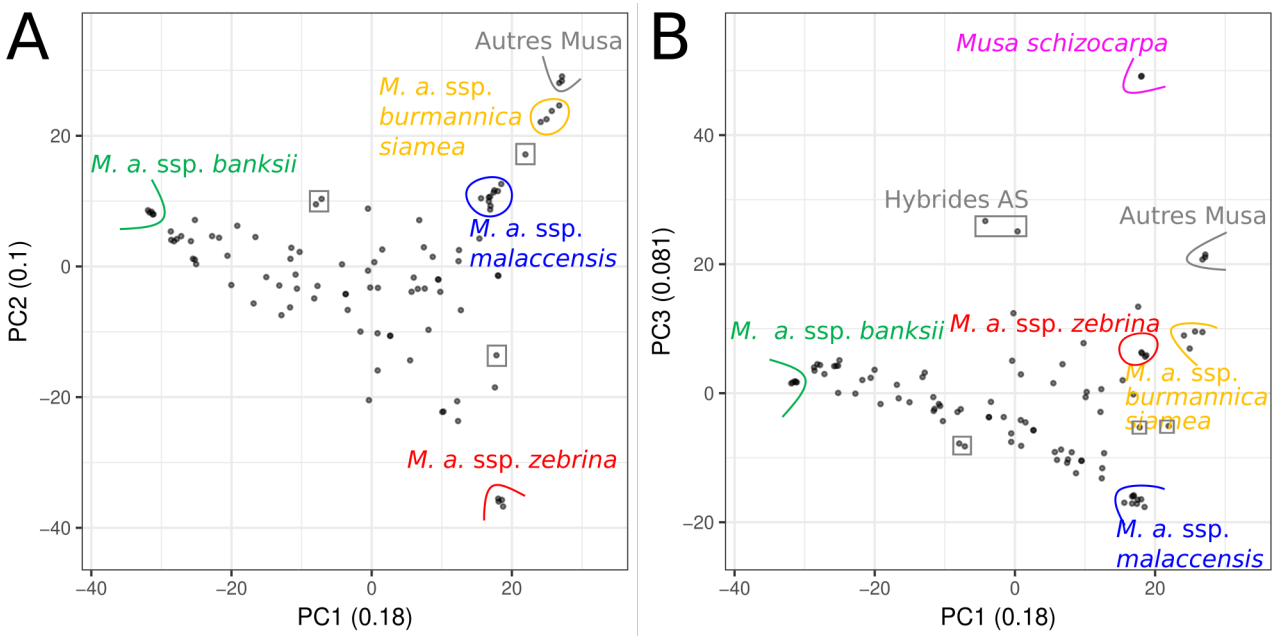

FIGURE 4.2 – ACP sur un groupe de 97 individus comprenant *M. laterita*, *M. rosea*, *M. schizocarpa*, *M. acuminata* et des individus hybrides (33 783 marqueurs SNP) (A) Projection des accessions sur les axes 1 et 2. (B) Projection des accessions sur les axes 1 et 3. La variance capturée par les axes est notée à côté de leur légende. Les 97 individus correspondent au groupe non annoté de la figure 4.1. Les groupes représentés sont *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *burmannica/siamea* (jaune orangé), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose), les hybrides AS et hybrides AA(encadré gris), et un second groupe « autres *Musa* » *Rhodochlamys* composé de *M. laterita* et *M. rosea*. Les 4 individus encadrés d'un rectangle gris sans annotations sont des hybrides F1 entre une accession *M. a.* ssp. *malaccensis* et une accession des sous-espèces *M. a.* ssp. *banksii*, (2 individus), *M. a.* ssp. *zebrina* (1 individu) et *M. a.* ssp. *burmannica* (1 individu), respectivement.

Une seconde ACP a été réalisée uniquement sur les 97 accessions groupées par les axes 1 et 2 de la première ACP (Figure 4.2). Le premier axe sépare un groupe *M. a.* ssp. *banksii* des autres sous-espèces de *M. acuminata* (*M. a.* ssp. *zebrina*, *M. a.* ssp. *malaccensis*, *M. a.* ssp. *burmannica/siamea*) et des deux autres espèces de l'ancienne section *Rhodochlamys* (*M. rosea* et *M. laterita*). Le second axe sépare un groupe *M. a.* ssp. *zebrina* et un groupe composé de *M. rosea* et *M. laterita* des autres accessions. Le groupe *M. a.* ssp. *burmannica/siamea* est aussi légèrement écarté du centre de l'axe. On note aussi que quatre individus F1 issus de croisements entre une accession de la sous-espèce *M. a.* ssp. *malaccensis* (PT-BA-00267) et une accession de chacune des trois sous-espèces *M. a.* ssp. *banksii* ('Banksii', PT-BA-00024), *M. a.* ssp. *zebrina* ('Maia Oa', PT-BA-00182) et *M. a.* ssp. *burmannica* (Calcutta 4, PT-BA-00051) ont des positions cohérentes avec leurs origines.

La liste d'individus représentant des groupes génétiques structurants sur base de l'ACP est présentée dans le Table S.4. Globalement, sont regroupées les 6 accessions *Australimusa*, 5 accessions *M. balbisiana*, 21 accessions *M. acuminata*, 2 *M. schizocarpa*, et trois groupes d'autres espèces, les *Rhodochlamys* en deux groupes et *M. coccinea* en un dernier. Les 21 accessions *M. acuminata* sont composées de 5 *M. a.* ssp. *banksii*, 4 *M. a.* ssp. *burmannica/siamea*, 8 *M. a.*

ssp. *malaccensis* (dont une accession précédemment notée *M. a.* ssp. *siamea* et une accession précédemment notée *M. acuminata* indéterminée) et 4 *M. a.* ssp. *zebrina*.

4.3.1.2 Inférence ancestrale globale avec `ADMIXTURE`

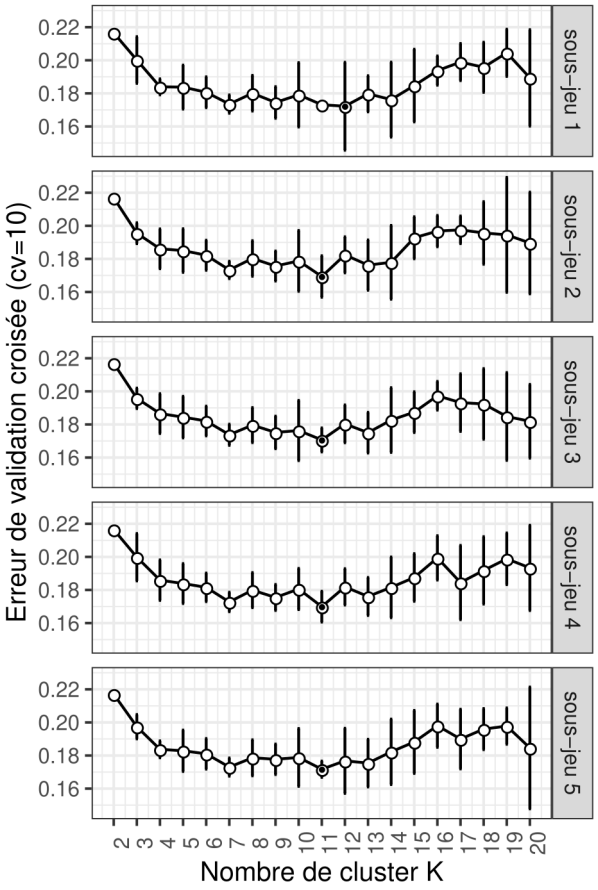

FIGURE 4.3 – Représentation des valeurs de validation croisée d'`ADMIXTURE` sur 5 sous jeu de données, avec 10 répétitions et une valeur de K allant de 2 à 20. L'axe des abscisses représente la valeur de K, l'axe des ordonnées le niveau d'erreur de validation croisée. Les barres d'erreurs représentent l'écart-type sur les 10 répétitions, et le point blanc la moyenne de ces répétitions. Le point noir indique la valeur d'erreur moyenne la plus faible de la courbe.

L'analyse par `ADMIXTURE` suggère un nombre de clusters compris entre 7 et 13. Les courbes de validation croisée (Figure 4.3) montrent que la tendance globale du clustering est semblable entre les différents sous-jeux de données. Un premier optimum local apparaît à $K = 7$, puis pour des valeurs de $K = 9$ et $K = 11$, avec des barres d'erreurs de plus faible amplitude, et un dernier infléchissement à $K = 13$.

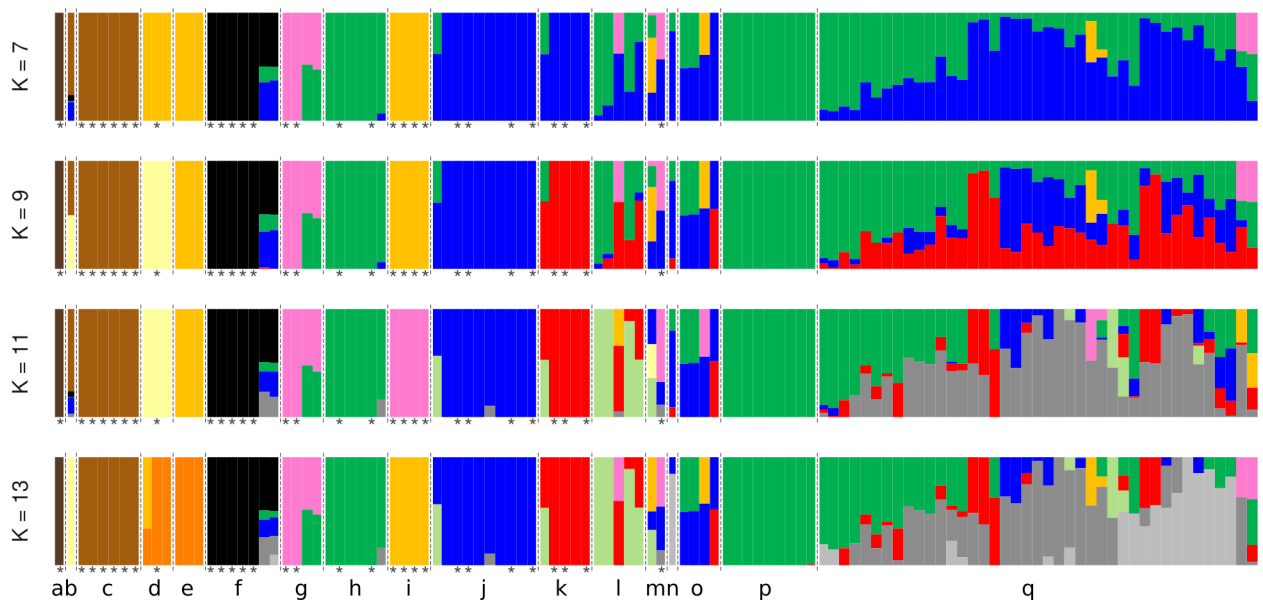

FIGURE 4.4 – Partitionnement de 112 individus du jeu de données par `ADMIXTURE` pour des valeurs de K de 7, 9, 11 et 13. Les trois individus manquants (112/115) correspondent à la sélection d'un individu sur deux pour les trois paires de clones naturels du jeu de données. L'inférence globale correspond à la première répétition d'`ADMIXTURE` sur le 3ème sous-jeu de 500 000 SNPs. Chaque individu est représenté par une colonne. Les couleurs représentent les différents groupes (clusters) inférés par `ADMIXTURE` et leurs proportions. Les barres verticales pointillées séparent des groupes annotés de 'a' à 'q'. (a) *Ensete*, (b) *M. coccinea*, (c) *Australimusa*, (d) *Rhodochlamys* (*M. ornata*, *M. velutina* et *M. sanguinea*), (e) *Rhodochlamys* (*M. rosea* et deux *M. laterita*), (f) *M. balbisiana* et deux hybrides AB, (g) *M. schizocarpa* et deux hybrides AS, (h) *M. a.* ssp. *banksii*, (i) *M. a.* ssp. *burmannica/siamea*, (j) *M. a.* ssp. *malaccensis*, (k) *M. a.* ssp. *zebrina*, (l) les sous-espèces *M. a.* ssp. *errans* et *M. a.* ssp. *microcarpa*, (m) les sous-espèces *M. a.* ssp. *truncata* et *M. a.* ssp. *sumatrana*, (n) un individu sauvage classifié *M. acuminata*, (o) les hybrides F1, qui sont des AA, (p) les cultivars AA trouvés non-hybrides, (q) les cultivars AA hybrides. Les individus marqués par des étoiles '*' sont utilisés comme référence pour l'analyse ARP

Globalement, pour les espèces et sous-espèces sauvages, les résultats sont cohérents avec nos connaissances de la diversité. Les groupes génétiques correspondant à *Ensete*, *Australimusa* y compris les cultivars F'éi, *M. balbisiana*, *M. a.* ssp. *banksii*, sont discriminés pour les 4 valeurs de K. Les groupes génétiques correspondant aux sous-espèces *M. a.* ssp. *malaccensis* ('j', Figure 4.4) et *M. a.* ssp. *zebrina* ('k', Figure 4.4) sont confondus à $K = 7$ mais discriminés pour les autres valeurs de K. Le groupe *M. a.* ssp. *burmannica/siamea* ('i', Figure 4.4) est associé aux *Rhodochlamys* ('d' et 'e' , Figure 4.4) à $K = 7$ et $K = 9$ puis (comme également observé par Sardos et al., 2016a) à *M. schizocarpa* à $K = 11$. Tous ces groupes ainsi que *M. coccinea* sont ici discriminés à $K = 13$ (Figure 4.4). Les inférences pour les accessions *M. a.* ssp. *microcarpa* et *M. a.* ssp. *errans* ne sont pas stables et suggèrent soit des accessions hybrides ou introgréssées avec une forte composante *M. a.* ssp. *banksii* soit la présence d'un cluster spécifique ('l', Figure 4.4). L'individu classifié *M. acuminata* ('2013-211-Ambihy_P1') est prédit hybride AA ('n', Figure 4.4). On note que les accessions représentant *M. a.* ssp. *truncata* et *M. a.* ssp. *sumatrana* sont trouvées hybrides (groupe 'p', Figure 4.4). Par ailleurs, des individus introgressés sont visibles chez *M. a.* ssp. *banksii*, *M. a.* ssp. *malaccensis* et *M. a.* ssp. *zebrina* ('h','j','k' respectivement,

Figure 4.4).

Pour les cultivars AAcv, on distingue deux catégories majeures ('p' et 'q', Figure 4.4) : un groupe de 9 cultivars ('p') prédits comme étant homogènes pour *M. a.* ssp. *banksii* et 41 cultivars AAcv prédits comme étant hybrides ou introgréssés ('q'). Le passage de $K = 7$ à $K = 13$, montre des structurations différentes de ces cultivars AAcv hybrides. Un groupe ancestral ($K = 11$, gris) puis deux groupes ($K = 13$, gris et gris clair) sont inférés chez plusieurs de ces hybrides ainsi que chez les hybrides AB ('q' et 'f', Figure 4.4) sans qu'il soit possible de les associer clairement à des groupes génétiques sauvages.

Les individus hybrides entre espèces comme les hybrides AB ('f', Figure 4.4), les hybrides AS ('g', Figure 4.4) et les quatre hybrides F1 ('o' , Figure 4.4) se comportent globalement comme attendu avec des proportions d'origine ancestrale autour de 50-50.

La liste d'individus homogènes pour des groupes génétiques des espèces et sous-espèces sauvages pouvant être dégagée des clusters d'ADMIXTURE présentée dans le Table S.4. Elle est globalement cohérente avec les résultats de l'ACP. Notamment, sont regroupés comme pour l'ACP 5 accessions *Australimusa*, 5 accessions *M. balbisiana*, 2 *M. schizocarpa*, 1 *M. coccinea*, et 1 *Ensete*. Les *Rhodochlamys* sont regroupés de façon variable en un ou deux groupes. Pour *M. acuminata*, 22 accessions sont trouvées homogènes pour 5 groupes, avec un individu en moins pour le groupe *M. a.* ssp. *malaccensis* (soit 7 accessions, l'accession 'DYN-ITC1348-Pisang_ serun' étant introgressée). Le statut du groupe *M. a.* ssp. *errans* et *M. a.* ssp. *microcarpa* ('l') qui apparaît à $K = 11$ et $K = 13$ sera inspecté avec l'approche ARP.

4.3.1.3 L'hétérozygotie des cultivars et des sauvages

L'hétérozygotie sur l'ensemble du jeu de données est comprise entre 0,07 % et 10,41 % (Table S.2). Les cultivars ont des taux d'hétérozygotie compris entre 1,05 % et 10,41 %, et les accessions sauvages un taux compris entre 0,07 % et 4,97 %. Les valeurs d'hétérozygotie les plus élevées sont observées chez les hybrides AB (moyenne $\approx$ 10%) et AS (moyenne $\approx$ 5%). L'hétérozygotie moyenne dans les groupes non-Acuminata est autour de 1,5 % (*M. balbisiana* 1,56 %, *Australimusa* 1,17 %, *Rhodochlamys* 1,47 %), sauf dans le cas de *M. schizocarpa* qui est peu hétérozygote (0,16 %) en lien avec l'autogamie chez cette espèce. Pour *M. acuminata*, l'hétérozygotie moyenne la plus élevée est observée chez *M. a.* ssp. *malaccensis* (2,46 %) et la plus faible est observée chez *M. a.* ssp. *banksii* (0,28 %) en accord avec l'autogamie préférentielle chez cette espèce. On note également une valeur très faible d'hétérozygotie pour l'accession représentant *M. a.* ssp. *sumatrana* (0,1 %). Les deux groupes *M. a.* ssp. *zebrina* et *M. a.* ssp. *burmannica/siamea* ont respectivement une hétérozygotie moyenne de 1,70 % et 1,55 %.

Les accessions les plus homozygotes par groupe génétique identifiées par les analyses glo-

bales sont 'DYN113-Hawain_2' pour *M. a.* ssp. *banksii*, 'DYN147-Khae-Phrae' pour *M. a.* ssp. *burmannica/siamea*, 'DYN-ITC0609-Pahang' pour *M. a.* ssp. *malaccensis*, 'DYN-Maia_Oa_Q10_2016-007' pour *M. a.* ssp. *zebrina*, 'DYN019-Balbisiana_Honduras' pour *M. balbisiana*, 'ITC0926-Schizocarpa' pour *M. schizocarpa*, 'DYN229-Musa_velutina' pour les *Rhodochlamys* et 'ITC0956-Musa_lolodensis' pour les *Australimusa*. Ces accessions seront utilisées comme point de départ pour l'exploration itérative des contributions ancestrales au niveau local présentée dans la partie suivante.

4.3.2 Identification des représentants des groupes génétiques sauvages par approche itérative et analyse exploratoire des cultivars par ARP

L'approche itérative s'est déroulée en trois étapes. Les deux premières avaient pour but de sélectionner puis d'affiner la sélection des individus représentatifs des groupes ancestraux. Une troisième analyse ARP est réalisée ensuite pour identifier les contributions ancestrales sur tout le jeu de données à partir des groupes ancestraux définis à l'étape 2.

La première analyse des profils de « chromosome painting » obtenus est réalisée à partir des individus les plus homozygotes de chaque pôle. Les accessions de *M. balbisiana*, *Australimusa* et *M. schizocarpa* ont été détectés homogènes pour leur groupe génétique (pas d'introgression ni d'origines inconnues) à partir de leur individu de départ. Ces groupes correspondent aux mêmes individus que ceux détectés par ACP et par `ADMIXTURE`. Les différentes espèces de *Rhodochlamys* ne portaient pas de signal fort correspondant à l'accession *M. velutina*, l'accession sera conservée mais aucun groupe ne sera formé par la suite à partir de celle-ci. À la fin de cette première analyse, un individu supplémentaire sans introgressions détectées a été ajouté par groupe pour les sous-espèces de *M. acuminata* (Table S.4).

La seconde analyse, avec au moins deux accessions par groupe permet de finaliser la sélection des représentants sous-espèces de *M. acuminata*. Toutes les accessions non ajoutées qui avaient été détectées comme potentiellement représentatives de groupes ancestraux par ACP et `ADMIXTURE` comportent des introgressions même si elles sont quelquefois de très petite taille. Nous avons ici choisi de ne pas inclure ces accessions (voir Figure S.6 pour des exemples d'individus candidats d'après l'ACP et `ADMIXTURE` qui portent des introgressions). A l'issue de ces étapes, deux accessions sur les cinq possibles ont été retenues pour *M. a.* ssp. *banksii* ('DYN113-Hawain_2' et 'DYN-Banksii_H09_2016-008'). Quatre accessions sont retenues pour *M. a.* ssp. *burmannica/siamea*, qui correspondent à celles détectées par les approches ACP et `ADMIXTURE` ('DYN178-Long_Tavoy', 'DYN-Calcutta_4_F08_2016-005', 'DYN147-Khae-Phrae' et 'DYN263-Pa_Rayong'). Le groupe *M. a.* ssp. *zebrina* est représenté par trois des quatre accessions iden-

tifiées par l'ACP et ADMIXTURE ('DYN-ITC0415-Pisang_Cici_Alas', 'DYN-Maia_Oa_Q10_2016-007' et 'DYN212-Monyet'). Les représentants de *M. a.* ssp. *malaccensis*, sont au nombre de quatre sur les sept attendues d'après l'ACP et ADMIXTURE ('DYN-ITC0609-Pahang', 'DYN363-Selangor', 'DYN454-Malaccensis_nain' et 'DYN262-Pa_Songkhla'). L'accession *M. a.* ssp. *sumatrana*, attendue hybride d'après ADMIXTURE ne porte pas de signal correspondant aux autres sous-espèces de *M. acuminata* et est très peu hétérozygote, elle sera donc utilisée pour déterminer si elle participe à la diversité des bananiers (Figure S.7). L'accession *M. a.* ssp. *truncata* attendue hybride est détectée partiellement sans signal d'origines ancestrales et partiellement *M. a.* ssp. *malaccensis* (Figure S.7). Les deux accessions *M. a.* ssp. *microcarpa* et *M. a.* ssp. *errans* trouvées homogènes par ADMIXTURE (Figure S.8) sont détectées avec un ratio de couverture allélique élevé correspondant à *M. a.* ssp. *banksii* pouvant s'interpréter comme une proximité forte avec cette sous-espèce, en cohérence avec des travaux précédents (Carreel et al., 1994; Perrier et al., 2009; Martin et al., 2020). Elles n'ont pas été retenues comme un groupe génétique différent. Au total, sur les 115 accessions du jeu de données, 26 accessions sont retenues comme représentatives de groupes (Table S.4).

Le nombre d'allèles présents dans un groupe et absents des autres groupes, sélectionnés par la méthode ARP (table S.3) est le plus élevé chez « l'out group » *Ensete* (≈ 2,7 M), suivi des *Australimusa* (≈ 2,5 M). Pour les sous-espèces de *M. acuminata*, c'est chez *M. a.* ssp. *burmannica/siamea* qu'il est le plus élevé (≈ 0,84 M) suivi par *M. a.* ssp. *malaccensis* (0,65 M). Le nombre total d'allèles du groupe *M. a.* ssp. *banksii* (≈ 0,16 M) diminue à chaque itération, probablement dû au faible nombre d'accessions et à l'hétérozygotie très basse. Le groupe *M. schizocarpa* est le groupe non *M. acuminata* possédant le plus faible nombre de marqueurs (≈ 0,34 M), ce qui comme pour *M. a.* ssp. *banksii* est probablement lié au nombre d'individus et à l'hétérozygotie bien plus faible que les autres espèces. Au total, 9 679 647 marqueurs sont assignés à des groupes sur les 16 891 756 marqueurs du jeu de données.

Enfin, une dernière analyse ARP est conduite avec tous les représentants de pôles sélectionnés après la seconde analyse. La figure 4.5 montre trois exemples de « painting » avec un individu source (4.5 A), un individu hybride avec des origines ancestrales identifiées (4.5 B) et un individu hybride avec sans doute un groupe ancestral inconnu (4.5 C). Des régions chromosomiques qui ne sont attribuées à aucun des groupes ancestraux définis ici ont en effet été observées chez 53 cultivars comme le cultivar 'Héva' (Figures 4.5 C, 4.6). Une accession sauvage 'DYN078-EN13-IDN075' est hybride avec de grandes régions d'origine indéterminée. *M. velutina* n'est pas détectée chez les cultivars testés et la contribution de *M. a.* ssp. *sumatrana* semble limitée.

Seuls 14 cultivars ne présentant pas de groupes manquants via l'approche ARP ont été identifiés (Figures 4.8, S.13 à S.25). Ils correspondent à des 12 cultivars AAcv et deux cultivars AS provenant de Nouvelle-Guinée à l'exception de 'DYN112-Guyod' (figure 4.6) provenant des

Philippines. Ils comportent majoritairement des origines *M. a.* ssp. *banksii*, puis *M. a.* ssp. *zebrina*, *M. a.* ssp. *malaccensis* et *M. schizocarpa*.

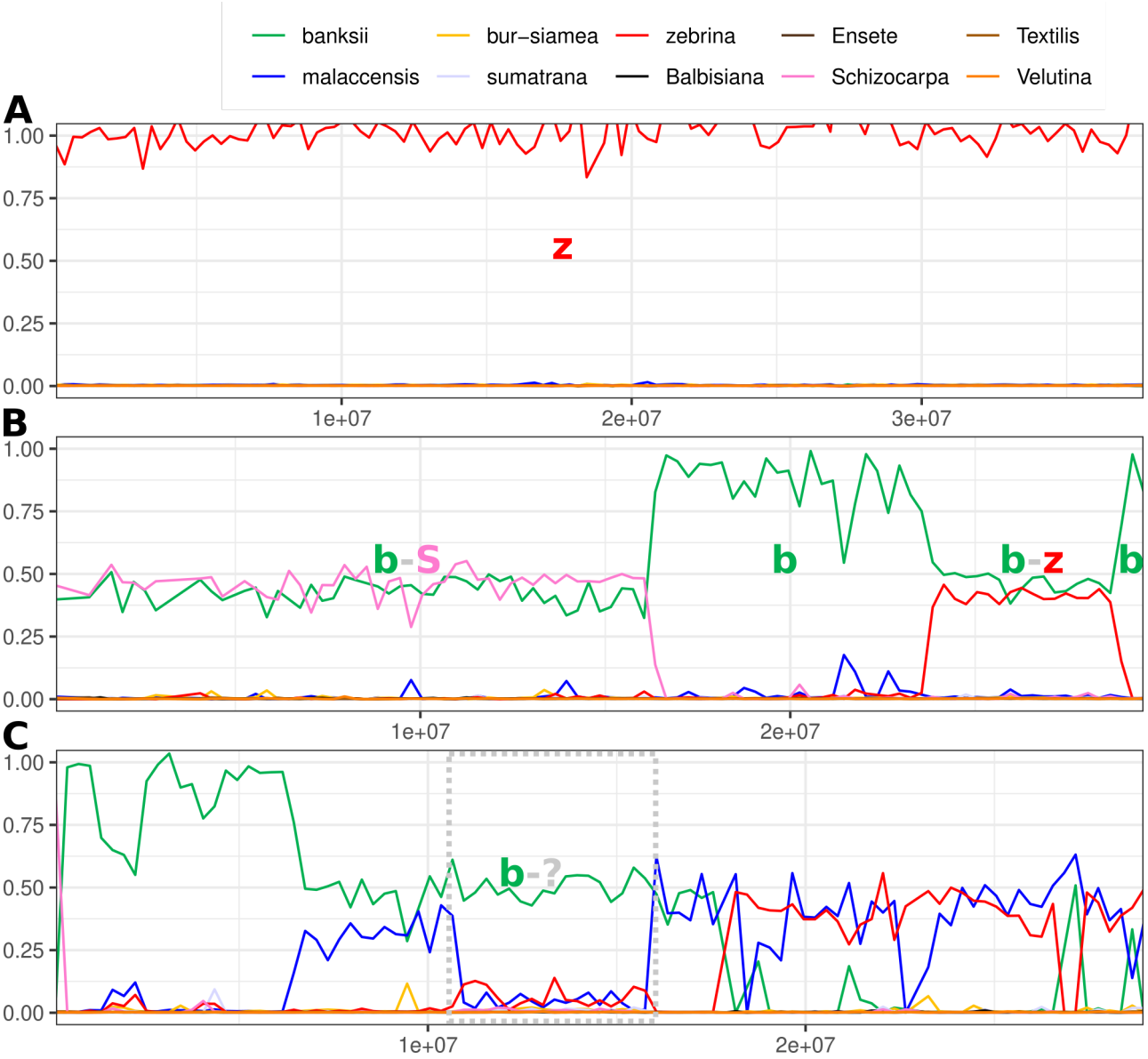

FIGURE 4.5 – Exemples de painting allélique d'un individu représentatif d'un groupe génétique sauvage, d'un cultivar dont les origines sont identifiées et d'un cultivar avec une origine manquante à la troisième itération. (A) Chromosome 6 de 'DYN-ITC0415-Pisang_Cici_Alas', exemple de sortie de l'ARP sur accession représentative de *M. a.* ssp. *zebrina* (rouge). (B) Chromosome 2 de 'ITC1211-Vudo_Beo', exemple de sortie de l'ARP sur un hybride portant des régions chromosomiques attribuées à *M. a.* ssp. *banksii* (vert), *M. schizocarpa* (rose) et *M. a.* ssp. *zebrina* (rouge). (C) Chromosome 1 de 'DYN114-Heva', exemple de sortie de l'ARP sur un hybride avec groupe manquant (encadré en pointillé). L'axe des abscisses représente les coordonnées génomiques des marqueurs utilisés pour le « painting ». L'axe des ordonnées représente la valeur du rapport du ratio allélique de l'individu sur le ratio attendu par marqueur, compris normalement entre 0 et 1. Chacune des courbes représente un groupe désigné lors de l'analyse. Les caractères 'z', 'b', 'S' et ' ?' indiquent respectivement *M. a.* ssp. *zebrina*, *M. a.* ssp. *banksii*, *M. schizocarpa* et une origine inconnue

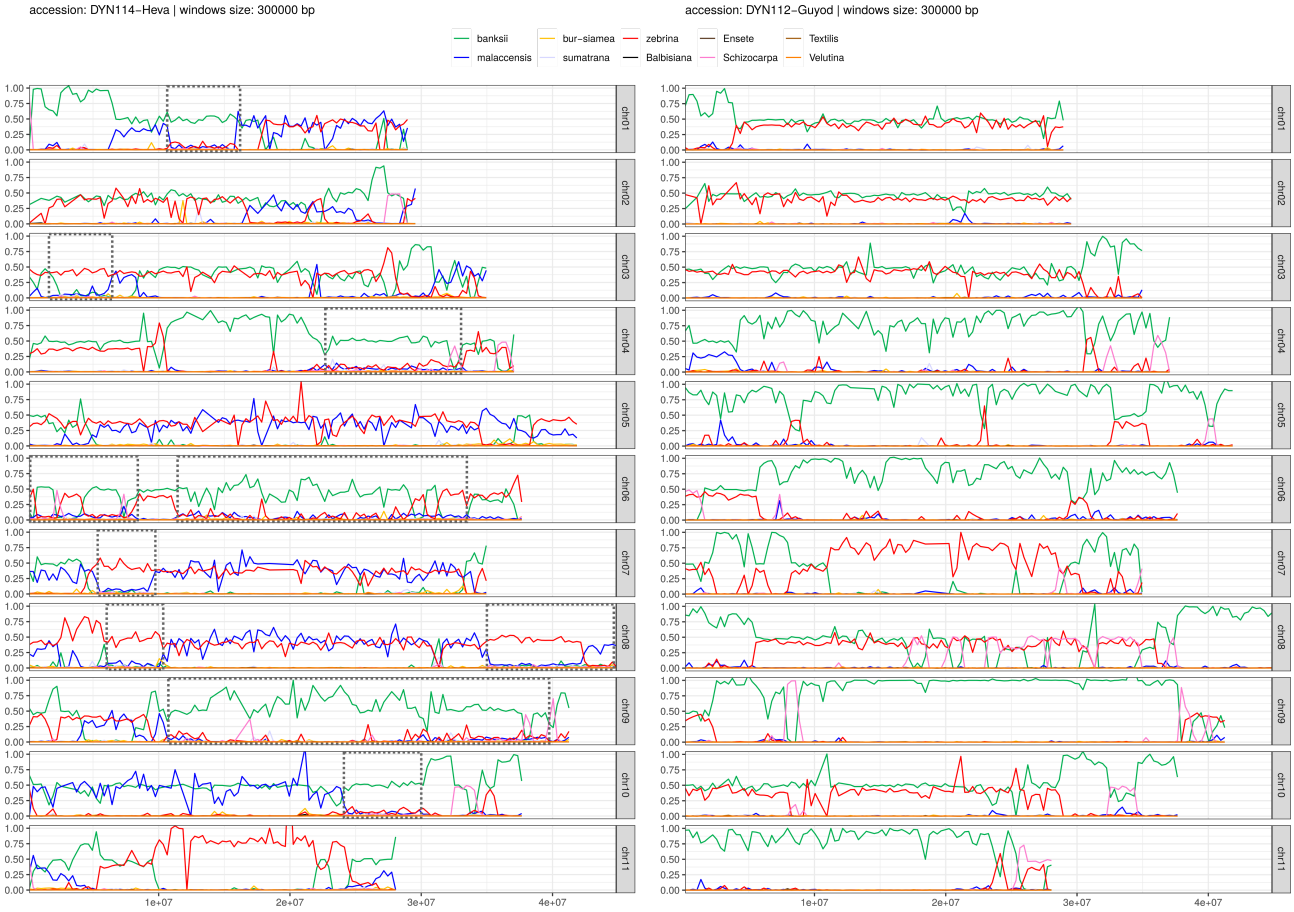

FIGURE 4.6 – Résultat de l'ARP pour les accessions 'DYN114_Heva' et 'DYN112_Guyod' à l'étape trois de l'analyse itérative. L'accession 'DYN114_Heva' est représentée à gauche et 'DYN112_Guyod' à droite. L'axe des abscisses représente les coordonnées génomiques des marqueurs retenus par l'ARP. L'axe des ordonnées représente la valeur du rapport du ratio de couverture allélique de l'individu sur le ratio attendu par marqueur. Chacune des courbes de couleur représente un groupe ancestral désigné lors de l'analyse. Chaque sous-graphe correspond à un chromosome noté sur la partie droite de la figure. Les régisions encadrées de 'DYN114_Heva' pourraient avoir une origine manquante en plus de l'origine assignée.

4.3.3 Mosaïques et cohérences des méthodes d'ILEA

Les méthodes ILEA sont utilisées sur un sous-jeu de données composé des 14 cultivars sans origines inconnues et 11 représentants de 4 sources (2 *M. a.* ssp. *banksii*, 4 *M. a.* ssp. *malaccensis*, 3 *M. a.* ssp. *zebrina*, 2 *M. schizocarpa*). Le niveau de différenciation mesurée par la statistique F_{ST} (table 4.1) montre une structuration forte entre ces groupes tels que représentés ici. La valeur de F_{ST} la plus faible reste supérieure à 0,2, ce qui est bien inclus dans les tests effectués dans le chapitre 2. Cependant, il faut souligner que des effectifs de 2 individus dans plusieurs populations sources n'ont pas été testés dans cette étude.

TABLE 4.1 – Valeurs de F_{ST} mesurées entre les groupes du sous-jeu de données.

Populations	*banksii*	*malaccensis*	*zebrina*	*Schizocarpa*
banksii		0.22	0.53	0.91
malaccensis	0.22		0.26	0.35
zebrina	0.53	0.26		0.61
Schizocarpa	0.91	0.35	0.61	

4.3.3.1 Cohérence entre méthodes d'ILEA

Les méthodes `ELAI`, `SABER` et `WINPOP` ont été utilisées pour analyser le jeu de données dans 9 conditions (3 sous-échantillons, 3 valeurs de temps depuis l'hybridation), ce qui a produit en tout 27 inférences locales. Elles ont été comparées deux à deux en se concentrant sur la différence entre les méthodes (figure 4.7), l'impact du temps depuis l'hybridation (figure S.12) et la reproductibilité entre sous-échantillons (figure S.11).

La corrélation entre les valeurs d'origines locales inférées par `ELAI` et `SABER` est importante ($> 0,75$), ce qui indique que les inférences produisent des structures similaires. En revanche, les résultats obtenus avec `WINPOP` montrent une corrélation inférieure à 0,5 avec les deux autres méthodes, indiquant une structure inférée s'écartant de celles des méthodes basées sur les HMM (figure 4.7). Visuellement (Figures 4.9, S.9, S.10), les blocs de l'inférence `WINPOP` sont plus étendus que ceux détectés par `ELAI` et `SABER`. De plus, `WINPOP` semble surestimer la proportion de *M. a.* ssp. *zebrina*, (particulièrement sur les chromosomes 4 et 11). Par ailleurs, les résultats de `SABER` sont bruités par rapport à ceux d'`ELAI` et de `WINPOP`.

La corrélation des inférences locales entre les sous-échantillons est très forte pour `ELAI` et `SABER` ($> 0,9$), avec une corrélation légèrement plus haute pour `ELAI`. Concernant `WINPOP`, une grande variabilité entre sous-échantillons est observée, avec des corrélations entre 0,25 et 0,75 (Figure S.11). Visuellement, la variabilité de l'inférence de `WINPOP` est visible avec des « structures en escaliers » sur toute la mosaïque moyennée entre sous-échantillons, qui forme des blocs de hauteur de 0,16 et 0,33 selon si les inférences par sous-échantillons sont différentes (Figure S.10).

La corrélation entre les inférences avec une variation sur le temps depuis l'hybridation est très forte chez `SABER` ($\approx 0,99$) et chez `ELAI` ($> 0,9$), ainsi que chez `WINPOP` dans une moindre mesure ($> 0,5$). Il y a une corrélation légérement plus faible entre les temps 5-50 chez `WINPOP` qu'entre 5-20 et 20-50, et qui n'est pas observée avec `SABER` ou `ELAI` (Figure S.12).

Enfin, on note également de fortes similitudes entre les résultats des inférences par les méthodes `ELAI` et `SABER` et les profils issus de l'approche ARP pour toutes les accessions étudiées (Figure 4.8 et Figures S.13 à S.25)

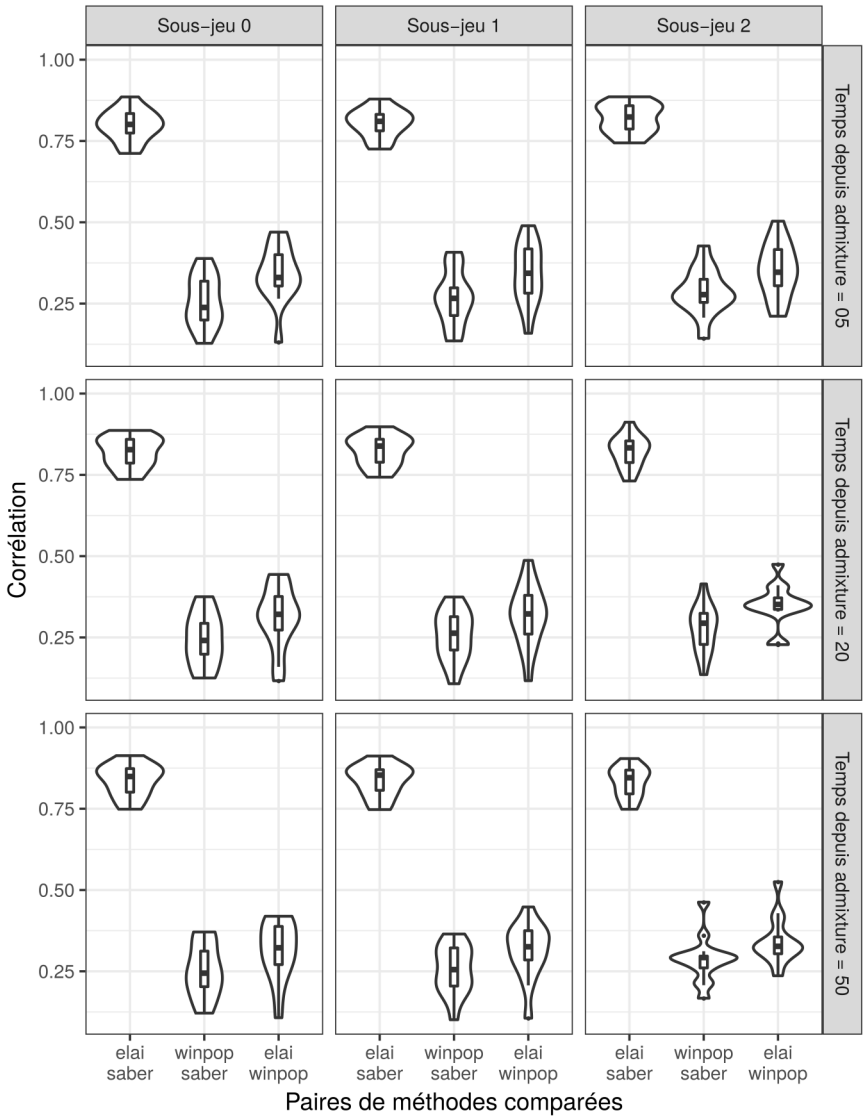

FIGURE 4.7 – Comparaison des inférences par paires de méthodes (`ELAI`, `SABER` et `WINPOP`) en fonction du paramètre de temps depuis l'hybridation et des sous-jeux de données utilisés. L'axe des abscisses indique la paire de méthode comparée sur l'ensemble des individus, pour un sous-jeu de données (en colonne de la grille) et une valeur de temps depuis l'hybridation (en ligne sur la grille). L'axe des ordonnées représente la valeur moyenne de la corrélation, comprise entre 0 et 1. Chaque point de données correspond à la corrélation moyenne entre les origines inférées des marqueurs des deux individus de la paire de condition comparée.

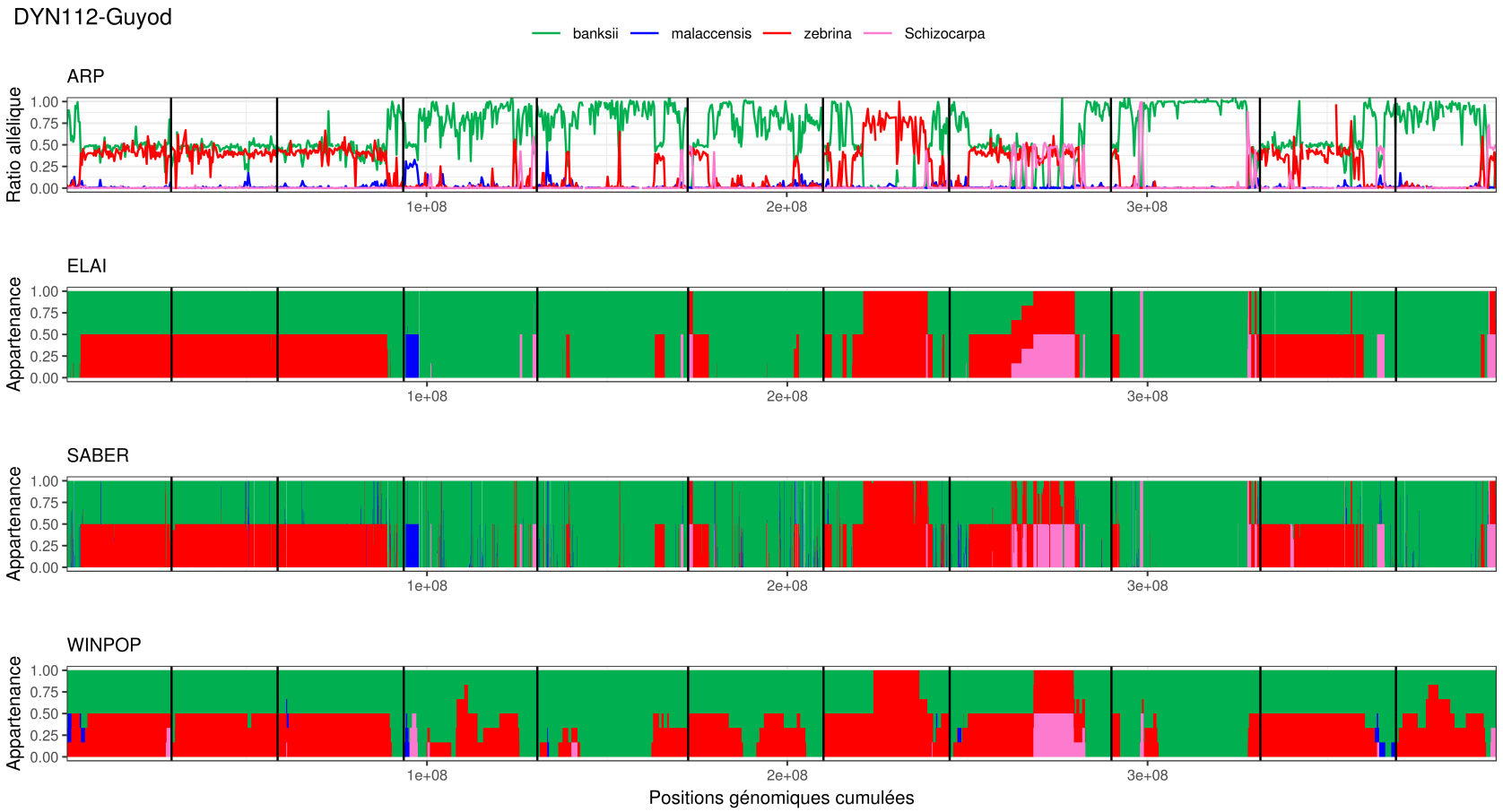

FIGURE 4.8 – Résultats de l'ILEA et « painting » allélique de l'accession 'Guyod'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500 M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

4.3.3.2 Mosaïque du sous jeu de données par ELAI

Il a été montré qu'entre les trois logiciels testés dans le chapitre 2, ELAI était la méthode la plus précise. Ici, ELAI a montré une plus grande stabilité que WINPOP (dans les comparaisons entre sous-échantillons) et un bruit plus faible que SABER, c'est donc son inférence qui fera l'objet d'une interprétation plus détaillée (figure 4.9).

Les mosaïques obtenues montrent chez tous les cultivars étudiés de larges segments chromosomiques voire des chromosomes entiers attribués à *M. a.* ssp. *banksii* (comme attendu, les individus provenant majoritairement de la zone d'origine de *M. a.* ssp. *banksii*). Des segments attribués à *M. a.* ssp. *zebrina* et à *M. schizocarpa* sont aussi prédits chez tous ces cultivars. De plus, des régions chromosomiques assignées à *M. a.* ssp. *malaccensis* sont retrouvées chez 6 individus, avec des segments de taille plus importante chez 5 d'entre eux ('DYN423-Wompa', 'ITC1211-Vudu_Beo', 'DYN-ITC0786-Katual_n2', 'DYN-ITC0785-Aivip', 'DYN-ITC0810-Sihir', Figure 4.9). Chez les deux hybrides AS 'DYN423-Wompa' et 'ITC0822-Tonton_Kepa', les inférences suggèrent la présence d'un haplotype entier de *M. schizocarpa* de même qu'un haplotype majoritairement *M. acuminata*. Parmi les 12 cultivars AAcv de ce jeu de données, trois avaient été prédits comme hybrides et neuf avaient été prédits comme homogènes pour *M. a.* ssp. *banksii* par ADMIXTURE (cluster 'p', Figure 4.4). Les mosaïques inférées montrent des segments d'origines différentes chez tous ces cultivars.

Les individus ont été triés de manière à faire ressortir les motifs de la mosaïque les plus marquants (figure 4.9). Des régions chromosomiques assignées à *M. schizocarpa* et *M. a.* ssp. *zebrina* semblent partagées entre plusieurs accessions en particulier sur les chromosomes 2, 3, 7, 8 et 10. Par exemple, une région assignée à *M. schizocarpa* en début de chromosome 2 est retrouvée chez 9 accessions. Les accessions 'DYN-ITC0589-Gulum', 'ITC1187-Tomolo' et 'ITC1245-Papat' partagent des mosaïques très similaires, avec quelques variations. Les deux individus 'DYN-ITC0785-Aivip' et 'ITC-Vudu_Beo' montrent également des profils de mosaïques similaires sur les chromosomes 5, 7 et 8.

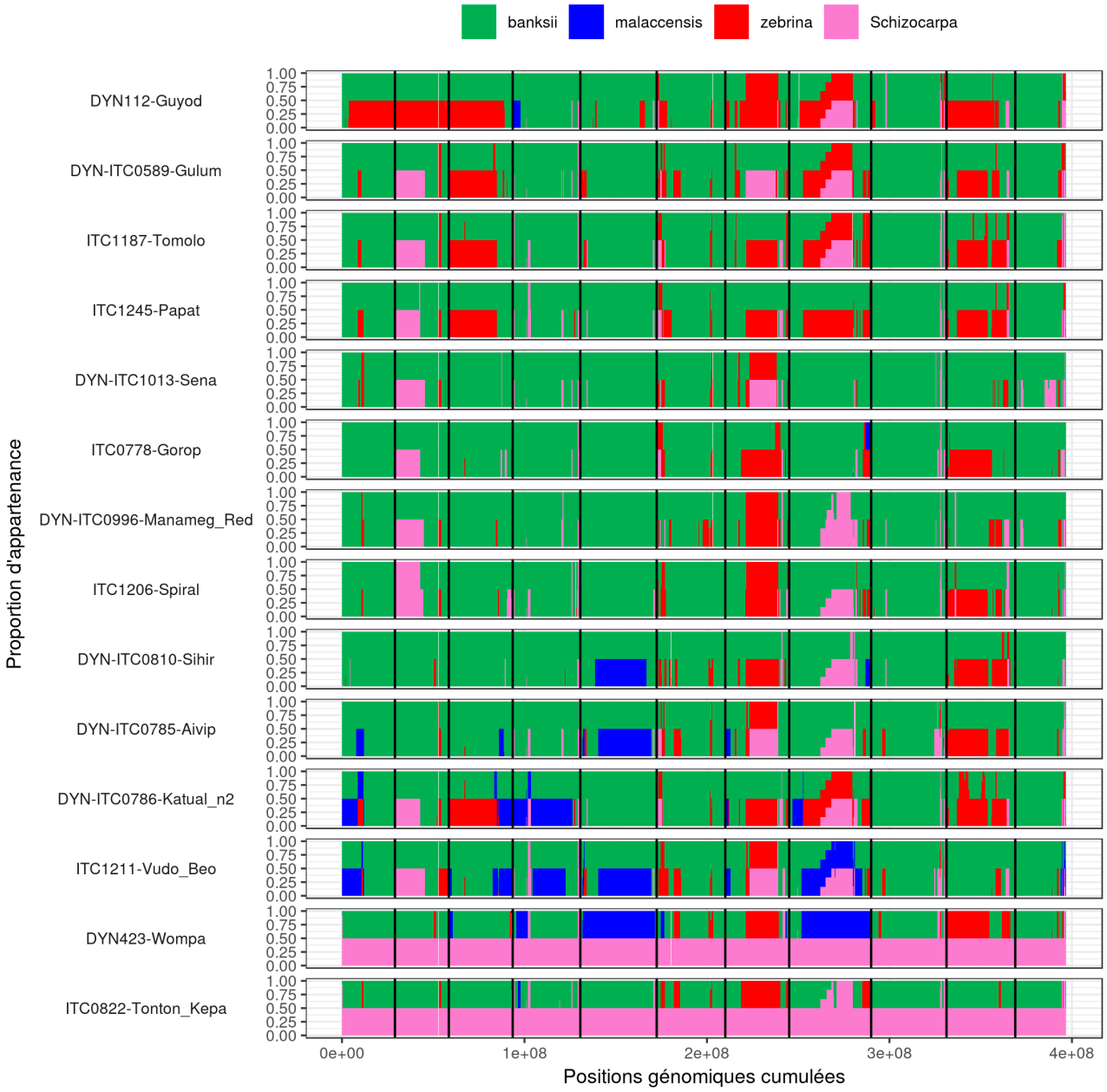

FIGURE 4.9 – Résultats de l'inférence locale d'`ELAI` sur les 14 individus hybrides. Les trois répétitions (sous-échantillons) sont représentées combinées ici, par la moyenne des origines ancestrales inferées par `ELAI`, avec un paramètre de temps depuis l'hybridation de 50 générations. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés sur les 11 chromosomes de bananiers. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500 M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Chaque sous-figure représente un individu et ses 11 chromosomes, séparés par des barres verticales noires.

4.4 Discussion

En combinant des analyses classiques de structure de population (ACP et `ADMIXTURE`) avec une approche empirique de « chromosome painting », nous avons identifié des individus non ou peu introgressés, représentants de groupes génétiques ancestraux connus des bananiers cultivés. L'approche de « chromosome painting » basée sur la détermination de ratios de couverture allélique a également permis la détection de zones du génome des bananiers qu'il n'a pas été possible d'assigner à une origine. Cela semble cohérent avec des travaux récemment publiés sur l'existence d'un ou deux groupes ancestraux pour lesquels des représentants sauvages homogènes n'ont pas encore été identifiés (Sardos et al., 2016b; Christelová et al., 2017; Martin et al., 2020). Cette analyse a rendu possible l'identification de 14 cultivars du jeu de données, sans ancêtre manquant qui ont pu être analysés avec trois méthodes d'ILEA. Deux de ces méthodes, `ELAI` et `SABER` ont donné des résultats reproductibles et corrélés, visuellement similaires à ceux de l'approche ARP. Les mosaïques obtenues montrent en particulier des régions attribuées à *M. a.* ssp. *zebrina* et à *M. schizocarpa* chez des cultivars de Papouasie-Nouvelle-Guinée prédits comme étant homogènes pour *M. a.* ssp. *banksii* par des approches globales.

4.4.1 Diversité des bananiers sauvages

Les résultats des analyses globales (ACP et `ADMIXTURE`) pour les bananiers sauvages sont cohérents avec ceux des études précédentes (Carreel et al., 1994; Perrier et al., 2011; Sardos et al., 2016a) et aussi avec les limites liées à l'échantillonnage de la diversité. Les groupes génétiques correspondant aux espèces *Australimusa*, à *M. balbisiana*, *M. schizocarpa* et aux sous-espèces de *M. acuminata* les plus représentées (*M. a.* ssp. *banksii*, *M. a.* ssp. *zebrina*, *M. a.* ssp. *malaccensis*, *M. a.* ssp. *burmannica/siamea*) ont bien été identifiés par les analyses globales. Pour les espèces de la section *Rhodochlamys* analysées ici (*M. laterita*, *M. rosea*, *M. ornata*, *M. sanguinea*, *M. velutina*) les structurations en un ou deux groupes observées pourraient être liées à une proximité de deux d'entre elles (*M. laterita*, *M. rosea*) avec *M. a.* ssp. *burmannica/siamea* suggérée par analyse phylogénétique (Janssens et al., 2016). Le statut des sous-espèces *M. a.* ssp. *microcarpa* et *M. a.* ssp. *errans* reste à clarifier, leurs accessions représentatives, en faible nombre, ont été trouvées soit proches de *M. a.* ssp. *banksii* soit hybrides comme prédits par des travaux précédents (Carreel et al., 1994; Perrier et al., 2009; Martin et al., 2020), soit regroupées en un groupe génétique spécifique. De même, la structure génétique des *M. a.* ssp. *truncata* et *M. a.* ssp. *sumatrana* représentées par une seule accession chacune dans les collections accessibles (`https://www.crop-diversity.org/mgis/`) reste encore à clarifier.

Les analyses globales ont aussi permis d'identifier des accessions hybrides ou introgréssées parmi les représentants des sous-espèces *M. a.* ssp. *banksii*, *M. a.* ssp. *malaccensis* et *M. a.* ssp.

zebrina. Les analyses locales avec la méthode ARP ont cependant été plus efficaces pour détecter des introgressions chez ces sous-espèces y compris chez des accessions considérées homogènes par l'analyse globale comme observé dans les travaux précédents (Martin et al., 2020) où l'approche locale affinait aussi les prédictions globales. Ces introgressions pouvaient être de petite taille ou concerner des régions de plusieurs centaines de kilobases visibles en particulier chez *M. a.* ssp. *malaccensis*. Certaines introgressions de petite taille détectées peuvent correspondre à des régions conservées entre les sous-espèces mais qui, du fait du faible nombre de représentants, sont attribuées à l'une ou l'autre de ces sous-espèces. Les introgressions de plus grande taille pourraient résulter de contacts entre accessions des différentes sous-espèces ou avec des cultivars ayant une fertilité résiduelle. On ne peut également pas exclure la possibilité d'hybridations dans les collections.

L'identification de représentants non (ou peu) introgressés est possible avec notre approche. Notre sélection de représentants les moins introgressés était basée sur la sélection d'un premier individu représentatif, sur un critère d'homozygotie. Il serait possible de vérifier également la présence possible d'introgressions locales chez ces individus à l'aide de profils d'hétérozygotie le long des chromosomes et en répétant les analyses ARP avec, en tant que première référence, un des autres représentants sélectionnés. Pour éviter des biais possibles dus aux introgressions dans les analyses ARP et ILEA, notre sélection d'accessions représentatives a été assez stringente et il en a résulté un nombre de représentants des groupes ancestraux réduit malgré un échantillonnage un peu plus important que l'étude de (Martin et al., 2020). Il pourrait aussi être envisagé d'utiliser plus d'accessions en masquant les zones introgressées. Cela permettrait de récupérer plus d'informations des groupes ancestraux sur toutes les régions non introgressées. En revanche, cela déséquilibrerait potentiellement l'analyse puisque les ratios de couvertures alléliques pour un groupe ancestral seraient calculés sur un nombre variable d'accessions représentatives le long du génome. Il ressort globalement que le faible nombre d'accessions sauvages disponibles est une des problématiques majeures des études de diversité des bananiers.

4.4.2 Diversité et structure des cultivars diploïdes

Pour les cultivars de bananiers analysés ici, les résultats de structuration avec `ADMIXTURE` ont montré que les deux cultivars F'éi ne semblent pas porter des contributions autres que *Australimusa* (ce qui a été confirmé par l'approche ARP) et ont permis de distinguer les hybrides interspécifiques AB, AS, un groupe de cultivars AAcv trouvés globalement homogènes pour le groupe génétique *M. a.* ssp. *banksii* et des cultivars AAcv hybrides.

Un à deux groupes génétiques ancestraux qui ne sont pas représentés par des accessions sauvages ont été prédits par `ADMIXTURE` chez des cultivars hybrides. Cela semble cohérent avec la mise en évidence récente de contributions d'origine inconnue, présentes seulement chez des

cultivars (Martin et al., 2020). De plus, ces prédictions de groupes génétiques par `ADMIXTURE` peuvent aussi être influencées par une parenté forte entre individus hybrides qui formeraient alors des clusters. L'échantillonnage de la diversité analysé par `ADMIXTURE` est plus large que celui de (Martin et al., 2020) incluant notamment 4 espèces de *Rhodochlamys*, *M. schizocarpa* et toutes les sous-espèces connues de *M. acuminata* mais cela n'a pas permis d'identifier un représentant sauvage homogène pour ces groupes ancestraux. Cette absence de représentants homogènes pour l'ensemble des groupes ancestraux structurant le jeu de données peut induire des biais dans les inférences globales en particulier pour les hybrides. Les profils observés pourraient être dus à la fois à la présence de composantes ancestrales non représentées par des accessions sauvages et une parenté forte entre hybrides. Les résultats de l'analyse ARP permettent de visualiser des régions qui ne sont attribuées à aucun des groupes ancestraux analysés ici et qui pourraient donc dériver de groupes génétiques encore inconnus.

Nous nous sommes focalisés ici sur les 14 cultivars du jeu de données qui ne semblaient pas porter de composantes ancestrales inconnues avec l'analyse ARP et qui pouvaient donc faire l'objet d'analyses par les méthodes ILEA du chapitre précédent. Les mosaïques de leurs génomes sont bien définies par quatre groupes ancestraux correspondant à trois sous-espèces de *M. acuminata* (*M. a.* ssp. *banksii*, *M. a.* ssp. *zebrina*, *M. a.* ssp. *malaccensis*) et *M. schizocarpa*. À l'exception d'une accession, toutes sont originaires de Papouasie-Nouvelle-Guinée. Cette région est un centre de diversité important pour les bananiers avec la présence endémique de *M. a.* ssp. *banksii*, de *M. schizocarpa* et d'une grande diversité de cultivars diploïdes. Les mosaïques des hybrides AS observées ici (Figures S.24 et S.25) précisent la structure d'hybrides parthénocarpiques entre *M. acuminata* et *M. schizocarpa* précédemment identifiés (Carreel et al., 1994). L'inférence suggère la présence d'un haplotype S complet et il est donc possible que ces accessions résultent d'un seul croisement entre des accessions majoritairement *M. acuminata* avec une accession sauvage *M. schizocarpa*, les hybrides formés étant stabilisés par multiplication végétative après l'hybridation A/S.

Pour les cultivars AAcv, leurs mosaïques ont montré des régions génomiques attribuées à *M. a.* ssp. *banksii*, *M. a.* ssp. *zebrina*, *M. schizocarpa* et quelquefois *M. a.* ssp. *malaccensis*, y compris pour les neuf accessions de Papouasie-Nouvelle-Guinée prédites comme homogènes pour *M. a.* ssp. *banksii* par `ADMIXTURE`. Des analyses globales de diversité de cultivars AAcv ont proposé l'existence de deux catégories de cultivars de Papouasie-Nouvelle-Guinée, des cultivars hybrides et des cultivars non hybrides dérivant de *M. a.* ssp. *banksii* (Sardos et al., 2016a,b). Six de nos neuf accessions ('Gulum', 'Gorop', 'Sihir', 'Sena', 'Tomolo', 'Spiral') sont communes avec l'étude de Sardos et al. (2016b) où l'analyse globale avaient classé aussi ces accessions comme étant non-hybrides. L'analyse ancestrale locale a donc permis d'identifier des composantes ancestrales qui n'étaient pas détectées précédemment chez ces accessions. Des structures de mosaïques qui semblent partagées entre certaines accessions suggèrent des origines communes et potentiellement un nombre limité d'individus introgressés impliqués dans la formation de ces cultivars.

L'analyse d'un plus grand nombre d'accessions sauvages et cultivées de Papouasie-Nouvelle-Guinée est nécessaire pour mieux comprendre le processus de formation de ces hybrides. Enfin, la présence de segments attribués à *M. schizocarpa* chez des cultivars diploïdes prédits comme étant dérivés de *M. acuminata* suggère une implication plus importante que prévu de cette espèce à la diversité des bananiers cultivés. Sa contribution à des cultivars triploïdes AAA est donc également possible comme suggéré par l'analyse de séquences ITS (Internal transcribed spacer region) de bananiers AAA d'Afrique de l'Est (Němečková et al., 2018).

4.4.3 Origine ancestrale locale selon le ratio de couverture allélique

Nous avons utilisé ici une approche basée sur les ratios de couverture alléliques qui a permis dans le cadre de l'analyse de la diversité des bananiers de détecter des zones d'origine inconnue dans les mosaïques et de valider des représentants des groupes ancestraux en détectant des introgressions. Les régions chromosomiques où il n'y pas a priori d'informations sur les origines ancestrales sont visibles dans les mosaïques produites par ARP. Cela peut couvrir plusieurs cas, à savoir une zone de faible différenciation entre les groupes ancestraux, une appartenance à un groupe mal échantillonné (dont une partie de la diversité ne serait pas représentée, diversité ayant pu contribuer à des hybrides) ou alors non échantillonné. Il est nécessaire de coupler cette visualisation à d'autres approches pour investiguer plus précisément ce à quoi ces zones correspondent. Par exemple, il devrait être possible de détecter des zones inconnues pour cause de faible différenciation en appliquant une mesure locale de la différenciation entre les groupes (F_{ST}). De la même manière, l'hétérozygotie locale permettrait d'éclairer potentiellement si une zone inconnue comporte une ou deux origines.

D'autres améliorations sont possibles, concernant l'ajustement des ratios de couverture et le lissage des courbes de ratio finales. Il est pour le moment possible que le rapport final du ratio de couverture observée chez un individu sur le ratio attendu soit supérieur à 1, si la couverture de l'allèle est plus grande chez cet individu que la moyenne de la couverture de l'allèle chez les représentants. A priori, mettre un seuil à 1 corrigerait la représentation, mais il reste encore à établir si cette information d'une couverture plus grande chez un non-représentant ne devrait pas être mieux exploitée (plutôt qu'un masquage). Dans l'état actuel de l'approche, l'interprétation des mosaïques est visuelle. La hauteur des courbes de ratio de couverture allélique pour l'interprétation est attendue à 0.5 pour deux origines et à 1 pour une seule origine, mais elle est dépendante de la composition en représentants du groupe inspecté.

On peut supposer que ces niveaux de courbes en deçà de l'attendu peuvent être liés à une hétérozygotie plus forte chez la population ancestrale concernée, et donc à des niveaux moins stables d'attendus de ratio de couverture allélique. Le lissage pourrait s'envisager en déterminant en chaque point l'appartenance (n'appartient pas, est hétérozygote ou homozygote) à un

groupe ancestral à partir du niveau de ratio de couverture allélique. Avec une sortie d'ARP lissée (pour chaque marqueur, son appartenance à aucun, un ou deux groupes ancestraux), il sera possible de chiffrer le taux d'origine inconnue par individu, et aussi de mesurer concrètement le niveau de ressemblance entre les sorties de l'ARP et les méthodes d'ILEA.

4.4.4 ILEA sur le modèle bananier

`ELAI` et `SABER` sont deux méthodes reposant sur des HMM, ce qui laisserait supposer un comportement globalement semblable, comme dans le chapitre 3. C'est ce qui est observé ici sur données réelles. Comme observé dans le chapitre 2 (figure 3.4), `SABER` produit un bruit en motif de code-barre le long de son inférence, ce qui explique ici que la corrélation entre les deux méthodes soit autour de 75% malgré des structures de mosaïques très proches. Visuellement (figures 4.9, S.9), la structure en mosaïque inter(sub)spécifique retrouvée par les deux méthodes est quasiment identique.

Malgré le bruit, les deux méthodes HMM ont une reproductibilité importante entre les sous-échantillons, et sont robustes au paramètre du temps depuis l'hybridation. Ce résultat confirme notre observation sur la robustesse des analyses à une surestimation possible du temps depuis l'hybridation, au sens où les inférences des méthodes HMM varient peu entre elles. De plus, la reproductibilité vient renforcer la confiance accordable à l'analyse qui par définition est sans référence.

L'inférence produite par `WINPOP` s'écarte des attendus des simulations du chapitre 2. La reproductibilité des analyses est bien en deçà de celle des méthodes HMM. Visuellement (figure S.10), la mosaïque comporte beaucoup de blocs situés hors des proportions 0/0,5/1, indiquant une différence entre les analyses des trois sous-échantillons. De la même manière, elle explique la faible corrélation entre `WINPOP` et les méthodes HMM. La mosaïque détectée suit globalement les mêmes motifs que ceux décrits par les méthodes HMM, avec des blocs plus larges comme ce qu'il était déjà possible de constater sur la figure 3.4 du chapitre précédent. De plus, `WINPOP` prédit une participation plus importante de *M. a.* ssp. *zebrina* que les méthodes `ELAI` et `SABER` (par exemple sur le chromosome 11). La méthode `WINPOP` est potentiellement plus sensible aux conditions de l'inférence que les méthodes HMM, ici avec moins de 5 individus représentatifs par groupe. L'inspection des fichiers de log indique que `WINPOP` sélectionne un nombre très faible de marqueurs, moins de 1000 sur des jeux de données compris en 60 000 et 90 000 marqueurs, soit 1,1 à 1,6 % des marqueurs ; il est possible que cela puisse affecter la précision de l'inférence le long des fenêtres considérées par `WINPOP`. Enfin, le taux de recombinaison utilisé ici (Martin et al., 2017) est probablement biaisé par rapport aux groupes ancestraux impliqués chez les individus du sous-jeu de données, étant donné qu'il a été calculé sur des descendants par autofécondation d'un individu issu du groupe *M. a.* ssp. *malaccensis*.

Globalement, les méthodes d'ILEA `ELAI` et `SABER` sont applicables sur le sous-jeu de données, malgré un nombre très réduit de représentants des groupes ancestraux. Les structures en mosaïques sont à la fois cohérentes entre elles entre ces deux méthodes, entre échantillons de SNPs et entre les variation du paramètre « temps depuis l'hybridation ». Elles sont de plus cohérentes avec les résultats de l'approche ARP (figures 4.8 à S.25). Il semble que le nombre élevé de SNPs du jeu de données ainsi que la différenciation forte entre les groupes (table S.3) permettent aux méthodes d'ILEA `ELAI` et `SABER` d'inférer une mosaïque de manière répétable et stable.

Chapitre 5

Discussion et Perspectives

L'objectif de la thèse était d'évaluer, dans des contextes de plantes cultivées non modèles, des méthodes d'inférence ancestrale locale développées dans le cadre de la génétique humaine, et d'appliquer ces méthodes à des données réelles de la diversité des bananiers. Les travaux de thèse ont d'abord visé à reprendre et finaliser un simulateur de données permettant de générer des données issues de scénarios s'approchant des problématiques qui pouvaient être rencontrées chez des plantes cultivées. Nous avons montré que malgré sa simplicité, ce simulateur permettait de produire des données portant des informations génétiques crédibles qui se diffusaient correctement dans les jeux de données simulées.

Dans un second temps, les performances de trois méthodes d'ILEA publiées ont été évaluées à l'aide de simulations. Ces évaluations ont montré que les modes de reproduction spécifiques des plantes ainsi que le nombre de populations ancestrales n'avaient pas un impact important sur les inférences, dans les conditions testées. En revanche, il n'était pas possible avec les méthodes testées de gérer les origines ancestrales inconnues.

Enfin, nous avons analysé un jeu de données de génotypage de 115 accessions de bananiers diploïdes sauvages et cultivés. Une approche empirique basée sur les couvertures en lecture des allèles spécifiques de groupes ancestraux et permettant de visualiser des origines ancestrales le long des chromosomes a été utilisée. Cela a rendu possible une exploration fine des structures mosaïques de représentants de différents groupes génétiques sauvages, ainsi qu'une visualisation de la mosaïque de cultivars et la sélection d'individus analysables par les méthodes ILEA. Les inférences par les méthodes ILEA ont pu être réalisées sur un jeu de données réduit de 25 accessions avec des hybrides comportant quatre origines ancestrales. Ces approches combinées ont permis d'affiner notre connaissance de la diversité des génomes de bananiers.

5.1 Usage des méthodes d'ILEA

5.1.1 Différences et points communs entre simulations et données réelles

Avec nos simulations, nous avons validé dans un premier temps que la précision des méthodes d'inférence était corrélée positivement au niveau de différenciation des groupes ancestraux et à leur nombre de représentants, et inversement corrélée au temps (en génération) depuis l'hybridation. Aux plus faibles niveaux, nous avons testé des jeux de données avec 5 représentants de chaque groupe ancestral, ou 2 représentants d'un groupe pour 20 représentants des autres groupes. Les effets de l'autofécondation (chez une population ancestrale) et de la propagation végétative (chez une population hybride) sont négligeables d'après nos simulations.

Cependant, le cas du sous-jeu de la diversité des bananiers reste plutôt un cas « limite », malgré la couverture de nos scénarios de simulations. Le nombre de représentants des populations ancestrales était globalement faible et l'autofécondation concernait les deux populations les plus faiblement représentées (*M. a.* ssp. *banksii* et *M. schizocarpa*). Il n'y a de plus aucune population atteignant les cinq représentants comme testé dans nos simulations. Seul le niveau de différenciation relativement élevé entre les populations est un élément positif. Concernant le nombre de générations depuis l'évènement d'hybridation, nous avons montré par simulations que même surestimé, ce nombre de générations, dans la gamme de temps évaluée n'a pas eu d'impact négatif fort sur la précision de l'inférence en particulier pour `ELAI` et `WINPOP`. Dans l'analyse des données de bananiers, et dans la gamme testée ici (5, 20 et 50 générations comme paramètre de temps depuis l'hybridation), les corrélations élevées pour les méthodes HMM vont dans le sens d'une robustesse à une surestimation possible de ce paramètre, même si dans ce cas les corrélations étaient meilleures entre `ELAI` et `SABER`. Notons que la vérité des inférences n'étant pas accessible pour les données réelles, la confiance que l'on peut accorder aux résultats est basée sur la cohérence forte retrouvée entre l'approche ARP et les méthodes d'ILEA basées sur des HMM. Les corrélations plus faibles entre `WINPOP` et les méthodes `ELAI` et `SABER` pourraient résulter d'une dépendance plus importante de `WINPOP` aux paramètres biologiques, comme les taux de recombinaisons. Il est aussi possible d'envisager que le filtre de DL de `WINPOP`, utilisé avec les paramètres par défaut nuise à la récupération d'un nombre suffisant de marqueurs informatifs du fait de la présence de populations autogames. Comme mentionné plus haut le jeu de données de bananiers combine plusieurs caractéristiques potentiellement limitantes : quatre sources ancestrales, un faible nombre de représentants des groupes ancestraux, l'autogamie pour deux des groupes ancestraux, l'incertitude sur le nombre de générations après hybridation qui doit de plus varier selon les accessions. Cela pourrait expliquer les différences observées entre les simulations et la réalité ainsi qu'entre les méthodes. Des simulations

plus précises permettraient de mieux prédire le comportement des méthodes d'ILEA et donc d'évaluer leur pertinence.

5.1.2 Limitations et perspectives des méthodes d'ILEA

Une limitation importante des méthodes d'ILEA testées (`SABER`, `ELAI` et `WINPOP`) est la non prise en charge des hybrides portant des composantes d'origines inconnues. Nous suggérions dans le chapitre 3 des approches récemment proposées quand certains groupes ancestraux n'étaient pas représentés. La méthode `MOSAIC` (Salter-Townshend and Myers, 2019) permet de par son modèle d'effectuer des inférences sans informations directes sur les groupes ayant contribué aux hybrides. La seconde approche proposée par Zhou et al. (2016) dans un cas d'usage spécifique, repose sur `ELAI`, qui permet lors de son apprentissage à partir des données (hybrides et représentants des groupes connus) d'apprendre les fréquences alléliques d'un groupe ancestral sans représentants. Cela nécessite cependant une représentation importante du dit groupe ancestral chez les hybrides, ainsi qu'une réduction de leur poids dans l'étape d'apprentissage pour compenser l'effet d'une sur-représentation par rapport aux groupes connus. Ces deux propositions reposent toutefois sur l'usage de données phasées et d'un grand nombre d'individus (800 et 2000 pour les tests réalisés avec `ELAI`, et entre 2700 et 3000 pour les tests réalisés avec `MOSAIC`). Malgré le fait qu'il semble peu probable que ces méthodes fonctionnent avec des échantillonnages réduits, les tester avec nos simulations pourrait permettre d'évaluer concrètement leur capacité à permettre la détection d'un groupe mal caractérisé ou sans représentant.

Les données phasées, qui ne sont pas nécessairement accessibles pour toutes les plantes cultivées, n'ont pas été prises en compte dans les travaux de thèse. Il est cependant démontré que les méthodes d'ILEA fonctionnant avec des données phasées de qualité sont plus précises que celles n'exploitant pas les phases (Guan, 2014). Il existe une grande diversité de méthodes de phasage (Browning and Browning, 2011; Klau and Marschall, 2017; Choi et al., 2018) qui couvrent à la fois des méthodes génétiques et moléculaires. Les techniques moléculaires visent essentiellement à exploiter les informations des lectures pour reconstituer les phases des individus, ce qui est d'autant plus efficace que les lectures sont longues ; ou à combiner différentes technologies (par exemple Edge et al., 2017) de séquençage et d'informations structurales (Klau and Marschall, 2017). Il existe de plus des techniques expérimentales plus spécifiques, comme des approches basées sur le Strand-seq (Porubský et al., 2016). Cette approche permet d'obtenir et de discriminer les séquences de brins d'ADN d'origine maternelle et paternelle et donc d'accéder aux haplotypes. Ces stratégies nécessitent cependant des ressources importantes dédiées au séquençage spécifiquement dans le but de produire des haplotypes. Les méthodes génétiques utilisent les informations d'un grand nombre d'individus (populations) pour produire des haplotypes en exploitant les informations génétiques partagées entre ces individus, par exemple

`BEAGLE` (Browning and Browning, 2007) ou `SHAPEIT` (Delaneau et al., 2012). Il existe de plus des méthodes génétiques reposant sur des trios parents-enfant combinant les informations pour permettre d'établir pour chaque marqueur si sa transmission est maternelle ou paternelle, puis de générer les haplotypes à partir de cette information (Marchini et al., 2006). Les méthodes reposant sur l'information des populations sont a priori plus difficile à mettre en place pour des jeux de données avec peu d'individus.

Une étude par simulations similaire à celle que nous avons réalisée pourrait permettre d'évaluer l'apport de données phasées sur la précision des méthodes d'ILEA comme `SABER` et `ELAI` (qui peuvent aussi traiter ce type de données), particulièrement pour les scénarios avec un grand nombre de populations ancestrales. Ce gain de précision potentiel devrait cependant aussi être évalué avec des données phasées imparfaites (en utilisant la simulation pour introduire des erreurs) par rapport à des données parfaitement phasées et des données non phasées Guan (2014). Enfin, l'étude des données phasées permettraient l'accès à d'autres méthodes d'ILEA basée uniquement sur ce type de données, dont les plus récentes LOTER (Dias-Alves et al., 2018) et `MOSAIC` (Salter-Townshend and Myers, 2019), cette dernière ayant l'avantage de tenir compte de la possibilité d'erreurs de phasage (de type « switching ») dans les données.

Nous avons constaté dans le jeu de données bananiers une forte présence d'individus appartenant majoritairement à une origine ancestrale avec des introgressions d'autres groupes. Ces individus sont porteurs d'une information génétique importante pour la détermination du groupe auquel ils appartiennent, mais nous avons cependant décidé de ne pas les utiliser. Une approche par simulations ajoutant chez chaque population représentative des groupes, des introgressions (de l'ordre de 0 à 10 %) permettrait d'évaluer l'impact sur la précision de l'ILEA des représentants introgressés.

De la même manière, nous avons proposé de répéter les analyses ILEA en échantillonnant le jeu de données de manière à tester la robustesse de l'inférence. Il pourrait être démontré par simulations la pertinence de cette approche, en faisant varier le nombre d'échantillons et leur taille (en marqueurs), en mesurant la précision de chaque répétition, la précision de la moyenne des répétitions et le niveau de corrélation entre les répétitions. Il serait alors possible de faire le lien entre la corrélation entre répétitions et la précision de l'inférence moyenne.

5.1.3 Extension du cadre de travail informatique

Nous avons mis en place un cadre de travail informatique pour permettre d'analyser efficacement les données simulées. L'approche est divisée en six étapes : i) génération des scénarios à simuler, ii) simulation du jeu de données à partir des fichiers du scénario, iii) formatage du fichier vcf et des paramètres en entrée pour les méthodes d'ILEA, iv) analyse par les méthodes

d'ILEA, v) formatage des fichiers de sortie vers un format commun et vi) comparaison entre les méthodes pour les mêmes jeux de données. L'approche a été conçue et automatisée dans le cadre de l'environnement de calcul du cluster de la plateforme Southgreen et n'est pas exportable sur d'autres systèmes sans modifications importantes sur les appels aux programmes nécessaires (entre autre, `vcftools`, `plink`, `sNMF`, `ADMIXTURE`, `SABER`, `ELAI`, `WINPOP`) et sur les lancements de calculs (qui sont écrits pour fonctionner avec le Sun Grid Engine uniquement). Différents efforts ont été réalisés pour utiliser et comparer des méthodes d'ILEA, comme `LAIT` (Hui et al., 2017) qui permet d'utiliser 4 méthodes en même temps à partir du même jeu de données. Très récemment, les auteurs de la revue des méthodes d'ILEA (Geza et al., 2018) ont proposé un outil, `FRANC` (Geza et al., 2019) pour utiliser 8 méthodes d'ILEA sur un même jeu de données (Geza et al., 2019). Cet outil nécessite cependant des étapes supplémentaires à l'installation pour s'assurer que chacune des 8 méthodes d'ILEA ait accès aux bibliothèques logicielles nécessaires à son fonctionnement. De plus, les programmes d'ILEA eux même finissent par ne plus être utilisables ou disponibles, avec par exemple `SABER` qui au début de cette thèse (2017) était téléchargeable, mais qui ne l'est plus au moment de la rédaction de la thèse. Enfin, `FRANC` n'est pas au moment de l'écriture de ce document capable de distribuer les calculs des inférences dans un environnement de calcul distribué. Les développements à suivre concernant les outils pour utiliser plusieurs méthodes d'ILEA devraient se focaliser sur la mise en place de solutions pour distribuer (The Bioconda Team et al., 2018) et conserver (da Veiga Leprevost et al., 2017) de manière pérenne les programmes d'ILEA, et de les inclure quand c'est possible dans des outils de gestion de « workflow » pour permettre un usage indépendant des plateformes de calculs et parallélisable (Di Tommaso et al., 2017, par exemple). Par ailleurs, l'optimisation de la manière de combiner des résultats obtenus par les différentes méthodes demeure une question ouverte et plus particulièrement dans les régions ou des désaccords persistent (plus particulièrement si des approches moins précises sont incluses dans la comparaison).

Le simulateur de données a été mis en place principalement pour permettre un suivi des origines ancestrales chez les individus hybrides simulés et permettre de contrôler l'autofécondation et la propagation végétative par génération. De récents développements dans les simulateurs en « forward » (Kelleher et al., 2018) ont introduit une structure de données permettant de suivre les origines des individus de manière efficace. Celle-ci a ensuite été implémentée dans un simulateur en « forward » populaire, `Slim3` (Haller and Messer, 2019; Haller et al., 2019), simulateur récent qui permet aussi un contrôle fin des modes de reproduction pendant les générations, tout en intégrant des modes de reproduction de type plante. Il semble être un bon candidat pour remplacer la partie « forward » de l'outil présenté dans ce document (`plmgg`), pour permettre de faire reposer nos simulations entièrement sur des outils dont les efforts de développement et les bases d'utilisateurs sont importantes.

5.2 Perspectives pour l'étude des mosaïques des génomes des bananiers

Ces travaux de thèse ont permis de proposer une approche dite ARP basée sur la sélection d'allèles diagnostics identifiés chez différents groupes ancestraux de bananiers et les ratios de couvertures de ces allèles chez les groupes ancestraux et les hybrides, permettant de visualiser des origines ancestrales le long des chromosomes et d'identifier des régions d'origines non déterminées dans la diversité à partir des références utilisées. Deux questions n'ont pas pu être abordées ici : à quoi correspondent les origines non déterminées, et quelle est la mosaïque intersubspécifique des bananiers triploïdes ?

5.2.1 Etude des groupes d'origines inconnues

Un grand nombre d'accessions du jeu de données de bananiers a été identifié comme portant des origines non identifiées, ce qui est cohérent avec l'existence d'une à deux origines inconnues des bananiers cultivés proposée par Martin et al. (2020).

Pour mieux caractériser ces origines inconnues, une stratégie possible est de repérer les accessions qui sont enrichies en ces composantes comme cela était le cas de l'accession 'Pisang Madu' dans l'étude de Martin et al. (2020). Idéalement, les accessions portant des haplotypes complets (ou quasi complets) d'origine inconnue seraient les plus utiles. Un haplotype complet non déterminé par ARP correspond à la situation où pour chaque position génomique représentée sur l'ARP, il y a au plus une seule courbe d'origine ancestrale avec un ratio de couverture proche de 0,5 et aucune avec un ratio proche de 1, le long des chromosomes. Il est donc supposé que parce que l'information manquante est répandue sur tout le génome de l'individu, celui-ci porte potentiellement l'équivalent d'un haplotype complet correspondant à une origine inconnue. La première étape de l'ARP (pour chaque individu de chaque groupe ancestral, sélection des allèles présents uniquement chez le groupe) peut être utilisée pour récupérer les allèles diagnostics de ces individus hybrides, en les assignant chacun leur tour comme un groupe ancestral supplémentaire à l'analyse comprenant les groupes déjà caractérisés (correspondant à *M. balbisiana*, *M. schizocarpa*, *Australimusa* et les sous-espèces de *M. acuminata*). Les listes d'allèles obtenues peuvent être croisées pour ne conserver que les points de données communs, qui seraient donc diagnostics de cette origine inconnue et utilisable pour une nouvelle analyse par ARP. Cette stratégie testée par l'équipe utilise notamment l'accession sauvage 'EN13' qui pourrait porter un haplotype quasi complet d'origine inconnue.

Il est important également de déterminer combien d'origines ancestrales sont manquantes et si elles correspondent à des sous-espèces non identifiées de *M. acuminata* ou à d'autres espèces

Musa. L'accès à des haplotypes (au niveau local) correspondant à ces origines permettrait d'envisager des analyses phylogénétiques. Le génotypage de trios parents-enfants est très utilisé notamment en génétique humaine pour accéder aux haplotypes (par exemple Chen et al., 2013).Cette approche en cours de développement repose sur des données de génotypage issues du reséquençage de trio parents-enfant résultant de différents croisements impliquant notamment l'accession 'Pisang Madu' et également des individus F1 issus de croisements entre les sous-espèces de *M. acuminata* (comme ceux présents dans l'analyse du chapitre précédent). Le cadre de transmission Mendélienne des allèles pour des trios parents-enfants permet de phaser les données SNP pour tous les sites où au moins un des membres des trios est homozygote. Il est donc possible d'obtenir des haplotypes locaux correspondant à des régions d'origines inconnues, ainsi que des séquences consensus pour représenter les groupes *M. a.* ssp. *banksii*, *M. a.* ssp. *burmannica/siamea*, *M. a.* ssp. *malaccensis*, *M. a.* ssp. *zebrina* et *M. schizocarpa* pour réaliser des analyses phylogénétiques.

Ces approches qui sont encore préliminaires suggèrent qu'il serait possible d'extraire via les accessions du projet DYNAMO des pistes plus précises sur ces potentiels groupes ancestraux inconnus. Au vu des très faibles nombres d'individus représentatifs des groupes connus disponibles dans les collections et de l'absence d'accession correspondant aux groupes inconnus, il apparaît qu'une des solutions permettant de mieux caractériser la diversité des bananiers est de pouvoir réaliser plus de prospections en Asie du Sud-Est. En ciblant les zones géographiques correspondant aux accessions porteuses de groupes inconnus, il serait possible d'obtenir de nouvelles accessions introgréssées, exploitables avec les approches par ratio allélique, voire d'identifier des accessions représentatives de ces groupes inconnus.

5.2.2 Le cas de la polyploïdie

Les bananiers triploïdes, qui comportent entre autres le groupe de cultivars 'Cavendish', n'ont pas été traités ici. Ils peuvent toutefois être analysés avec l'approche ARP, puisque l'usage du ratio de couverture allélique n'est pas directement dépendant du niveau de ploïdie. Toutefois, l'interprétation des mosaïques produites par ARP dépend des ratios observés, qui sont liés à la profondeur de séquençage des marqueurs utilisés. Dans le cas du jeu de données DYNAMO, la profondeur moyenne pour les individus triploïdes est de 50X, ce qui permet de bien différencier les ratios possibles de 1/3 ou 2/3. L'utilisation de l'approche ARP permettra donc d'analyser les accessions triploïdes.

La polyploïdie reste un cas problématique pour les approches ILEA. La méthode `ANCESTRYHMM` (Corbett-Detig and Nielsen, 2017) propose une approche par HMM basée sur les données des lectures alignées sur un génome de référence, fonctionnant théoriquement quel que soit le niveau de ploïdie et estimant le paramètre du temps depuis l'hybridation. En utilisant les sorties

de l'approche ARP comme référence, il serait possible de tester `ANCESTRYHMM` et son apport possible à la définition des mosaïques.

Dans l'ensemble, pour des contextes complexes comme celui des bananiers, une approche comme l'ARP nous paraît adaptée car elle permet avec des données non phasées d'obtenir des informations plus précises par rapport aux approches d'inférence globale, de pouvoir visualiser des origines ancestrales le long des chromosomes si ces origines sont représentées, d'identifier des régions potentiellement d'origine inconnue et de gérer différents niveaux de ploïdie (diploïde et triploïde). Il reste cependant à développer une méthode pour discrétiser ses sorties, ainsi qu'à évaluer sa précision, probablement dépendante de la différenciation des populations et du nombre de marqueurs des jeux de données. L'usage des approches ILEA est pour le moment restreint par notre représentation incomplète de la diversité des bananiers dans nos jeux de données et par la polyploïdie. Cependant, l'usage de ces méthodes en plus de l'approche ARP a permis de soutenir nos résultats sur les contributions de l'espèce *M. schizocarpa* aux bananiers cultivés.

La culture des bananiers est affectée par plusieurs maladies et ravageurs et est particulièrement menacée par la fusariose causée par le champignon phytopathogène *Fusarium oxysporum* f. sp. *cubense*. Les stratégies d'amélioration basées sur la reconstruction de triploïdes intègrent maintenant des stratégies d'amélioration au niveau diploïde avec ensuite la reconstruction de triploïdes par des croisements 4x X 2x. Ces processus d'amélioration pourraient bénéficier d'une meilleure connaissance de l'origine et de la structure mosaïque des génomes des bananiers cultivés. Il pourrait également être possible de croiser les informations de mosaïques avec les résultats de recherche de QTLs actuellement en cours pour d'identifier l'origine de caractères d'intérêt.

Bibliographie

D. Ahmed, A. Comte, F. Curk, G. Costantino, F. Luro, A. Dereeper, P. Mournet, Y. Froelicher, and P. Ollitrault. Genotyping by sequencing can reveal the complex mosaic genomes in gene pools resulting from reticulate evolution : a case study in diploid and polyploid citrus. *Annals of Botany*, 123(7) :1231–1251, July 2019. ISSN 0305-7364, 1095-8290. doi : 10.1093/aob/mcz029. URL https://academic.oup.com/aob/article/123/7/1231/5423114.

D. H. Alexander and K. Lange. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12(1) :246, Dec. 2011. ISSN 1471-2105. doi : 10.1186/1471-2105-12-246. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-246.

D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9) :1655–1664, Jan. 2009. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.094052.109. URL http://genome.cshlp.org/content/19/9/1655.

L. Alhusain and A. M. Hafez. Nonparametric approaches for population structure analysis. *Human Genomics*, 12(1), Dec. 2018. ISSN 1479-7364. doi : 10.1186/s40246-018-0156-4. URL https://humgenomics.biomedcentral.com/articles/10.1186/s40246-018-0156-4.

M. L. Arnold. Natural hybridization and the evolution of domesticated, pest and disease organisms. *Molecular Ecology*, 13(5) :997–1007, 2004. ISSN 1365-294X. doi : 10.1111/j.1365-294X.2004.02145.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2004.02145.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-294X.2004.02145.x.

Y. Baran, B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux, C. Eng, W. Rodriguez-Cintron, R. Chapela, J. G. Ford, P. C. Avila, J. Rodriguez-Santana, E. G. Burchard, and E. Halperin. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10) :1359–1367, May 2012. ISSN 1460-2059, 1367-4803. doi : 10.1093/bioinformatics/bts144. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts144.

J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview : analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2) :263–265, Jan. 2005. ISSN 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/bth457. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth457.

S. C. H. Barrett, R. Arunkumar, and S. I. Wright. The demography and population genomics of evolutionary transitions to self-fertilization in plants. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 369(1648), Aug. 2014. ISSN 0962-8436. doi : 10.1098/rstb.2013.0344. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4071518/.

F.-C. Baurens, G. Martin, C. Hervouet, F. Salmon, D. Yohomé, S. Ricci, M. Rouard, R. Habas, A. Lemainque, N. Yahiaoui, and A. D'Hont. Recombination and Large Structural Variations Shape Interspecific Edible Bananas Genomes. *Molecular Biology and Evolution*, 36(1) :97–111, Jan. 2019. ISSN 0737-4038. doi : 10.1093/molbev/msy199. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6340459/.

C. Belser, B. Istace, E. Denis, M. Dubarry, F.-C. Baurens, C. Falentin, M. Genete, W. Berrabah, A.-M. Chèvre, R. Delourme, G. Deniot, F. Denoeud, P. Duffé, S. Engelen, A. Lemainque, M. Manzanares-Dauleux, G. Martin, J. Morice, B. Noel, X. Vekemans, A. D'Hont, M. Rousseau-Gueutin, V. Barbe, C. Cruaud, P. Wincker, and J.-M. Aury. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, 4(11) :879–887, Nov. 2018. ISSN 2055-0278. doi : 10.1038/s41477-018-0289-4. URL http://www.nature.com/articles/s41477-018-0289-4.

R. Boonruangrod, S. Fluch, and K. Burg. Elucidation of origin of the present day hybrid banana cultivars using the 5'ETS rDNA sequence information. *Molecular Breeding*, 24(1) : 77–91, Aug. 2009. ISSN 1380-3743, 1572-9788. doi : 10.1007/s11032-009-9273-z. URL http://link.springer.com/10.1007/s11032-009-9273-z.

Y. Brandvain, T. Slotte, K. M. Hazzouri, S. I. Wright, and G. Coop. Genomic Identification of Founding Haplotypes Reveals the History of the Selfing Species Capsella rubella. *PLoS Genetics*, 9(9), Sept. 2013. ISSN 1553-7390. doi : 10.1371/journal.pgen.1003754. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3772084/.

J. V. Bredeson, J. B. Lyons, S. E. Prochnik, G. A. Wu, C. M. Ha, E. Edsinger-Gonzales, J. Grimwood, J. Schmutz, I. Y. Rabbi, C. Egesi, P. Nauluvula, V. Lebot, J. Ndunguru, G. Mkamilo, R. S. Bart, T. L. Setter, R. M. Gleadow, P. Kulakow, M. E. Ferguson, S. Rounsley, and D. S. Rokhsar. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology*, 34(5) :562–570, May 2016. ISSN 1087-0156, 1546-1696. doi : 10.1038/nbt.3535. URL http://www.nature.com/articles/nbt.3535.

Broad Institute. Picard toolkit. *Broad Institute, Github repository*, 2019. URL http://broadinstitute.github.io/picard/.

S. R. Browning and B. L. Browning. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, 81(5) :1084–1097, Nov. 2007. ISSN

00029297. doi : 10.1086/521987. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929707638828.

S. R. Browning and B. L. Browning. Haplotype phasing : existing methods and new developments. *Nature Reviews Genetics*, 12(10) :703–714, Oct. 2011. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg3054. URL http://www.nature.com/articles/nrg3054.

C. Burgarella, A. Barnaud, N. A. Kane, F. Jankowski, N. Scarcelli, C. Billot, Y. Vigouroux, and C. Berthouly-Salazar. Adaptive Introgression : An Untapped Evolutionary Mechanism for Crop Adaptation. *Frontiers in Plant Science*, 10 :4, Feb. 2019. ISSN 1664-462X. doi : 10.3389/fpls.2019.00004. URL https://www.frontiersin.org/article/10.3389/fpls.2019.00004/full.

H. M. Cann. A Human Genome Diversity Cell Line Panel. *Science*, 296(5566) :261b–262, Apr. 2002. ISSN 00368075, 10959203. doi : 10.1126/science.296.5566.261b. URL https://www.sciencemag.org/lookup/doi/10.1126/science.296.5566.261b.

F. Carreel, S. Fauré, D. G. de León, P. Lagoda, X. Perrier, F. Bakry, H. T. du Montcel, C. Lanaud, and J. Horry. Evaluation de la diversité génétique chez les bananiers diploïdes (Musa sp). *Genetics Selection Evolution*, 26(Suppl 1) :S125, 1994. ISSN 1297-9686. doi : 10.1186/1297-9686-26-S1-S125. URL http://www.gsejournal.org/content/26/S1/S125.

F. Carreel, D. G. de Leon, P. Lagoda, C. Lanaud, C. Jenny, J. P. Horry, and H. T. du Montcel. Ascertaining maternal and paternal lineage within Musa by chloroplast and mitochondrial DNA RFLP analyses. *Genome*, 45(4) :679–692, Aug. 2002. ISSN 0831-2796. doi : 10.1139/g02-033. URL https://www.nrcresearchpress.com/doi/10.1139/g02-033.

L.-Y. Chen, R. VanBuren, M. Paris, H. Zhou, X. Zhang, C. M. Wai, H. Yan, S. Chen, M. Alonge, S. Ramakrishnan, Z. Liao, J. Liu, J. Lin, J. Yue, M. Fatima, Z. Lin, J. Zhang, L. Huang, H. Wang, T.-Y. Hwa, S.-M. Kao, J. Y. Choi, A. Sharma, J. Song, L. Wang, W. C. Yim, J. C. Cushman, R. E. Paull, T. Matsumoto, Y. Qin, Q. Wu, J. Wang, Q. Yu, J. Wu, S. Zhang, P. Boches, C.-W. Tung, M.-L. Wang, G. Coppens d'Eeckenbrugge, G. M. Sanewski, M. D. Purugganan, M. C. Schatz, J. L. Bennetzen, C. Lexer, and R. Ming. The bracteatus pineapple genome and domestication of clonally propagated crops. *Nature Genetics*, 51(10) :1549–1558, Oct. 2019. ISSN 1061-4036, 1546-1718. doi : 10.1038/s41588-019-0506-8. URL http://www.nature.com/articles/s41588-019-0506-8.

W. Chen, B. Li, Z. Zeng, S. Sanna, C. Sidore, F. Busonero, H. M. Kang, Y. Li, and G. R. Abecasis. Genotype calling and haplotyping in parent-offspring trios. *Genome Research*, 23 (1) :142–151, Jan. 2013. ISSN 1088-9051. doi : 10.1101/gr.142455.112. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.142455.112.

Y. Choi, A. P. Chan, E. Kirkness, A. Telenti, and N. J. Schork. Comparison of phasing strategies for whole human genomes. *PLOS Genetics*, 14(4) :e1007308, Apr. 2018. ISSN 1553-7404. doi : 10.1371/journal.pgen.1007308. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007308. Publisher : Public Library of Science.

P. Christelová, E. De Langhe, E. Hřibová, J. Čížková, J. Sardos, M. Hušáková, I. Van den houwe, A. Sutanto, A. K. Kepler, R. Swennen, N. Roux, and J. Doležel. Molecular and cytological characterization of the global *Musa* germplasm collection provides insights into the treasure of banana diversity. *Biodiversity and Conservation*, 26(4) :801–824, Apr. 2017. ISSN 0960-3115, 1572-9710. doi : 10.1007/s10531-016-1273-9. URL http://link.springer.com/10.1007/s10531-016-1273-9.

C. Churchhouse and J. Marchini. Multiway Admixture Deconvolution Using Phased or Unphased Ancestral Panels. *Genetic Epidemiology*, 37(1) :1–12, Jan. 2013. ISSN 07410395. doi : 10.1002/gepi.21692. URL http://doi.wiley.com/10.1002/gepi.21692.

P. Civáň, S. Ali, R. Batista-Navarro, K. Drosou, C. Ihejieto, D. Chakraborty, A. Ray, P. Gladieux, and T. A. Brown. Origin of the Aromatic Group of Cultivated Rice (Oryza sativa L.) Traced to the Indian Subcontinent. *Genome Biology and Evolution*, 11(3) :832–843, Feb. 2019. ISSN 1759-6653. doi : 10.1093/gbe/evz039. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6427689/.

R. Corbett-Detig and R. Nielsen. A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy. *PLOS Genetics*, 13(1) :e1006529, Jan. 2017. ISSN 1553-7404. doi : 10.1371/journal.pgen.1006529. URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006529.

O. E. Cornejo, M.-C. Yee, V. Dominguez, M. Andrews, A. Sockell, E. Strandberg, D. Livingstone, C. Stack, A. Romero, P. Umaharan, S. Royaert, N. R. Tawari, P. Ng, O. Gutierrez, W. Phillips, K. Mockaitis, C. D. Bustamante, and J. C. Motamayor. Population genomic analyses of the chocolate tree, Theobroma cacao L., provide insights into its domestication process. *Communications Biology*, 1, Oct. 2018. ISSN 2399-3642. doi : 10.1038/s42003-018-0168-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6191438/.

A. Cornille, P. Gladieux, M. J. M. Smulders, I. Roldán-Ruiz, F. Laurens, B. Le Cam, A. Nersesyan, J. Clavel, M. Olonova, L. Feugey, I. Gabrielyan, X.-G. Zhang, M. I. Tenaillon, and T. Giraud. New Insight into the History of Domesticated Apple : Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties. *PLoS Genetics*, 8(5) :e1002703, May 2012. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002703. URL https://dx.plos.org/10.1371/journal.pgen.1002703.

F. Curk. *Organisation du complexe d'espèce et décryptage des structures des génomes en mosaïque interspécifiques chez les agrumes cultivés*. Thesis, UM2, Montpellier, 2014. URL http://agritrop.cirad.fr/575172/.

F. Curk, G. Ancillo, F. Ollitrault, X. Perrier, J.-P. Jacquemoud-Collet, A. Garcia-Lor, L. Navarro, and P. Ollitrault. Nuclear Species-Diagnostic SNP Markers Mined from 454 Amplicon Sequencing Reveal Admixture Genomic Structure of Modern Citrus Varieties. *PLOS ONE*, 10(5) :e0125628, May 2015. ISSN 1932-6203. doi : 10.1371/journal.pone.0125628. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0125628.

F. Curk, F. Ollitrault, A. Garcia-Lor, F. Luro, L. Navarro, and P. Ollitrault. Phylogenetic origin of limes and lemons revealed by cytoplasmic and nuclear markers. *Annals of Botany*, 117(4) :565–583, Apr. 2016. ISSN 0305-7364. doi : 10.1093/aob/mcw005. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817432/.

F. da Veiga Leprevost, B. A. Grüning, S. Alves Aflitos, H. L. Röst, J. Uszkoreit, H. Barsnes, M. Vaudel, P. Moreno, L. Gatto, J. Weber, M. Bai, R. C. Jimenez, T. Sachsenberg, J. Pfeuffer, R. Vera Alvarez, J. Griss, A. I. Nesvizhskii, and Y. Perez-Riverol. BioContainers : an open-source and community-driven framework for software standardization. *Bioinformatics*, 33 (16) :2580–2582, Aug. 2017. ISSN 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/btx192. URL https://academic.oup.com/bioinformatics/article/33/16/2580/3096437.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27 (15) :2156–2158, Aug. 2011. ISSN 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/btr330. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330.

E. De Langhe. Relevance of Banana Seeds in Archaeology. *Ethnobotany Research and Applications*, 7 :271, July 2009. ISSN 1547-3465. doi : 10.17348/era.7.0.271-281. URL http://journals.sfu.ca/era/index.php/era/article/view/355.

E. De Langhe, L. Vrydaghs, P. d. Maret, X. Perrier, and T. Denham. Why Bananas Matter : An introduction to the history of banana domestication. *Ethnobotany Research and Applications*, 7(0) :165–177, July 2009. ISSN 1547-3465. URL http://journals.sfu.ca/era/index.php/era/article/view/356.

O. Delaneau, J. Marchini, and J.-F. Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2) :179–181, Feb. 2012. ISSN 1548-7091, 1548-7105. doi : 10.1038/nmeth.1785. URL http://www.nature.com/articles/nmeth.1785.

P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4) :316–319, Apr. 2017. ISSN 1087-0156, 1546-1696. doi : 10.1038/nbt.3820. URL `http://www.nature.com/articles/nbt.3820`.

T. Dias-Alves, J. Mairal, and M. G. B. Blum. Loter : A Software Package to Infer Local Ancestry for a Wide Range of Species. *Molecular Biology and Evolution*, 35(9) :2318–2326, Sept. 2018. ISSN 0737-4038. doi : 10.1093/molbev/msy126. URL `https://academic.oup.com/mbe/article/35/9/2318/5040668`.

K. S. Dodds. Musa Fehi, the Indigenous Banana of Fiji. *Nature*, 157(3996) :729–730, June 1946. ISSN 0028-0836, 1476-4687. doi : 10.1038/157729c0. URL `http://www.nature.com/articles/157729c0`.

J. Dubcovsky, M. Luo, and J. Dvorak. Differentiation between homoeologous chromosomes 1A of wheat and 1Am of Triticum monococcum and its recognition by the wheat Ph1 locus. *Proceedings of the National Academy of Sciences*, 92(14) :6645–6649, July 1995. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.92.14.6645. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.92.14.6645`.

M. Dupouy, F.-C. Baurens, P. Derouault, C. Hervouet, C. Cardi, C. Cruaud, B. Istace, K. Labadie, C. Guiougou, L. Toubi, F. Salmon, P. Mournet, M. Rouard, N. Yahiaoui, A. Lemainque, G. Martin, and A. D'Hont. Two large reciprocal translocations characterized in the disease resistance-rich burmannica genetic group of Musa acuminata. *Annals of Botany*, 124 (2) :319–329, Sept. 2019. ISSN 0305-7364, 1095-8290. doi : 10.1093/aob/mcz078. URL `https://academic.oup.com/aob/article/124/2/319/5523264`.

A. D'Hont, A. Paget-Goy, J. Escoute, and F. Carreel. The interspecific genome structure of cultivated banana, Musa spp. revealed by genomic DNA in situ hybridization. *Theoretical and Applied Genetics*, 100(2) :177–183, Jan. 2000. ISSN 0040-5752, 1432-2242. doi : 10.1007/s001220050024. URL `http://link.springer.com/10.1007/s001220050024`.

A. D'Hont, F. Denoeud, J.-M. Aury, F.-C. Baurens, F. Carreel, O. Garsmeur, B. Noel, S. Bocs, G. Droc, M. Rouard, C. Da Silva, K. Jabbari, C. Cardi, J. Poulain, M. Souquet, K. Labadie, C. Jourda, J. Lengellé, M. Rodier-Goud, A. Alberti, M. Bernard, M. Correa, S. Ayyampalayam, M. R. Mckain, J. Leebens-Mack, D. Burgess, M. Freeling, D. Mbéguié-A-Mbéguié, M. Chabannes, T. Wicker, O. Panaud, J. Barbosa, E. Hribova, P. Heslop-Harrison, R. Habas, R. Rivallan, P. Francois, C. Poiron, A. Kilian, D. Burthia, C. Jenny, F. Bakry, S. Brown, V. Guignon, G. Kema, M. Dita, C. Waalwijk, S. Joseph, A. Dievart, O. Jaillon, J. Leclercq, X. Argout, E. Lyons, A. Almeida, M. Jeridi, J. Dolezel, N. Roux, A.-M. Risterucci, J. Weissenbach, M. Ruiz, J.-C. Glaszmann, F. Quétier, N. Yahiaoui, and P. Wincker. The

banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, 488(7410) :213–217, Aug. 2012. ISSN 0028-0836. doi : 10.1038/nature11241. URL http://www.nature.com/nature/journal/v488/n7410/full/nature11241.html.

P. Edge, V. Bafna, and V. Bansal. HapCUT2 : robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, 27(5) :801–812, May 2017. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.213462.116. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.213462.116.

M. El Baidouri, F. Murat, M. Veyssiere, M. Molinier, R. Flores, L. Burlot, M. Alaux, H. Quesneville, C. Pont, and J. Salse. Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytologist*, 213(3) :1477–1486, Feb. 2017. ISSN 0028646X. doi : 10.1111/nph.14113. URL http://doi.wiley.com/10.1111/nph.14113.

D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. *Genetics*, 164(4) :1567–1587, Aug. 2003. ISSN 0016-6731.

S. Fauré, J.-L. Noyer, F. Carreel, J.-P. Horry, F. Bakry, and C. Lanaud. Maternal inheritance of chloroplast genome and paternal inheritance of mitochondrial genome in bananas (Musa acuminata). *Current Genetics*, 25(3) :265–269, Mar. 1994. ISSN 0172-8083, 1432-0983. doi : 10.1007/BF00357172. URL http://link.springer.com/10.1007/BF00357172.

E. Frichot, F. Mathieu, T. Trouillon, G. Bouchard, and O. François. Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics*, 196(4) :973–983, Apr. 2014. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.113.160572. URL http://www.genetics.org/content/196/4/973.

X. Gao and J. Starmer. Human population structure detection via multilocus genotype clustering. *BMC genetics*, 8 :34, June 2007. ISSN 1471-2156. doi : 10.1186/1471-2156-8-34.

O. Garsmeur, G. Droc, R. Antonise, J. Grimwood, B. Potier, K. Aitken, J. Jenkins, G. Martin, C. Charron, C. Hervouet, L. Costet, N. Yahiaoui, A. Healey, D. Sims, Y. Cherukuri, A. Sreedasyam, A. Kilian, A. Chan, M.-A. Van Sluys, K. Swaminathan, C. Town, H. Bergès, B. Simmons, J. C. Glaszmann, E. van der Vossen, R. Henry, J. Schmutz, and A. D'Hont. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature Communications*, 9(1) :2638, Dec. 2018. ISSN 2041-1723. doi : 10.1038/s41467-018-05051-5. URL http://www.nature.com/articles/s41467-018-05051-5.

E. Geza, J. Mugo, N. J. Mulder, A. Wonkam, E. R. Chimusa, and G. K. Mazandu. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Briefings in Bioinformatics*, June 2018. ISSN 1467-5463, 1477-4054. doi : 10.1093/bib/bby044. URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby044/5047382.

E. Geza, N. J. Mulder, E. R. Chimusa, and G. K. Mazandu. FRANC : a unified framework for multi-way local ancestry deconvolution with high density SNP data. *Briefings in Bioinformatics*, page bbz117, Nov. 2019. ISSN 1467-5463, 1477-4054. doi : 10.1093/bib/bbz117. URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz117/5612160.

S. Glémin, C. Scornavacca, J. Dainat, C. Burgarella, V. Viader, M. Ardisson, G. Sarah, S. Santoni, J. David, and V. Ranwez. Pervasive hybridizations in the history of wheat relatives. *Science Advances*, 5(5) :eaav9188, May 2019. ISSN 2375-2548. doi : 10.1126/sciadv.aav9188. URL https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aav9188.

Y. Guan. Detecting structure of haplotypes and local ancestry. *Genetics*, 196(3) :625–642, Mar. 2014. ISSN 1943-2631. doi : 10.1534/genetics.113.160697.

B. C. Haller and P. W. Messer. SLiM 3 : Forward Genetic Simulations Beyond the Wright-Fisher Model. *Molecular Biology and Evolution*, 36(3) :632–637, 2019. ISSN 1537-1719. doi : 10.1093/molbev/msy228.

B. C. Haller, J. Galloway, J. Kelleher, P. W. Messer, and P. L. Ralph. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, 19(2) :552–566, Mar. 2019. ISSN 1755-098X, 1755-0998. doi : 10.1111/1755-0998.12968. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12968.

P. W. Hedrick. *Genetics of populations*. Jones and Bartlett Publishers, Sudbury, Mass, 4th ed edition, 2011. ISBN 978-0-7637-5737-3. OCLC : ocn369303950.

G. Hellenthal, G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. A Genetic Atlas of Human Admixture History. *Science*, 343(6172) :747–751, Feb. 2014. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1243518. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1243518.

W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6) :226–231, June 1968. ISSN 0040-5752, 1432-2242. doi : 10.1007/BF01245622. URL http://link.springer.com/10.1007/BF01245622.

I. Hippolyte, C. Jenny, L. Gardes, F. Bakry, R. Rivallan, V. Pomies, P. Cubry, K. Tomekpe, A. M. Risterucci, N. Roux, M. Rouard, E. Arnaud, M. Kolesnikova-Allen, and X. Perrier. Foundation characteristics of edible Musa triploids revealed from allelic distribution of SSR markers. *Annals of Botany*, 109(5) :937–951, Apr. 2012. ISSN 1095-8290, 0305-7364. doi : 10.1093/aob/mcs010. URL https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcs010.

S. Hoban, G. Bertorelle, and O. E. Gaggiotti. Computer simulations : tools for population and evolutionary genetics. *Nature Reviews Genetics*, 13(2) :110–122, Feb. 2012. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg3130. URL http://www.nature.com/articles/nrg3130.

R. R. Hudson. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2) :337–338, Feb. 2002. ISSN 1367-4803. doi : 10.1093/bioinformatics/18.2.337. URL https://academic.oup.com/bioinformatics/article/18/2/337/225783.

M. B. Hufford, P. Lubinksy, T. Pyhäjärvi, M. T. Devengenzo, N. C. Ellstrand, and J. Ross-Ibarra. The Genomic Signature of Crop-Wild Introgression in Maize. *PLoS Genetics*, 9(5) : e1003477, May 2013. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003477. URL https://dx.plos.org/10.1371/journal.pgen.1003477.

D. Hui, Z. Fang, J. Lin, Q. Duan, Y. Li, M. Hu, and W. Chen. LAIT : a local ancestry inference toolkit. *BMC Genetics*, 18(1) :83, Sept. 2017. ISSN 1471-2156. doi : 10.1186/s12863-017-0546-y. URL https://doi.org/10.1186/s12863-017-0546-y.

M. Häkkinen. Reappraisal of sectional taxonomy in *Musa* (*Musaceae*). *Taxon*, 62(4) :809–813, Aug. 2013. ISSN 0040-0262. doi : 10.12705/624.3. URL http://doi.wiley.com/10.12705/624.3.

A. Intarapanich, P. J. Shaw, A. Assawamakin, P. Wangkumhang, C. Ngamphiw, K. Chaichoompu, J. Piriyapongsa, and S. Tongsima. Iterative pruning PCA improves resolution of highly structured populations. *BMC Bioinformatics*, 10(1) :382, 2009. ISSN 1471-2105. doi : 10.1186/1471-2105-10-382. URL http://www.biomedcentral.com/1471-2105/10/382.

S. B. Janssens, F. Vandelook, E. D. Langhe, B. Verstraete, E. Smets, I. Vandenhouwe, and R. Swennen. Evolutionary dynamics and biogeography of Musaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. *New Phytologist*, 210(4) :1453–1465, June 2016. ISSN 1469-8137. doi : 10.1111/nph.13856. URL https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.13856.

T. Jombart. adegenet : a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11) :1403–1405, June 2008. ISSN 1460-2059, 1367-4803. doi : 10.1093/bioinformatics/btn129. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn129.

J. Kelleher, K. R. Thornton, J. Ashander, and P. L. Ralph. Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, 14(11) :e1006581, Nov. 2018. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1006581. URL http://dx.plos.org/10.1371/journal.pcbi.1006581.

J. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3) :235–248, Sept. 1982. ISSN 03044149. doi : 10.1016/0304-4149(82)90011-4. URL https://linkinghub.elsevier.com/retrieve/pii/0304414982900114.

G. W. Klau and T. Marschall. A Guided Tour to Computational Haplotyping. In J. Kari, F. Manea, and I. Petre, editors, *Unveiling Dynamics and Complexity*, volume 10307, pages 50–63. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58740-0 978-3-319-58741-7. doi : 10.1007/978-3-319-58741-7_6. URL http://link.springer.com/10.1007/978-3-319-58741-7_6.

R. Lacy. VORTEX : a computer simulation model for population viability analysis. *Wildlife Research*, 20(1) :45, 1993. ISSN 1035-3712. doi : 10.1071/WR9930045. URL http://www.publish.csiro.au/?paper=WR9930045.

D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8(1) :e1002453, Jan. 2012. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002453. URL http://dx.plos.org/10.1371/journal.pgen.1002453.

D. J. Lawson, L. van Dorp, and D. Falush. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9(1) :3258, Dec. 2018. ISSN 2041-1723. doi : 10.1038/s41467-018-05257-7. URL http://www.nature.com/articles/s41467-018-05257-7.

T. Lescot. Banane. Diversité génétique. *Fruitrop (Ed. Française)*, (256) :92–96, 2018. ISSN 1256-544X. URL http://www.fruitrop.com/media/Magazines-FruiTrop/2018/fruitrop-256. Place : France Section : CIRAD-PERSYST-UPR Systèmes bananes et ananas (FRA).

H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv :1303.3997 [q-bio]*, May 2013. URL http://arxiv.org/abs/1303.3997. arXiv : 1303.3997.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14) :1754–1760, July 2009. ISSN 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/btp324. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16) :2078–2079, Aug. 2009. ISSN 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/btp352. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4) :2213–2233, Dec. 2003. ISSN 0016-6731. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462870/.

T. Limpiti, C. Amornbunchornvej, A. Intarapanich, A. Assawamakin, and S. Tongsima. iNJclust : Iterative Neighbor-Joining Tree Clustering Framework for Inferring Population Structure. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11 (5) :903–914, Sept. 2014. ISSN 1545-5963. doi : 10.1109/TCBB.2014.2322372. URL http://ieeexplore.ieee.org/document/6811209/.

N. Liu and H. Zhao. A non-parametric approach to population structure inference using multilocus genotypes. *Human Genomics*, 2(6) :353, 2006. ISSN 1479-7364. doi : 10.1186/1479-7364-2-6-353. URL http://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-2-6-353.

Y. Liu, G. Athanasiadis, and M. E. Weale. A survey of genetic simulation software for population and epidemiological studies. *Human Genomics*, 3(1) :79, 2008. ISSN 1479-7364. doi : 10.1186/1479-7364-3-1-79. URL http://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-3-1-79.

Y. Liu, T. Nyunoya, S. Leng, S. A. Belinsky, Y. Tesfaigzi, and S. Bruse. Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics*, 7(1) :1, 2013. ISSN 1479-7364. doi : 10.1186/1479-7364-7-1. URL http://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-7-1.

P.-R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, and B. Berger. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, 193(4) :1233–1254, Apr. 2013. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.112.147330. URL http://www.genetics.org/lookup/doi/10.1534/genetics.112.147330.

H. Ma, Q. Pan, L. Wang, Z. Li, Y. Wan, and X. Liu. *Musella lasiocarpa* var. *rubribracteata* (Musaceae), a New Variety from Sichuan, China. *Novon : A Journal for Botanical Nomenclature*, 21(3) :349–353, Sept. 2011. ISSN 1055-3177. doi : 10.3417/2010125. URL http://www.bioone.org/doi/abs/10.3417/2010125.

B. Maples, S. Gravel, E. Kenny, and C. Bustamante. RFMix : A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, 93(2) :278–288, Aug. 2013. ISSN 0002-9297. doi : 10.1016/j.ajhg.2013.06.020. URL http://www.sciencedirect.com/science/article/pii/S0002929713002899.

J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, G. R. Abecasis, and P. Donnelly. A Comparison of Phasing Algorithms for Trios and Unrelated Individuals. *The American Journal of Human Genetics*, 78(3) :437–450, Mar. 2006. ISSN 00029297. doi : 10.1086/500808. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929707623830.

G. Martin, F.-C. Baurens, G. Droc, M. Rouard, A. Cenci, A. Kilian, A. Hastie, J. Doležel, J.-M. Aury, A. Alberti, F. Carreel, and A. D'Hont. Improvement of the banana "Musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics*, 17(1) :243, Dec. 2016. ISSN 1471-2164. doi : 10.1186/s12864-016-2579-4. URL http://www.biomedcentral.com/1471-2164/17/243.

G. Martin, F. Carreel, O. Coriton, C. Hervouet, C. Cardi, P. Derouault, D. Roques, F. Salmon, M. Rouard, J. Sardos, K. Labadie, F.-C. Baurens, and A. D'Hont. Evolution of the Banana Genome (Musa acuminata) Is Impacted by Large Chromosomal Translocations. *Molecular Biology and Evolution*, 34(9) :2140–2152, Sept. 2017. ISSN 0737-4038, 1537-1719. doi : 10.1093/molbev/msx164. URL https://academic.oup.com/mbe/article/34/9/2140/3852084.

G. Martin, C. Cardi, G. Sarah, S. Ricci, C. Jenny, E. Fondi, X. Perrier, J. Glaszmann, A. D'Hont, and N. Yahiaoui. Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana. *The Plant Journal*, page tpj.14683, Feb. 2020. ISSN 0960-7412, 1365-313X. doi : 10.1111/tpj.14683. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.14683.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9) :1297–1303, Sept. 2010. ISSN 1088-9051. doi : 10.1101/gr.107524.110. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.107524.110.

G. A. T. McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2) : 987–991, Oct. 2002. ISSN 0016-6731.

A. J. Miller and B. L. Gross. From forest to field : Perennial fruit crop domestication. *American Journal of Botany*, 98(9) :1389–1414, Sept. 2011. ISSN 00029122. doi : 10.3732/ajb.1000522. URL http://doi.wiley.com/10.3732/ajb.1000522.

S. G. Milner, M. Jost, S. Taketa, E. R. Mazón, A. Himmelbach, M. Oppermann, S. Weise, H. Knüpffer, M. Basterrechea, P. König, D. Schüler, R. Sharma, R. K. Pasam, T. Rutten, G. Guo, D. Xu, J. Zhang, G. Herren, T. Müller, S. G. Krattinger, B. Keller, Y. Jiang, M. Y. González, Y. Zhao, A. Habekuß, S. Färber, F. Ordon, M. Lange, A. Börner, A. Graner, J. C. Reif, U. Scholz, M. Mascher, and N. Stein. Genebank genomics highlights the diversity of a global barley collection. *Nature Genetics*, 51(2) :319–326, Feb. 2019. ISSN 1061-4036, 1546-1718. doi : 10.1038/s41588-018-0266-x. URL http://www.nature.com/articles/s41588-018-0266-x.

A. Moreno-Estrada, S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes, C. R. Gignoux, P. A. Ortiz-Tello, R. J. Martínez, D. J. Hedges, R. W. Morris, C. Eng, K. Sandoval, S. Acevedo-

Acevedo, P. J. Norman, Z. Layrisse, P. Parham, J. C. Martínez-Cruzado, E. G. Burchard, M. L. Cuccaro, E. R. Martin, and C. D. Bustamante. Reconstructing the Population Genetic History of the Caribbean. *PLoS Genetics*, 9(11) :e1003925, Nov. 2013. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003925. URL http://dx.plos.org/10.1371/journal.pgen.1003925.

A. Němečková, P. Christelová, J. Čížková, M. Nyine, I. Van den houwe, R. Svačina, B. Uwimana, R. Swennen, J. Doležel, and E. Hřibová. Molecular and Cytogenetic Study of East African Highland Banana. *Frontiers in Plant Science*, 9 :1371, Oct. 2018. ISSN 1664-462X. doi : 10.3389/fpls.2018.01371. URL https://www.frontiersin.org/article/10.3389/fpls.2018.01371/full.

B. Padhukasahasram. Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics*, 5 :204, July 2014. ISSN 1664-8021. doi : 10.3389/fgene.2014.00204. URL http://journal.frontiersin.org/article/10.3389/fgene.2014.00204/abstract.

N. Patterson, N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. O'Brien, D. Altshuler, M. J. Daly, and D. Reich. Methods for High-Density Admixture Mapping of Disease Genes. *The American Journal of Human Genetics*, 74(5) :979–1000, May 2004. ISSN 00029297. doi : 10.1086/420871. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929707643638.

N. Patterson, A. L. Price, and D. Reich. Population Structure and Eigenanalysis. *PLOS Genetics*, 2(12) :e190, Dec. 2006. ISSN 1553-7404. doi : 10.1371/journal.pgen.0020190. URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190.

B. Paşaniuc, S. Sankararaman, G. Kimmel, and E. Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12) :i213–i221, June 2009. ISSN 1367-4803. doi : 10.1093/bioinformatics/btp197. URL https://academic.oup.com/bioinformatics/article/25/12/i213/188095.

X. Perrier, F. Bakry, F. Carreel, C. Jenny, J.-P. Horry, V. Lebot, and I. Hippolyte. Combining Biological Approaches to Shed Light on the Evolution of Edible Bananas. *Ethnobotany Research and Applications*, 7 :199, July 2009. ISSN 1547-3465. doi : 10.17348/era.7.0.199-216. URL http://journals.sfu.ca/era/index.php/era/article/view/362.

X. Perrier, E. De Langhe, M. Donohue, C. Lentfer, L. Vrydaghs, F. Bakry, F. Carreel, I. Hippolyte, J.-P. Horry, C. Jenny, V. Lebot, A.-M. Risterucci, K. Tomekpe, H. Doutrelepont, T. Ball, J. Manwaring, P. de Maret, and T. Denham. Multidisciplinary perspectives on banana (*Musa spp.*) domestication. *Proceedings of the National Academy of Sciences*, 108(28) : 11311–11318, July 2011. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1102001108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1102001108.

X. Perrier, C. Jenny, F. Bakry, D. Karamura, M. Kitavi, C. Dubois, C. Hervouet, G. Philippson, and E. De Langhe. East African diploid and triploid bananas : a genetic complex transported from South-East Asia. *Annals of Botany*, 123(1) :19–36, Jan. 2019. ISSN 0305-7364, 1095-8290. doi : 10.1093/aob/mcy156. URL `https://academic.oup.com/aob/article/123/1/19/5104470`.

L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, A. Carracedo, and M. V. Lareu. An overview of STRUCTURE : applications, parameter settings, and supporting software. *Frontiers in Genetics*, 4, 2013. ISSN 1664-8021. doi : 10.3389/fgene.2013.00098. URL `http://journal.frontiersin.org/article/10.3389/fgene.2013.00098/abstract`.

D. Porubský, A. D. Sanders, N. van Wietmarschen, E. Falconer, M. Hills, D. C. Spierings, M. R. Bevova, V. Guryev, and P. M. Lansdorp. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Research*, 26(11) :1565–1574, Nov. 2016. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.209841.116. URL `http://genome.cshlp.org/lookup/doi/10.1101/gr.209841.116`.

A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics*, 5(6) :e1000519, June 2009. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000519. URL `https://dx.plos.org/10.1371/journal.pgen.1000519`.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945–959, June 2000. ISSN 0016-6731.

S. J. Puechmaille. The program structure does not reliably recover the correct population structure when sampling is uneven : subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, 16 (3) :608–627, May 2016. ISSN 1755098X. doi : 10.1111/1755-0998.12512. URL `http://doi.wiley.com/10.1111/1755-0998.12512`.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3) :559–575, Sept. 2007. ISSN 00029297. doi : 10.1086/519795. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929707613524`.

M. D. Purugganan. Evolutionary Insights into the Nature of Plant Domestication. *Current Biology*, 29(14) :R705–R714, July 2019. ISSN 0960-9822. doi : 10.1016/j.cub.2019.05.053. URL `https://www.cell.com/current-biology/abstract/S0960-9822(19)30623-2`. Publisher : Elsevier.

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, Feb. 1989. ISSN 00189219. doi : 10.1109/5.18626. URL http://ieeexplore.ieee.org/document/18626/.

L.-M. Raboin, F. Carreel, J.-L. Noyer, F.-C. Baurens, J.-P. Horry, F. Bakry, H. T. D. Montcel, J. Ganry, C. Lanaud, and P. J. Lagoda. Diploid Ancestors of Triploid Export Banana Cultivars : Molecular Identification of 2n Restitution Gamete Donors and n Gamete Donors. *Molecular Breeding*, 16(4) :333–341, Nov. 2005. ISSN 1380-3743, 1572-9788. doi : 10.1007/s11032-005-2452-7. URL http://link.springer.com/10.1007/s11032-005-2452-7.

A. Raj, M. Stephens, and J. K. Pritchard. fastSTRUCTURE : Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, 197(2) :573–589, June 2014. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.114.164350. URL http://www.genetics.org/lookup/doi/10.1534/genetics.114.164350.

M. Rouard, G. Droc, G. Martin, J. Sardos, Y. Hueber, V. Guignon, A. Cenci, B. Geigle, M. S. Hibbins, N. Yahiaoui, F.-C. Baurens, V. Berry, M. W. Hahn, A. D'Hont, and N. Roux. Three new genome assemblies support a rapid radiation in Musa acuminata (wild banana). *Genome Biology and Evolution*, Oct. 2018. ISSN 1759-6653. doi : 10.1093/gbe/evy227. URL https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evy227/5129088.

M. Ruas, V. Guignon, G. Sempere, J. Sardos, Y. Hueber, H. Duvergey, A. Andrieu, R. Chase, C. Jenny, T. Hazekamp, B. Irish, K. Jelali, J. Adeka, T. Ayala-Silva, C. Chao, J. Daniells, B. Dowiya, B. Effa effa, L. Gueco, L. Herradura, L. Ibobondji, E. Kempenaers, J. Kilangi, S. Muhangi, P. Ngo Xuan, J. Paofa, C. Pavis, D. Thiemele, C. Tossou, J. Sandoval, A. Sutanto, G. Vangu Paka, G. Yi, I. Van den houwe, N. Roux, and M. Rouard. MGIS : managing banana (Musa spp.) genetic resources information and high-throughput genotyping data. *Database*, 2017, Jan. 2017. ISSN 1758-0463. doi : 10.1093/database/bax046. URL https://academic.oup.com/database/article/doi/10.1093/database/bax046/3866796.

M. Salter-Townshend and S. Myers. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, 212(3) :869–889, July 2019. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.119.302139. URL http://www.genetics.org/lookup/doi/10.1534/genetics.119.302139.

S. Sankararaman, G. Kimmel, E. Halperin, and M. I. Jordan. On the inference of ancestries in admixed populations. *Genome Research*, 18(4) :668–675, Mar. 2008a. ISSN 1088-9051. doi : 10.1101/gr.072751.107. URL http://www.genome.org/cgi/doi/10.1101/gr.072751.107.

S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics*, 82(2) :290–303, Feb. 2008b. ISSN 0002-9297. doi : 10.1016/j.ajhg.2007.09.022. URL http://www.sciencedirect.com/science/article/pii/S0002929708000797.

J. D. Santos, D. Chebotarov, K. L. McNally, J. Bartholomé, G. Droc, C. Billot, and J. C. Glaszmann. Fine Scale Genomic Signals of Admixture and Alien Introgression among Asian Rice Landraces. *Genome Biology and Evolution*, 11(5) :1358–1373, Apr. 2019. ISSN 1759-6653. doi : 10.1093/gbe/evz084. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6499253/.

J. Sardos, X. Perrier, J. Doležel, E. Hřibová, P. Christelová, I. Van den houwe, A. Kilian, and N. Roux. DArT whole genome profiling provides insights on the evolution and taxonomy of edible Banana (*Musa* spp.). *Annals of Botany*, 118(7) :1269–1278, Dec. 2016a. ISSN 0305-7364, 1095-8290. doi : 10.1093/aob/mcw170. URL https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcw170.

J. Sardos, M. Rouard, Y. Hueber, A. Cenci, K. E. Hyma, I. van den Houwe, E. Hribova, B. Courtois, and N. Roux. A Genome-Wide Association Study on the Seedless Phenotype in Banana (Musa spp.) Reveals the Potential of a Selected Panel to Detect Candidate Genes in a Vegetatively Propagated Crop. *PLOS ONE*, 11(5) :e0154448, May 2016b. ISSN 1932-6203. doi : 10.1371/journal.pone.0154448. URL https://dx.plos.org/10.1371/journal.pone.0154448.

P. Scheet and M. Stephens. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data : Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics*, 78(4) :629–644, Apr. 2006. ISSN 00029297. doi : 10.1086/502802. URL https://linkinghub.elsevier.com/retrieve/pii/S000292970763701X.

M. F. Seldin, B. Pasaniuc, and A. L. Price. New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 12(8) :523–528, Aug. 2011. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg3002. URL http://www.nature.com/articles/nrg3002.

M. Semon, R. Nielsen, M. P. Jones, and S. R. McCouch. The Population Structure of African Cultivated Rice *Oryza glaberrima* (Steud.) : Evidence for Elevated Levels of Linkage Disequilibrium Caused by Admixture with *O. sativa* and Ecological Adaptation. *Genetics*, 169 (3) :1639–1647, Mar. 2005. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.104.033175. URL http://www.genetics.org/lookup/doi/10.1534/genetics.104.033175.

K. Shepherd. Translocations and inversions in diploid *Musa acuminata*. In *Cytogenetics of the genus Musa*, pages 15–38. INIBAP, Montpellier, 1999. URL http://www.musalit.org/seeMore.php?id=5123.

N. Simmonds. *The Evolution of the Bananas*. Tropical science series. Longmans, 1962. URL https://books.google.fr/books?id=HpI_AAAAYAAJ.

N. W. Simmonds and K. Shepherd. The taxonomy and origins of the cultivated bananas. *Journal of the Linnean Society of London, Botany*, 55(359) :302–312, Dec. 1955. ISSN 03682927. doi : 10.1111/j.1095-8339.1955.tb00015.x. URL https://academic.oup.com/botlinnean/article-lookup/doi/10.1111/j.1095-8339.1955.tb00015.x.

P. R. Staab and D. Metzler. Coala : an R framework for coalescent simulation. *Bioinformatics*, 32(12) :1903–1904, June 2016. ISSN 1367-4803. doi : 10.1093/bioinformatics/btw098. URL https://academic.oup.com/bioinformatics/article/32/12/1903/1744397.

P. R. Staab, S. Zhu, D. Metzler, and G. Lunter. scrm : efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, 31 (10) :1680–1682, May 2015. ISSN 1460-2059, 1367-4803. doi : 10.1093/bioinformatics/btu861. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu861.

M. Stift, F. Kolář, and P. G. Meirmans. Structure is more robust than other clustering methods in simulated mixed-ploidy populations. *Heredity*, 123(4) :429–441, Oct. 2019. ISSN 0018-067X, 1365-2540. doi : 10.1038/s41437-019-0247-6. URL http://www.nature.com/articles/s41437-019-0247-6.

A. Suarez-Gonzalez, C. A. Hefer, C. Christe, O. Corea, C. Lexer, Q. C. B. Cronk, and C. J. Douglas. Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P . trichocarpa* (black cottonwood). *Molecular Ecology*, 25(11) :2427–2442, June 2016. ISSN 09621083. doi : 10.1111/mec.13539. URL http://doi.wiley.com/10.1111/mec.13539.

A. Sundquist, E. Fratkin, C. B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4) :676–682, Apr. 2008. ISSN 1088-9051. doi : 10.1101/gr.072850.107. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2279255/.

H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture : Analytical and study design considerations. *Genetic Epidemiology*, 28(4) :289–301, May 2005. ISSN 0741-0395, 1098-2272. doi : 10.1002/gepi.20064. URL http://doi.wiley.com/10.1002/gepi.20064.

H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. *The American Journal of Human Genetics*, 79(1) :1–12, July 2006. ISSN 0002-9297. doi : 10.1086/504302. URL http://www.sciencedirect.com/science/article/pii/S0002929707600135.

The 3,000 rice genomes project. The 3,000 rice genomes project. *GigaScience*, 3(1) :7, Dec. 2014. ISSN 2047-217X. doi : 10.1186/2047-217X-3-7. URL https://academic.oup.com/gigascience/article-lookup/doi/10.1186/2047-217X-3-7.

The Bioconda Team, B. Grüning, R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, and J. Köster. Bioconda : sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7) :475–476, July 2018. ISSN 1548-7091, 1548-7105. doi : 10.1038/s41592-018-0046-7. URL http://www.nature.com/articles/s41592-018-0046-7.

The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311) :52–58, Sept. 2010. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature09298. URL http://www.nature.com/articles/nature09298.

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437 (7063) :1299–1320, Oct. 2005. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature04226. URL http://www.nature.com/articles/nature04226.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145) :661–678, June 2007. ISSN 1476-4687. doi : 10.1038/nature05911. URL https://www.nature.com/articles/nature05911.

H. Väre and M. Häkkinen. Typification and check-list of *Ensete* Horan. names (Musaceae) with nomenclatural notes. *Adansonia*, 33(2) :191–200, Dec. 2011. ISSN 1280-8571, 1639-4798. doi : 10.5252/a2011n2a3. URL http://www.bioone.org/doi/abs/10.5252/a2011n2a3.

H. Wang, F. G. Vieira, J. E. Crawford, C. Chu, and R. Nielsen. Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. *Genome Research*, 27(6) :1029–1038, June 2017. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.204800.116. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.204800.116.

Z. Wang, H. Miao, J. Liu, B. Xu, X. Yao, C. Xu, S. Zhao, X. Fang, C. Jia, J. Wang, J. Zhang, J. Li, Y. Xu, J. Wang, W. Ma, Z. Wu, L. Yu, Y. Yang, C. Liu, Y. Guo, S. Sun, F.-C. Baurens, G. Martin, F. Salmon, O. Garsmeur, N. Yahiaoui, C. Hervouet, M. Rouard, N. Laboureau, R. Habas, S. Ricci, M. Peng, A. Guo, J. Xie, Y. Li, Z. Ding, Y. Yan, W. Tie, A. D'Hont, W. Hu, and Z. Jin. Musa balbisiana genome reveals subgenome evolution and functional divergence. *Nature Plants*, 5(8) :810–821, Aug. 2019. ISSN 2055-0278. doi : 10.1038/s41477-019-0452-6. URL http://www.nature.com/articles/s41477-019-0452-6.

B. S. Weir and C. C. Cockerham. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6) :1358, Nov. 1984. ISSN 00143820. doi : 10.2307/2408641. URL https://www.jstor.org/stable/2408641?origin=crossref.

E.-M. Willing, C. Dreyer, and C. van Oosterhout. Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. *PLoS ONE*, 7(8) :e42649, Aug. 2012. ISSN 1932-6203. doi : 10.1371/journal.pone.0042649. URL `https://dx.plos.org/10.1371/journal.pone.0042649`.

G. A. Wu, S. Prochnik, J. Jenkins, J. Salse, U. Hellsten, F. Murat, X. Perrier, M. Ruiz, S. Scalabrin, J. Terol, M. A. Takita, K. Labadie, J. Poulain, A. Couloux, K. Jabbari, F. Cattonaro, C. Del Fabbro, S. Pinosio, A. Zuccolo, J. Chapman, J. Grimwood, F. R. Tadeo, L. H. Estornell, J. V. Muñoz-Sanz, V. Ibanez, A. Herrero-Ortega, P. Aleza, J. Pérez-Pérez, D. Ramón, D. Brunel, F. Luro, C. Chen, W. G. Farmerie, B. Desany, C. Kodira, M. Mohiuddin, T. Harkins, K. Fredrikson, P. Burns, A. Lomsadze, M. Borodovsky, G. Reforgiato, J. Freitas-Astúa, F. Quetier, L. Navarro, M. Roose, P. Wincker, J. Schmutz, M. Morgante, M. A. Machado, M. Talon, O. Jaillon, P. Ollitrault, F. Gmitter, and D. Rokhsar. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology*, 32(7) :656–662, July 2014. ISSN 1546-1696. doi : 10.1038/nbt.2906. URL `https://www.nature.com/articles/nbt.2906`.

G. A. Wu, J. Terol, V. Ibanez, A. López-García, E. Pérez-Román, C. Borredá, C. Domingo, F. R. Tadeo, J. Carbonell-Caballero, R. Alonso, F. Curk, D. Du, P. Ollitrault, M. L. Roose, J. Dopazo, F. G. Gmitter, D. S. Rokhsar, and M. Talon. Genomics of the origin and evolution of *Citrus*. *Nature*, 554(7692) :311–316, Feb. 2018. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature25447. URL `http://www.nature.com/articles/nature25447`.

X. Yuan, D. J. Miller, J. Zhang, D. Herrington, and Y. Wang. An Overview of Population Genetic Data Simulation. *Journal of Computational Biology*, 19(1) :42–54, Jan. 2012. ISSN 1066-5277, 1557-8666. doi : 10.1089/cmb.2010.0188. URL `http://www.liebertpub.com/doi/10.1089/cmb.2010.0188`.

K. Zhao, M. Wright, J. Kimball, G. Eizenga, A. McClung, M. Kovach, W. Tyagi, M. L. Ali, C.-W. Tung, A. Reynolds, C. D. Bustamante, and S. R. McCouch. Genomic Diversity and Introgression in *O. sativa* Reveal the Impact of Domestication and Breeding on the Rice Genome. *PLOS ONE*, 5(5) :e10780, May 2010. ISSN 1932-6203. doi : 10.1371/journal.pone.0010780. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010780`.

Q. Zhou, L. Zhao, and Y. Guan. Strong Selection at MHC in Mexicans since Admixture. *PLOS Genetics*, 12(2) :e1005847, Feb. 2016. ISSN 1553-7404. doi : 10.1371/journal.pgen.1005847. URL `https://dx.plos.org/10.1371/journal.pgen.1005847`.

Figures et tables supplémentaires

File S1 : Number of SNP for all simulations.

File S2 : Accuracy of LAI methods with varying number of generations, differentiation and source-representative sample size (**DiffGenSam** simulation)

File S3 : Accuracy of LAI methods with varying sample size of the third source-representatives.

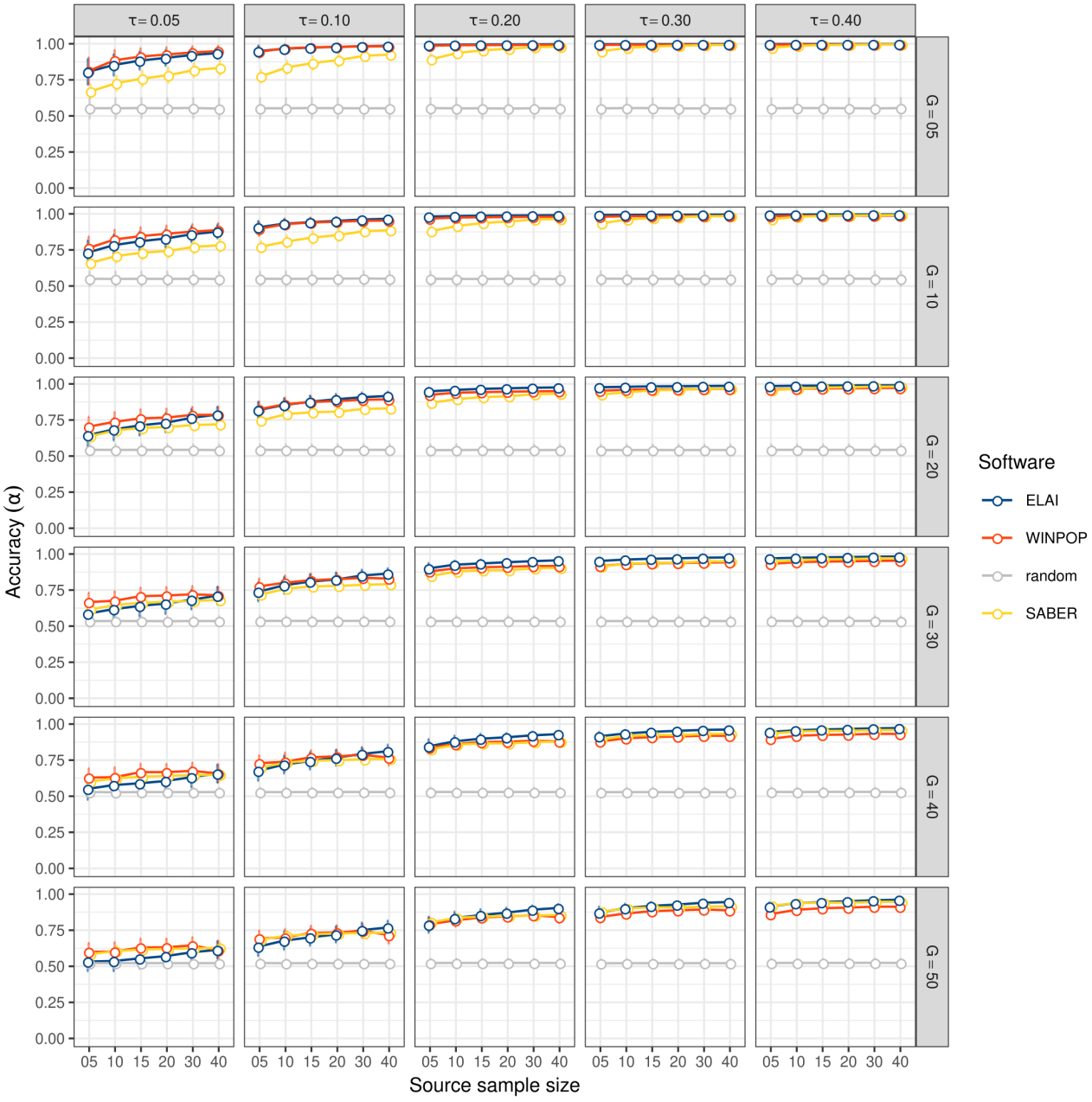

FIGURE S.1 – Accuracy of LAI methods with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation). The accuracy of the LAI methods (y-axis) is plotted for different levels of differentiation that vary from 0.05 to 0.4 (vertical tiles), number of generations after admixture that varies from 5 to 50 (horizontal tiles) and size of source-representative sample varies from 5 to 40 individuals (x-axis). Each dot is the mean value of 50 repetitions of each simulation. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` scores are plotted in blue, red and yellow, respectively. Accuracy of random inference (proportion of ancestry fixed at 1/3) is plotted in gray.

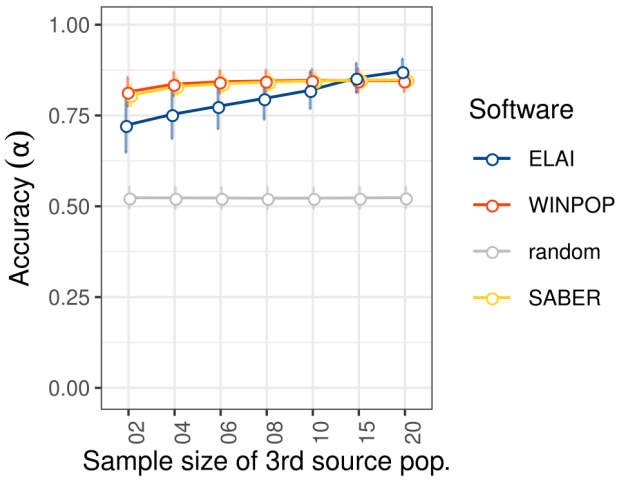

FIGURE S.2 – Accuracy of LAI methods with varying sample size of the third source-representatives (**SamBal** simulation). The accuracy of the LAI methods (y-axis) is plotted for different sample size of the third source-representative population that vary from 2 to 20 (x-axis). The sample size for the two others source-representative populations is set to 20, the differentiation set to 0.2 and the total number of generations after the admixture event set to 50. Each dot is the mean value of 50 repetitions of each simulation. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` scores are plotted in blue, red and yellow, respectively. Accuracy of random inference (proportion of ancestry fixed at 1/3) is plotted in gray.

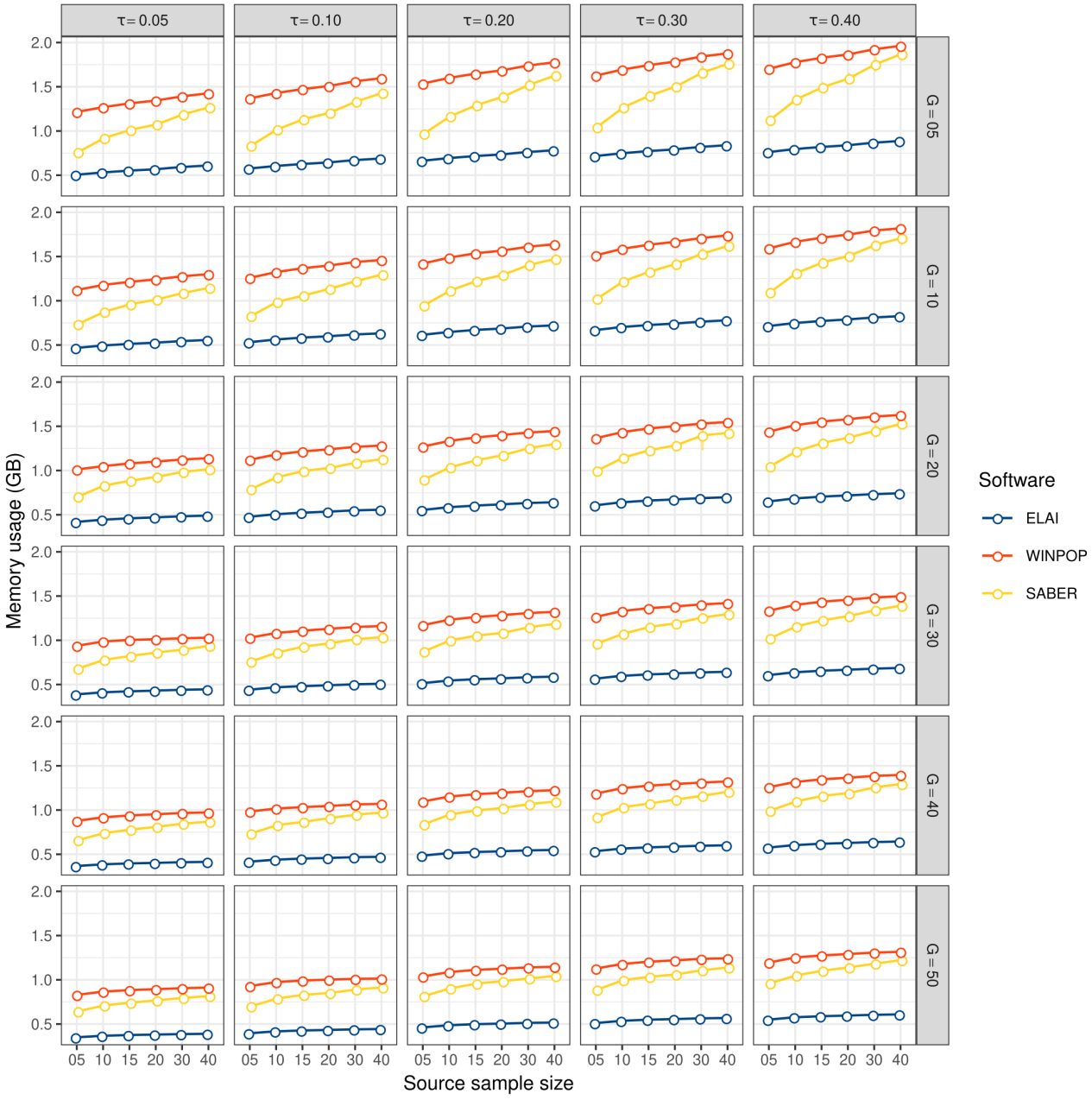

FIGURE S.3 – Memory usage of LAI methods with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation). The memory usage of LAI methods in giga bytes (y-axis), is plotted for different levels of differentiation that vary from 0.05 to 0.4 (vertical tiles), number of generations after admixture that varies from 5 to 50 (horizontal tiles) and size of source-representative sample that varies from 5 to 40 individuals (x-axis). Each dot is the mean value of 50 repetitions of each simulation. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` performance in memory usage are plotted in blue, red and yellow, respectively.

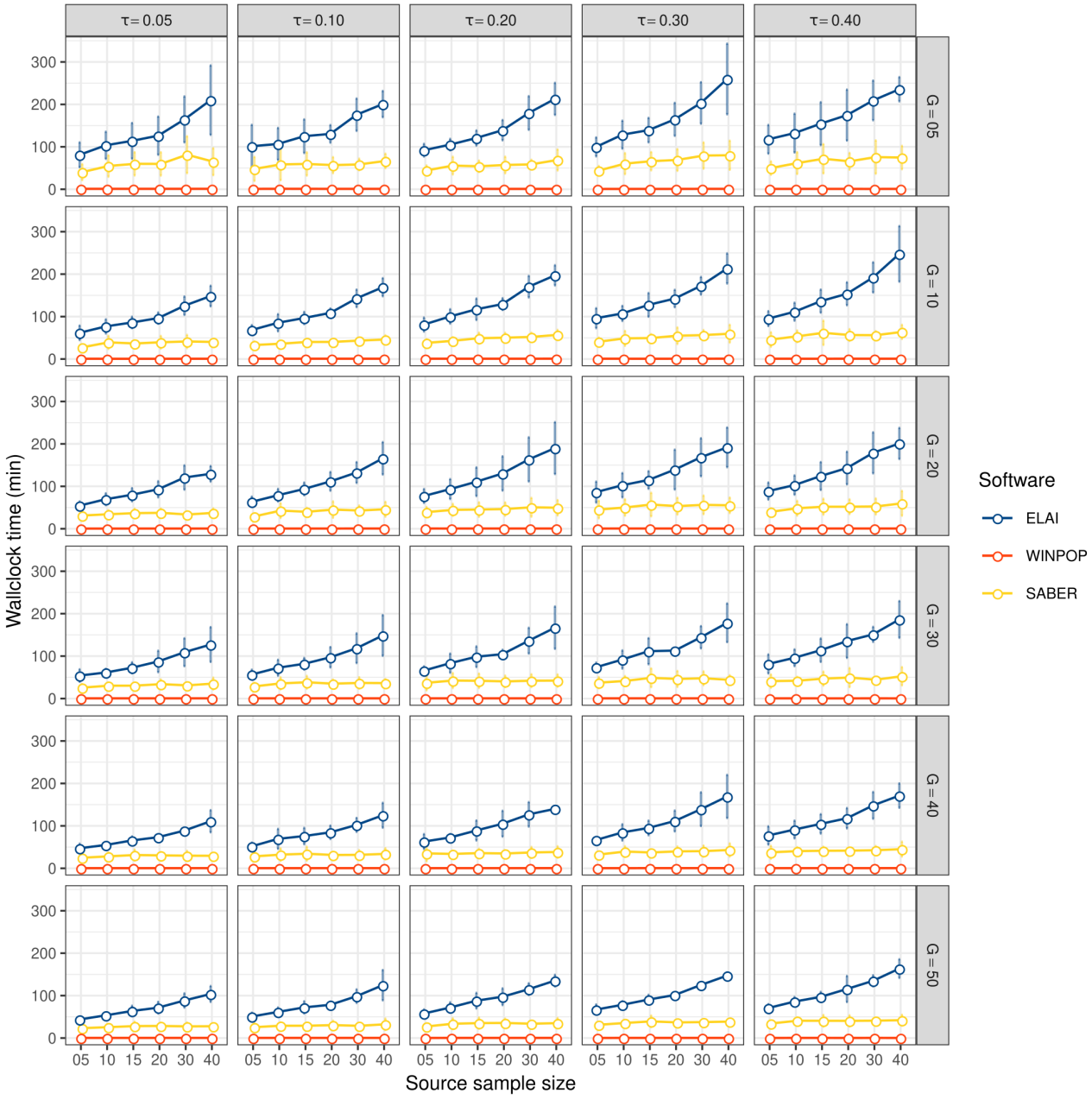

FIGURE S.4 – Computation time of LAI methods with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation). The wallclock time of LAI methods in minutes (y-axis), is plotted for different levels of differentiation that vary from 0.05 to 0.4 (vertical tiles), number of generations after admixture that varies from 5 to 50 (horizontal tiles) and size of source-representative sample that varies from 5 to 40 individuals (x-axis). Each dot is the mean value of 50 repetitions of each simulation. Error bars indicate the standard deviation. `ELAI`, `WINPOP` and `SABER` performance in computation time are plotted in blue, red and yellow, respectively.

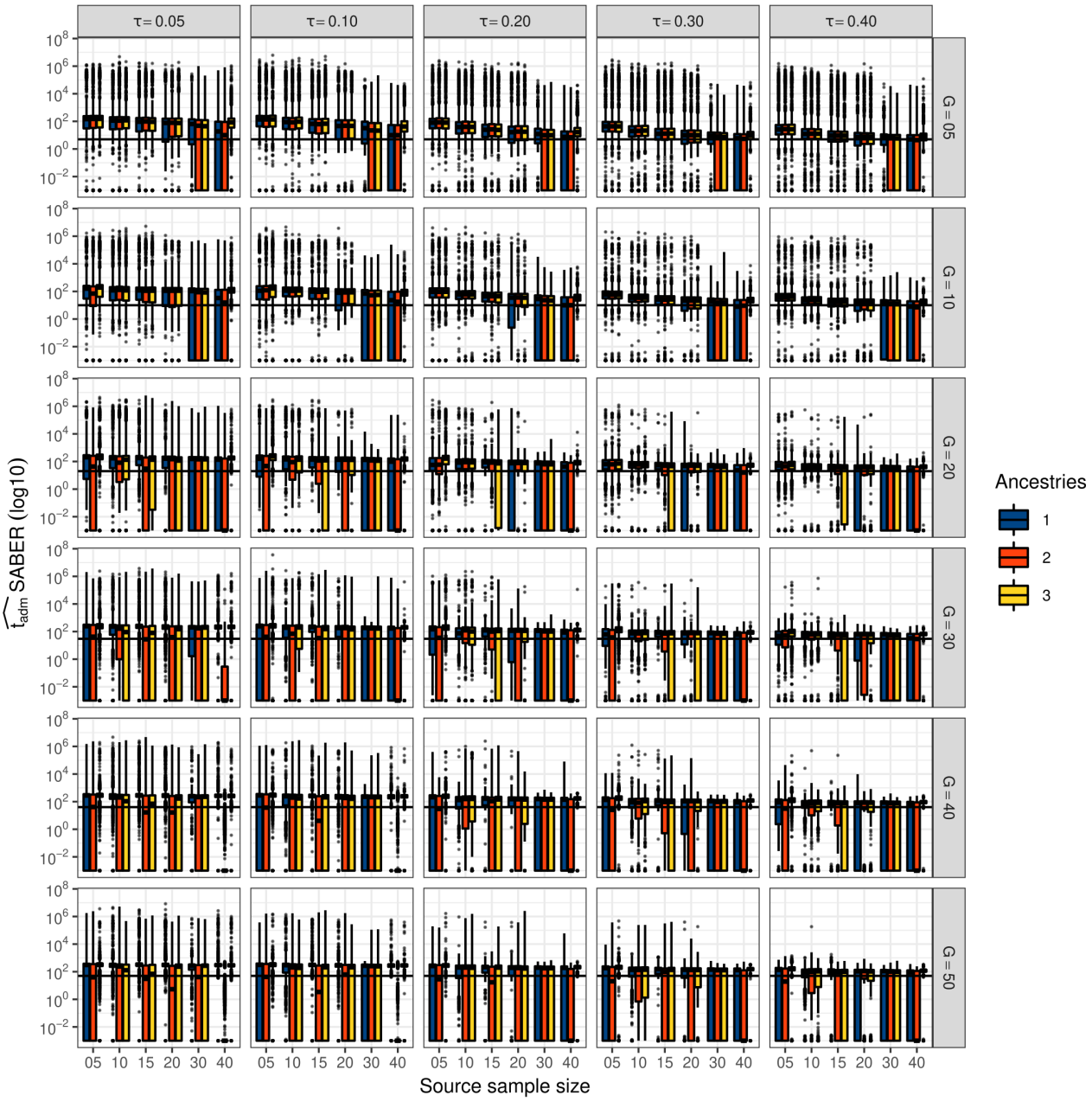

FIGURE S.5 – Time since admixture estimation ($\widehat{t_{adm}}$) of `SABER` with varying levels of differentiation, number of generations and source-representative sample size (**DiffGenSam** simulation). The distribution of estimated time since admixture per ancestry and across all repetitions (y-axis, log10 scale) is plotted for different levels of differentiation that vary from 0.05 to 0.4 (vertical tiles), number of generations after admixture that vary from 5 to 50 (horizontal tiles) and size of source-representative sample that vary from 5 to 40 individuals (x-axis). Outlier values are represented as black transparent dots. The three ancestries are represented using blue, red and yellow, respectively. Expected time since admixture is represented by a horizontal line on each subplot.

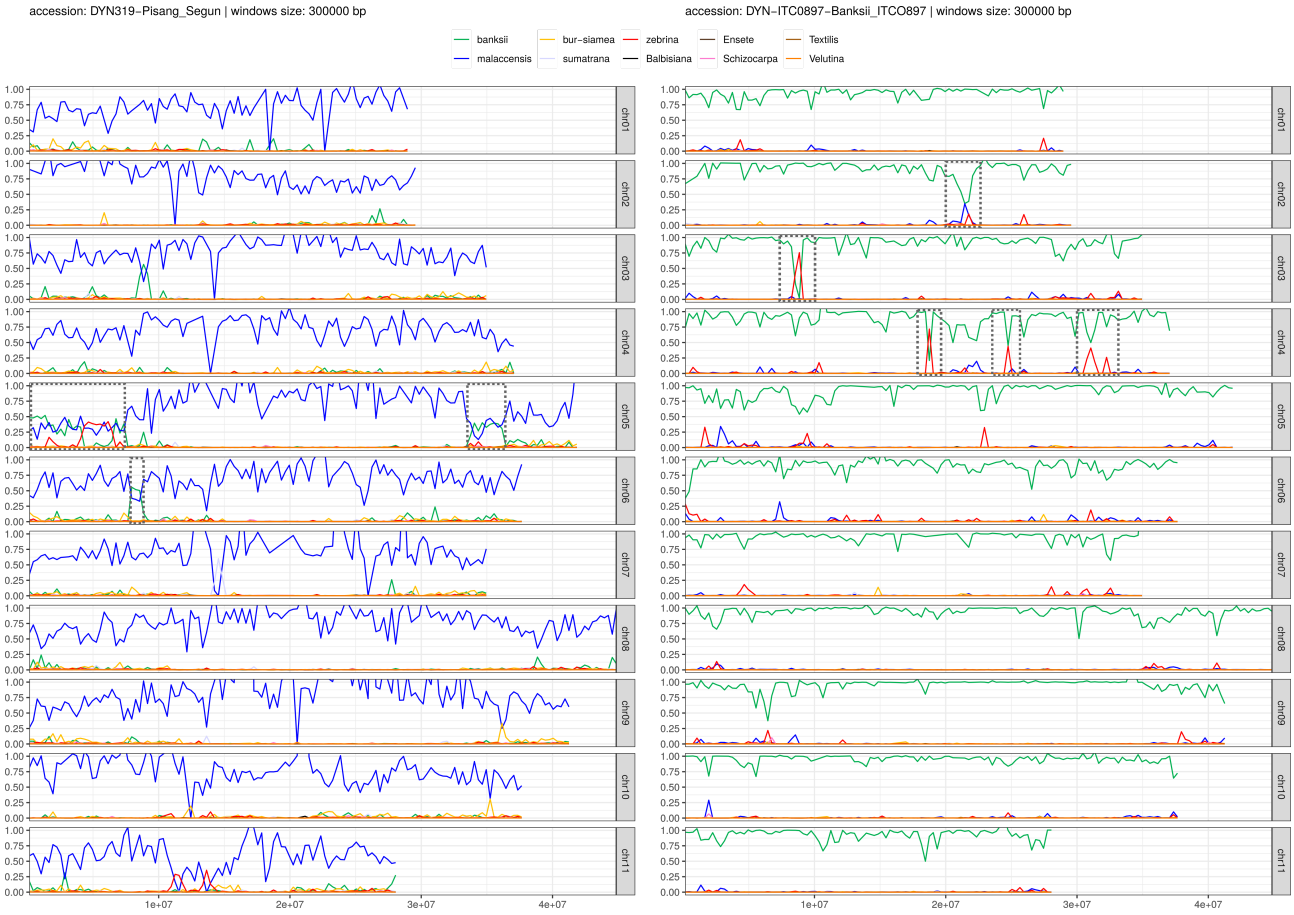

FIGURE S.6 – Résultat de l'ARP pour les accessions 'DYN319-Pisang_Segun' et 'DYN-ITC0897-Banksii_ITCO897' à l'étape trois de l'analyse itérative. L'accession 'DYN319-Pisang_Segun' (type *M. a.* ssp. *malaccensis*) montrant de grandes régions introgréssées (encadré) est représentée à gauche et 'DYN-ITC0897-Banksii_ITCO897' (type *M. a.* ssp. *banksii*) montrant quelques introgressions de petite taille à droite. L'axe des abscisses représente les coordonnées génomiques des marqueurs retenues par l'ARP. L'axe des ordonnées représente la valeur du rapport du ratio de couverture allélique de l'individu sur le ratio attendu par marqueur. Chacune des courbes représente un groupe désigné lors de l'analyse. Chaque sous-graphe correspond à un chromosome noté sur la partie droite de la figure.

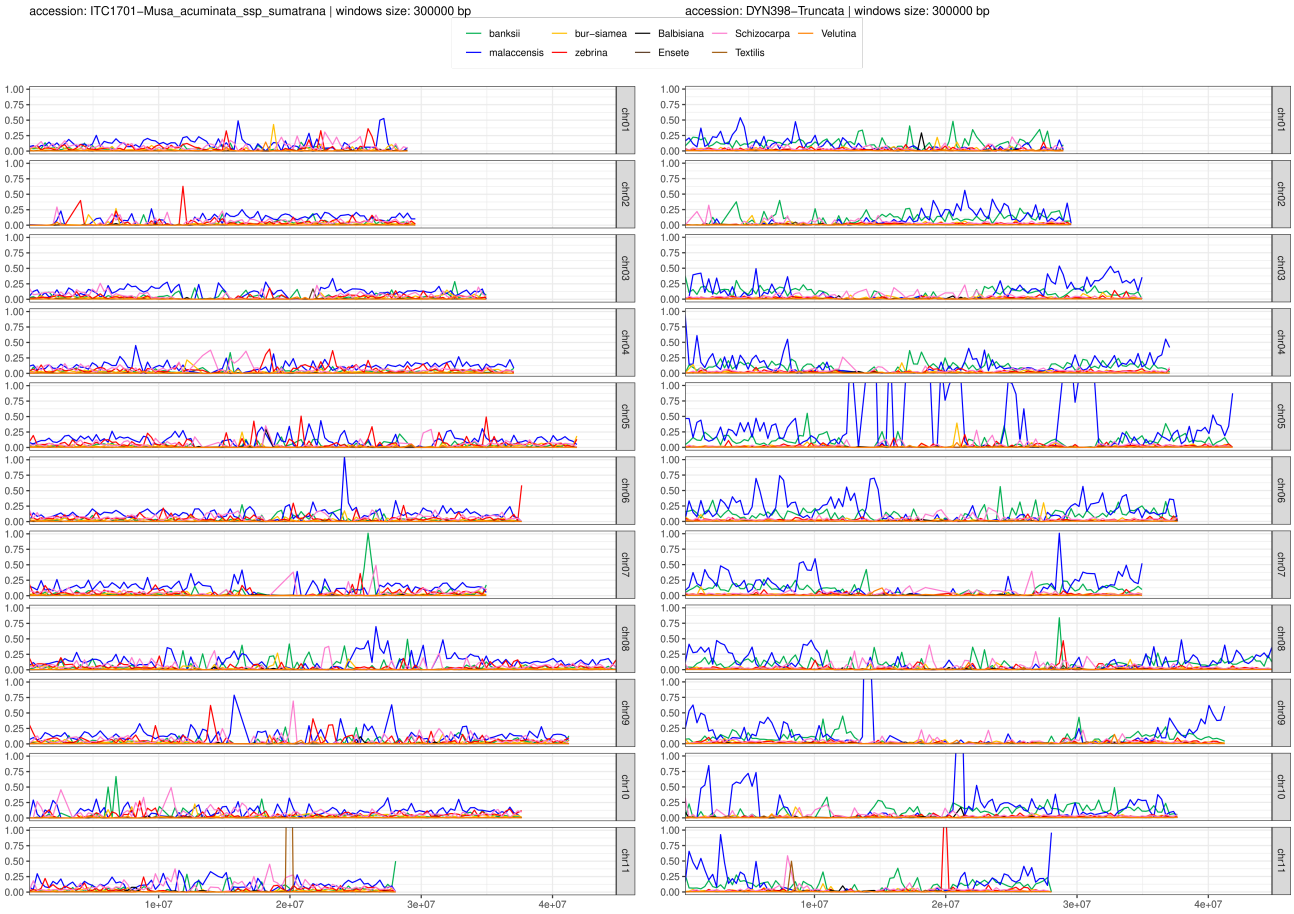

FIGURE S.7 – Résultat de l'ARP pour les accessions 'ITC1701-Musa_acuminata_ssp_sumatrana' et 'DYN398-Truncata' à l'étape deux de l'analyse itérative. L'accession 'ITC1701-Musa_acuminata_ssp_sumatrana' (*M. a.* ssp. *sumatrana*) est représentée à gauche et 'DYN398-Truncata' (*M. a.* ssp. *truncata*) à droite. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés sur les 11 chromosomes de bananiers. L'axe des ordonnées représente la valeur du rapport du ratio de couverture allélique de l'individu sur le ratio attendu par marqueur. Chacune des courbes représente un groupe désigné lors de l'analyse. Chaque sous-graphe correspond à un chromosome noté sur la partie droite de la figure.



FIGURE S.8 – Résultat de l'ARP pour les accessions 'DYN-ITC1028-Agutay' et 'DYN040-Borneo' à l'étape deux de l'analyse itérative. L'accession 'DYN-ITC1028-Agutay' (*M. a.* ssp. *errans*) est représentée à gauche et 'DYN040-Borneo' (*M. a.* ssp. *microcarpa*) à droite. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés sur les 11 chromosomes de bananiers. L'axe des ordonnées représente la valeur du rapport du ratio de couverture allélique de l'individu sur le ratio attendu par marqueur. Chacune des courbes représente un groupe désigné lors de l'analyse. Chaque sous-graphe correspond à un chromosome noté sur la partie droite de la figure.

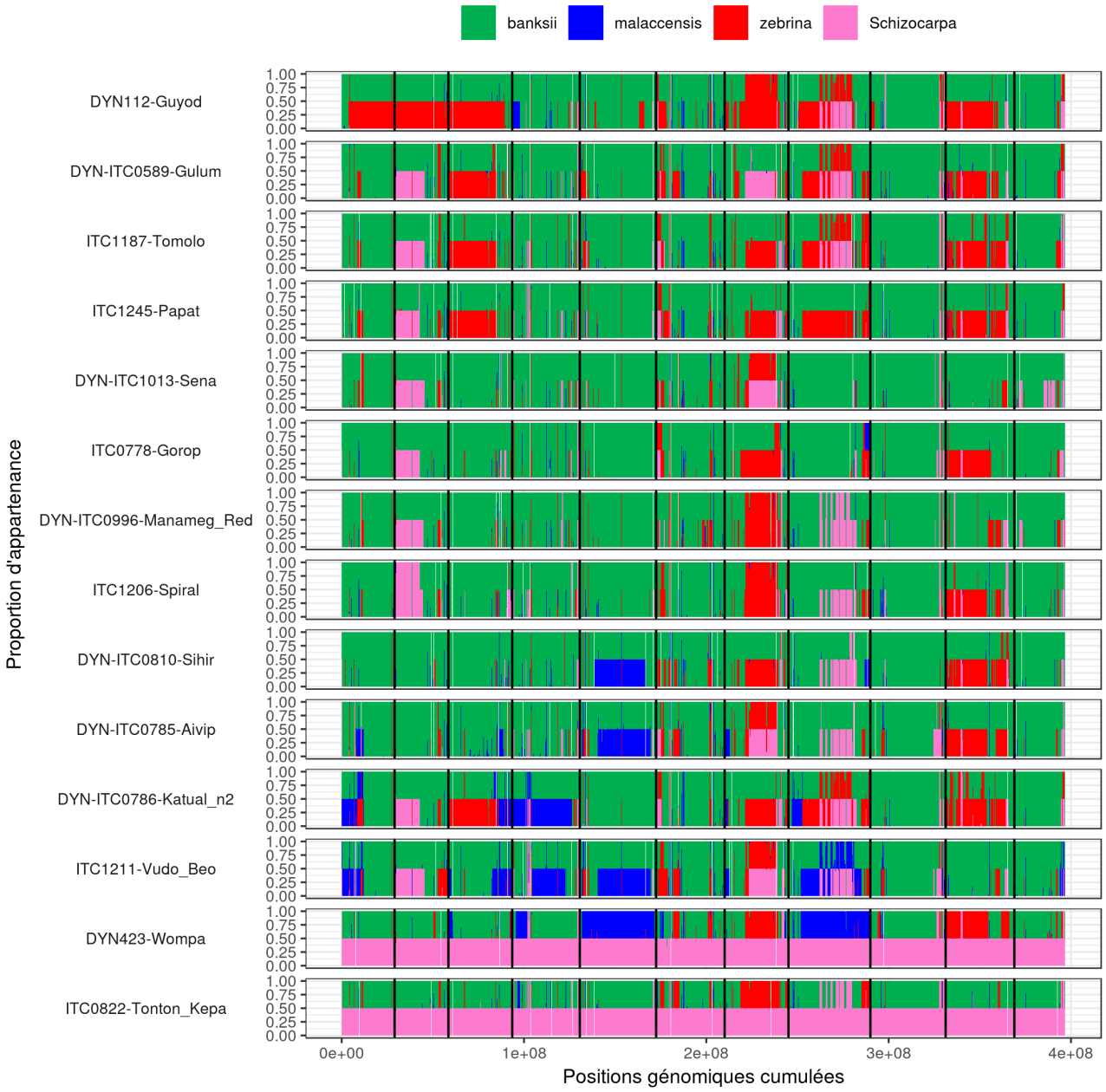

FIGURE S.9 – Résultats de l'inférence locale de `SABER` sur les 14 individus hybrides. Les trois répétitions (sous-échantillons) sont représentées combinées ici, par la moyenne des origines ancestrales inférées par `SABER`, avec un paramètre de temps depuis l'hybridation de 50 générations. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des pôles. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Chaque sous-figure représente un individu et ses 11 chromosomes, séparés par des barres verticales noires.

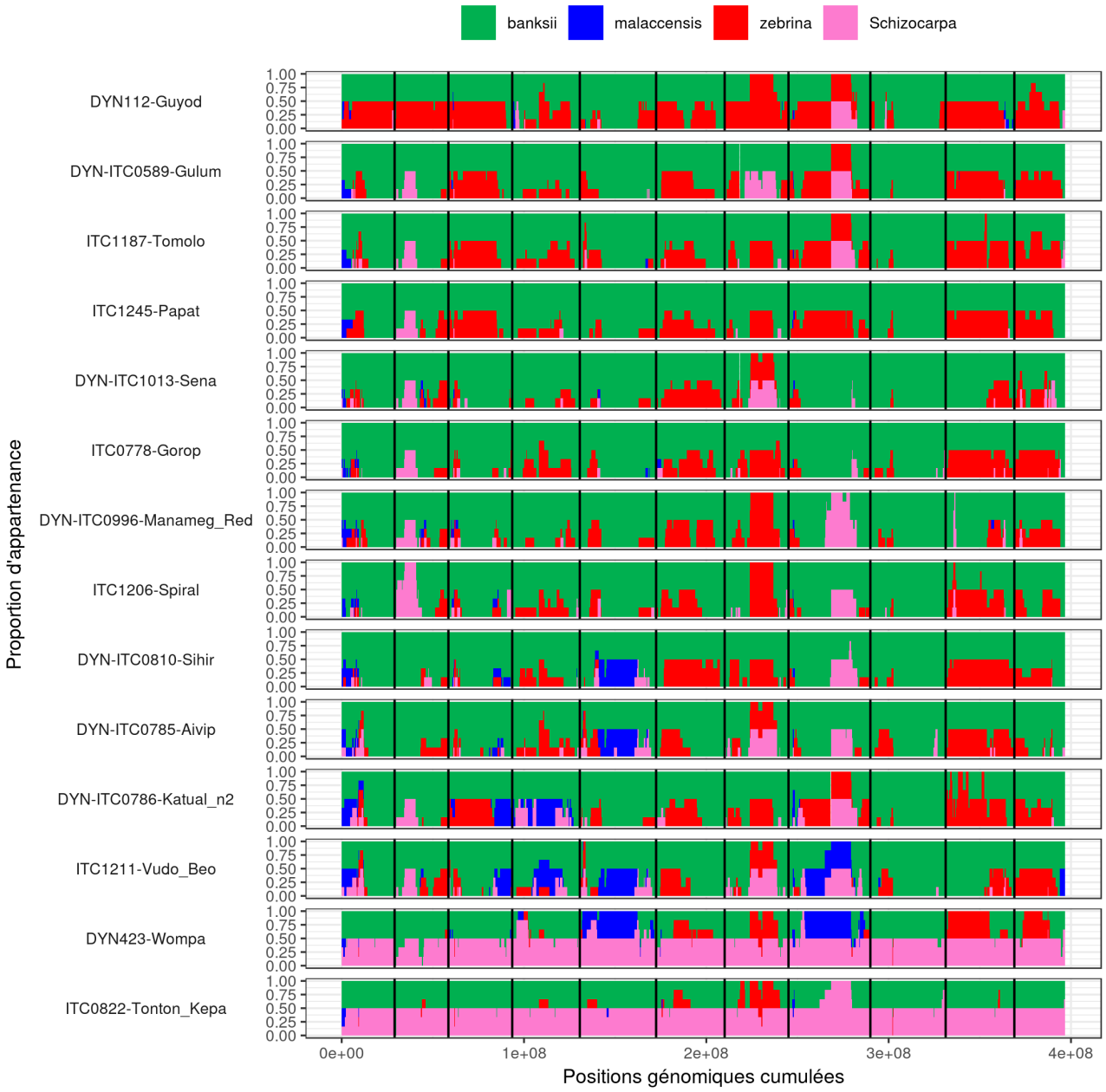

FIGURE S.10 – Résultats de l'inférence locale de `WINPOP` sur les 14 individus hybrides. Les trois répétitions (sous-échantillons) sont représentées combinées ici, par la moyenne des origines ancestrales inférées par `WINPOP`, avec un paramètre de temps depuis l'hybridation de 50 générations. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des pôles. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Chaque sous-figure représente un individu et ses 11 chromosomes, séparés par des barres verticales noires.

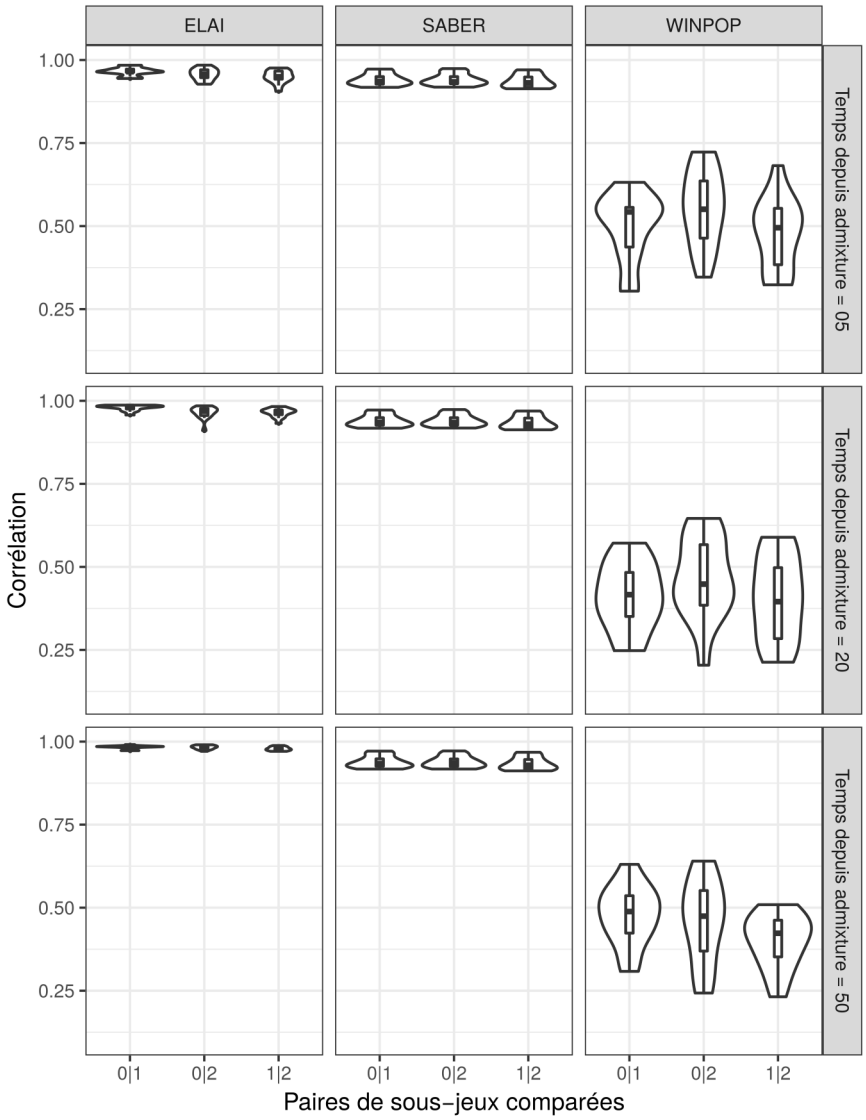

FIGURE S.11 – Stabilité des ILEA entre les différents sous-jeux de données dans des comparaisons par paires. L'axe des abscisses indique la paire de sous-jeu comparée sur l'ensemble des individus, pour chaque méthode (en colonne de la grille) et chaque valeur du paramètre temps depuis l'hybridation (en ligne sur la grille). L'axe des ordonnées représente la valeur moyenne de la corrélation, comprise entre 0 et 1. Chaque point de données correspond à la corrélation moyenne entre les origines inférées des marqueurs des deux individus de la paire de condition comparée.

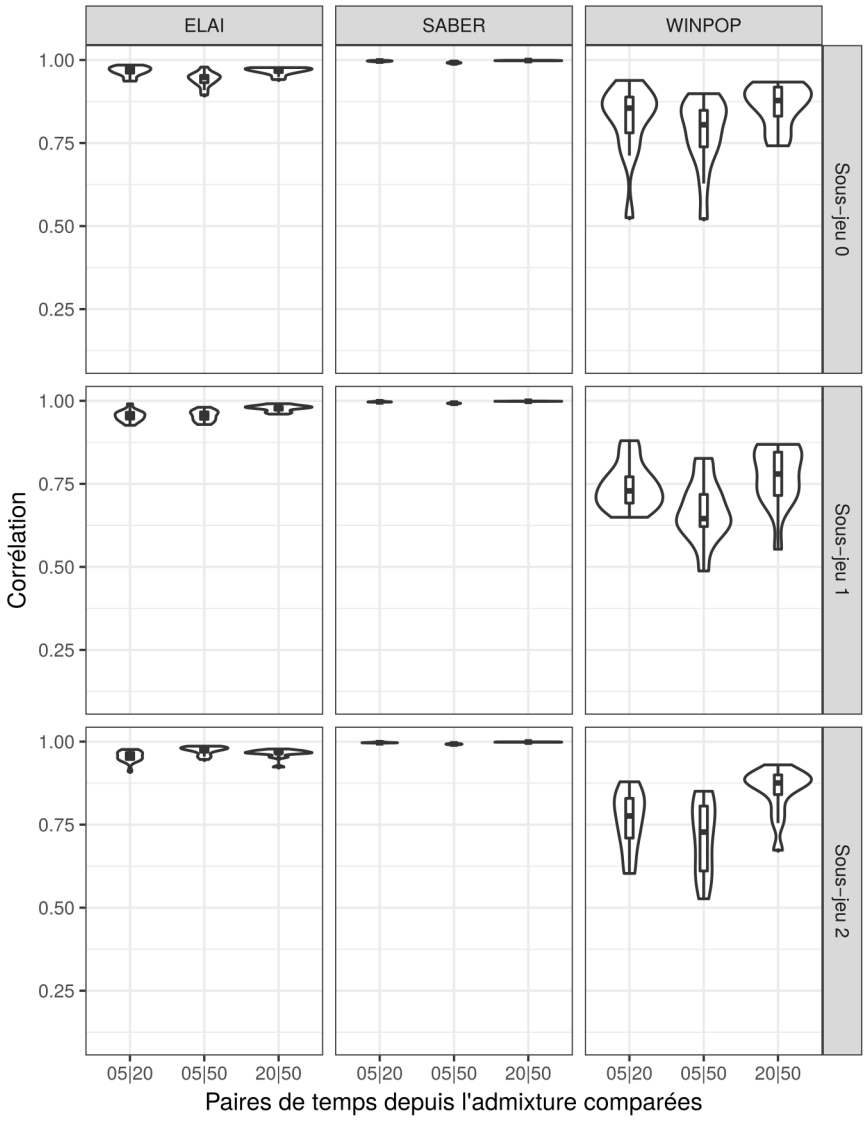

FIGURE S.12 – Stabilité des ILEA par rapport a paramètre du temps depuis l'hybridation dans des comparaisons par paires. L'axe des abscisses indique la paire de sous-jeu comparée sur l'ensemble des individus, pour chaque méthode (en colonne de la grille) et chaque sous-jeu de données (en ligne sur la grille). L'axe des ordonnées représente la valeur moyenne de la corrélation, comprise entre 0 et 1. Chaque point de données correspond à la corrélation moyenne entre les origines inférées des marqueurs des deux individus de la paire de condition comparée.

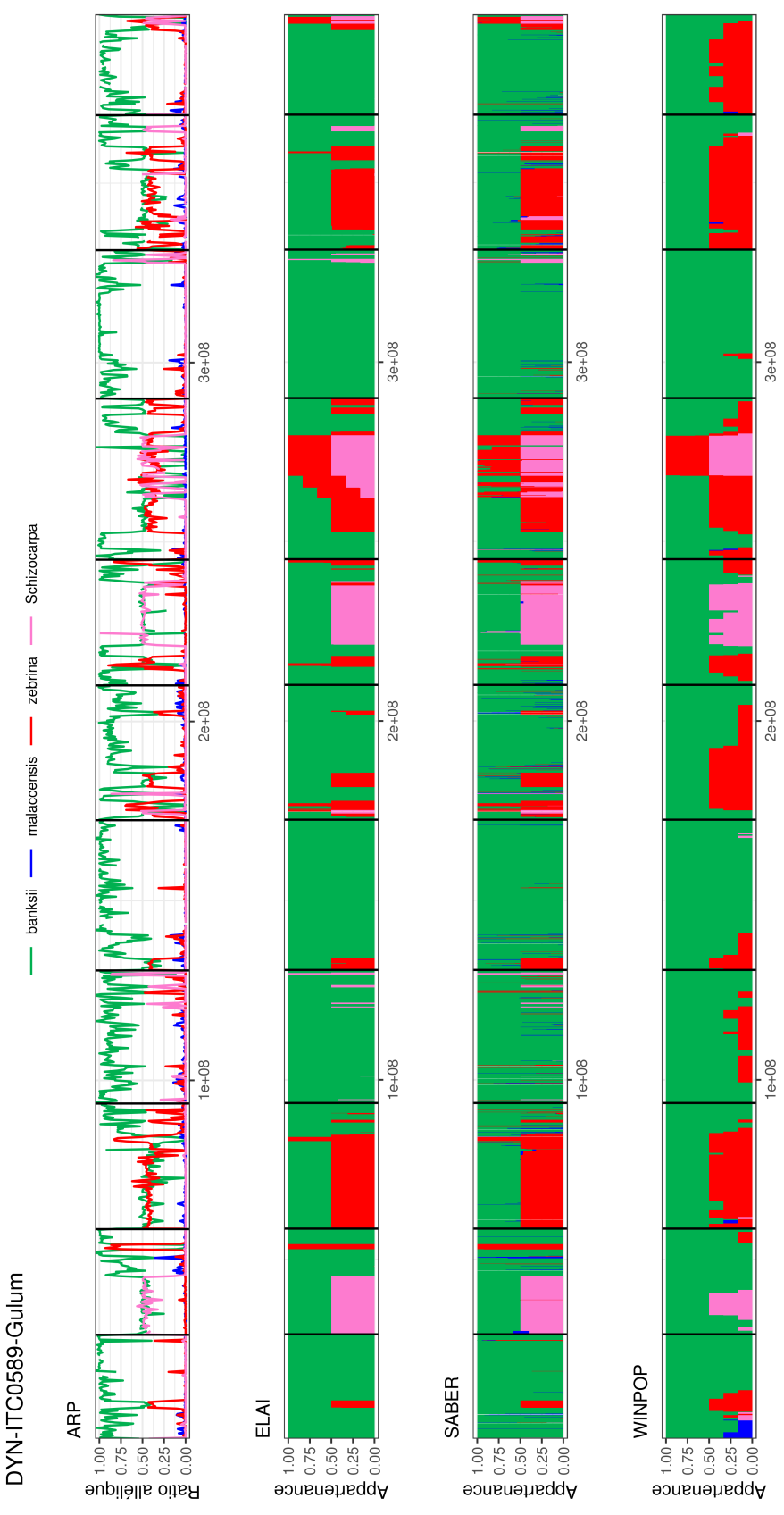

FIGURE S.13 – Résultats de l'ILEA et de l'ARP de l'accession 'Gulum'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

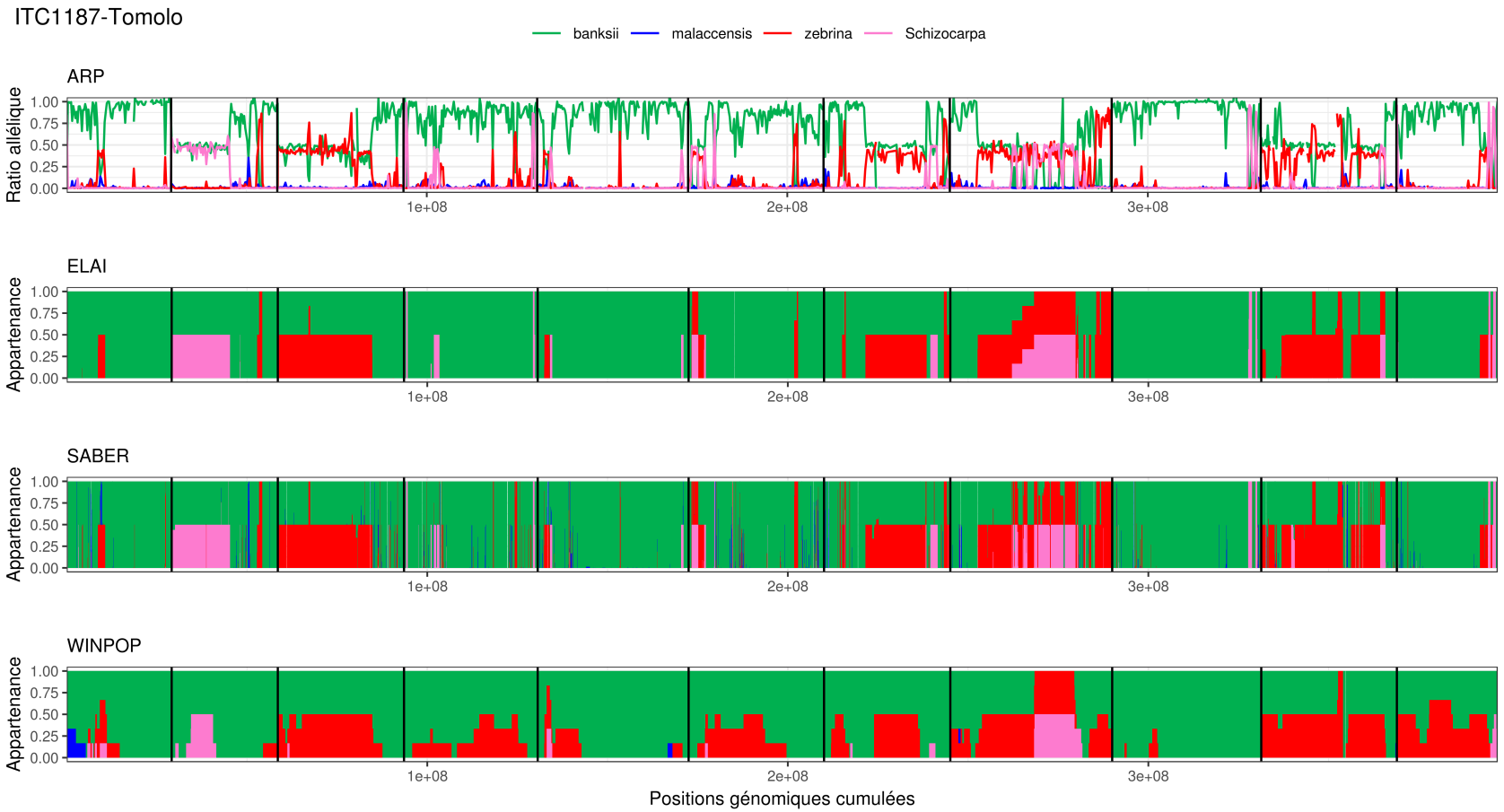

FIGURE S.14 – Résultats de l'ILEA et de l'ARP de l'accession 'Tomolo'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

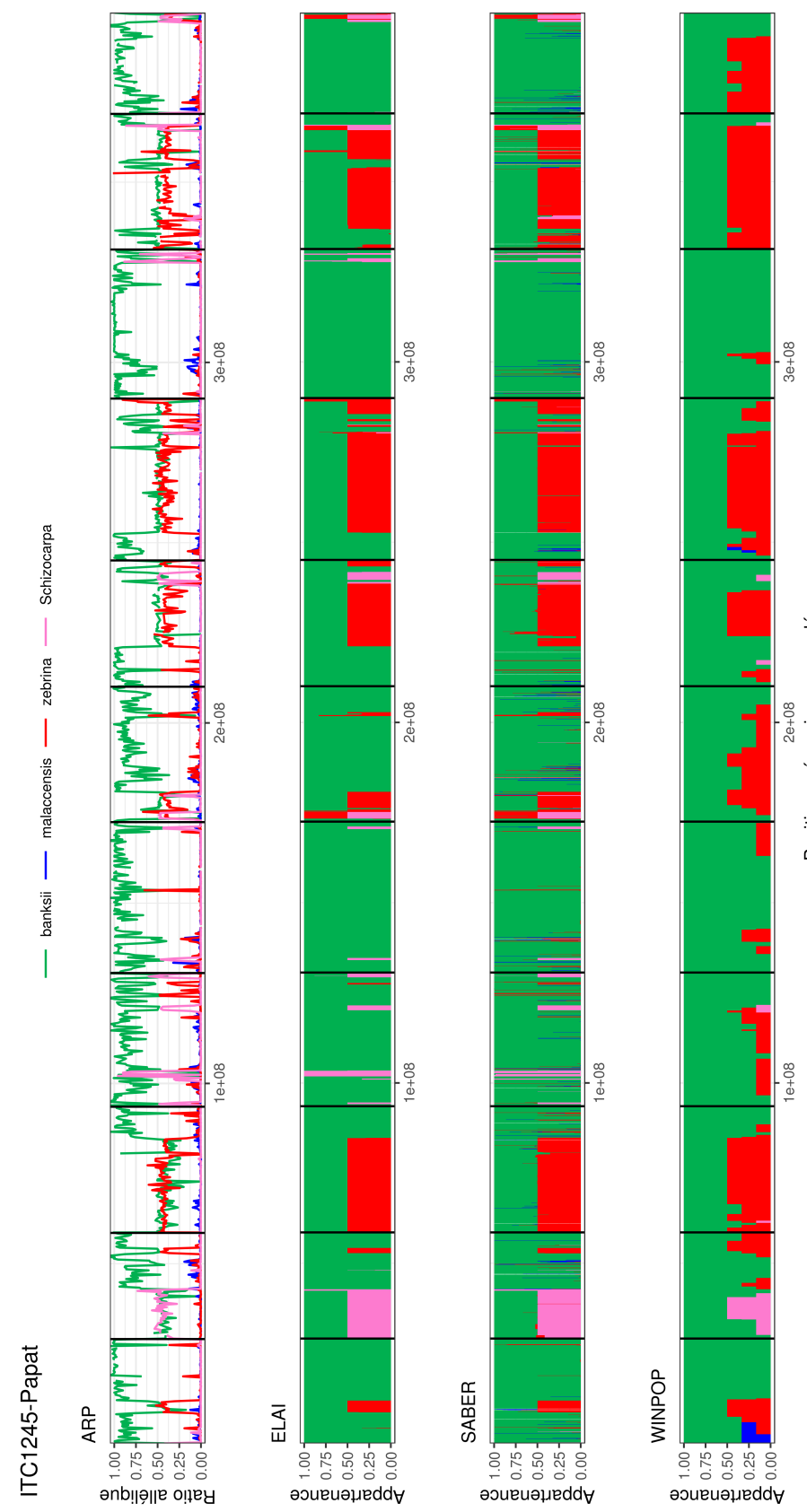

FIGURE S.15 – Résultats de l'ILEA et de l'ARP de l'accession 'Papat'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.



FIGURE S.16 – Résultats de l'ILEA et de l'ARP de l'accession 'Sena'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

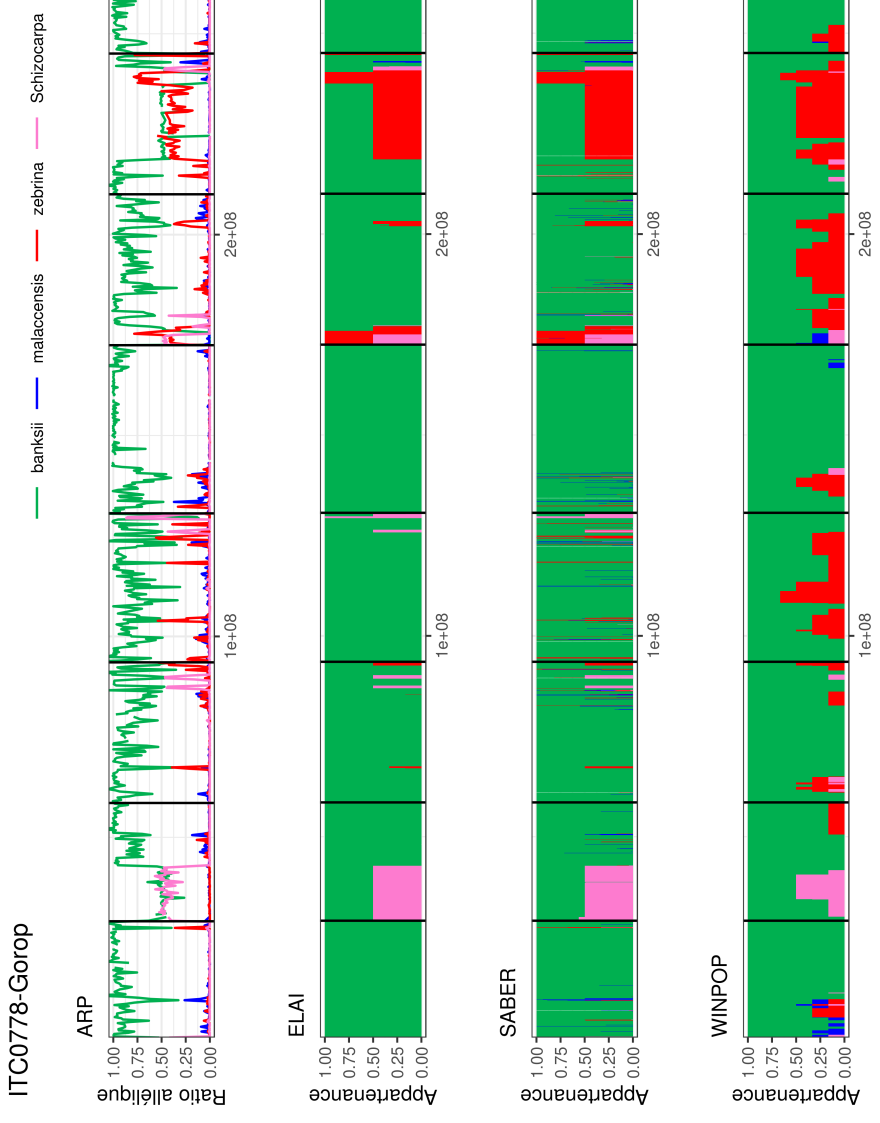

FIGURE S.17 – Résultats de l'ILEA et de l'ARP de l'accession 'Gorop'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

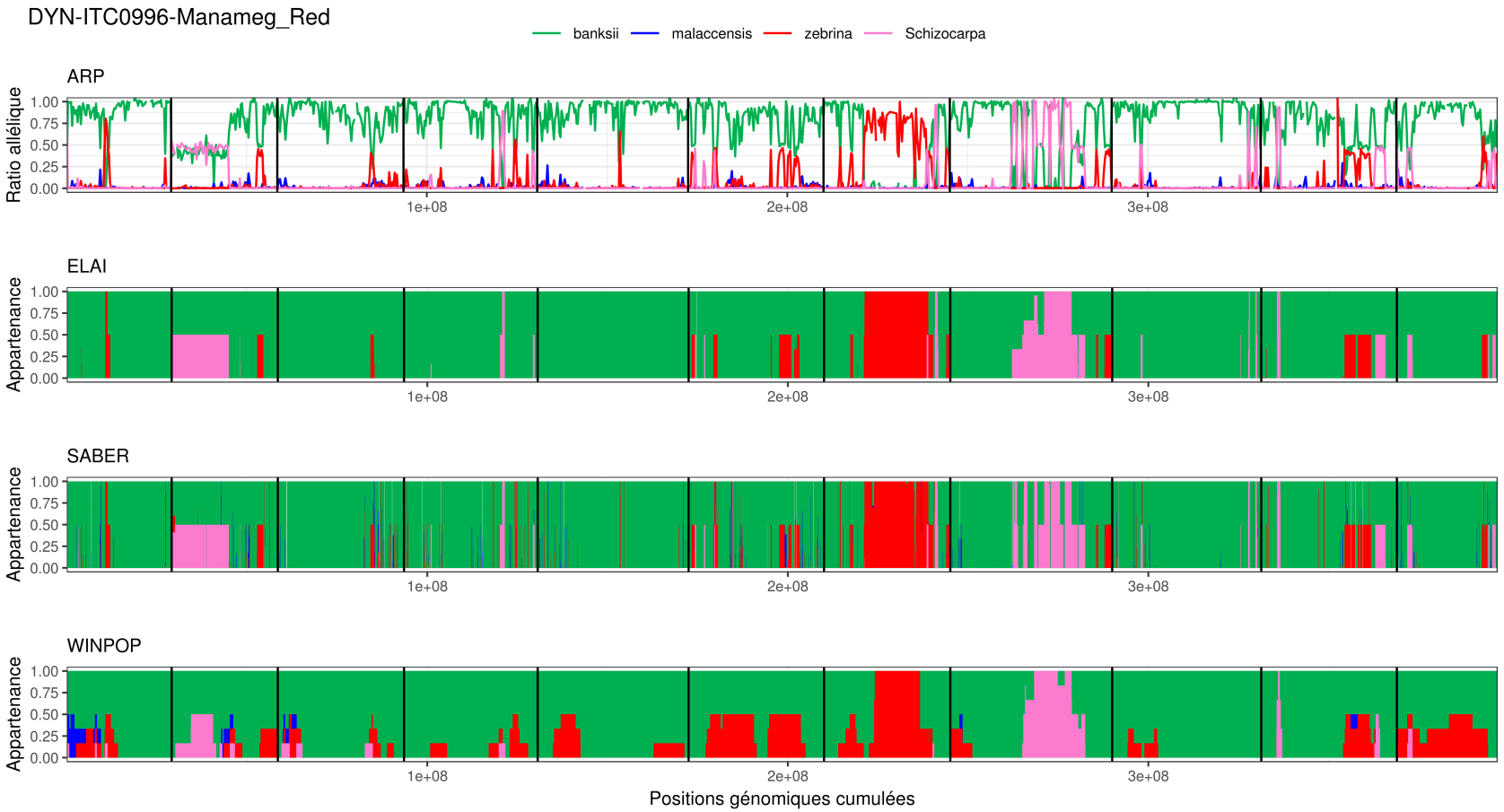

FIGURE S.18 – Résultats de l'ILEA et de l'ARP de l'accession 'Manameg Red'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

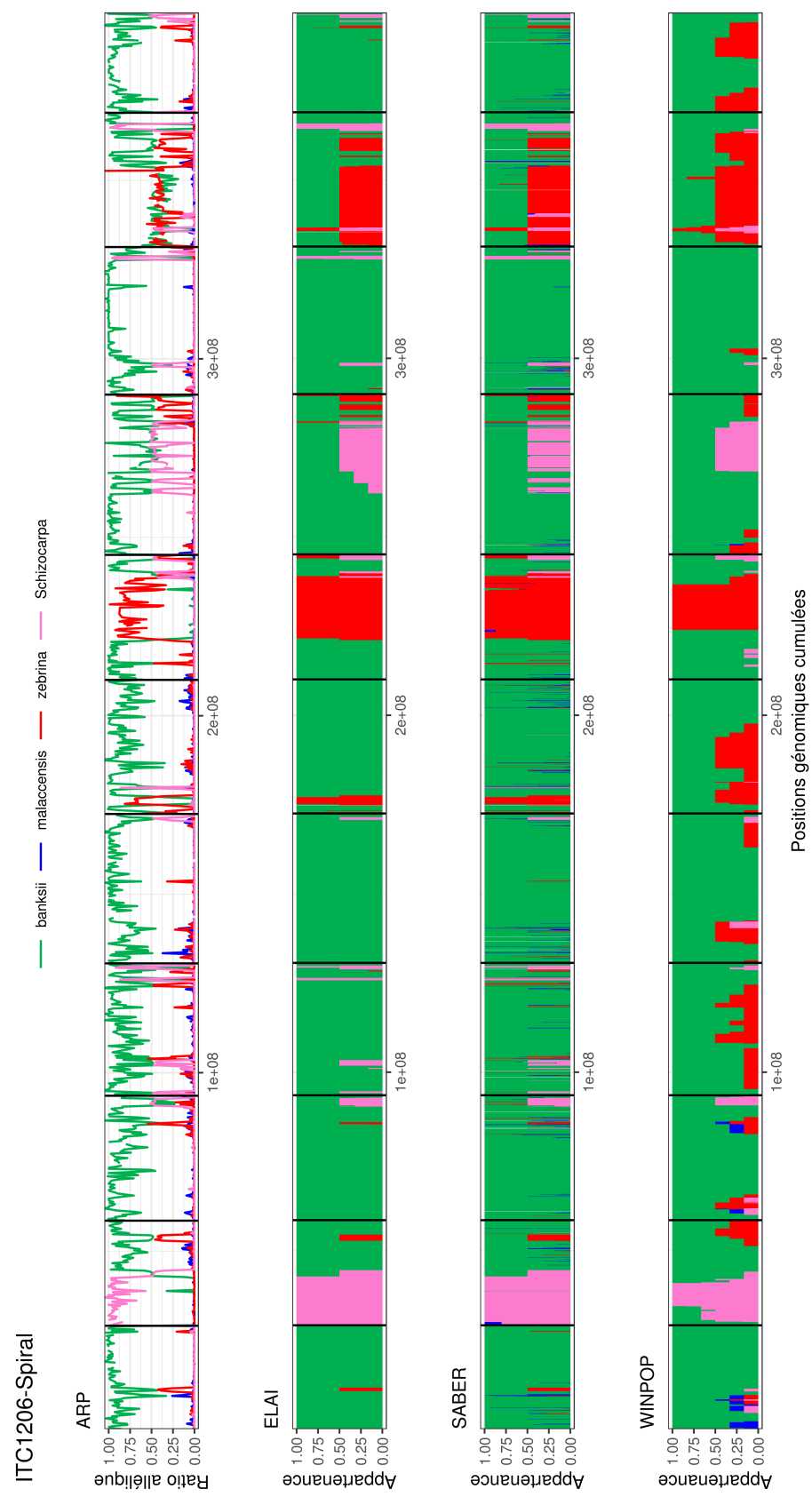

FIGURE S.19 – Résultats de l'ILEA et de l'ARP de l'accession 'Spiral'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.



FIGURE S.20 – Résultats de l'ILEA et de l'ARP de l'accession 'Sihir'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

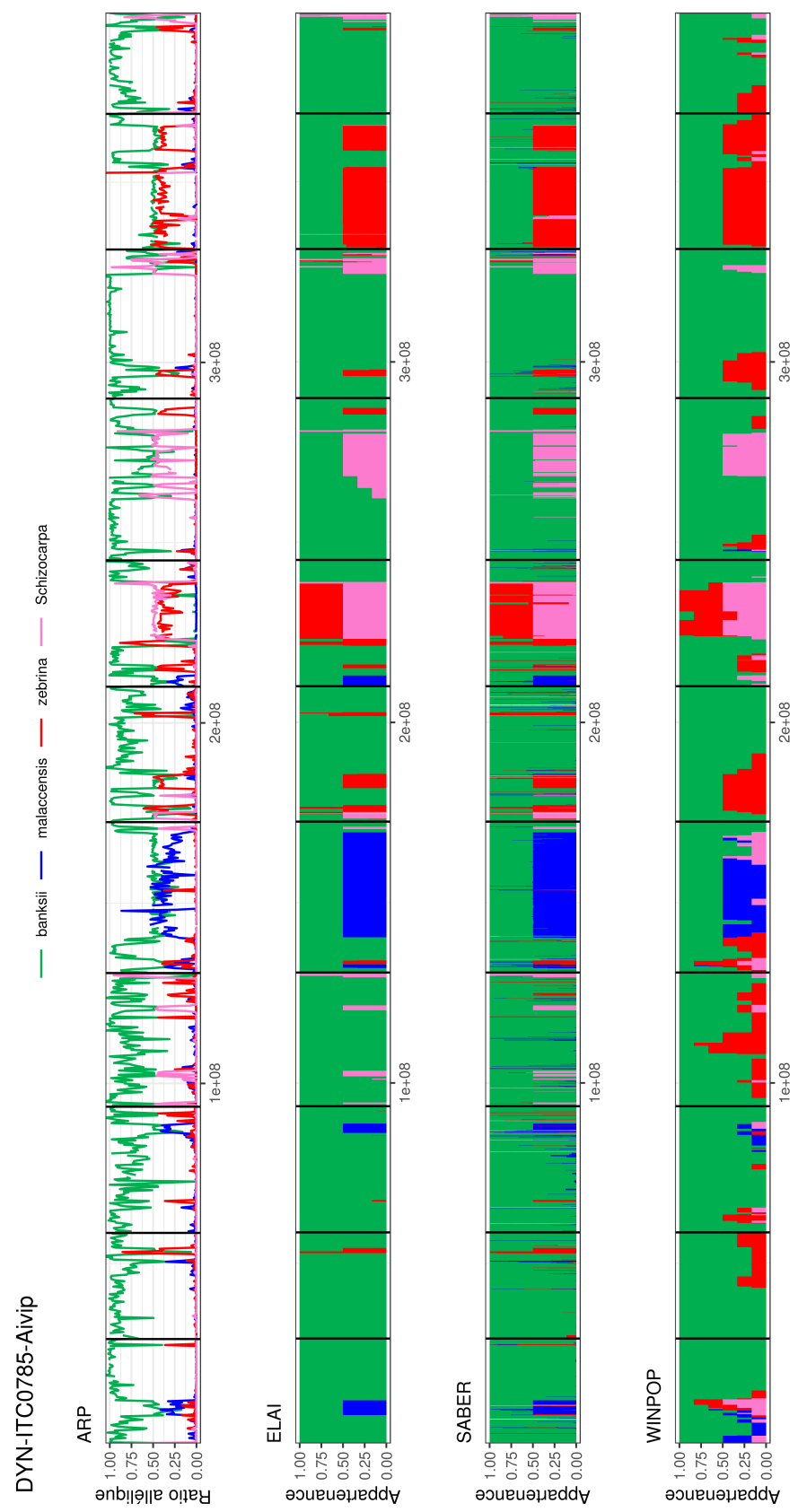

FIGURE S.21 – Résultats de l'ILEA et de l'ARP de l'accession 'Aivip'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

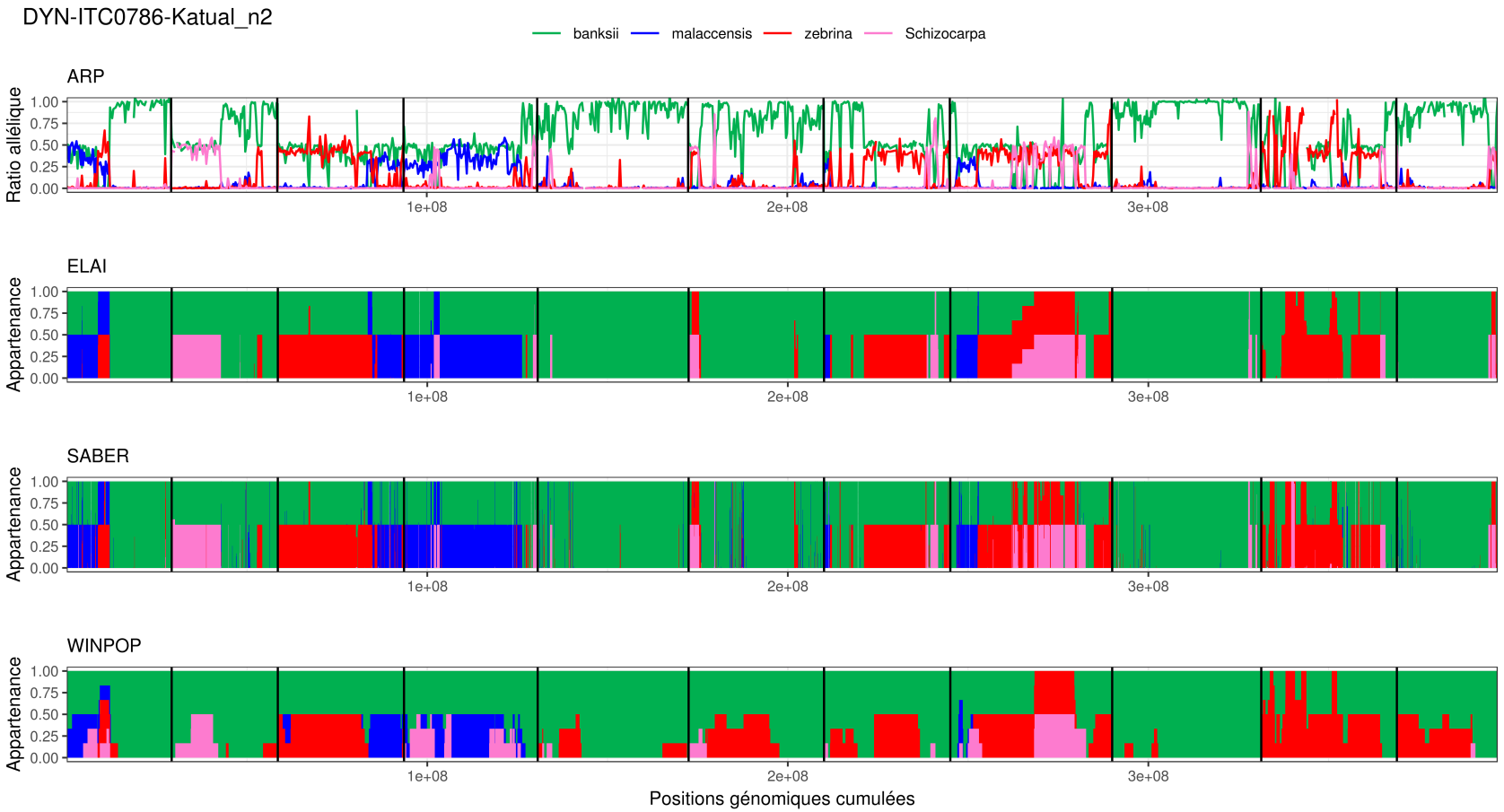

FIGURE S.22 – Résultats de l'ILEA et de l'ARP de l'accession 'Katual n2'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

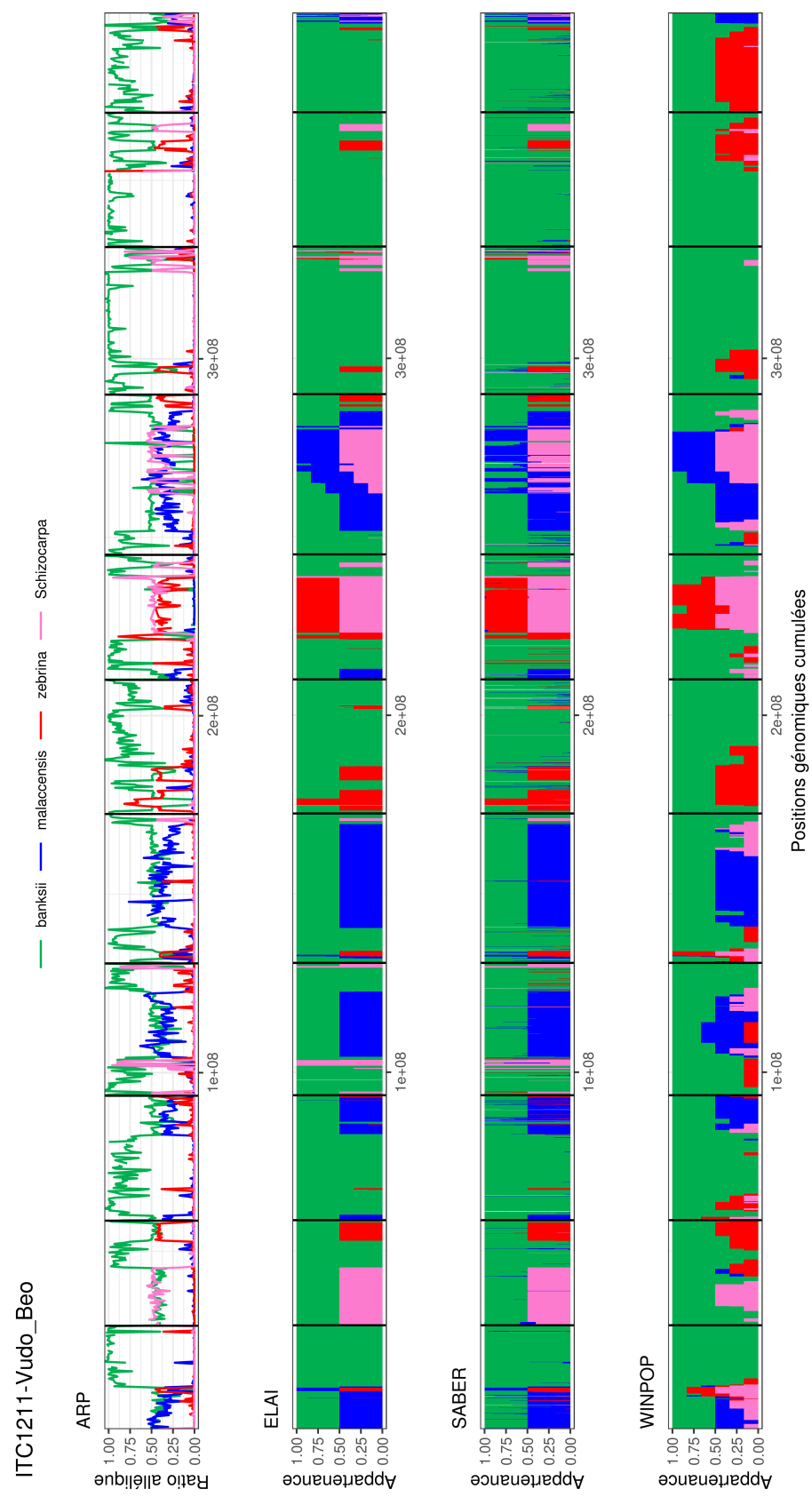

FIGURE S.23 – Résultats de l'ILEA et de l'ARP de l'accession 'Vudo Beo'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.



FIGURE S.24 – Résultats de l'ILEA et de l'ARP de l'accession 'Wompa'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

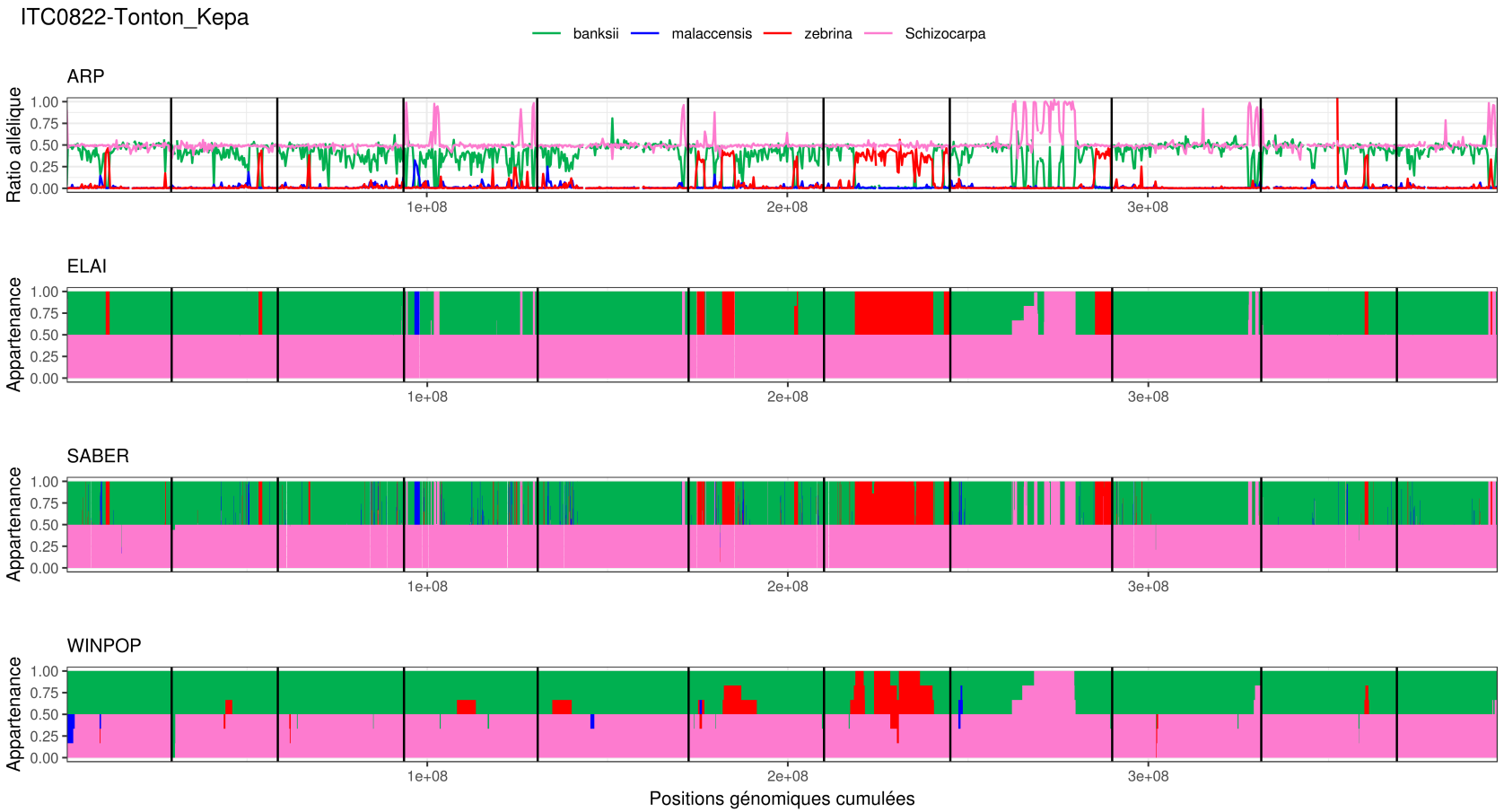

Figure S.25 – Résultats de l'ILEA et de l'ARP de l'accession 'Tonton Kepa'. Les trois répétitions sur les sous-jeux différents sont représentées combinées ici, par la moyenne des trois sur chacun des marqueurs pour chaque méthode d'ILEA. L'axe des abscisses représente les coordonnées génomiques cumulées des marqueurs utilisés. Le graphe n'est donc pas à l'échelle du génome entier ($\approx$ 500M) mais à celle du premier au dernier marqueur. L'axe des ordonnées représente la proportion d'appartenance à chacun des groupes ancestraux pour les ILEA, et le ratio de couverture allélique pour l'ARP. Les couleurs représentent les 4 groupes ancestraux de l'analyse, *M. a.* ssp. *banksii* (vert), *M. a.* ssp. *malaccensis* (bleu), *M. a.* ssp. *zebrina* (rouge), *M. schizocarpa* (rose). Les résultats d'ILEA présentés correspondent aux inférences avec un paramètre de temps depuis l'hybridation de 50 générations. Chaque sous-figure représente l'individu analysé par les différentes méthodes et ses 11 chromosomes séparés par des barres verticales noires.

TABLE S.1 – Origines, status et identifiants des 115 accessions de bananiers utilisées

Code DYNAMO	Section	*Groupe/Espèce*	Génome	Sauvage Cultivar	Noms accessions	Origine	Numéro d'accession	Origine géographique ou (collection)
DYN-ITC1387-Ensete	_	*Ensete ventricosum*	_	_	Ensete ventricosum	ITC Bioversity Int.	ITC1387	_
DYN005-Aata	AUSTRALIMUSA	*Fe'i*	_	cv	Aata	CRB Plantes Tropicales	PT-BA-00005	_
DYN009-Aiori	AUSTRALIMUSA	*Fe'i*	_	cv	Aiori	CRB Plantes Tropicales	PT-BA-00009	Polynésie française
ITC0956-Musa_lolodensis	AUSTRALIMUSA	*Musa lolodensis*	_	w	Musa lolodensis	ITC Bioversity Int.	ITC0956	PNG
DYN124-Hung_Si	AUSTRALIMUSA	*Musa maclayi*	_	w	Hung Si	CRB Plantes Tropicales	PT-BA-00124	PNG
ITC0917-Musa_peekelii_ ssp_peekelii	AUSTRALIMUSA	*Musa peekelii ssp. peekelii*	_	w	Musa peekelii ssp. peekelii	ITC Bioversity Int.	ITC0917	PNG
DYN228-Musa_textilis	AUSTRALIMUSA	*Musa textilis*	TT	w	Musa textilis	CRB Plantes Tropicales	PT-BA-00228	_
DYN222-Musa_coccinea	CALLIMUSA	*Musa coccinea*	_	w	Musa coccinea	CRB Plantes Tropicales	PT-BA-00222	_
DYN225-Musa_laterita	RHODOCHLAMYS	*Musa laterita*	_	w	Musa laterita	CRB Plantes Tropicales	PT-BA-00225	_
DYN-ITC1076-Musa_laterita	RHODOCHLAMYS	*Musa laterita*	_	w	Musa laterita	ITC Bioversity Int.	ITC1076	_
DYN226-Musa_ornata	RHODOCHLAMYS	*Musa ornata*	_	w	Musa ornata	CRB Plantes Tropicales	PT-BA-00226	_
DYN-ITC1591-Musa_rosea	RHODOCHLAMYS	*Musa rosea*	_	w	Musa rosea	ITC Bioversity Int.	ITC1591	_
DYN227-Musa_sanguinea	RHODOCHLAMYS	*Musa sanguinea*	_	w	Musa sanguinea	CRB Plantes Tropicales	PT-BA-00227	_
DYN229-Musa_velutina	RHODOCHLAMYS	*Musa velutina*	_	w	Musa velutina	CRB Plantes Tropicales	PT-BA-00229	_
DYN018-Balbisiana_-CMR-	EUMUSA	*Musa balbisiana*	BB	w	Balbisiana (CMR)	CRB Plantes Tropicales	PT-BA-00018	_
DYN019-Balbisiana_ Honduras	EUMUSA	*Musa balbisiana*	BB	w	Balbisiana Honduras	CRB Plantes Tropicales	PT-BA-00019	_
DYN049-Butuhan	EUMUSA	*Musa balbisiana*	BB	w	Butuhan	CRB Plantes Tropicales	PT-BA-00049	Philippines
DYN172-Lal_Velchi	EUMUSA	*Musa balbisiana*	BB	w	Lal Velchi	CRB Plantes Tropicales	PT-BA-00172	_
DYN302-PKW	EUMUSA	*Musa balbisiana*	BB	w	Pisang Klutuk Wulung	CRB Plantes Tropicales	PT-BA-00302	(Indonésie)
ITC0599-Schizocarpa	EUMUSA	*Musa schizocarpa*	SS	w	Schizocarpa	ITC Bioversity Int.	ITC0599	PNG
ITC0926-Schizocarpa	EUMUSA	*Musa schizocarpa*	SS	w	Schizocarpa	ITC Bioversity Int.	ITC0926	PNG
DYN319-Pisang_Segun	EUMUSA	*ind.*	AA	w	Pisang Segun	CRB Plantes Tropicales	PT-BA-00319	Malaisie ?
2013-211-Ambihy_P1	EUMUSA	*Musa acuminata*	AA	w	Ambihy_P1	Comores	_	Comores
DYN113-Hawain_2	EUMUSA	*M.a. ssp. banksii*	AA	w	Hawain 2	CRB Plantes Tropicales	PT-BA-00113	PNG
DYN412-Waigu	EUMUSA	*M.a. ssp. banksii*	AA	w	Waigu	CRB Plantes Tropicales	PT-BA-00412	_
DYN-Banksii_H09_2016-008	EUMUSA	*M.a. ssp. banksii*	AA	w	Banksii	CRB Plantes Tropicales	PT-BA-00024	_
DYN-ITC0464-Higa_BS464	EUMUSA	*M.a. ssp. banksii*	AA	w	Higa (ITC0464)	ITC Bioversity Int.	ITC0464	_
DYN-ITC0897-Banksii_ ITCO897	EUMUSA	*M.a. ssp. banksii*	AA	w	Musa acuminata ssp. banksii (ITCO897)	ITC Bioversity Int.	ITC0897	PNG
ITC0885-Banksii_ITC0885	EUMUSA	*M.a. ssp. banksii*	AA	w	Musa acuminata ssp. banksii (ITC0885)	ITC Bioversity Int.	ITC0885	PNG
DYN178-Long_Tavoy	EUMUSA	*M.a. ssp. burmannica*	AA	w	Long Tavoy	CRB Plantes Tropicales	PT-BA-00178	_
DYN-Calcutta_4_F08_ 2016-005	EUMUSA	*M.a. ssp. burmannicoïdes*	AA	w	Calcutta 4	CRB Plantes Tropicales	PT-BA-00051	Jardin Botanique Inde
DYN-ITC1028-Agutay	EUMUSA	*M.a. ssp. errans*	AA	w	Agutay	ITC Bioversity Int.	ITC1028	Philippines
DYN363-Selangor	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Sélangor	CRB Plantes Tropicales	PT-BA-00363	Malaisie
DYN454-Malaccensis_nain	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Malaccensis nain	CRB Plantes Tropicales	PT-BA-00454	_
DYN-ITC0609-Pahang	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Pahang	ITC Bioversity Int.	ITC0609	Malaisie
DYN-ITC1345-Pisang_Kra	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Pisang Kra	ITC Bioversity Int.	ITC1345	Malaisie
DYN-ITC1346-Malaccensis-Pisang_Karok_391	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Pisang Karok 391	ITC Bioversity Int.	ITC1346	Malaisie
DYN-ITC1348-Pisang_ serun_404	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Pisang serun 404	ITC Bioversity Int.	ITC1348	Malaisie
DYN-ITC1349-Pisang_ serun_400	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Pisang serun 400	ITC Bioversity Int.	ITC1349	Malaisie
DYN-Pahang_B07_2016-006	EUMUSA	*M.a. ssp. malaccensis*	AA	w	Pahang	CRB Plantes Tropicales	PT-BA-00267	_
DYN040-Borneo	EUMUSA	*M.a. ssp. microcarpa*	AA	w	Bornéo	CRB Plantes Tropicales	PT-BA-00040	Borneo

Code DYNAMO	Section	*Groupe/Espèce*	Génome	Sauvage Cultivar	Noms accessions	Origine	Numéro d'accession	Origine géographique ou (collection)
DYN078-EN13-IDN075	EUMUSA	*M.a. ssp. microcarpa*	AA	w	EN 13 (IDN075)	CRB Plantes Tropicales	PT-BA-00078	(Indonésie)
DYN004-AAs-IDN113	EUMUSA	*M.a. ssp. microcarpa der.*	AA	w	AAs-IDN113	CRB Plantes Tropicales	PT-BA-00004	(Indonésie)
DYN204-Microcarpa	EUMUSA	*M.a. ssp. microcarpa hyb.*	AA	w	Microcarpa	CRB Plantes Tropicales	PT-BA-00204	_
DYN147-Khae-Phrae	EUMUSA	*M.a. ssp. siamea*	AA	w	Khae (Phrae)	CRB Plantes Tropicales	PT-BA-00147	Thailande
DYN263-Pa_Rayong	EUMUSA	*M.a. ssp. siamea*	AA	w	Pa Rayong	CRB Plantes Tropicales	PT-BA-00263	Thailande
DYN262-Pa_Songkhla	EUMUSA	*M.a. ssp. siamea ?*	AA	w	Pa Songkhla	CRB Plantes Tropicales	PT-BA-00262	Thailande
ITC1701-Musa_acuminata_ssp_sumatrana	EUMUSA	*M.a. ssp. sumatrana*	AA	w	Musa acuminata sp.sumatrana	ITC Bioversity Int.	ITC1701	Indonésie
DYN398-Truncata	EUMUSA	*M.a. ssp. truncata*	AA	w	Truncata	CRB Plantes Tropicales	PT-BA-00398	_
DYN046-Buitenzorg	EUMUSA	*M.a. ssp. zebrina*	AA	w	Buitenzorg	CRB Plantes Tropicales	PT-BA-00046	Jardin Botanique Java
DYN059-Cici-Bresil	EUMUSA	*M.a. ssp. zebrina*	AA	w	Cici (Brésil)	CRB Plantes Tropicales	PT-BA-00059	_
DYN212-Monyet	EUMUSA	*M.a. ssp. zebrina*	AA	w	Monyet	CRB Plantes Tropicales	PT-BA-00212	(Indonésie)
DYN-ITC0415-Pisang_Cici_Alas	EUMUSA	*M.a. ssp. zebrina*	AA	w	Pisang Cici Alas	ITC Bioversity Int.	ITC0415	Indonésie
DYN-Maia_Oa_Q10_2016-007	EUMUSA	*M.a. ssp. zebrina*	AA	w	Maia Oa	CRB Plantes Tropicales	PT-BA-00182	Martinique
DYN-F1-048	EUMUSA	*F1 Pahang x Banksii*	AA	_	F1-048	hybride CIRAD	_	_
DYN-F1-050	EUMUSA	*F1 Pahang x Banksii*	AA	_	F1-050	hybride CIRAD	_	_
DYN-F1-061	EUMUSA	*F1 Pahang x Calcutta4*	AA	_	F1-061	hybride CIRAD	_	_
DYN-F1-075	EUMUSA	*F1 Pahang x Maia 0a*	AA	_	F1-075	hybride CIRAD	_	_
2013-098-Mlali_Mshia_Wa_Komba	EUMUSA	*AAcv*	AA	cv	Mlali Mshia Wa Komba	Mayotte	_	Mayotte
DYN010-Akondro_mainty	EUMUSA	*AAcv*	AA	cv	Akondro mainty	CRB Plantes Tropicales	PT-BA-00010	Madagascar
DYN027-Bebek	EUMUSA	*AAcv*	AA	cv	Bebek	CRB Plantes Tropicales	PT-BA-00027	Nouvelle Guinée, Indonésie
DYN031-Beram	EUMUSA	*AAcv*	AA	cv	Beram	CRB Plantes Tropicales	PT-BA-00031	Nouvelle Guinée, Indonésie
DYN067-Dibit	EUMUSA	*AAcv*	AA	cv	Dibit	CRB Plantes Tropicales	PT-BA-00067	Nouvelle Guinée, Indonésie
DYN097-Galeo	EUMUSA	*AAcv*	AA	cv	Galeo	CRB Plantes Tropicales	PT-BA-00097	PNG
DYN112-Guyod	EUMUSA	*AAcv*	AA	cv	Guyod	CRB Plantes Tropicales	PT-BA-00112	Philippines
DYN114-Heva	EUMUSA	*AAcv*	AA	cv	Heva	CRB Plantes Tropicales	PT-BA-00114	PNG
DYN120-Hom	EUMUSA	*AAcv*	AA	cv	Hom	CRB Plantes Tropicales	PT-BA-00120	Thailande
DYN127-IDN_077	EUMUSA	*AAcv*	AA	cv	IDN 077	CRB Plantes Tropicales	PT-BA-00127	(Indonésie)
DYN131-IDN_110	EUMUSA	*AAcv*	AA	cv	IDN 110	CRB Plantes Tropicales	PT-BA-00131	(Indonésie)
DYN148-Khai_Nai_on	EUMUSA	*AAcv*	AA	cv	Khai Nai on	CRB Plantes Tropicales	PT-BA-00148	Thailande
DYN154-Kirun	EUMUSA	*AAcv*	AA	cv	Kirun	CRB Plantes Tropicales	PT-BA-00154	PNG
DYN160-Kumburgh	EUMUSA	*AAcv*	AA	cv	Kumburgh	CRB Plantes Tropicales	PT-BA-00160	PNG
DYN190-Manang	EUMUSA	*AAcv*	AA	cv	Manang	CRB Plantes Tropicales	PT-BA-00190	Philippines
DYN233-N110-THA052	EUMUSA	*AAcv*	AA	cv	N°110 / THA 052	CRB Plantes Tropicales	PT-BA-00233	Thailande
DYN242-Niyarma_Yik	EUMUSA	*AAcv*	AA	cv	Niyarma Yik	CRB Plantes Tropicales	PT-BA-00242	(PNG)
DYN261-Pa-Patthalong	EUMUSA	*AAcv*	AA	cv	Pa (Patthalong)	CRB Plantes Tropicales	PT-BA-00261	Thailande
DYN270-Paka	EUMUSA	*AAcv*	AA	cv	Paka	CRB Plantes Tropicales	PT-BA-00270	_
DYN273-Pallenberry	EUMUSA	*AAcv*	AA	cv	Pallen Berry	CRB Plantes Tropicales	PT-BA-00273	Thailande
DYN281-Pisang_Bangkahulu	EUMUSA	*AAcv*	AA	cv	Pisang Bangkahulu	CRB Plantes Tropicales	PT-BA-00281	Indonésie
DYN284-Pisang_Berlin_IDN074	EUMUSA	*AAcv*	AA	cv	Pisang Berlin IDN074	CRB Plantes Tropicales	PT-BA-00284	(Indonésie)
DYN285-Pisang_Buntal	EUMUSA	*AAcv*	AA	cv	Pisang Buntal	CRB Plantes Tropicales	PT-BA-00285	Malaisie
DYN292-Pisang_Jaran	EUMUSA	*AAcv*	AA	cv	Pisang Jaran	CRB Plantes Tropicales	PT-BA-00292	(Indonésie)
DYN293-Pisang_jari_Buaya_BS312	EUMUSA	*AAcv*	AA	cv	Pisang Jari Buaya BS312	CRB Plantes Tropicales	PT-BA-00293	Malaisie
DYN304-Pisang_Madu	EUMUSA	*AAcv*	AA	cv	Pisang Madu	CRB Plantes Tropicales	PT-BA-00304	Malaisie

Code DYNAMO	Section	*Groupe/Espèce*	Génome	Sauvage Cultivar	Noms accessions	Origine	Numéro d'accession	Origine géographique ou (collection)
DYN305-Pisang_Mas	EUMUSA	*AAcv*	AA	cv	Pisang Mas	CRB Plantes Tropicales	PT-BA-00305	_
DYN310-Pisang_Pipit	EUMUSA	*AAcv*	AA	cv	Pisang Pipit	CRB Plantes Tropicales	PT-BA-00310	Indonésie
DYN316-Pisang_Sapon	EUMUSA	*AAcv*	AA	cv	Pisang Sapon	CRB Plantes Tropicales	PT-BA-00316	Indonésie
DYN367-SF265	EUMUSA	*AAcv*	AA	cv	S.F.265	CRB Plantes Tropicales	PT-BA-00367	(PNG)
DYN371-Sinwobogi	EUMUSA	*AAcv*	AA	cv	Sinwobogi	CRB Plantes Tropicales	PT-BA-00371	PNG
DYN377-Sowmuk	EUMUSA	*AAcv*	AA	cv	Sowmuk	CRB Plantes Tropicales	PT-BA-00377	PNG
DYN391-Thong_Det	EUMUSA	*AAcv*	AA	cv	Thong Det	CRB Plantes Tropicales	PT-BA-00391	Thailande
DYN393-Tjau_Lagada	EUMUSA	*AAcv*	AA	cv	Tjau Lagada	CRB Plantes Tropicales	PT-BA-00393	_
DYN443-Gwanhour	EUMUSA	*AAcv*	AA	cv	Gwanhour	CRB Plantes Tropicales	PT-BA-00443	(PNG)
DYN-ITC0063-Pisang_ Tongat	EUMUSA	*AAcv*	AA	cv	Pisang Tongat	ITC Bioversity Int.	ITC0063	_
DYN-ITC0589-Gulum	EUMUSA	*AAcv*	AA	cv	Gulum	ITC Bioversity Int.	ITC0589	PNG
DYN-ITC0785-Aivip	EUMUSA	*AAcv*	AA	cv	Aivip	ITC Bioversity Int.	ITC0785	PNG
DYN-ITC0786-Katual_n2	EUMUSA	*AAcv*	AA	cv	Katual no.2	ITC Bioversity Int.	ITC0786	PNG
DYN-ITC0996-Manameg_ Red	EUMUSA	*AAcv*	AA	cv	Manameg Red	ITC Bioversity Int.	ITC0996	PNG
DYN-ITC1013-Sena	EUMUSA	*AAcv*	AA	cv	Sena	ITC Bioversity Int.	ITC1013	PNG
DYN-ITC1031-Veinte_Cohol	EUMUSA	*AAcv*	AA	cv	Veinte Cohol	ITC Bioversity Int.	ITC1031	Philippines
DYN-ITC1704-Kole	EUMUSA	*AAcv*	AA	cv	Kole	ITC Bioversity Int.	ITC1704	(Indonésie)
ITC0434-Racadag	EUMUSA	*AAcv*	AA	cv	Racadag	ITC Bioversity Int.	ITC0434	Philippines
ITC0568-Malaysian_Blood	EUMUSA	*AAcv*	AA	cv	Malaysian Blood	ITC Bioversity Int.	ITC0568	_
ITC0778-Gorop	EUMUSA	*AAcv*	AA	cv	Gorop	ITC Bioversity Int.	ITC0778	PNG
ITC1187-Tomolo	EUMUSA	*AAcv*	AA	cv	Tomolo	ITC Bioversity Int.	ITC1187	PNG
ITC1206-Spiral	EUMUSA	*AAcv*	AA	cv	Spiral	ITC Bioversity Int.	ITC1206	PNG
ITC1211-Vudo_Beo	EUMUSA	*AAcv*	AA	cv	Vudu Beo	ITC Bioversity Int.	ITC1211	PNG
ITC1245-Papat	EUMUSA	*AAcv*	AA	cv	Papat	ITC Bioversity Int.	ITC1245	PNG
DYN-ITC0810-Sihir	EUMUSA	*AAcv*	AA	cv	Sihir	ITC Bioversity Int.	ITC0810	PNG
DYN061-Colatina_Ouro	EUMUSA	*AAcv ?*	AA	cv	Colatina Ouro	CRB Plantes Tropicales	PT-BA-00061	_
DYN303-Pisang_Lilin	EUMUSA	*AAcv.malaccencis der.*	AA	cv	Pisang Lilin	CRB Plantes Tropicales	PT-BA-00303	_
DYN163-Kunnan	EUMUSA	*ABcv Kunnan*	AB	cv	Kunnan	CRB Plantes Tropicales	PT-BA-00163	_
DYN359-Safet_Velchi	EUMUSA	*ABcv Ney Poovan*	AB	cv	Safet Velchi	CRB Plantes Tropicales	PT-BA-00359	_
DYN423-Wompa	EUMUSA	*AScv*	AS	cv	Wompa	CRB Plantes Tropicales	PT-BA-00423	PNG
ITC0822-Tonton_Kepa	EUMUSA	*AScv*	AS	cv	Tonton Kepa	ITC Bioversity Int.	ITC0822	PNG

TABLE S.2 – Pourcentage de site hétérozygote par accession.

Code DYNAMO	Section	*Groupe/Espèce*	Génome	Sauvage/Cultivar	Heterozygotie
DYN-ITC1387-Ensete	_	*Ensete ventricosum*	_	_	0,88
DYN005-Aata	AUSTRALIMUSA	*Fe'i*	_	cv	1,53
DYN009-Aiori	AUSTRALIMUSA	*Fe'i*	_	cv	1,52
ITC0956-Musa_lolodensis	AUSTRALIMUSA	*Musa lolodensis*	_	w	0,9
DYN124-Hung_Si	AUSTRALIMUSA	*Musa maclayi*	_	w	1,31
ITC0917-Musa_peekelii_ssp_peekelii	AUSTRALIMUSA	*Musa peekelii ssp. peekelii*	_	w	1,1
DYN228-Musa_textilis	AUSTRALIMUSA	*Musa textilis*	TT	w	1,35
DYN222-Musa_coccinea	CALLIMUSA	*Musa coccinea*	_	w	0,75
DYN225-Musa_laterita	RHODOCHLAMYS	*Musa laterita*	_	w	1,64
DYN-ITC1076-Musa_laterita	RHODOCHLAMYS	*Musa laterita*	_	w	1,22
DYN226-Musa_ornata	RHODOCHLAMYS	*Musa ornata*	_	w	0,6
DYN-ITC1591-Musa_rosea	RHODOCHLAMYS	*Musa rosea*	_	w	1,17
DYN227-Musa_sanguinea	RHODOCHLAMYS	*Musa sanguinea*	_	w	3,96
DYN229-Musa_velutina	RHODOCHLAMYS	*Musa velutina*	_	w	0,22
DYN018-Balbisiana_-CMR-	EUMUSA	*Musa balbisiana*	BB	w	1,26
DYN019-Balbisiana_Honduras	EUMUSA	*Musa balbisiana*	BB	w	1,06
DYN049-Butuhan	EUMUSA	*Musa balbisiana*	BB	w	1,46
DYN172-Lal_Velchi	EUMUSA	*Musa balbisiana*	BB	w	2,2
DYN302-PKW	EUMUSA	*Musa balbisiana*	BB	w	1,83
ITC0599-Schizocarpa	EUMUSA	*Musa schizocarpa*	SS	w	0,2
ITC0926-Schizocarpa	EUMUSA	*Musa schizocarpa*	SS	w	0,12
DYN319-Pisang_Segun	EUMUSA	*ind.*	AA	w	1,8
2013-211-Ambihy_P1	EUMUSA	*Musa acuminata*	AA	w	2,26
DYN113-Hawain_2	EUMUSA	*M.a. ssp. banksii*	AA	w	0,07
DYN412-Waigu	EUMUSA	*M.a. ssp. banksii*	AA	w	0,91
DYN-Banksii_H09_2016-008	EUMUSA	*M.a. ssp. banksii*	AA	w	0,08
DYN-ITC0464-Higa_BS464	EUMUSA	*M.a. ssp. banksii*	AA	w	0,12
DYN-ITC0897-Banksii_ITCO897	EUMUSA	*M.a. ssp. banksii*	AA	w	0,38
ITC0885-Banksii_ITC0885	EUMUSA	*M.a. ssp. banksii*	AA	w	0,1
DYN178-Long_Tavoy	EUMUSA	*M.a. ssp. burmannica*	AA	w	1,86
DYN-Calcutta_4_F08_2016-005	EUMUSA	*M.a. ssp. burmannicoïdes*	AA	w	2,3
DYN-ITC1028-Agutay	EUMUSA	*M.a. ssp. errans*	AA	w	0,5
DYN363-Selangor	EUMUSA	*M.a. ssp. malaccensis*	AA	w	1,67
DYN454-Malaccensis_nain	EUMUSA	*M.a. ssp. malaccensis*	AA	w	2,39
DYN-ITC0609-Pahang	EUMUSA	*M.a. ssp. malaccensis*	AA	w	1,28
DYN-ITC1345-Pisang_Kra	EUMUSA	*M.a. ssp. malaccensis*	AA	w	4,04
DYN-ITC1346-Malaccensis-Pisang_Karok_391	EUMUSA	*M.a. ssp. malaccensis*	AA	w	2,5
DYN-ITC1348-Pisang_serun_404	EUMUSA	*M.a. ssp. malaccensis*	AA	w	2,37
DYN-ITC1349-Pisang_serun_400	EUMUSA	*M.a. ssp. malaccensis*	AA	w	2,85
DYN-Pahang_B07_2016-006	EUMUSA	*M.a. ssp. malaccensis*	AA	w	3,03
DYN040-Borneo	EUMUSA	*M.a. ssp. microcarpa*	AA	w	1,96
DYN078-EN13-IDN075	EUMUSA	*M.a. ssp. microcarpa*	AA	w	4,97
DYN004-AAs-IDN113	EUMUSA	*M.a. ssp. microcarpa der.*	AA	w	1,4
DYN204-Microcarpa	EUMUSA	*M.a. ssp. microcarpa hyb.*	AA	w	3,73
DYN147-Khae-Phrae	EUMUSA	*M.a. ssp. siamea*	AA	w	0,46
DYN263-Pa_Rayong	EUMUSA	*M.a. ssp. siamea*	AA	w	1,59
DYN262-Pa_Songkhla	EUMUSA	*M.a. ssp. siamea ?*	AA	w	1,96
ITC1701-Musa_acuminata_ssp_sumatrana	EUMUSA	*M.a. ssp. sumatrana*	AA	w	0,1
DYN398-Truncata	EUMUSA	*M.a. ssp. truncata*	AA	w	1,83
DYN046-Buitenzorg	EUMUSA	*M.a. ssp. zebrina*	AA	w	4,37
DYN059-Cici-Bresil	EUMUSA	*M.a. ssp. zebrina*	AA	w	1,05
DYN212-Monyet	EUMUSA	*M.a. ssp. zebrina*	AA	w	1,29
DYN-ITC0415-Pisang_Cici_Alas	EUMUSA	*M.a. ssp. zebrina*	AA	w	1,03
DYN-Maia_Oa_Q10_2016-007	EUMUSA	*M.a. ssp. zebrina*	AA	w	0,76
DYN-F1-048	EUMUSA	*F1 Pahang x Banksii*	AA	_	3,83
DYN-F1-050	EUMUSA	*F1 Pahang x Banksii*	AA	_	3,82
DYN-F1-061	EUMUSA	*F1 Pahang x Calcutta4*	AA	_	4,59
DYN-F1-075	EUMUSA	*F1 Pahang x Maia 0a*	AA	_	4,42
2013-098-Mlali_Mshia_Wa_Komba	EUMUSA	*AAcv*	AA	cv	3,52
DYN010-Akondro_mainty	EUMUSA	*AAcv*	AA	cv	3,53
DYN027-Bebek	EUMUSA	*AAcv*	AA	cv	2,83
DYN031-Beram	EUMUSA	*AAcv*	AA	cv	3,33
DYN067-Dibit	EUMUSA	*AAcv*	AA	cv	3,19
DYN097-Galeo	EUMUSA	*AAcv*	AA	cv	3,43
DYN112-Guyod	EUMUSA	*AAcv*	AA	cv	2,32
DYN114-Heva	EUMUSA	*AAcv*	AA	cv	3,63
DYN120-Hom	EUMUSA	*AAcv*	AA	cv	4,89
DYN127-IDN_077	EUMUSA	*AAcv*	AA	cv	3,66
DYN131-IDN_110	EUMUSA	*AAcv*	AA	cv	3,83
DYN148-Khai_Nai_on	EUMUSA	*AAcv*	AA	cv	3,74
DYN154-Kirun	EUMUSA	*AAcv*	AA	cv	3,44
DYN160-Kumburgh	EUMUSA	*AAcv*	AA	cv	3,04
DYN190-Manang	EUMUSA	*AAcv*	AA	cv	3,9
DYN233-N110-THA052	EUMUSA	*AAcv*	AA	cv	4,02
DYN242-Niyarma_Yik	EUMUSA	*AAcv*	AA	cv	2,97

Code DYNAMO	Section	*Groupe/Espèce*	Génome	Sauvage/Cultivar	Heterozygotie
DYN261-Pa-Patthalong	EUMUSA	*AAcv*	AA	cv	4
DYN270-Paka	EUMUSA	*AAcv*	AA	cv	3,25
DYN273-Pallenberry	EUMUSA	*AAcv*	AA	cv	3,7
DYN281-Pisang_Bangkahulu	EUMUSA	*AAcv*	AA	cv	3,66
DYN284-Pisang_Berlin_IDN074	EUMUSA	*AAcv*	AA	cv	3,78
DYN285-Pisang_Buntal	EUMUSA	*AAcv*	AA	cv	3,35
DYN292-Pisang_Jaran	EUMUSA	*AAcv*	AA	cv	3,82
DYN293-Pisang_jari_Buaya_BS312	EUMUSA	*AAcv*	AA	cv	5,24
DYN304-Pisang_Madu	EUMUSA	*AAcv*	AA	cv	5,13
DYN305-Pisang_Mas	EUMUSA	*AAcv*	AA	cv	3,45
DYN310-Pisang_Pipit	EUMUSA	*AAcv*	AA	cv	3,78
DYN316-Pisang_Sapon	EUMUSA	*AAcv*	AA	cv	3,51
DYN367-SF265	EUMUSA	*AAcv*	AA	cv	2,63
DYN371-Sinwobogi	EUMUSA	*AAcv*	AA	cv	1,86
DYN377-Sowmuk	EUMUSA	*AAcv*	AA	cv	3,31
DYN391-Thong_Det	EUMUSA	*AAcv*	AA	cv	3,68
DYN393-Tjau_Lagada	EUMUSA	*AAcv*	AA	cv	3,61
DYN443-Gwanhour	EUMUSA	*AAcv*	AA	cv	2,32
DYN-ITC0063-Pisang_Tongat	EUMUSA	*AAcv*	AA	cv	3,25
DYN-ITC0589-Gulum	EUMUSA	*AAcv*	AA	cv	1,61
DYN-ITC0785-Aivip	EUMUSA	*AAcv*	AA	cv	1,54
DYN-ITC0786-Katual_n2	EUMUSA	*AAcv*	AA	cv	2,36
DYN-ITC0996-Manameg_Red	EUMUSA	*AAcv*	AA	cv	1,18
DYN-ITC1013-Sena	EUMUSA	*AAcv*	AA	cv	1,1
DYN-ITC1031-Veinte_Cohol	EUMUSA	*AAcv*	AA	cv	3,38
DYN-ITC1704-Kole	EUMUSA	*AAcv*	AA	cv	3,81
ITC0434-Racadag	EUMUSA	*AAcv*	AA	cv	3,96
ITC0568-Malaysian_Blood	EUMUSA	*AAcv*	AA	cv	3,42
ITC0778-Gorop	EUMUSA	*AAcv*	AA	cv	1,05
ITC1187-Tomolo	EUMUSA	*AAcv*	AA	cv	1,52
ITC1206-Spiral	EUMUSA	*AAcv*	AA	cv	1,19
ITC1211-Vudo_Beo	EUMUSA	*AAcv*	AA	cv	2,2
ITC1245-Papat	EUMUSA	*AAcv*	AA	cv	1,66
DYN-ITC0810-Sihir	EUMUSA	*AAcv*	AA	cv	1,35
DYN061-Colatina_Ouro	EUMUSA	*AAcv ?*	AA	cv	3,71
DYN303-Pisang_Lilin	EUMUSA	*AAcv.malaccencis der.*	AA	cv	3,7
DYN163-Kunnan	EUMUSA	*ABcv Kunnan*	AB	cv	10,33
DYN359-Safet_Velchi	EUMUSA	*ABcv Ney Poovan*	AB	cv	10,41
DYN423-Wompa	EUMUSA	*AScv*	AS	cv	5,41
ITC0822-Tonton_Kepa	EUMUSA	*AScv*	AS	cv	5,31

TABLE S.3 – Nombre d'allèles associés aux groupes pour les 3 itérations de l'approche « ARP »

Groupes	Itération 1	Itération 2	Itération 3
banksii	204 960	188 660	157 603
burmannica/siamea	331 758	560 778	840 397
malaccensis	260 333	413 973	650 576
sumatrana	_	_	207 250
zebrina	280 829	333 137	421 514
Balbisiana	901 140	1 264 425	1 248 229
Ensete	2 760 188	2 711 046	2 685 417
Schizocarpa	380 749	376 470	342 192
Australimusa	1 911 477	2 571 963	2 544 385
Velutina	609 024	592 259	582 084
TOTAL	7 877 784	9 233 638	9 679 647

TABLE S.4 – Résumé de la séléction d'individu représentatif des pôles par ACP, `Admixture` et « ARP ».

Code Dynamo	Groupe	ACP	ADMX $K = 7$	ADMX $K = 9$	ADMX $K = 11$	ADMX $K = 13$	ARP 1	ARP 2	ARP 3	Résumé
DYN009-Aiori	*Fe'i*	T	T	T	T	T	_	T	T	T
DYN005-Aata	*Hybrid Eumusa x Australimusa*	T	T	T	T	T	_	T	T	T
ITC0956-Musa_lolodensis	*M. lolodensis*	T	T	T	T	T	_	T	T	T
DYN124-Hung_Si	*M. maclayi*	T	T	T	T	T	_	T	T	T
ITC0917-Musa_peekelii_ssp_peekelii	*M. peekelii ssp. peekelii*	T	T	T	T	T	_	T	T	T
DYN228-Musa_textilis	*M. textilis*	T	T	T	T	T	T	T	T	T
DYN222-Musa_coccinea	*M. coccinea*	c1	_	_	_	c1	_	_	_	_
DYN319-Pisang_Segun	*indeterminé*	_	m	m	m	m	_	_	_	_
DYN-Banksii_H09_2016-008	*M. a. ssp. banksii*	b	b	b	b	b	_	b	b	b
DYN-ITC0464-Higa_BS464	*M. a. ssp. banksii*	b	b	b	b	b	_	_	_	_
DYN-ITC0897-Banksii_ITCO897	*M. a. ssp. banksii*	b	b	b	b	b	_	_	_	_
DYN113-Hawain_2	*M. a. ssp. banksii*	b	b	b	b	b	b	b	b	b
DYN412-Waigu	*M. a. ssp. banksii*	_	_	_	_	_	_	_	_	_
ITC0885-Banksii_ITC0885	*M. a. ssp. banksii*	b	b	b	b	b	_	_	_	_
DYN178-Long_Tavoy	*M. a. ssp. burmannica*	s	s,c2,c3	s,c3	S,s	s	_	_	s	s
DYN-Calcutta_4_F08_2016-005	*M. a. ssp. burmannicoïdes*	s	s,c2,c3	s,c3	S,s	s	_	_	s	s
DYN-ITC1028-Agutay	*M. a. ssp. errans*	_	_	_	c4	c4	_	_	_	_
DYN-ITC0609-Pahang	*M. a. ssp. malaccensis*	m	m,z	m	m	m	m	m	m	m
DYN-ITC1345-Pisang_Kra	*M. a. ssp. malaccensis*	_	_	_	_	_	_	_	_	_
DYN-ITC1346-Malaccensis-Pisang_Karok_391	*M. a. ssp. malaccensis*	m	m,z	m	m	m	_	_	_	_
DYN-ITC1348-Pisang_serun_404	*M. a. ssp. malaccensis*	m	m,z	m	m	m	_	_	_	_
DYN-ITC1349-Pisang_serun_400	*M. a. ssp. malaccensis*	m	m,z	m	_	_	_	_	_	_
DYN-Pahang_B07_2016-006	*M. a. ssp. malaccensis*	m	m,z	m	m	m	_	_	_	_
DYN363-Selangor	*M. a. ssp. malaccensis*	m	m,z	m	m	m	_	m	m	m
DYN454-Malaccensis_nain	*M. a. ssp. malaccensis*	m	m,z	m	m	m	_	_	m	m
DYN040-Borneo	*M. a. ssp. microcarpa*	_	_	_	c4	c4	_	_	_	_
DYN078-EN13-IDN075	*M. a. ssp. microcarpa*	_	_	_	_	_	_	_	_	_
DYN004-AAs-IDN113	*M. a. ssp. microcarpa der.*	_	_	_	_	_	_	_	_	_
DYN204-Microcarpa	*M. a. ssp. microcarpa hyb.*	_	_	_	_	_	_	_	_	_
DYN147-Khae-Phrae	*M. a. ssp. siamea*	s	s,c2,c3	s,c3	S,s	s	s	s	s	s
DYN263-Pa_Rayong	*M. a. ssp. siamea*	s	s,c2,c3	s,c3	S,s	s	_	s	s	s
DYN262-Pa_Songkhla	*M. a. ssp. siamea ?*	m	m	m	m	m	_	_	m	m
ITC1701-Musa_a._ssp_sumatrana	*M. a. ssp. sumatrana*	_	_	_	_	_	_	u	u	u
DYN398-Truncata	*M. a. ssp. truncata*	_	_	_	_	_	_	_	_	_
DYN-ITC0415-Pisang_Cici_Alas	*M. a. ssp. zebrina*	z	m,z	z	z	z	_	_	z	z

Code Dynamo	Groupe	ACP	ADMX $K = 7$	ADMX $K = 9$	ADMX $K = 11$	ADMX $K = 13$	ARP 1	ARP 2	ARP 3	Résumé
DYN-Maia_Oa_Q10_2016-007	*M. a. ssp. zebrina*	z	m,z	z	z	z	z	z	z	z
DYN046-Buitenzorg	*M. a. ssp. zebrina*	_	_	_	_	_	_	_	_	_
DYN059-Cici-Bresil	*M. a. ssp. zebrina*	z	m,z	z	z	z	_	_	_	_
DYN212-Monyet	*M. a. ssp. zebrina*	z	m,z	z	z	z	_	z	z	z
DYN018-Balbisiana_-CMR-	*M. balbisiana*	B	B	B	B	B	_	B	B	B
DYN019-Balbisiana_Honduras	*M. balbisiana*	B	B	B	B	B	B	B	B	B
DYN049-Butuhan	*M. balbisiana*	B	B	B	B	B	_	B	B	B
DYN172-Lal_Velchi	*M. balbisiana*	B	B	B	B	B	_	B	B	B
DYN302-PKW	*M. balbisiana*	B	B	B	B	B	_	B	B	B
ITC0599-Schizocarpa	*M. schizocarpa*	S	S	S	S,s	S	S	S	S	S
ITC0926-Schizocarpa	*M. schizocarpa*	S	S	S	S,s	S	_	S	S	S
DYN-ITC1387-Ensete	*Ensete ventricosum*	E	E	E	E	E	_	_	_	_
DYN-ITC1076-Musa_laterita	*M. laterita*	c3	s,c2,c3	s,c3	S,c3	c2,c3	_	_	_	_
DYN225-Musa_laterita	*M. laterita*	c3	s,c2,c3	s,c3	S,c3	c2,c3	_	_	_	_
DYN226-Musa_ornata	*M. ornata*	c2	s,c2,c3	c2	c2	_	_	_	_	_
DYN-ITC1591-Musa_rosea	*M. rosea*	c3	s,c2,c3	s,c3	S,c3	c2,c3	_	_	_	_
DYN227-Musa_sanguinea	*M. sanguinea*	c2	s,c2,c3	c2	c2	c2,c3	_	_	_	_
DYN229-Musa_velutina	*M. velutina*	c2	s,c2,c3	c2	c2	c2,c3	_	V	V	V

Les 2 premières colonnes contiennent le code de l'accession et son groupe présumé. La colonne « Code » contient un codage court correspondant aux groupes potentiels des accessions. Ceux ci sont : 'b' pour *M. a.* ssp. *banksii*, 'm' pour *M. a.* ssp. *malaccensis*, 's' pour *M. a.* ssp. *burmannica/siamea*, 'z' pour *M. a.* ssp. *zebrina*, 'B' pour *M. balbisiana*, 'E' pour *Ensete*, 'T' pour *Australimusa* (basé sur *M. textilis*). 'c' désigne des clusters mélangeant différentes espèces ou sous-espèces, 'c1' pour *M. coccinea*, 'c2' pour *M. ornata* ; *M. sanguinea* ; *M. velutina* (*Rhodochlamys*), 'c3' pour *M. laterita* ; *M. rosea* (*Rhodochlamys*) et 'c4' pour *M. a.* ssp. *errans* et *M. a.* ssp. *microcarpa*. *M. velutina* est représenté par 'V' dans l'analyse ARP (puisqu'il n'agrège pas de groupes). Les colonnes suivantes résument les groupes proposés par les différentes méthodes, ADMX K correspondant à `ADMIXTURE` avec la valeur de K de 7 à 13, ARP correspondant aux individus non hybrides des trois itérations notées 1, 2 et 3. La colonne « résumé » contient la sélection finale des individus représentatifs des groupes.