Etude de la domestication et de l'adaptation de l'igname (Dioscorea spp) en Afrique par des approches génomiques

Roland Akakpo

université PARIS-SACLAY

NNT2018SACLS124

Thèse de doctorat

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Etude de la domestication et de l'adaptation de l'igname (*Dioscorea spp*)en Afrique par des approches génomiques

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°567 Sciences du Végétal : du gène à l'écosystème
Spécialité de doctorat : Biologie

Thèse présentée et soutenue le 16 Mai 2018, à Gif-sur-Yvette, par

Roland AKAKPO

Composition du Jury :

Mme	Ronfort	Joëlle	DR INRA, UMR AGAP, Montpellier	Rapporteur
M.	Castric	Vincent	DR CNRS, UMR EEP, Lille	Rapporteur
M.	Choulet	Frédéric	IR INRA, UMR GDEC, Clermont-Ferrand	Examinateur
Mme	Shykoff	Jacqui	DR CNRS, UMR ESE, Orsay	Président
Mme	Alix	Karine	MC AgroParisTech, UMR GQE, Gif s/Yvette	Co-directrice de Thèse
M.	Vigouroux	Yves	DR IRD, UMR DIADE, Montpellier	Co-directeur de Thèse

« La chose la plus importante est de ne pas s'arrêter de s'interroger.
La curiosité a sa propre raison d'exister. »

Albert Einstein

A ma mère feue Albertine Bouco,
Maman, ne te lasse pas de veiller sur nous…

« L'Homme le plus heureux est celui qui fait le bonheur d'un plus grand nombre d'autres ».
C'est de cette pensée de Denis Diderot que je m'inspire pour dire un Grand Merci à Karine Alix et Yves Vigouroux pour avoir accepté d'encadrer ce travail. Vos qualités scientifiques et humaines m'ont été d'une très grande utilité pour ces trois années passées à vos côtés. Honneur aux femmes dit-on, je vais commencer par Karine. Karine, tu n'as pas hésité à accepter la main tendue de Yves quand il s'est agi pour lui de me trouver un encadrant de l'Université Paris-Sud, sur exigence de la BID qui a financé mon travail. Et, ensuite, au-delà de la relation de travail qui s'est installée entre nous, tu as été pour moi une conseillère, une confidente, une mère tout simplement. Puisse Dieu te le rendre au centuple. Comme l'a dit Diderot, soit heureuse Karine, car tu as fait le bonheur d'un de plus (Moi). Yves, trouves par ces mots l'expression de ma gratitude pour ces moments passés ensemble. Tu as su, par ta rigueur me pousser à donner le maximum de moi-même. Que dire ? Si j'étais une orange, tu m'as pressé jusqu'à vider tout mon contenu. Mais, tout comme les reptiles qui se muent, il fallait me vider de ce contenu pour que je devienne tout neuf. Nous-y sommes. Yves, sois l'un des heureux dont parle Diderot, car tu m'as procuré le bonheur.

Puisque l'enfant qui naît ne fait pas que le bonheur de ses deux parents, je voudrais ici formuler ma gratitude à l'endroit de certaines personnes qui n'ont ménagé aucun effort quant à l'heureuse issue de mes travaux de thèse.

A Nora Scarcelli et à Marie Courdec, pour avoir accepté de partager le même local avec moi pendant ces trois années, je voudrais vous témoigner ma reconnaissance en ces lignes. Particulièrement à toi Nora, je te suis reconnaissant pour tes disponibilités en temps voulu. J'aurais souhaité passer tout ce temps avec toi au Cameroun plutôt qu'à Montpellier mais hélas ! Merci également pour tes multiples lectures de mon manuscrit.

A tous les membres de l'équipe Dynadiv : Philippe, Cédric, Leila, Anne-Céline, Bénédicte, Cécile, Thomas, Valérie, Adeline, Jérôme, Kevin, merci pour vos précieux conseils.
A tous les membres de l'UMR DIADE, en l'occurrence à son Directeur Alain Guesquière avec qui j'avais déjà des relations personnelles avant d'atterrir dans son UMR. Merci Alain pour avoir facilité mon travail.
Je remercie également Bruno et Carole pour leur disponibilité et accompagnement administratif.
Mes remerciements vont particulièrement à l'endroit des membres de la plateforme de Bioinformatique : je reconnais que j'ai été parfois trop demandeur, mais rassurez-vous, c'était pour la bonne cause, et nous y voici. Ndomassi Tando, merci pour ta disponibilité ; François Sabot, merci pour tes conseils en tout genre. Ah ! Christine Trachant : tu m'as toujours taxé de quelqu'un qui fait deux thèses à la fois, et voilà, elles sont finies ! Merci Christine pour ta disponibilité. Je venais te voir avec des problèmes toujours neufs et j'ai toujours trouvé des

solutions. Dieu te bénisse Christine. Comme l'a dit Diderot, soit heureuse Christine, car tu as contribué à mon bonheur.
Enfin dans cette rubrique sur Montpellier, mes remerciements vont à l'endroit de tous les collègues doctorants, ceux qui sont passés et ceux qui y sont toujours. Je voudrais remercier particulièrement Rémi Tournebize, qui n'a jamais ménagé ses efforts pour se mettre à mon service, avec ses connaissances combien importantes en analyse de données. Merci Rémi !

Mes remerciements vont également à l'endroit des membres de mon équipe DyGAP de l'UMR GQE – Le Moulon et également, au Directeur de l'UMR M. Olivier Martin. Eh oui, j'ai toujours les choses en double ! A vous tous du Moulon qui avez contribué de près ou de loin à ce travail, trouvez ici l'expression de ma reconnaissance : Khalid, Valérie et Frédéric, Zeineb, je vous adresse mes remerciements.
Je n'oublie pas les collaborateurs du Bénin, mon co-encadrant du Bénin, Prof Alexandre Dansi, pour toute sa disponibilité et tout son soutien ; M. Gustave Djedatin également pour son soutien scientifique non négociable.

Je voudrais enfin remercier mes parents, mes frères et mes amis, et tous mes proches pour leur soutien moral et spirituel. Mes gratitudes vont également à l'endroit de mon père, que Dieu te bénisse papa ! J'ai une pensée toute particulière pour ma compagne, Noemie pour tous ses sacrifices durant ces trois années passées loin du corps, mais près du cœur. Je t'en suis reconnaissant, partenaire, comme je t'appelle affectueusement !

Je pense également à vous, chers membres de mon jury. Malgré vos agendas assez chargés, vous avez réussi chacun à consacrer du temps pour mon travail. Merci à Madame Joëlle Ronfort et Monsieur Vincent Castric pour avoir accepté de rapporter mon travail ; et à Monsieur Frédéric Choulet pour avoir accepté de l'examiner. Ah ! Madame Jacqui Shykoff, Directrice de l'Ecole Doctorale, vous n'avez ménagé aucun effort dans votre rôle administratif lors de la mise en place de mon financement. Et vous voici encore membre et examinateur de mon travail. Trouvez ici l'expression de ma profonde gratitude.

Je m'en voudrais de ne pas reconnaître le mérite de ces personnes qui ont très tôt accepté de jouer le rôle d'éclaireur à ce travail de doctorat. A Maud Tenaillon, à Kader Aïnouche, et à Florian Maumus : Merci infiniment pour vos multiples conseils, remarques, recommandations et orientation à mon endroit lors de mes deux comités de thèse. Merci particulièrement à Florian pour son rôle, au combien important, dans l'annotation des premiers éléments transposables de l'igname !

A tous ceux dont je n'ai pas cité ici les noms, ce n'est pas un oubli, je vous porte dans mon cœur, comme le dit Hans Christian Andersen « *La reconnaissance est la mémoire du cœur* ».

Résumé

L'igname (*Dioscorea spp*) est un aliment de base de plus de 100 millions de personnes en Afrique. Cette espèce est multipliée de manière végétative, avec un nombre de variétés cultivées limité. Des études suggèrent que de nouvelles variétés pourraient être créées par l'exploitation de plantes hybrides cultivée/sauvage ou même de plantes sauvages. Mais cela nécessite de comprendre la diversité des ressources génétiques disponibles et notamment d'identifier des cibles d'amélioration liées à l'adaptation. L'objectif de cette thèse était de comprendre les bases génétiques et génomiques de la domestication de l'igname en Afrique, et de mener la caractérisation génomique de son adaptation à différentes zones climatiques. L'étude du processus de domestication de l'igname a été menée par une approche de génomique comparée entre l'espèce cultivée *D. rotundata* et deux espèces sauvages apparentées *D. praehensilis* et *D. abyssinica*, en utilisant des données de séquençage NGS et en exploitant les variants SNP identifiés. Nous avons mis en évidence, comme déjà identifié chez certaines céréales, que la voie de biosynthèse et de stockage de l'amidon a été particulièrement sélectionnée au cours de la domestication de l'igname. De manière plus spécifique, des gènes impliqués dans la morphologie des tubercules ou l'aptitude au phototropisme, ainsi que des gènes du complexe NADH deshydrogenase ont également été sélectionnés. En lien avec des fonctions associées à la photosynthèse, cette sélection de NADH DH traduirait le passage d'une adaptation des ignames sauvages inféodées à des milieux sombres de forêt ou de savanes, à une adaptation de l'igname cultivée à des environnements de plein soleil de culture au champ. Ce même complexe NADH-DH a également été identifié lors de la recherche de gènes associés à la distribution d'une collection d'ignames selon la variabilité climatique. Des analyses complémentaires permettront de vérifier si l'étude de l'association gènes-climat ne permettrait pas aussi d'étudier les bases génétiques de l'adaptation liée à la domestication. L'étude que nous avons menée sur les éléments répétés (ER) du génome de l'igname nous a permis d'identifier une forte corrélation entre la variabilité des abondances relatives d'un grand nombre d'ER et la variabilité climatique. Après la création *de novo* d'une banque d'éléments transposables (ET) de l'igname, nous avons démontré un enrichissement en ET relativement similaire entre les trois espèces étudiées. Nous avons pu identifier quelques ET différentiellement abondants au sein des 180 individus étudiés, montrant des corrélations significatives avec des données bioclimatiques. Notre étude suggère ainsi une association significative entre les composants répétés du génome de l'igname, voire la taille du génome, et l'adaptation à l'environnement. Enfin, nous avons pu faire une première hypothèse quant à l'origine de l'igname cultivée *D. rotundata*, par exploitation des données génomiques disponibles et tests de scénarios de domestication. Cette hypothèse placerait l'origine de l'igname cultivée en région de forêt, à partir de l'espèce *D. praehensilis*. Sans avoir déterminé exactement la localisation probable de la domestication de l'igname, ces résultats remettent en cause l'hypothèse d'une origine stricte de savane pour les espèces cultivées et l'agriculture en Afrique.

Mots-clés : Adaptation – *Dioscorea* spp. – Domestication – Eléments transposables – Gènes sous sélection - Génétique d'association – Génomique des populations – Inférence démographique – NGS – SNP

Abstrat

Yam (*Dioscorea* spp) is a major staple for more than 100 million people in Africa. This species is vegetatively propagated from a relatively limited number of cultivated varieties. Several studies suggest that new varieties could be created using cultivated × wild hybrids or wild plants. However, this requires the characterization of the genetic resources available, as well as the identification of breeding targets for yam adaptation. The main objectives of the present PhD project were to understand the genetic and genomic bases of yam domestication in Africa, and to characterize the genomic determinism of its adaptation to different climatic zones. We investigated the genetic basis of yam domestication in a comparative genomic approach between the cultivated species *D. rotundata* and two wild close relatives *D. praehensilis* and *D. abyssinica*, by exploiting NGS sequencing data and the SNP variants identified. We demonstrated that the starch biosynthesis and storage pathway was selected during yam domestication, similarly to several cereals. More specifically to yam domestication, genes related to tuber morphology or phototropism ability, as well as genes of the NADH dehydrogenase complex were also under selection. With respect to an enrichment in photosynthesis-related functions, selection of the NADH DH complex suggests changes in adaptation with a transfer from shading environments of forest / savannah of wild yams to full sunlight environments in the field specific of cultivated yam. Interestingly, we also detected the same NADH-DH complex in the study that aimed at identifying significant associations between genetic variation and climate variability. Additional analyses are now necessary to confirm the possibility of studying adaptation during domestication through association studies between genes and environment. The study we performed on the repeat elements (REs) of the yam genome highlighted a strong correlation between the variability in relative abundances of numerous REs and climatic variability. We created a *de novo* database of yam transposable elements (TEs) and demonstrated quite similar TE contents for the three species. Nevertheless, we could identify some TEs with differential genomic abundances between the 180 genotypes surveyed, showing significant correlations with bioclimatic variables. Our study thus suggests the significant association of the repeat fraction of the yam genome, and eventually of the genome size, with adaptation to environment. Finally, we were able to make a first hypothesis on the origin of the cultivated yam *D. rotundata,* using the genomic data available to test for various domestication scenarios. Our hypothesis identifies the origin of yam in the forest areas, with the species *D. praehensilis* as the putative progenitor. Even if the precise geographical origin of yam domestication could not be established, our results question the generally admitted hypothesis of savannah origins for crops and agriculture in Africa.

Keywords: Adaptation – Association genetics – *Dioscorea* spp. – Demographic inferring – Domestication – Genes under selection – NGS – Population genomics – SNPs – Transposable elements – Yam

Table des matières

Listes des Figures

Liste des Tableaux

Liste des Abréviations

ACP	Analyse en Composantes Principales
ADP-glucose	Adénosine diphospho-glucose
AFLP	Amplified Fragment-Length Polymorphism
BP	Before Present ; avant le présent
DL	Déséquilibre de Liaison
EMMA	Efficient Mixed Model Analysis
ER	Eléments Répétés
ET	Eléments Transposables
FAO	Food & Agriculture Organisation
FDR	False Discovery Rate
GBS	Genotype By Sequencing
IITA	International Institute of Tropical Agriculture
IRD	Institut de Recherche pour le Développement
JIRCAS	Japan International Research Center for Agricultural Sciences
LFMM	Latent Factor Mixed Models
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MKT	Test de McDonald-Kreitman
NADPH	Nicotinamide Adénine Dinucléotide Phosphate
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
PROG1	*Prostrate growth 1*
QTL	Quantitative Trait Loci
SFS	Spectre de Fréquence
SINE	Short Interspersed Nuclear Elements
SNP	Single Nucleotide Polymorphism
SPS	Sucrose-Phosphate Synthase
SUS	Sucrose Synthase
tb1	Teosinte Branched 1
TRIM	Terminal-Repeat Retrotransposons in Miniature

Introduction générale

Les hypothèses sur l'origine de la vie sur Terre supposent l'existence de molécules d'ARN capables de s'auto-répliquer (Haldane 1965). Ces molécules auraient ainsi été porteuses d'une information génétique transmissible. La plupart des organismes supérieurs possèdent aujourd'hui une molécule d'ADN comme support de l'hérédité (Griffith 1928). Si cette simple phrase semble banale, elle est quand même le résultat de plus d'une centaine d'années de recherche. Même sans une idée encore très claire des bases de l'hérédité, Darwin (1859) a pu formuler une théorie générale de l'évolution dans un ouvrage intitulé « L'origine des espèces », et dont le premier chapitre fut dédié à l'étude de la domestication des plantes et des animaux. Plus tard en 1868, Darwin y consacrera tout un ouvrage intitulé « The variation of Animals and Plants under domestication » (Darwin et Gray 1868).

L'étude de l'origine des espèces et de l'agriculture est devenue une discipline dont les précurseurs sont Alphonse De Candolle (1806-1893) ; Nikolaï Ivanovitch Vavilov (1887-1943) et Jack Harlan (1917-1998). Leurs recherches et observations ont permis de distinguer au moins sept régions d'origine de l'agriculture à travers le monde. Toutes ces régions ont fait objet d'études approfondies sauf la région de l'Afrique qui demeurent encore peu ou mal connue. On pense que l'origine de l'agriculture en Afrique se situerait dans les régions de savanes chaudes (Harlan 1976), et des restes archéologiques des principales céréales d'origine africaine que sont le mil et le sorgho, ont été découverts dans cette région (Oumar *et al.* 2008; Winchell *et al.* 2017).

Le but de la présente thèse est de comprendre le processus de domestication d'une plante non-modèle d'origine africaine, l'igname, une plante à racine et tubercule du genre *Dioscorea*. Les plantes à racines et tubercules en général et l'igname en particulier, sont multipliées végétativement par l'homme. Les variétés sont donc propagées à l'identique d'une année à l'autre. Cependant, ce système de reproduction n'est pas totalement fermé, et, des études ont montré que les populations sauvages sont utilisées par les agriculteurs pour développer de nouvelles accessions cultivées (Wilkes 1977; Manu-Aduening *et al.* 2005; Scarcelli *et al.* 2006). Chez l'igname, ces études soulèvent de nombreuses questions sur le rôle des populations sauvages lors de la dispersion de la culture en zones sèche et humide. En effet, l'espèce principalement cultivée en Afrique, *Dioscorea rotundata,* possède deux parents sauvages présumés : *D. abyssinica* et *D. praehensilis,* respectivement inféodées aux milieux de savane et de forêts (Hamon *et al.*1995*)*. Des flux de gènes entre les trois espèces ont été démontrés (Scarcelli *et al.* 2006). La question qui se pose est de savoir si ces flux de gènes n'ont pas favorisé dans le passé, des adaptations des ignames des forêts humides aux zones plus sèches de savane et vice-versa.

Une autre question que posent ces échanges de gènes cultivés/sauvages est la nature génétique du caractère « domestiqué » chez l'igname. En d'autres termes, une question toujours d'actualité est de conclure quant à l'existence d'un syndrome de domestication pour cette espèce cultivée. Le syndrome de domestication, associé à des modifications génétiques majeures, reflète des changements de traits d'histoire de vie pour une meilleure valeur sélective (*fitness*) des individus en réponse à la sélection exercée par l'Homme (Harlan 1971). Le résultat de cette sélection humaine, même inconsciente, s'oppose généralement à ceux issus de la sélection naturelle que connaissent les individus au sein du compartiment sauvage. Le système de reproduction, souvent de nature multiple chez les individus sauvages, est ainsi particulièrement ciblé par l'homme chez les individus cultivés en lien avec une stratégie de propagation du matériel à cultiver (Meyer *et al.* 2012). Ceci aboutit généralement à un niveau important d'isolement reproducteur entre sauvages et cultivés. Dans le cas de l'igname, la persistance des échanges entre espèces cultivées et sauvages pose question, et il serait envisageable que ces phénomènes d'échange cultivés/sauvages se soient au contraire renforcés avec la transition d'un système de reproduction d'allofécondation vers le système majoritaire actuel de multiplication végétative mis en place par les agriculteurs. Il reste aujourd'hui largement méconnu, quand et comment ces transitions reproductives associées à la sélection humaine ont eu lieu.

Afin de répondre à ces interrogations, deux principaux axes de recherche ont été abordés dans le cadre de mon travail de thèse de doctorat : l'axe 1 a porté sur la compréhension des bases génétiques de la domestication chez l'igname et l'axe 2 s'est focalisé sur l'analyse des mécanismes génétiques de l'adaptation des ignames. Le présent manuscrit est structuré en six chapitres.

Le premier chapitre présente une synthèse bibliographique et méthodologique sur l'histoire de la domestication des plantes. Les quatre chapitres qui suivent présentent différentes approches d'étude de la domestication sur mon modèle d'étude. Les chapitres 2 et 3, en lien avec l'axe 1 de mon travail de thèse, présentent : une approche de détection de sélection pour identifier les bases moléculaires de la domestication chez l'igname (chapitre 2) ; et une approche de modélisation démographique pour déterminer l'origine biologique de l'igname cultivée (chapitre 3). Les chapitres 4 et 5 se rapportent au deuxième axe de mon travail et abordent la question de la domestication comme processus d'adaptation, par des approches de génétique d'association. Le chapitre 4 présente une étude menée sur l'implication des éléments transposables du génome dans le processus d'adaptation de l'igname aux variations climatiques. Le chapitre 5 présente la détection de polymorphismes génétiques associés aux variations environnementales. Enfin, le chapitre 6 concerne la discussion générale de mes travaux sur l'igname pour décrire l'histoire évolutive des populations d'ignames d'Afrique et apporte quelques perspectives de poursuite de des recherches sur cette étude.

Chapitre 1. La domestication: définition, origines et méthodes d'étude

La domestication des plantes est vue comme « *un processus historique, évolutif et continu, sous l'impulsion de la volonté humaine, qui produit une fixation d'allèles, qui fournit des plantes alimentaires sauvages et cultivées à caractères favorables pour l'homme, mais qui diminue ou élimine la capacité de survie dans les conditions naturelles, rendant une population domestiquée dépendante de l'homme* » (Harlan 1992).

1. Histoire de la domestication

La domestication peut être considérée comme un processus rapide d'évolution par sélection récurrente sous l'action humaine. Les modifications phénotypiques subordonnées à cette évolution entraînent des modifications génétiques importantes, transmissibles de génération en génération. Cela aboutit à la fixation de caractères répondant à l'intervention de l'Homme et aux critères de sélection de ce dernier, dans un but de satisfaire ses besoins d'alimentation, de protection, de transport, de compagnie, etc. Ces changements vont jusqu'à la création de formes très différentes des plantes et des animaux sauvages (Darwin 1859). La domestication constitue donc un excellent exemple de modification majeure de la morphologie (Gracia 2014) et illustre très bien le processus d'évolution des espèces. De plus, la domestication représente une étape déterminante de l'agriculture. La domestication des espèces végétales et animales, associée à l'émergence des sociétés agricoles constitue un des tournants décisifs de l'histoire humaine d'un point de vue alimentaire, démographique, comportemental, économique et socio-culturel (Diamond 2002). L'étude du processus de la domestication des espèces végétales contribue aujourd'hui, à appréhender comment une plante sauvage à production relativement modeste, a pu être transformée en plante cultivée inféodée au champ.

1.1. Les principaux foyers de domestication dans le monde

L'origine de l'agriculture remonte à environ 10 000 ans BP (BP : Before Present ; avant le présent) (Diamond 2002). Cependant, les idées demeurent toujours divergentes dans le rang des anthropologues et des historiens, en ce qui concerne les principales raisons de la domestication des premiers animaux par l'homme (revue dans Mazoyer et Roudart 2002). Toutefois, la plupart de ceux-ci reconnaissent que la motivation des premiers éleveurs ne se résume pas au seul besoin d'alimentation, d'autant plus que la chasse fournissait alors suffisamment de viandes. Certains pensent que le début de la sédentarisation des hommes serait à l'origine de la domestication animale (Driscoll *et al.* 2009). Pour d'autres, cette domestication serait liée à des croyances selon lesquelles l'Homme soumis à la domination des dieux, aurait pris le pouvoir sur l'animal (Grosseteste 1997). Dans l'un ou l'autre de ces cas, les recherches actuelles présentent

le chien comme étant le premier animal à avoir été domestiqué entre 10 000 et 15 000 ans BP en Asie de l'Ouest (Israel, Iraq) (Germonpré *et al.* 2012). Le chien est considéré comme un animal de protection, également de compagnie mais aussi comme un auxiliaire pour la chasse (Trut *et al.* 2009). L'élevage des caprins au Moyen-Orient, des bovins en Inde, des ovins et des porcins aurait débuté il y a environ 11000 ans BP. L'aquaculture serait quant à elle apparue il ya 4 000 ans BP en Égypte et en Chine.

En ce qui concerne les plantes, les réflexions concernant l'origine des espèces cultivées ont commencé dans les années 1800. C'est Alphonse De Candolle (1806-1893), un botaniste phytogéographe suisse, qui a été le premier à proposer des centres d'origine des plantes cultivées. Dans son ouvrage intitulé «*Origine des plantes cultivées*» (1886), Alphonse De Candolle a tenté d'énumérer les lieux géographiques probables d'origine de diverses espèces cultivées à travers le monde sur la base d'observations botaniques, historiques, archéologiques et linguistiques. Par la suite, Nikolaï Ivanovitch Vavilov (1887-1943), un botaniste et généticien russe, a exploré le monde entier pour comprendre la diversité des plantes sauvages et cultivées. Il fut le premier scientifique à mener de si nombreuses campagnes de prospection à travers les continents (Demol 2002). Les observations de Vavilov lui ont permis de comprendre que les espèces cultivées ainsi que leurs effectifs respectifs n'étaient pas uniformément distribués dans le monde (Demol 2002). Les zones de distribution des espèces observées par Vavilov l'ont amené à caractériser les différentes régions du globe en fonction de leur biodiversité végétale. Ainsi, le botaniste retient qu'il existe des zones précises, plus diverses que d'autres, et pour lui, ces zones correspondraient aux centres d'origine des espèces cultivées et aussi de l'agriculture (Demol 2002). Ces centres d'origine seraient caractérisés par une hétérogénéité marquée, une diversité d'espèces cultivées et sauvages importante voire maximale, et l'existence de processus d'introgression matérialisée par la présence d'individus hybrides. Sur cette base, en 1926, Valvilov a répertorié au moins huit grandes régions distinctes de l'Ancien et du Nouveau monde qui constitueraient des centres indépendants d'origine des espèces cultivées et donc de l'agriculture (Vavilov et Dorofeev 1992) (Figure 1). Plus tard, l'agronome américain Jack Harlan a complété la théorie de l'origine des espèces cultivées de Vavilov par sa théorie des centres et non-centres[1] (Harlan 1971) : il a ainsi proposé, notamment un «non-centre» d'Amérique latine comprenant les Andes, l'Amazonie et la région du Venezuela et des Guyanes, d'où proviendraient de nombreuses espèces domestiquées (Métailié 2016).

[1] Un non-centre pour Harlan est une zone d'origine de plantes cultivées dont l'étendu est tellement vaste que le terme de « centre » n'apparaît plus approprié.

Il est à présent clair que la domestication des plantes s'est faite de manière indépendante les unes des autres dans le monde (Diamond 2002). Au total, six principaux foyers de domestication ont été décrits (Figure 2).

Si des études ont permis de mieux comprendre l'évolution qu'ont connue la plupart de ces foyers, le foyer africain fait encore objet de discussions. En partculier, l'identification de l'Abyssinie, région savanicole relativement chaude, comme centre de domestication, reste à démontrer dans la mesure où de récents développements scientifiques ont identifié l'origine de certaines plantes en zones forestières (Scarcelli *et al.* en Préparation).

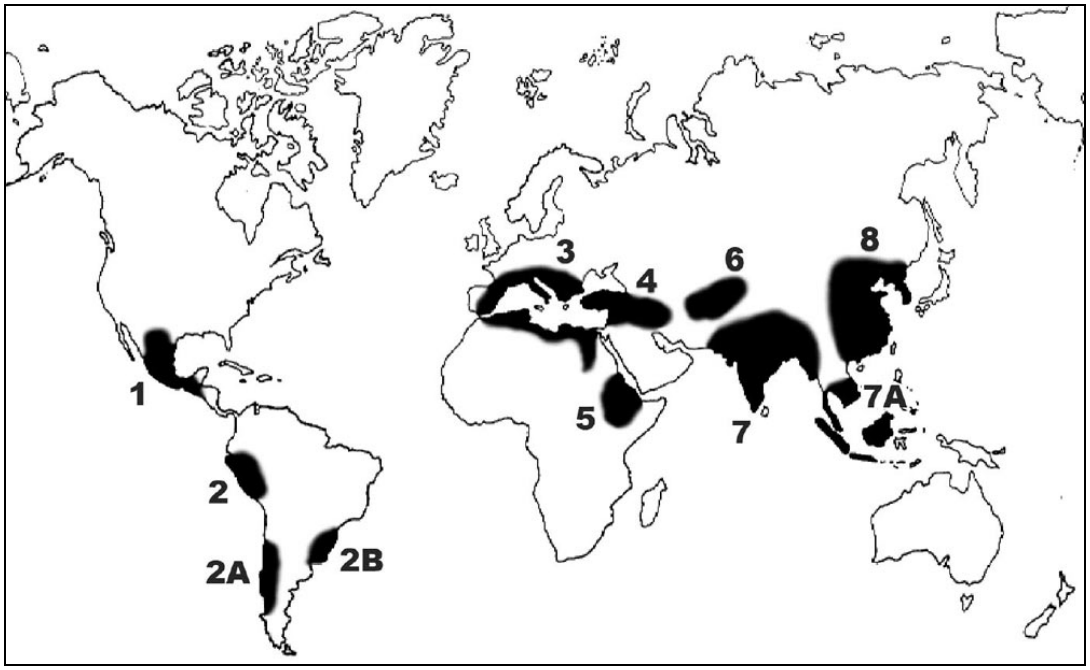


Figure 1. Centres de domestication indépendants dans le monde selon Vavilov.
(1) Mexique-Guatemala, (2) Pérou-Equateur-Bolivie, (2A) Sud Chili, (2B) Sud Brésil, (3) Bassin Méditerranéen, (4) Moyen-Orient, (5) Ethiopie, (6) Asie Centrale, (7) Indo-birmanie, (7A) Siam-Malaisie-Java, (8) Chine-Corée. (Adaptée de Vavilov (1951) par R. W. Schery, Plants for Man, Prentice Hall, Englewood Cliffs, NJ, 1972).

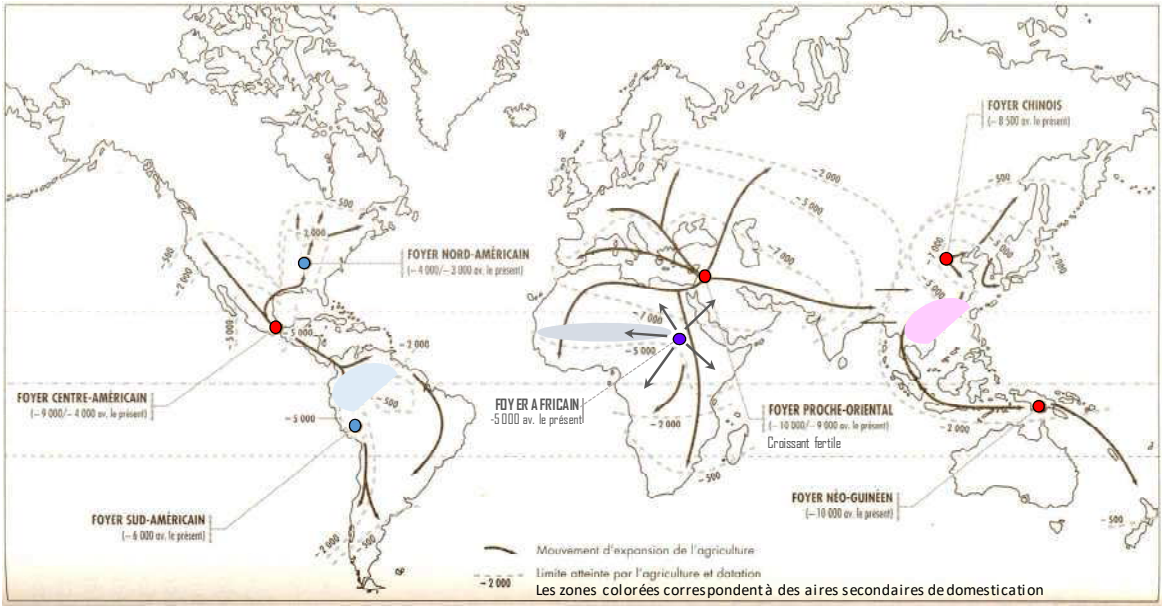


Figure 2. Foyers d'origine de l'agriculture associés aux principaux centres de domestication des plantes cultivées et aires secondaires de domestication
(Adaptée de Mazoyer & Roudart, 2002).

1.2. Le processus de domestication en Afrique

Le centre de domestication africain a été moins étudié que les autres en raison des conditions climatiques qui limitent la conservation des fossiles. Pourtant, ce centre de domestication présente des particularités qu'il est important de souligner. Selon une étude sur le sujet, les modèles de changements économiques associés à la mise en place de l'agriculture en Afrique diffèreraient de ceux de la plupart des autres continents, en ce sens que l'élevage aurait précédé la culture des plantes (Fuller et Hildebrand 2013). On pense que la domestication de plantes indigènes se serait produite assez tardivement au cours des 5 000 dernières années (Marshall et Hildebrand 2002). Les « foyers d'origine de l'agriculture » suspectés en Afrique seraient situés dans les zones de savane (Harlan 1976) de l'empire d'Ethiopie, anciennement connues sous le nom de l'Abyssinie. Ces régions englobent l'actuelle Ethiopie, l'Est du Soudan et l'Érythrée. C'est après domestication, que les plantes cultivées auraient par la suite colonisé les zones de forêt. Cette hypothèse est soutenue par le fait que les formes sauvages des principales céréales africaines, à savoir le mil (*Pennisetum glaucum*) et le sorgho (*Sorghum bicolour*) sont originaires des zones de savane d'Afrique (Harlan 1976; Portères 1976; Wetterstrom 1998; Oumar *et al.* 2008). Si l'hypothèse de l'origine de la domestication du mil en Afrique a fait l'unanimité, cela n'a pas été le cas avec le sorgho. En effet, en 1989, Haardland a émis l'hypothèse selon laquelle le sorgho sauvage aurait été exporté vers l'Inde, où il aurait été domestiqué avant d'être ensuite réintroduit en Afrique. Cette hypothèse s'oppose cependant à certaines hypothèses issues de différents résultats de recherche. Par exemple, il a été établi que

le sorgho cultivé proviendrait de l'espèce sauvage *S. bicolor arundinaceum* originaire d'Abyssinie (de Wet 1978). Cette hypothèse est notamment en accord avec les résultats d'une étude comparative menée sur les chloroplastes de la forme cultivée et des formes sauvages du sorgho, mettant en évidence une similarité plus marquée avec le groupe génétique sauvage de l'Afrique du Centre-Nord (Aldrich et Doebley 1992). Les restes archéologiques les plus anciens du sorgho ont ensuite été découverts près de Kassala dans l'Est du Soudan, datés de 5500 BP, permettant de corroborer cette hypothèse (Winchell *et al.* 2017). De la même manière, l'origine africaine et savanicole du mil a pu être confirmée par la découverte du plus ancien reste archéobotanique à ce jour, dans le nord du Mali et daté à 4500 BP (Manning *et al.* 2011).

2. Domestication et diffusion post-domestication comme un processus d'adaptation récente

2.1. Définition de l'adaptation

L'adaptation chez les plantes s'effectue dès que la constitution génétique de la population change sous l'action de la sélection naturelle (Merilä et Hendry 2014). Les facteurs de sélection qui entaînent l'adaptation peuvent être de nature biotique, par exemple en réponse à une pression parasitaire (Gladieux *et al.* 2010) ou abiotique, par exemple en réponse à une faible intensité lumineuse (Briggs et Christie 2002; Quiles et López 2004). Dès lors que l'adaptation a lieu, elle s'exprime par une différenciation génétique entre les populations présentes dans différents environnements. Des études de génétique d'association permettent de mettre en évidence ces adaptations, qu'elles soient d'origines biotiques ou abiotiques comme le sont les contraintesclimatiques (Louthan et Kay 2011; Frichot *et al.* 2013; Franks *et al.* 2014). De plus en plus d'algorithmes basés sur la génétique des populations, la modélisation écologique et les statistiques sont développés pour cribler les génomes entiers afin de rechercher de signatures génétiques d'adaptation (Leimu et Fischer 2008; Hancock *et al.* 2011; De Mita *et al.* 2013; Frichot *et al.* 2013; Luu *et al.* 2017). Ces méthodes se fondent sur la différentiation génétique entre populations, ainsi que l'existence de corrélations génétiques entre les génotypes des populations avec des variables d'intérêt. Par exemple, une étude sur *Arabidopsis thaliana* a identifié des loci fortement corrélés au climat ; parmi eux, certains sont particulièrement enrichis en variants correspondant à des mutations sélectives dans les gènes (avec changement d'acides aminés de la protéine associée) (Hancock *et al.* 2011). Plus récemment, une étude sur la téosinte a révélé une colocalisation fréquente entre des régions génétiques candidates et des loci impliqués dans la variation des caractères associés aux interactions plante-sol, telles que la morphologie racinaire ou la tolérance à une forte teneur en aluminium ou au phosphore (Fustier *et al.* 2017).

Nous tenons à insister sur le fait que si la domestication correspond à un processus

d'adaptation, l'adaptation représente un processus bien plus général, car toutes les adaptations ne sont pas liées à la domestication. Une illustration de cet état de fait est bien documentée chez le maïs (Fustier *et al.* 2017; Li *et al.* 2017). En effet, le maïs a été domestiqué au Mexique (première phase d'adaptation de la nouvelle espèce ''maïs''). Il a ensuite diffusé vers d'autres régions, où l'espèce s'est adaptée progressivement à des contraintes environnementales locales (adaptations multiples de l'espèce ''maïs'' à des contraintes variées) sans avoir besoin d'une autre forme de domestication (Tenaillon et Charcosset 2011).

2.2. Des changements phénotypiques majeurs

La réponse aux modifications génétiques modelées par l'action humaine peut se traduire par des changements phénotypiques chez les plantes (Allaby 2014). L'ensemble de ces modifications morphologiques liées à la domestication constituent ce qu'il convient d'appeler le syndrome de domestication (Hammer 1984). Chez les plantes, les changements phénotypiques ont lieu généralement au niveau de l'architecture, de la morphologie et de la qualité des produits de récoltes, mais aussi au niveau du système de reproduction (Salamini *et al.* 2002). Des études ont révélé que ces caractères qui émergent chez les formes domestiquées sont le résultat d'une sélection lors de la domestication (Doebley *et al.* 2006; Sang et Li 2013). Nous présentons ici des exemples de modifications qui ont lieu au niveau de l'architecture de la plante et des organes de stockage, en nous basant sur les connaissances validées chez les céréales.

2.2.1. L'Architecture de la plante

Les modifications de la structure du système végétatif chez les graminées passent par la réduction du nombre de talles ou de branches. La mise en culture des plantes peut entrainer un phénomène de compétition entre les pieds, pour l'accès à la lumière par exemple, ce qui sélectionne des talles plus allongées (Sang et Li 2013). La figure 3 illustre la réduction drastique du nombre de talles entre le maïs et son ancêtre la téosinte au cours de la domestication. D'ailleurs, le maïs représente l'espèce qui a subi les changements les plus remarquables lors de la domestication (Sang et Li 2013). Des études génétiques ont démontré que la réduction du nombre de talles chez le maïs est sous contrôle d'un gène appelé le *tb1* (teosinte branched1) (Doebley *et al.* 1995; Wang *et al.* 1999). Le gène *tb1* induit une dominance apicale chez le maïs, en empêchant l'excroissance des méristèmes axillaires et l'élongation des ramifications (Doebley *et al.* 1997, 2006). Il a aussi été démontré que la modification de l'architecture de la plante chez le riz est sous le contrôle d'un gène : le *PROG1* (*Prostrate growth 1)* (Jin *et al.* 2008). Le gène *PROG1* code pour un facteur de transcription à doigts de zinc et est exprimé de façon prédominante dans les méristèmes axillaires où se forment les bourgeons. Chez le blé, un gène à effet pléiotropique, le gène *Q*, a été également identifié comme impliqué dans le contrôle de la structuration de l'architecture des plantes (Simons *et al.* 2006).

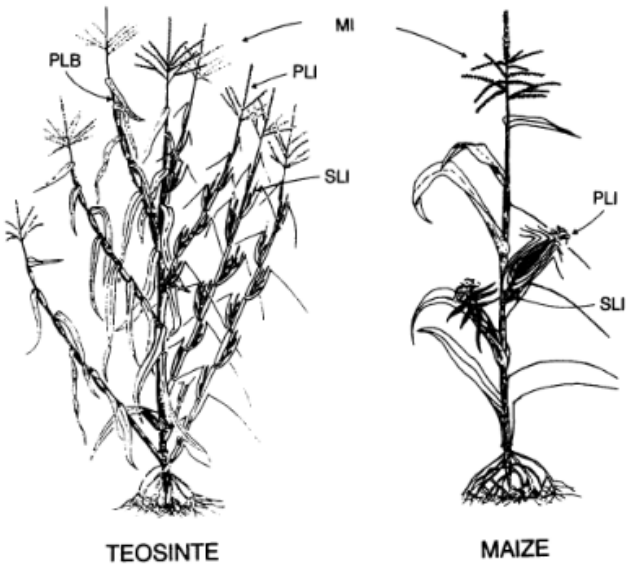


Figure 3. Différences morphologiques entre l'architecture de la plante chez la téosinte et le maïs.
On observe que l'architecture est plus ramifiée chez la téosinte que chez le maïs. (Doebley 1992).

2.2.2. Morphologie et qualité des organes de stockage

Les graminées sauvages évoluent sous une forte sélection pour leur capacité à disperser leurs graines à la maturité. Au contraire, la domestication a favorisé l'obtention de plantes à épis compacts jusqu'à maturité, de plantes à rachis soudés et solides et avec une synchronisation de la maturité des graines (Olsen 2012).

La base moléculaire de ces sélections a été démontrée chez plusieurs céréales. C'est le cas chez le sorgho (*Sorghum bicolor*) avec le gène *Shattering1* (*Sh1*) qui contrôle le caractère de la dispersion des graines. Durant la domestication du sorgho, le locus *Sh1* a subi trois mutations différentes impliquant une contre-sélection de l'égrenage chez la forme cultivée (Olsen 2012; Yeh *et al.* 2012). Chez le riz asiatique (*Oriza sativa*), des études de génétique quantitative ont aussi mis en évidence des signatures de sélection liées à la domestication pour le caractère de la dispersion des graines (Lam *et al.* 2010; He *et al.* 2011). Chez le maïs (*Zea mays mays*), des orthologues de *Sh1* sont co-localisés au niveau de deux QTLs associés au caractère de la dispersion des graines au sein d'une population hybride maïs-téosinte (Olsen 2012). Une autre étude a révélé une faible diversité à un locus correspondant au gène *Si037789m* responsable de la dispersion des graines chez le petit mil (*Foxtail millet*) (Jia *et al.* 2013). La région génomique de ce gène est synténique à celle du gène *Sh1,* confirmant avec les cas du riz, du maïs et du sorgho, l'assertion d'une sélection parallèle et indépendante sur le même gène pour un même

caractère de domestication (dispersion des graines) chez les céréales (Olsen 2012; Yeh *et al.* 2012).

Il a également été démontré que la domestication des plantes est associée à l'augmentation de la quantité d'amidon dans les graines au détriment de la teneur en protéines (Campbell *et al.* 2016). Les céréales contribuent substantiellement à l'alimentation humaine par leur richesse en amidon. Cet amidon est stocké dans les graines et présente un intérêt pour le choix des génotypes depuis l'avènement de l'agriculture. Chez le sorgho par exemple, il a été démontré que des gènes impliqués dans la voie d'initiation de la biosynthèse de l'amidon, présentaient une importante réduction de la diversité nucléotidique chez les variétés locales cultivées comparées aux formes sauvages (Campbell *et al.* 2016). Chez le maïs, l'accroissement de la synthèse du saccharose serait sous le contrôle d'un gène associé à la domestication, le gène *SUS,* qui favorise l'augmentation du taux d'amidon et d'une enzyme régulatrice de la production d'amidon (ADP-glucose : l'Adénosine diphospho-glucose) (Li *et al.* 2013).

Ces deux exemples de sélection des caractères « non dispersion des graines » et « biosynthèse de l'amidon » confirment la définition du « syndrome de domestication ». Des caractères importants (voire nécessaires) à la mise en culture de plantes sauvages et donc à la domestication ont abouti à des presssions de sélection identiques exercées par l'Homme pour amener à un idéotype commun. Les modifications génétiques qui ont accompagné cette réponse à la sélection ont concerné souvent des loci orthologues, quand on compare le processus entre espèces (comme illustré ci-dessus). Cependant, ce n'est pas systématique. Par exemple, le gène *SH4* qui code pour la protéine d'égrenage chez le riz a été sélectionné au cours de la domestication chez le riz africain *Oryza glaberrima*, mais ne l'a pas été chez le riz asiatique *Oryza sativa* (Cubry et all., soumis). Ces types de sélection peuvent également avoir lieu chez des espèces plus éloignées, avec des systèmes de reproduction et de stockage différents.

Chez les plantes à racines et tubercules, c'est en particulier cette capacité à produire de grosses racines ou tubercules riches en amidon qui présente un intérêt agronomique. Chez la pomme de terre, une approche comparative de la diversité génomique entre une population sauvage de l'Amérique du sud et une population constituée de variétés locales nord-américaines, a été utilisée pour détecter un gène sous séletion, codant pour la Sucrose-Phosphate Synthase (*SPS*) (Hardigan *et al.* 2017).Cette étude a révélé que le gène *SPS* est en lien avec la domestication, et impliqué dans le métabolisme glucidique via la mobilisation et le transport de l'amidon synthétisé au niveau des feuilles et mobilisé vers les tubercules. Or, le rôle prépondérant et déterminant du saccharose dans la grosseur des tubercules chez la pomme de terre cultivée avait été mis en évidence des années plus tôt (Fernie et Willmitzer 2001). Chez le manioc (*Manihot esculenta*), des gènes impliqués dans le transport et la synthèse du saccharose, et la

production d'enzymes de ramification de l'amidon ont été identifiés comme des gènes clés fortement associés au développement des racines de stockage (Wang *et al.* 2014b). Macko-Podgórni *et al.* (2017) ont, quant à eux, démontré que le gène *DcAHLc1,* impliqué dans le développement racinaire chez la carotte (*Daucus carota*), présentait une faible diversité dans un pool cultivé comparé à une population sauvage. La différenciation génétique entre les carottes sauvages et cultivées est plus élevée au niveau du gène *DcAHLc1* par rapport à la moyenne du génome entier, signature d'une sélection liée à la domestication.

2.3. Méthodes de détection de signature de sélection chez les plantes cultivées

Il existe différentes formes décrites de sélection. La sélection balancée caractérise le maintien des allèles multiples dans un pool de gènes d'une population (Charlesworth 2006). La sélection négative ou sélection purificatrice caractérise une purge des mutations plus ou moins délétères (Page et Holmes 1998). Enfin, la sélection positive ou sélection directionnelle caractérise une augmentation en fréquence jusqu'à fixation d'allèles bénéfiques dans une population concernée (Page et Holmes 1998). Ce dernier type de sélection, si il intervient sur une nouvelle mutation entraine des balayages sélectifs forts correspondent à la signature génomique d'une sélection positive sur un allèle avantageux (Smith et Haigh 1974; Wallace 1975). Ainsi cette signature de sélection s'étend aux variations neutres dans des régions génomiques voisines de cet allèle favorisé. Ces polymorphismes ont été « entrainés » au cours de la fixation de l'allèle bénéfique (Figure 4-a et Figure 4-b). Le balayage sélectif engendre (1) la réduction voire l'élimination du polymorphisme dans des sites neutres génétiquement liés à l'allèle sélectionné (Nielsen 2005) ; et (2) l'augmentation du déséquilibre de liaison (DL). La réduction de la diversité génétique de part et d'autre de la mutation positivement sélectionnée peut être détectée au travers d'un enrichissement des variants de fréquences extrêmes dans le spectre de fréquence alléliques. Cet effet du balayage sélectif diminue au fur et à mesure que la distance génétique au locus sélectionné s'accroit.

Les méthodes d'identification de ces balayages sélectifs recourent donc à l'analyse de la diversité génétique et/ou du déséquilibre de liaison (LD) (Wright *et al.* 2005). L'histoire (neutre) des populations peut parfois ressembler à ces balayages sélectifs. Aussi, la prise en compte de l'histoire démographique des populations permet une meilleure caractérisation de la diversité génétique neutre et un accroissement du pouvoir d'identification des gènes sélectionnés au cours du processus de domestication ou lors des événements de sélection ultérieurs (Vigouroux *et al.* 2002; Wright *et al.* 2005). La domestication est souvent associée à un goulot d'étranglement ou « bottleneck » de la diversité sans doute associé à une taille efficace réduite lors de la domestication (Doebley 1989). La figure 5 présente une illustration du principe des méthodes basées sur la diversité. Le tableau 1 présente une liste non exhaustive des méthodes de détections de sélection existantes. Plus de détails dans le Chapitre 2 et dans la revue de (Vitti *et al.* (2013).

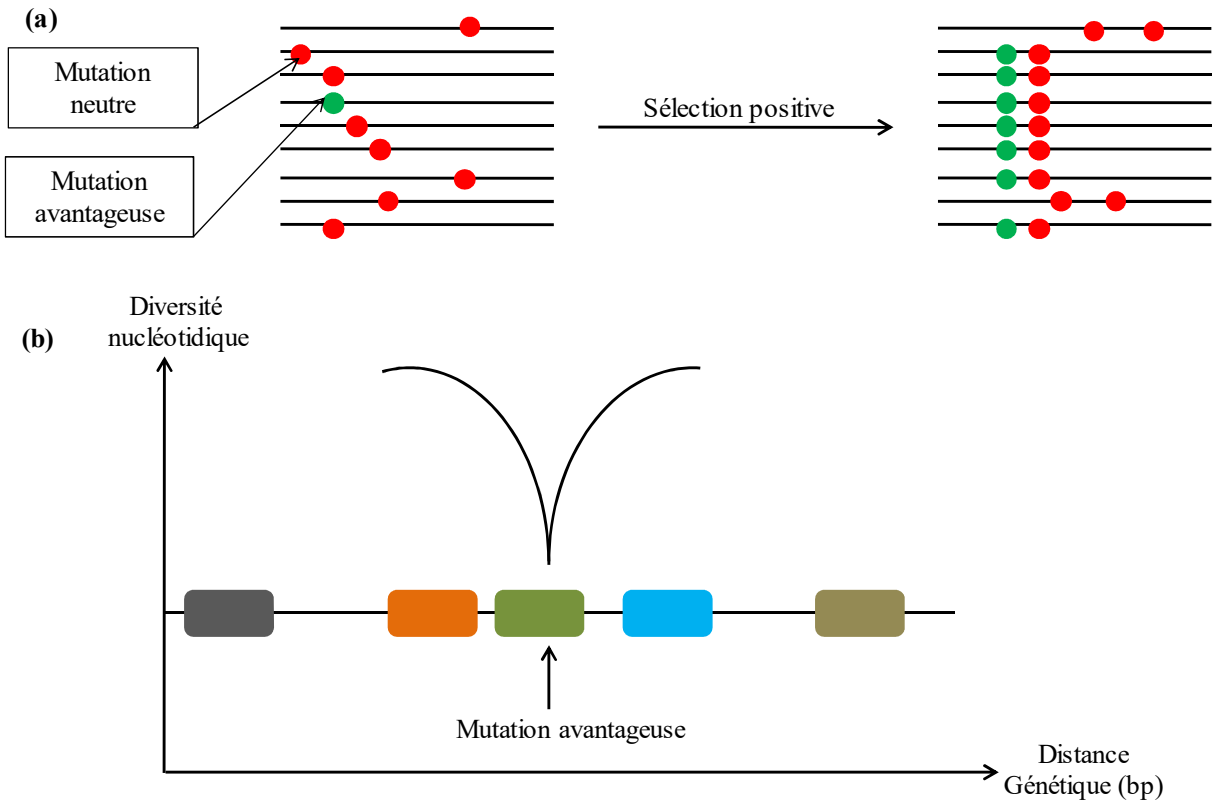


Figure 4. Illustration du balayage sélectif.
(a) Fixation d'un allèle avantageux et d'un allèle neutre physiquement liés entre eux, (b) Décroissance du balayage sélectif en fonction de la distance : la diversité nucléotidique est réduite au maximum à l'endroit de la mutation avantageuse, (c) Impact du balayage sélectif sur différentes statistiques génétiques (D de Tajima, déséquilibre de liaison et nombre de sites en ségrégation) de part et d'autre de la mutation avantageuse (Nielsen 2005).

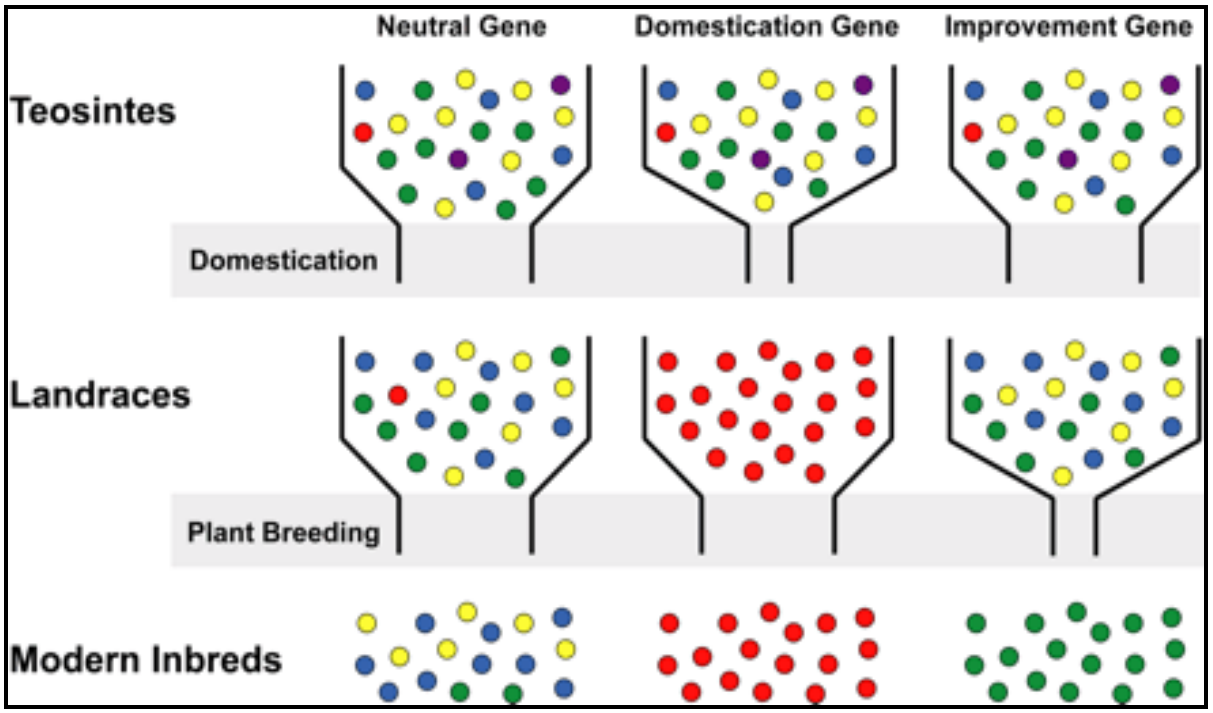


Figure 5. Mise en évidence des effets de la domestication sur la diversité génétique des populations.
(Yamasaki *et al.* 2005).

Tableau 1. Synthèse de quelques approches de détection de sélection

Méthodes	**Tests**	**Principes**	**Références**
Méthode d'analyse des codons	d_N/d_S	Compare le taux de substitutions non-synonymes (d_N) d'une séquence codante au taux de substitutions synonymes (d_S) de la même séquence d'une espèce par rapport à une référence.	Hurst (2002) Graur et Li (2000)
	Test de McDonald-Kreitman (MKT)	Compare d_N/d_S interspécifique et intra-spécifique.	
Méthode d'analyse de la diversité génétique	Test D de Tajima	Un excès d'allèles rares en comparaison du nombre de différences génétiques moyennes deux à deux entre individus au nombre total de sites polymorphes en ségrégation. Un D de Tajima négatif révèle un excès d'allèles rares, ce qui peut être causé par une expansion démographique ou une sélection positive.	(Tajima 1983, 1989) (Fu et Li 1993; Fu 1997)
	Différentiation Fst	Compare la différenciation entre populations en tenant en compte la variabilité génétique intra- et inter-population. On attend qu'un gène sélectionné au cours de la domestication présente une différenciation extrême entre populations sauvages et populations cultivées.	Holsinger et Weir (2009)

Tableau adapté de Vitti *et al.* (2013).

2.4. Exemple de domestication suivie d'adaptation à de nouveaux environnement: cas du maïs (*Zea may mays*)

Le maïs cultivé (*Zea mays*) est la première céréale cultivée dans le monde devant le blé et le riz. En Afrique, il représente environ 40% de la production céréalière. Des restes archéologiques et des données génétiques ont mis en évidence un événement unique de domestication (Matsuoka *et al.* 2002) il y a 9 000 ans dans la vallée de la rivière Balsas au Mexique (Piperno *et al.* 2009; Ranere *et al.* 2009). Le maïs est le produit de la domestication de la sous-espèce de téosinte *Zea mays spp. parviglumis* (Matsuoka *et al.* 2002), forme sauvage endémique des régions de moyenne à basse terre du sud-ouest du Mexique. Deux autres téosintes annuelles sont classées dans la même espèce *Zea mays* : il s'agit des sous-espèces *Zea mays spp. mexicana* et *Zea mays spp. huehuetenangensis* (Iltis et Doebley 1980).

Les téosintes mexicaines *parviglumis* et *mexicana* sont inféodées à deux niches écologiques distinctes avec une structuration spatiale inter- et intra-population marquée (Ross-Ibarra *et al.* 2009) présentant des aires de répartition en mosaïque (Figure 6). Les téosintes *mexicana* occupent les environnements semi-tempérés arides, aux altitudes sèches et fraîches comme au nord et au centre du Mexique (1600-2700 m) tandis que les *parviglumis* sont inféodées aux milieux tropicaux, aux altitudes moyennement plus chaudes comme dans le sud-ouest du Mexique (<1800 m) (Hufford *et al.* 2012). La séparation spatiale des deux sous-espèces est une conséquence d'une divergence intervenue entre *mexicana* et *parviglumis* il y a environ 50 000 ans (Ross-Ibarra *et al.* 2009). La sous-espèce *Zea mays mexicana* a pu coloniser les hauts-plateaux mexicains plus froids grâce à des événements d'adaptation locale (Yu et Buckler 2006; Pyhäjärvi *et al.* 2013; Fustier *et al.* 2017).

Des études comparatives de diversité révèlent que le maïs cultivé présente une diversité génétique plus faible que les téosintes, avec une perte d'environ 25% de la diversité présente chez les téosintes actuelles (Eyre-Walker *et al.* 1998). Dans une étude menée sur 774 gènes, l'analyse de la diversité nucléotidique indique que 2 à 4% de ces gènes ont subi une sélection artificielle (Wright *et al.* 2005). Ces gènes sélectionnés sont impliqués dans des fonctions de croissance des plantes et sont localisés près des loci impliqués dans la variation de caractères quantitatifs (QTL) qui contribuent aux différences phénotypiques entre le maïs et la téosinte. Pour ces auteurs, les gènes non sélectionnés témoignent d'un goulot d'étranglement associé à la domestication. Les maïs ont en effet subi deux goulots d'étranglement distincts lors de la domestication et plus récemment, lors de l'amélioration variétale (Fustier 2016). La domestication et l'amélioration ont engendré la sélection d'allèles spécifiques de gènes contrôlant des caractères morphologiques et agronomiques clés, avec comme conséquence une réduction de la diversité génétique par rapport à des gènes non sélectionnés (Yamasaki *et al.*

2007). Il a été démontré que les flux de gènes ne se sont pas interrompus entre le maïs et ses deux apparentés sauvages *parviglumis* et *mexicana* (Ross-Ibarra *et al.* 2009). Les introgressions de *mexicana* dans le maïs sont plus importantes que celle de *parviglumis* (van Heerwaarden *et al.* 2011). Les flux de gènes de la forme sauvage *mexicana* vers le maïs cultivé ont été à la base de l'adaptation du maïs aux hautes altitudes (Takuno *et al.* 2015).

Si l'hybridation entre les maïs cultivés et la téosinte *mexicana* a favorisé le maintien des capacités adaptatives des maïs à des environnements nouveaux, ce ne serait pas le cas pour des gènes majeurs de différenciation entre espèces cultivée et sauvage tel que le gène *tb1* (Doebley *et al.* 1995). Pour ces gènes leur diversité est davantage réduite, du fait de la sélection artificielle (Yamasaki *et al.* 2007).

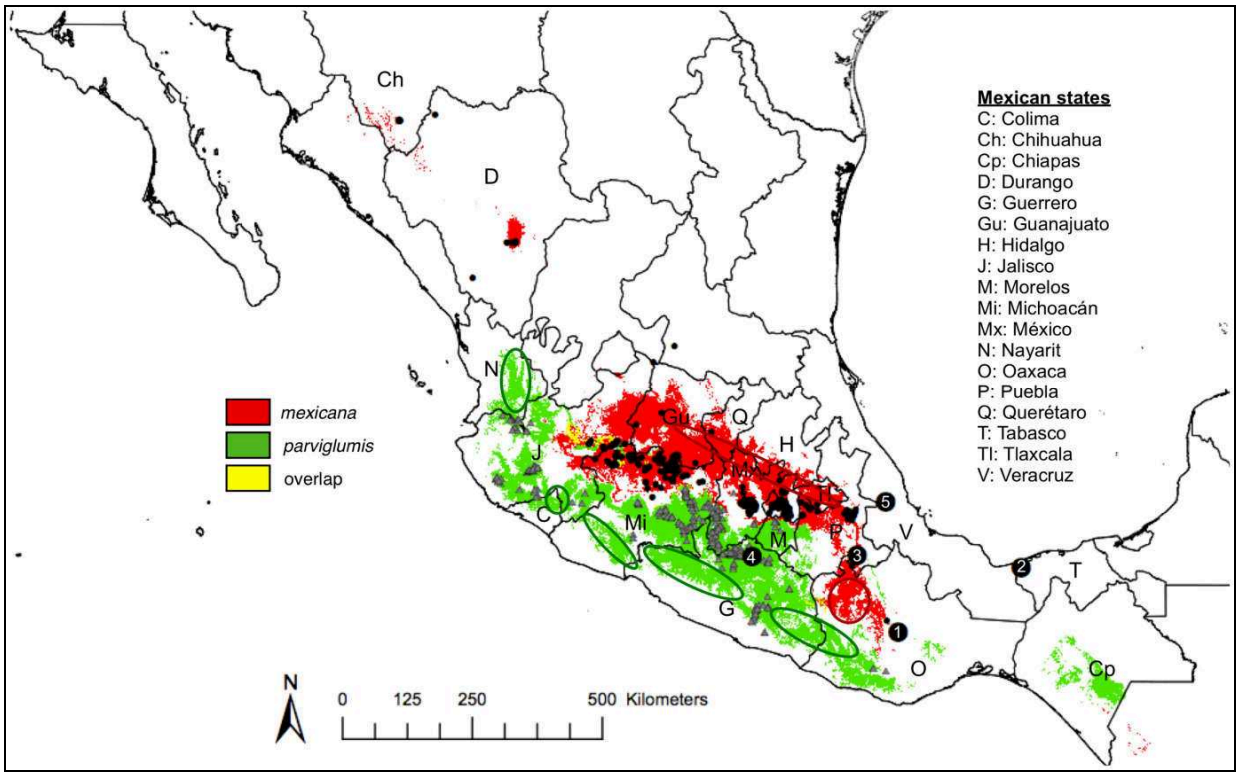


Figure 6. Prédiction des aires de répartions actuelles des téosintes *Zea mays mexicana* et *Zea mays parviglumis*.
(Hufford *et al.* 2012).

S'il y a des gènes liés à la domestication et à l'adaptation, d'autres composants du reste du génome peuvent aussi intervenir dans ces processus. Tout le génome est sous pression de sélection, notamment en lien avec la dérive associée à la domestication. Les éléments transposables font parties de ces composantes, et de plus en plus d'études y sont consacrées pour comprendre leur implication dans les processus de domestications et d'adaptation

(Tenaillon *et al.* 2010; Britten 2010; Díez *et al.* 2013; Jian *et al.* 2017). Une étude menée sur les chiens de la race « loup gris » a révélé qu'un rétroélément de type SINE est associé à la domestication de ces animaux (Gray *et al.* 2010). Chez le maïs, une insertion du rétroélément *Hopscotch* dans les séquences régulatrices du gène *tb1* a été sélectionné durant la domestication (Studer *et al.* 2011). Une étude comparative entre populations tropicales et tempérées de Drosophiles a mis en évidence l'implication de rétrotransposons à LTR dans l'adaptation au climat tempéré (González *et al.* 2010). Dans une autre étude, l'insertion d'un élément transposable dans la séquence régulatrice du gène *ZmCCT,* associé à la date de floraison, a été sélectionnée après la domestication et est fortement impliquée dans l'adaptation des maïs aux milieux tempérés (Yang *et al.* 2013). Nous présentons ici quelques caractéristiques des éléments transposables du génome.

2.5. Le rôle des éléments transposables dans la domestication et l'adaptation

L'intérêt de ce paragraphe n'est pas de faire une bibliographie exhaustive sur les éléments transposables et de leur implication dans la domestication et l'adaptation. Il s'agit plutôt de faire une brève généralité sur la question, le chapitre 4 étant esclusivement consacré à une étude plus approfondie sur les éléments répétés du génome.

Dans les années 1940, la cytogénéticienne Barbara McClintock a mis en évidence l'existence de fragments d'ADN capables de se déplacer au sein du génome du maïs : ce sont les éléments transposables (ETs), nommés ainsi près de 30 années plus tard avec l'avènement de la génétique moléculaire (McClintock 1984). Ils constituent une fraction importante du génome des plantes, proportionnellement à la taille du génome (Figure 7) (Bennetzen 2000). Ces éléments contribuent à la diversité des génomes végétaux, et jouent un rôle majeur dans la régulation de l'expression des gènes par leur dynamique d'insertion ou par contrôle épigénétique (Vicient et Casacuberta 2017). Selon leurs mécanismes de transposition, les éléments transposables sont subdivisés en deux classes : classe I et classe II. La classe I regroupe des éléments dont le mécanisme de transposition est réplicatif via un intermédiaire ARN : on les appelle des rétrotransposons. Pour ces éléments, la séquence originale est gardée intacte, et une nouvelle copie identique est réinsérée à un autre emplacement dans le génome, par mécanisme *"copier-coller"*. Contrairement aux éléments de la classe I, les éléments de la classe II ont une transposition non réplicative. Ils codent une transposase leur permettant un mécanisme de *"couper-coller"*. Chez ces éléments, la même copie est exisée de sa position initiale puis réinsérée à un autre emplacement du génome. Chacune des deux classes d'éléments transposables est organisée en « ordres » (Wicker *et al.* 2007) (plus de détails dans le chapitre 4 du présent manuscrit, et les revues de Wicker *et al.* (2007) et Vitte *et al.* (2014)).

La transposition réplicative des éléments peut avoir de nombreux impacts sur la structure, les fonctions et l'évolution des génomes : le déplacement régulier des ETs à travers le génome peut causer des mutations par insertion (Horváth *et al.* 2017), et d'importantes modifications structurales (Saxena *et al.* 2014). Au cours de leur transposition, les ETs peuvent perturber une séquence fonctionnelle par ajout de séquences régulatrices ou de marqueurs épigénétiques pouvant affecter la fonction d'un gène avec l'apparition de mutations préjudiciables allant jusqu'à l'inactivation du gène (Bennetzen 2000; Rebollo *et al.* 2012). Parfois des modifications bénéfiques dues aux activités des ETs peuvent être sélectionnées. Au cours de la domestication des plantes, des cas de recrutement d'ETs ont été mis en évidence (Feschotte 2008; Sinzelle *et al.* 2009; Tenaillon *et al.* 2011; Hermann 2011).

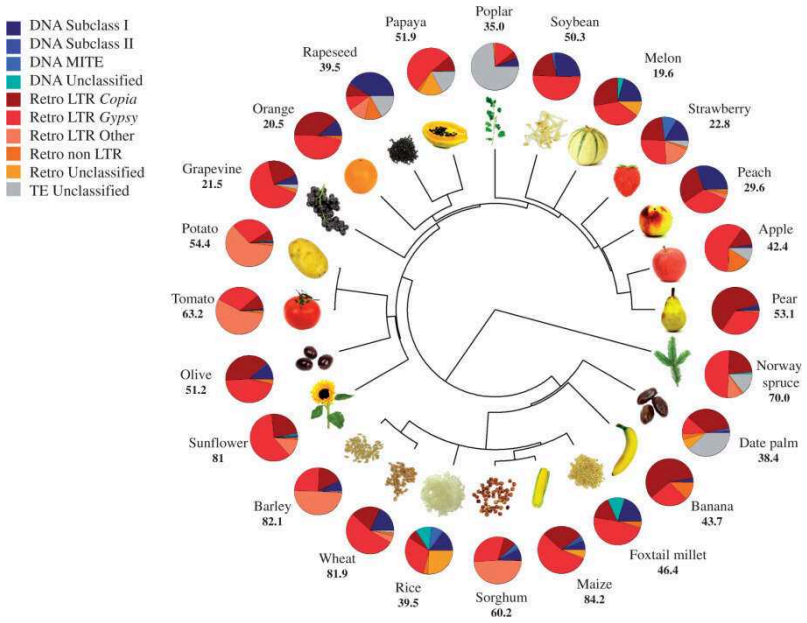


Figure 7. Composition relative de la fraction en ETs dans le génome de 24 espèces de plantes cultivées.
On remarque une nette prépondérance des éléments de classe I (couleur rouge au sein des génomes végétaux) (Vitte *et al.* 2014).

3. L'igname africaine, notre modèle d'étude

3.1. Généralités sur l'igname africaine

L'igname est une liane du genre *Dioscorea,* de la famille des Dioscoreaceae (Lebot 2008). Plus de 600 espèces ont été décrites (Ayensu et Coursey 1972; Govaerts *et al.* 2007). C'est une plante à tubercules, monocotylédone dioïque (Bradshaw 2010) et occasionnellement hermaphrodite (Lebot 2008). Elle est cultivée pour la consommation de ses tubercules amylacés (Kwon *et al.* 2015). L'igname est cultivée dans les régions tropicales et tempérées chaudes :

principalement dans les zones chaudes d'Afrique, d'Amérique, d'Asie et en Océanie, où certaines espèces présentent un intérêt économique important (Idumah *et al.* 2014).

Historiquement, l'exploitation des tubercules d'ignames sauvages en Afrique coïnciderait avec la découverte et le développement des outils tranchants préhistoriques (Ayensu et Coursey 1972). L'igname joue un très grand rôle dans le développement local, en assurant la sécurité alimentaire, et c'est particulièrement en Afrique de l'ouest que les ignames présentent un intérêt économique important (FAO, 2018). En Afrique de l'ouest, les ignames contribuent à la stratégie de subsistance des ménages (Mignouna et Dansi 2003). Les ignames ont aussi une grande importance médicinale. Elles ont longtemps été utilisées en médecine traditionnelle orientale, pour leur richesse nutritionnelle et leurs caractéristiques anti-diarrhéique et anti-inflammatoire (Yang *et al.* 2009; Chandrasekara et Kumar 2016; Kumar *et al.* 2017). L'industrie pharmaceutique leur a accordé très tôt un intérêt tout particulier pour leur contenance en alcaloïdes, en tanins, en phytostérols, en glycosides, en saponines stéroïdiennes ; mais surtout en raison de leur forte teneur en diosgénine ($\leq$ 3,5%), une substance naturelle chimiquement très proche de la progestérone (Sautour *et al.* 2007; Dhont 2010). Depuis les années 1950, ces substances contenues dans les racines de *Dioscorea villosa* étaient utilisées dans l'industrie cosmétique pour la fabrication des pommades, de poudre cosmétiques, de lotions et surtout dans la fabrication des molécules contraceptives et des molécules utilisées pour ralentir le vieillissement post-ménopause (Int J Toxicol 2004). Les ignames ont également un intérêt horticole ; le caudex de l'espèce *Dioscorea elephantipes* est utilisé comme ornement (Everett 1981).

Les espèces d'ignames ont divergé sur trois continents séparés par la formation de l'océan Atlantique et la dessiccation du Moyen-Orient, conduisant à la domestication de trois espèces cultivées du genre *Dioscorea* il y a environ 5000 ans BP (Coursey 1976). Il s'agit des espèces *D. rotundata* d'Afrique ; *D. trifida* d'Amérique ; et de l'espèce *D. alata* d'Asie.

3.2. Les espèces sauvages d'Afrique

Il existe une centaine d'espèces d'ignames sauvages en Afrique dont une quarantaine au Madagascar (Ayensu et Coursey 1972; Govaerts *et al.* 2007) ; mais seulement quelques unes ont été exploitées avec la sédentarisation des hommes autrefois chasseurs-cueilleurs. La récolte des tubercules sauvages d'ignames peut nécessiter un lourd investissement en temps et en énergie. Pour récolter un tubercule de *Dioscorea praehensilis*, il faut parfois fouiller le sol jusqu'à plus d'un mètre de profondeur (Figure 8).

En Afrique et plus particulièrement dans la zone au sud du Sahara, les ignames sauvages occupent les écologies forestières et savanicoles (Hamon *et al.* 1995). Ce sont les espèces *D. abyssinica* (Fig 9-a-b) et *D. praehensilis* (Fig 10-a-b) qui présentent plus d'intérêt en raison de

leur très grande utilisation locale (Hamon *et al.* 1995). L'espèce *D. praehensilis* se différencie de *D. abyssinica* par la présence d'un amas épineux à la base de la tige, et dont le rôle est de protéger le tubercule (Figure 11). C'est la distribution géographique qui différencie principalement les deux espèces : *D. abyssinica* est inféodée aux environnements de savane ; et *D. praehensilis* est native des milieux de forêt (Hamon *et al.* 1995). Il existe par endroit des régions mosaïques ou les deux espèces vivent en sympatrie (Figure 12) (Coursey 1976).

Figure 8. Illustration de la profondeur pouvant être atteinte par un tubercule d'igname sauvage *D. praehensilis*.
A gauche N. Scarcelli enfonce un bâton dans le trou creusé pour extraire le tubercule. A droite N. Scarcelli montre le bâton qui mesure la profondeur du trou. © R. Akakpo, N. Scarcelli, 2015**.**

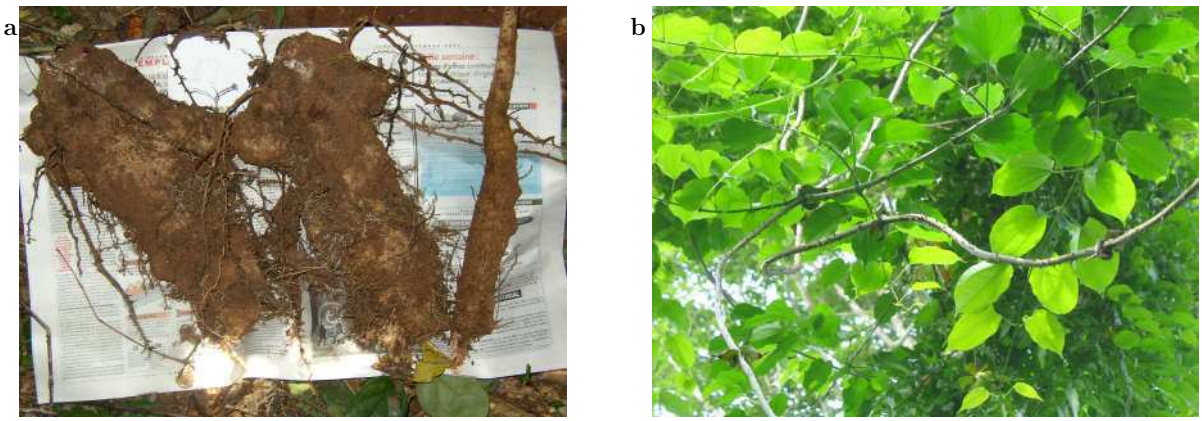


Figure 9 Caractériqtiques morphologiques des racines (a) et feuilles (b) de *D. abyssinica.*
© Crédit photo : IRD, N. Scarcelli

Figure 10. Caractériqtiques morphologiques des racines (a) et feuilles (b) de *D. praehensilis.*
Crédit photo : IRD, N. Scarcelli

Figure 11. Illustration de l'amas épineux de *D. praehensilis.*
Crédit Photo : IRD, N. Scarcelli

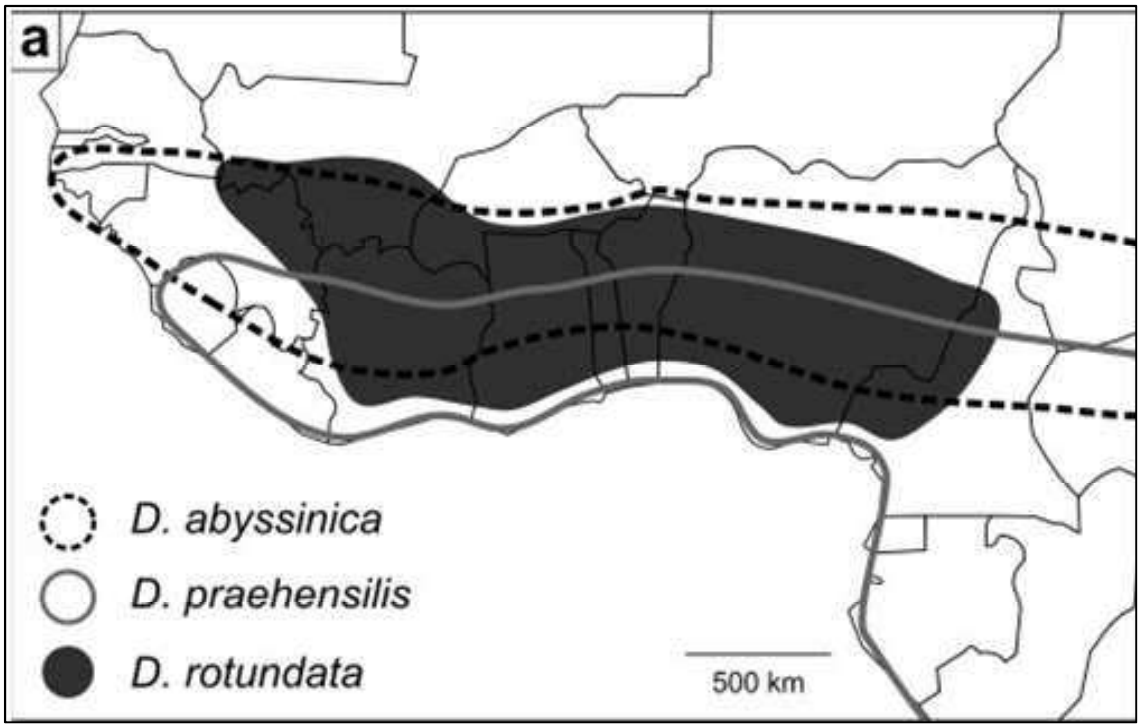


Figure 12. Aires de répartition des espèces d'ignames d'Afrique.
D. abyssinica = Sauvage de savane ; *D. praehensilis* = sauvage de forêt; *D. rotundata* = Cultivé des deux environnements. (Scarcelli *et al.* 2017).

3.3. Les espèces d'ignames cultivées d'Afrique

Le genre *Dioscorea* est organisé en sections et les trois principales espèces cultivées *D. rotundata*, *D. trifida* et *D. alata* sont réparties dans deux sections (Lebot 2008) (tableau 2). Presque toute la production mondiale d'ignames provient de l'Afrique ; soit environ 97% en 2016 (Figure 13) (FAO, 2018). La région dite de la ceinture de l'igname, comprenant la Côte d'Ivoire, la Guinée, le Ghana, le Togo, le Bénin, le Nigéria et le Cameroun, est la principale zone de production d'igname dans le monde (Figure 13) (FAO, 2018). Toute la production de cette zone est consommée localement (Bricas et Attaie 1998). L'espèce la plus cultivée et qui contribue à l'économie locale et régionale est *D. rotundata* (Figure 14, 15) (Dansi *et al.* 1999; Mignouna et Dansi 2003). Mais il existe d'autres espèces d'igname cultivées à importance locale non négligeable en Afrique (Dumont *et al.* 2010). Ce sont les espèces peu cultivées telles que *D. dumetorum*, d'origine africaine, surtout cultivée au Cameroun occidental ; et *D. alata*, d'origine asiatique, surtout cultivée en Côte d'Ivoire (Sartie et Robert 2011). D'autres espèces mineures telles que *D. bulbifera* d'origine africaine et/ou asiatique et *D. esculenta* d'origine asiatique sont également rencontrées en Afrique.

L'igname africaine *D. rotundata,* encore appelée « igname guinéenne » ou « igname blanche », sur laquelle a porté notre étude, présente une grande diversité morphologique (Dansi *et al.*

1999) qui est associée à une grande variabilité d'appellations vernaculaires. Au Bénin par exemple, Baco (2003) a répertorié 112 noms différents utilisés dans 4 villages uniquement ; Dansi *et al.* (1999) ont quant à eux dénombré plus de 300 noms différents pour seulement 10 ethnies. Au Cameroun, Mignouna *et al.* (2002) ont identifié et regroupé 45 noms différents qu'ils ont finalement classés en 6 groupes de cultivars sur la base de leurs similarités morphologiques. Il existe certainement un niveau de synonymie qui augmenterait artificiellement la diversité perçue. En effet, les noms des variétés diffèrent d'une localité à une autre, suivant les variabilités linguistique et ethnique, mais aussi du fait des considérations endogènes (des croyances traditionnelles locales) (Scarcelli 2005; Baco 2007). Par exemple, Baco (2007), dans une étude au Bénin, a fermement identifié, pour une même variété d'igname, l'appellation « kpouna » chez les Bariba, « laboko » chez les Nago, « déiboko » chez les Gando, et « kpounadjè » chez les Peulh.

Plusieurs études sur la diversité morphologique et génétique ont révélé la proximité entre *D. rotundata* et les deux espèces sauvages *D. abyssinica* et *D. praehensilis* (Dansi *et al.* 2000; Mignouna et Dansi 2003; Mignouna *et al.* 2005; Malapa *et al.* 2005; Scarcelli 2005; Tostain *et al.* 2007; Girma *et al.* 2014). Les trois espèces sont d'ailleurs diploïdes à 40 chromosomes (Hamon 1987; Scarcelli *et al.* 2005; Tamiru *et al.* 2017). Il existerait aussi des formes polyploïdes chez *D. rotundata*, mais cela reste à être clairement démontré. Dans une étude récente menée par la technique de cytométrie en flux, des individus de l'espèce *D. rotundata* à 3x chromosomes avec de faibles niveaux d'hétérozygotie ont été identifiés (Girma *et al.* 2014).

Tableau 2. Classification et localisation géographique de quelques ignames cultivées

Espèce	Sections	Noms communs	Origine Probable
D. alata	Enantiophyllum	Greater, water, winged yam	Asie du Sud-Est, Mélanésie
D. rotundata		White guinea yam	Afrique de l'Ouest
D. bulbifera	Opsophyton	Aerial, bulbil-bearing yam	Amérique du Sud, Afrique, Asie, Mélanésie
D. Trifida	Macrogynodium	Aja, aje, cush-cush, yampi	Amérique du Sud
D. esculenta	Combilium	Lesser Yam, Asiatic yam	Asie du Sud-Est, Mélanésie
D. dumetorum	Lasiophyton	trifoliate yam, bitter yam	Afrique de l'Ouest

Source: Lebot (2008); Sahore (2011)

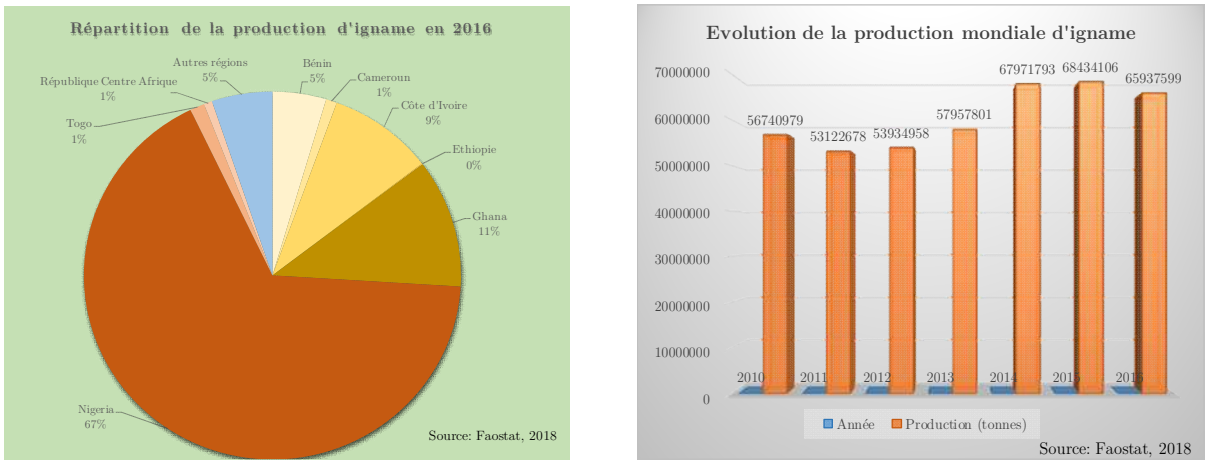


Figure 13. Données de production d'igname dans le monde en 2016
Source : Faostat, 2018 (http://www.fao.org/faostat/fr/#data/QC/visualize)

Figure 14. Tubercule de *Dioscorea rotundata*
Crédit photo : IRD, N. Scarcelli

Figure 15. Champ d'igname cultivé avec les buttes
Crédit photo : IRD, M. Donnat

3.4. Hypothèse sur la domestication de *D. rotundata*

Sur la base d'observations morphologiques, trois principales hypothèses ont été émises sur l'origine de la domestication de l'igname cultivée d'Afrique.

✓ **Origine hybride**

Coursey (1976) formule la première hypothèse selon laquelle *D. rotundata* serait le produit d'une hybridation entre *D. praehensilis* (des milieux de forêts) et *D. abyssinica* (des régions de savanes). L'auteur justifie son hypothèse sur la base de l'existence des zones mosaïques de sympatrie forêt/savane. La possibilité d'hybridation a été mise en évidence entre les deux espèces sauvages (Zoundjihekpon *et al.* 1994; Scarcelli *et al.* 2006, 2013). Une étude récente utilisant des technologies de séquençage de nouvelle génération a également soutenu l'implication à la fois des deux espèces sauvages dans l'origine de *D. rotundata* (Girma *et al.* 2014). Cette étude a notamment utilisé une approche de séquençage GBS pour analyser la structuration et les liens phylogénétiques existant entre sept espèces d'ignames ; et a révélé une forte proximité entre les deux espèces sauvages *D. praehensilis* et *D. abyssinica* ainsi que l'espèce cultivée *D. rotundata*. Il a également été démontré que les trois espèces *D. abyssinica, D. praehensilis* et *D. rotundata* sont interfécondes (Akoroda 1985; Mignouna et Dansi 2003; Scarcelli *et al.* 2006). En effet, la culture de l'igname nécessite souvent des terres neuves, jamais cultivées ou en jachère. Ceci amène généralement les paysans à détruire une partie des forêts u

des savanes pour installer leurs champs, créant ainsi la situation de sympatrie cultivé-sauvages. Cet environnement de sympatrie facilite des flux de gènes dans le sens cultivés-sauvages (Scarcelli *et al.* 2006). En conséquence, on assiste à une pollution des génomes sauvages. Une étude comparative entre la diversité des génomes chloroplastiques des ignames sauvages et cultivées du Bénin a mis en évidence l'ampleur de l'hybridation cultivé-sauvages et des flux de gènes qui polluent le génome des ignames sauvages (Scarcelli *et al.* 2017). Ces auteurs ont révélé que 43% des 65 ignames sauvages analysées possédaient soit un chloroplaste cultivé, soit une introgression nucléaire cultivé-sauvages.

✓ **Origine unique : *D. praehensilis***

La seconde hypothèse sur l'origine de la domestication de *D. rotundata* suggère que l'espèce cultivée dériverait uniquement de *D. praehensilis* (Coursey 1976). Coursey a émis son hypothèse sur la base de similarités morphologiques entre les ignames cultivés *D. rotundata* et les ignames sauvages *D. praehensilis*, surtout au niveau des feuilles. Ceratains plant d'igname cultivé portaient également des tiges épineuses, caractéristique semblable à l'amas épineux retrouvé à la partie basale de la tige de *D. praehensilis*. D'autres auteurs ont aussi observé qu'en pratiquant l'ennoblissement, un processus assimilé à tord à la domestication, des D. praehensilis donnaient au fil des années, des ignames qui ressemblent à *D. rotundata* (Vernier *et al.* 2003; Dumont *et al.* 2010). L'ennoblissement était également pratiqué sur les ignames sauvages *D. abyssinica*. Nous reviendrons sur cette pratique dans le paragraphe suivant.

✓ **Origine unique : *D. abyssinica***

La troisième hypothèse sur l'origine de *D. rotundata* propose *D. abyssinica* comme ascendant sauvage. Cette hypothèse se fonde sur la théorie de l'origine des espèces cultivées en Afrique Harlan (1976). Par ailleurs, il a été rapporté que c'est en savane que *D. rotundata* s'est imposé dans l'agriculture (Dumont *et al.* 2005). Est-ce du fait d'une parfaite adaptation locale? Ou parce qu'il s'agit de son centre de domestication ? D'autres auteurs ont aussi démontré que la mise en culture des plants de *D. abyssinica* dans une pratique dite d'ennoblissement, confondue à tord à la domestication, donne au cours du temps des ignames qui présentent des traits de *D. rotundata*. En effet, chez l'igname, certains auteurs ont avancé que la domestication est toujours en cours (Vernier *et al.* 2003; Dumont *et al.* 2010). Mais en réalité, la pratique en question consiste à replanter, exclusivement par multiplication végétative, des ignames sauvages dans les champs et à pratiquer différents stress pour que la plante s'adapte aux contraintes de l'agriculture (Scarcelli 2005). Le terme d'ennoblissement avait été proposé à la place de « domestication », car la domestication proprement dite implique généralement des changements génétiques majeurs dans le matériel végétal, ce qui n'est pas le cas dans cette pratique dite d'ennoblissement (Mignouna et Dansi 2003). Les résultats de cette pratique sont

similaires à ce qui s'observe chez d'autre plantes à racines et tubercules telles que le manioc (Elias *et al.* 2000; Manu-Aduening *et al.* 2005) et la pomme de terre (Quiros *et al.* 1992). La différence ici est que les graines germent spontanément dans les champs sans l'intervention humaine, et aucun stress n'est opéré. Ces pratiques permettent entre autres, de maintenir et d'enrichir la diversité génétique (Scarcelli *et al.* 2006).

La question d'actualité sur le complexe des trois espèces *D. rotundata*, *D. abyssinica* et *D. praehensilis* est de savoir quelle est l'implication des deux sauvages à l'origine de l'espèce cultivée. Des études génétiques ont permis de séparer clairement les deux espèces sauvages. Comme exemple, dansune étude basée sur l'utilisation des marqueurs AFLP (Tostain *et al.* 2002; Scarcelli *et al.* 2006) une autre où l'utilisation de marqueurs microsatellites a permis de structurer nettement les trois espèces du complexe (Scarcelli *et al.* 2017). L'utilisation de marqueurs chloroplastiques n'a cependant pas permis de différencier clairement les deux espèces sauvages de la forme cultivée (Scarcelli *et al.* 2017). Cette situation suggère donc un fort niveau d'apparentement entre l'espèce cultivée et les formes sauvages, confortant les trois hypothèses sur l'origine de *D. rotundata* ; hypothèses que nous avons testé au cours de nos travaux.

Des études de génétique des populations peuvent aujourd'hui, permettre de confirmer ou d'infirmer chacune de ces hypothèses sur l'origine de l'igname africaine. Les parents sauvages présumés de *D. rotundata* étant inféodés à deux environnements différents, la possibilité de rencontrer les deux espèces sauvages dans des régions mosaïques de transition savane/forêt et l'existence d'hybridations offrent aujourd'hui une opportunité d'étude comparée de la diversité entre les formes sauvages et la forme cultivée. Cela permettra de comprendre en partie le processus de domestication ayant conduit à la formation de *D. rotundata*, mais aussi d'aboutir à de nouvelles perspectives quant à l'origine de l'agriculture en Afrique.

3.4. Les ressources génomiques disponibles chez l'igname africaine

Les études de biodiversité, de phylogéographie et de génétique des populations ont connu une révolution grâce à l'accès à de grandes masses de données génétiques rendues disponibles par les méthodes de séquençage haut débit (NGS) (Mardis 2008). La disponibilité de ces ressources génomiques a favorisé de nouvelles approches pour étudier les divergences inter- et intra-spécifiques des espèces, et permet de préciser la structuration de la diversité qui existe au sein d'un complexe d'espèces.

Pour l'igname, les nouvelles technologies de séquençage NGS ont permis d'accéder en premier lieu, à un chloroplaste de référence de l'espèce cultivée *D. rotundata.* Ce premier génome chloroplastique a été assemblé en 2014 dans le cadre d'un projet financé par le programme

«Investissements d'avenir» de la Fondation Agropolis. Il est en libre téléchargement et référencé dans NCBI sous le numéro NC_024170.1 (Mariac *et al.* 2014).

Plus tard en 2015, un transcriptome de référence a été assemblé (Sarah *et al.* 2016). Les séquences de référence ainsi que les annotations sont en libre téléchargement soit sur le site du projet (http://arcad-bioinformatics.southgreen.fr/node/26), soit sur NCBI sous le BioProject PRJNA326055. C'est également un projet financé par la Fondation Agropolis sous la référence ARCAD 0900-001, qui a servi de support pour ce projet.

Tout récemment en 2017, le premier génome de référence de *D. rotundata* a été publié (Tamiru *et al.* 2017). C'est une collaboration internationale principalement entre le « Japan International Research Center for Agricultural Sciences » (JIRCAS) au Japon et le « International Institute of Tropical Agriculture » (IITA) au Nigéria, qui a servi de cadre pour la mise en œuvre du projet. Toutes les données génomiques relatives au projet sont disponibles en libre téléchargement sur NCBI sous le BioProject PRJDB3383.

Les données génomiques d'environ 200 individus des trois espèces, utilisées dans les différents travaux de ma thèse ont été générées dans le cadre du projet AfriCrop ANR-13-BSV7–0017.

4. Questions de recherche

Pour mener l'étude de la domestication sur notre modèle, l'igname africaine, j'ai formulé trois questions de recherche.

4.1. Quelle est la base moléculaire de la domestication chez l'igname ?

Pour répondre à cette question de recherche, j'ai testé l'hypothèse des conséquences de la domestication concernant les événements de goulot d'étranglement. Pour ce faire, j'ai développé des approches d'analyse comparative de la diversité génétique entre l'espèce cultivée et les deux espèces sauvages apparentées, sur la base de données de séquençage de génomes entiers afin de détecter des signatures de sélection associées à la domestication. Mes réponses à cette première question de recherche sont présentées dans le chapitre 2 du présent manuscrit, sous forme d'un article scientifique soumis, accepté et publié dans BMC Genomics (https://doi.org/10.1186/s12864-017-4143-2).

4.2. Quelle est l'origine de la domestication de l'igname africaine *D. rotundata* ?

Il s'agit ici de mettre en évidence la (les) probable (s) espèce (s) sauvage (s) à l'origine de l'espèce cultivée *D. rotundata.* Pour cela, j'ai développé deux approches : la première a consisté

en l'analyse de la structuration de la diversité génétique sur la base de données de variants génétiques SNPs et la deuxième a concerné une analyse d'inférence démographique des populations en utilisant des données de spectre de fréquences alléliques. Pour cette question, j'ai présenté les résultats sous le format d'un article scientifique qui sera publié ensemble avec d'autres résultats obtenus par Nora Scarcelli (IRD), Anne-Célinne Tuillet (IRD) et Phillipe Cubry (IRD). Cet article est en cours de préparation et a pour objectif de reconstruire l'histoire évolutive des ignames d'afrique. Je serai donc co-auteur de cet article.

4.3. Quelle est la variabilité génétique et génomique associée à l'adaptation de *D. rotundata* à différentes zones climatiques ?

J'ai tout d'abord procédé ici à l'annotation *de novo* des éléments répétés du génome des ignames sauvage et cultivé, et ai utilisé cette annotation pour analyser la corrélation entre la variabilité de la fraction répétée du génome et des données environnementales. J'ai cherché à mettre en évidence des familles d'éléments répétés, potentiellement associées à la dynamique d'adaptation de l'igname. Cette partie de mes travaux fait actuellement l'objet d'un article scientifique qui sera soumis à « Frontier in Plants Science » pour publication. Je serai premier auteur de cet article.

Une deuxième approche d'étude de l'adaptation est celle de la génétique d'association, à l'échelle du génome entier. L'analyse que j'ai développée a consisté en l'étude d'éventuelles associations entre la variabilité génomique des SNPs et les données climatiques caractéristiques des différents environnements. L'interprétation des signatures de sélection a permis de tester s'il existait une base génétique de l'adaptation chez l'igname. Cette dernière partie fait aussi objet de prépartion d'un article scientifique pour lequel je serai premier auteur.

Chapitre 2 : Détection de signatures de sélection associées à la domestication chez l'igname Africaine

1. Contexte et objectifs

L'homme a transformé les plantes des milieux naturels en plantes cultivées comestibles. Ce processus est associé à de nombreuses adaptations à l'environnement humain. Quelle est la base génétique de ces adaptations ? De telles études sont peu réalisées chez les plantes à racines et tubercules. Pourtant, ces plantes constituent la deuxième source de calories de l'alimentation humaine après les céréales (Meyer and Purugganan 2013). Les gènes associés à la domestication des principales céréales sont les plus documentés. La voix de synthèse de l'amidon a mis en évidence la sélection du gène *SUS* chez le maïs (Li *et al.* 2013) et chez le blé (Hou *et al.* 2014). Qu'en est-il chez les plantes à racines et tubercules ? Les organes de réserve que sont les racines et les tubercules sont particulièrement développés chez ces plantes ; comment cette domestication s'est-elle alors traduite en termes de sélection ? Sur la voix de biosynthèse de l'amidon, sur la morphologie de la plante et de ses racines ? Nous essayons de répondre à ces questions sur l'igname. Nous avons essayé d'identifier des signatures de sélection associées à la domestication.

2. Méthodes

Les génomes de 30 accessions d'igname collectées au Bénin ont été séquencés par technologie Illumina. Un total de 162 millions de reads pairés (2x 100pb) a été généré. La détection de traces de sélection associée à la domestication a concerné le compartiment exprimé du génome de l'igname, en utilisant comme référence le transcriptome de *D. rotundata* (Sarah *et al.* 2016). L'alignement des données de séquençage contre le transcriptome a permis de générer un jeu de 300K variants SNPs. Quatre tests statistiques de détection de sélections ont été effectués : 1) le ratio entre la diversité cultivée et celle sauvage ; 2) la statistique D de Tajima ; 3) le test de différenciation Fst et 4) une analyse de différenciation basé sur une approche d'ACP.

3. Principaux résultats

La diversité nucléotidique était plus faible pour le groupe cultivé par rapport aux espèces sauvages. Nous avons identifié des contigs candidats à la sélection au cours de la domestication qui montrent une différenciation extrême et des valeurs extrêmes de la statistique D de Tajima, signature de sélections positives avec notamment des gènes associés à la voix de biosynthèse de l'amidon, la morphologie racinaire et le développement aérien. Une série de gènes, associés à

des activités de déshydrogénase et d'oxydoréductase et aux gènes du complexe NADPH DH de la photosynthèse, a aussi été identifiée.

Ces résultats ont fait l'objet de la rédaction et de la publication de l'article scientifique intitulé : **Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes**.
Roland Akakpo, Nora Scarcelli, Hana Chaïr, Alexandre Dansi, Gustave Djedatin, Anne-Céline Thuillet, Bénédicte Rhoné, Olivier François, Karine Alix and Yves Vigouroux. 2017. BMC Genomics 18:782 DOI 10.1186/s12864-017-4143-2.

Cet article est ici présenté dans la suite du présent manuscrit de doctorat.

Akakpo et al. BMC Genomics (2017) 18:782
DOI 10.1186/s12864-017-4143-2

BMC Genomics

RESEARCH ARTICLE Open Access

Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes

Roland Akakpo[1,2,3†], Nora Scarcelli[1†], Hana Chaïr[4], Alexandre Dansi[3], Gustave Djedatin[3], Anne-Céline Thuillet[1], Bénédicte Rhoné[1,5], Olivier François[6], Karine Alix[2] and Yves Vigouroux[1*]

Abstract

Background: After cereals, root and tuber crops are the main source of starch in the human diet. Starch biosynthesis was certainly a significant target for selection during the domestication of these crops. But domestication of these root and tubers crops is also associated with gigantism of storage organs and changes of habitat.

Results: We studied here, the molecular basis of domestication in African yam, *Dioscorea rotundata*. The genomic diversity in the cultivated species is roughly 30% less important than its wild relatives. Two percent of all the genes studied showed evidences of selection. Two genes associated with the earliest stages of starch biosynthesis and storage, the sucrose synthase 4 and the sucrose-phosphate synthase 1 showed evidence of selection. An adventitious root development gene, a *SCARECROW-LIKE* gene was also selected during yam domestication. Significant selection for genes associated with photosynthesis and phototropism were associated with wild to cultivated change of habitat. If the wild species grow as vines in the shade of their tree tutors, cultivated yam grows in full light in open fields.

Conclusions: Major rewiring of aerial development and adaptation for efficient photosynthesis in full light characterized yam domestication.

Keywords: Domestication, *Dioscorea spp.*, Adaptation, Population genomics, selection, Root development, Starch biosynthesis, Plant development

Background

One of the major changes in human history was the emergence of agricultural societies [1]. About 13,000 years ago, farmers began to domesticated plants and animals for agriculture. Domestication was done by selecting plants and animals with suitable traits for farming like increased yield. As a result, the morphology of our cultivated plants was reshaped by human selection for a period certainly spanning thousands of years [2–4]. The domestication process offers an interesting glimpse of the broad adaptation process and of the genetic basis of morphological and physiological traits [5, 6]. It helps understand how a relatively lowly productive wild relative can be transformed into a high yielding cultivated variety. Insights into crop domestication have primarily come from cereals [5]. Root and tuber crops are also a major contributor of starch to the human diet. These crops have the particularity of very often being vegetatively propagated [7]. The domestication process increased their ability to store starch in their roots or tubers and other specialized storage organs as well as the size of these organs [7]. Today it is not clear if the knowledge we have of the process of domestication of cereal crops can be extrapolated to root and tuber crops. For example, selection on several genes responsible for starch biosynthesis has been documented in maize [8, 9]. So, one would expect that domestication also allows more efficient production and/or storage of starch in root and

* Correspondence: yves.vigouroux@ird.fr
†Equal contributors
[1]Institut de Recherche pour le Développement, Université de Montpellier, Unité Mixte de Recherche Diversité Adaptation et Développement des Plantes (UMR DIADE), 911, avenue Agropolis, 34394 Montpellier, France
Full list of author information is available at the end of the article

tuber crops. One would also expect that domestication reshaped the formation and development of roots as a support for efficient starch storage.

The most widely grown root and tuber crops in Africa are cassava and yam. The two main species of yam, *Dioscorea spp*., were domesticated independently, *D. rotundata* in Africa and *D. alata* in Asia. *D. rotundata*, the most widely cultivated yam species in Africa is a staple food for over 100 million people [10]. This species has two close wild relatives *D. abyssinica* and *D. praehensilis* [11–14]. The three species are diploid and have 20 chromosomes [2n = 40] [14–16]. The African cultivated yam and its closest wild relatives are compulsory out-crossers because they are dioecious. However, *D. rotundata* is preferentially propagated through vegetative multiplication [17]. Interestingly, the two wild species have distinct ecological distribution: *D. abyssinica* is found in the wooded savanna areas while *D. praehensilis* is found in tropical forested areas [18]. The diploid African yam is cultivated in both ecological areas, thereby allowing gene flow between cultivated and the two wild species [13]. Several key phenotypes differentiate cultivated varieties from their wild relatives. Cultivated yams are characterized by larger and less ramified roots than their wild relatives, and some cultivated varieties do not develop inflorescences [19]. Finally, the wild relatives of yam are vines which grow partly in the shade of their tutor tree, while cultivated yams grow in full sunlight. This change of habitat might be associated with major adaptation.

Our objective was to uncover the molecular basis of yam domestication. To find what genes and specific functions were selected during yam domestication, we sequenced the genome of wild and cultivated African yams. Using this dataset, we then scanned for selection signature to pinpoint genes associated with domestication.

Methods

Plant material and DNA sequencing

Thirty plants were collected in 15 villages in Benin (Additional file 2: Table S1). Sampling included 10 individuals belonging to the cultivated species *D. rotundata,* and 10 individuals belonging to each of its two closest wild relatives, *D. abyssinica* and *D. praehensilis*. Plants were identified by Serge Tostain (yam specialist, IRD), Nora Scarcelli (yam specialist, IRD) and local yam farmers. DNA was extracted as previously described using a standard protocol [16]. Genomic libraries were constructed using a recent protocol [20]. The genomic libraries were 2 × 100 bp paired-end sequenced by sample multiplexing using the Illumina HiSeq 2000 technology (GeT_Genotoul, Toulouse, France).

Bioinformatics analysis and SNP detection

Raw data were first filtered using a previously described pipeline [21]. Briefly, we performed a demultiplexing python script demuladapt (https://github.com/Maillol/demultadapt). Adaptors and low-quality bases were eliminated using cutadapt 1.2.1 [22]. Reads with a mean quality score < 30 were removed using a free perl script https://github.com/SouthGreenPlatform/arcad-hts/blob/master/scripts/arcad_hts_2_Filter_Fastq_On_Mean_Quality.pl. Mapping was performed using default options of BWA aln-sampe V0.7.5a–r405 [23], and using the *D. rotundata* transcriptome reference [24]. We validated by modelling that the mapping of genomic DNA reads on a transcriptome reference did not lead to major bias of SNP identification (Additional file 1: Table S1).

We estimated the genotype likelihood (GL) for each site using the option "-GL 3" (SOAPsnp model) implemented in angsd 0.700 [25]. We also performed SNP calling using the HaplotypeCaller in the Genome Analysis Toolkit (GATK) V-3.4-46 [26]. Default options of GATK and the "-rf BadCigar" options were used. SNPs were filtered for low missing rate < 5% and a mean depth ≥ 4. The complete script from the raw data to the GL or SNP data analysis is available as a Additional file 1: Table S1.

Analysis of diversity, population structure and linkage disequilibrium

Genetic structure was assessed using a least-squares optimization approach implemented in the sNMF program [27]. This approach is based on SNP calling and consists in estimating admixture coefficients based on sparse non-negative matrix factorization [27]. We assessed a number of K populations varying from 1 to 6 clusters. Ten replications were performed for each K value. To select the best K value, we used the minimum value of the cross entropy criterion [27]. We also used the maximum likelihood structure approach implemented in the NgsAdmix program [28]. This approach directly uses the genotype likelihood given by angsd, without calling genotypes. The most relevant K number of population was selected by comparing the results obtained with NgsAdmix and sNMF. Genetic diversity was estimated using nucleotide diversity π [29] and nucleotide polymorphism θ [30] computed using the option "-doThetas" implemented in angsd 0.700 [31]. We calculated the ratio of diversity between the cultivated species *D. rotundata* and each of the wild species *D. praehensilis* and *D. abyssinica* using the R package. Pairwise linkage disequilibrium (LD) was calculated with the squared allele frequency correlation r^2 [32] using the R packages SNPRelate [33] and LDcorSV [34]. A set of contigs corresponding to 1% of all contigs was randomly selected and used as reference. Intra-contig LDs within

these contigs were performed for pairs of SNPs with minor allele frequencies (MAF) higher than 0.01.

Identifying candidate genomic regions for selection in yam

We used four different approaches to identify regions under selection: two methods allowing identifying a reduction of diversity for the selected genes, two methods allowing identifying an excess of differentiation. The diversity reduction was assessed using Tajima's D and by the ratio of cultivated to wild diversity. The excess of differentiation was assessed using the F_{ST} between cultivated and wild populations and a principal component based analysis. Tajima's D value of each contig was calculated for the species using vcftools v0.1.13 [35]. (1) We plotted the distribution of Tajima's D values and then used a 1% threshold to identify extremely low values. (2) The ratio of the cultivated genetic diversity divided by the mean diversity of the two wild relative species using π [29] and θ [30]. We used a 1% threshold to identify outlier contigs with extremely low ratios. (3) We estimated the differentiation index F_{ST} [36] between the cultivated group and each of the two wild groups for each contig using vcftools v0.1.13 [35]. Using the cutoff of the 1% top values, contigs with extreme F_{ST} between the cultivated and both two wild relatives were selected as candidates. (4) Based on principal component analysis at the SNP level we used the program Pcadapt V2.2 [37] to identify SNPs with extreme differentiation between the three species. The Mahalanobis distance [38] was calculated and we used the 5% threshold of the false discovery rate (FDR) [39] to detect candidate SNPs. The four selection tests were compared using a Venn diagram [40] to reveal the most likely candidate regions for selection. The annotation of the candidate selected genes was retrieved from a previous study [24].

Enrichment analysis for annotated candidate contigs

First, all the candidate contigs annotated in the reference transcriptome were tested for enrichment of gene ontology (GO) molecular function terms. Standard Fisher's exact tests implemented in the R package TopGO [41] were performed. A minimum of five annotated genes were required per term in order to limit statistical artifacts of GO terms with less annotated genes. Then, to control for false positive effects, only candidate contigs identified by at least two different selection tests were chosen, and the enrichment of GO terms analysis was rerun.

Results

Diversity structuration supports the three major species

We generated 162 million 100-bp paired-end reads. The yam transcriptome size has been estimated to be approximately 64 Mb [24] and the genome size to be 550 Mb. We obtained an average mapping rate of ~ 12.6% of our genomic reads i.e. close to the expected 12.4% based on the relative transcriptome size compared to the whole genome (Additional file 2: Table S2). We identified a total of 308,840 SNPs. These SNPs were found in 23,136 contigs with a mean contig length of 1316 bp (ranging from 250 to 15,691). A low correlation was observed between the length of the contigs and the number of SNPs detected ($r = 0.34$, $p < 0.001$).

Analysis of the population structure using sNMF led to three major genetic groups (Additional file 2: Figure S1), corresponding to the three species (Fig. 1-a). We identified four individuals (A420, P599, A433 and P624) as interspecific hybrids. One individual (A3085) was certainly misclassified in the field: it was recorded as *D. abyssinica* in the field but was genetically close to the *D. praehensilis* group. The exact structuration was similarly found using the NgsAdmix approach, with only minor differences in the estimated proportion of admixture (Fig. 1-b). As hybrids could bias the calculation of diversity; the differentiation tests; and Tajima's D statistics, we removed the four hybrids for further analysis. Departures for neutrality or extreme differentiation were consequently assessed on 26 individuals.

We compared nucleotide diversity π and the nucleotide polymorphism θ between the cultivated species and each of the wild species. First, the cultivated diversity π was 26% and 36% respectively lower than *D. abyssinica* and *D. praehensilis* (Additional file 2: Table S3 a and b). Secondly, the cultivated diversity θ was 28% and 44% lower than *D. abyssinica* and *D. praehensilis* respectively. Linkage disequilibrium (LD) computed between 400,760 pairs of SNP decreased rapidly at $r^2 = 0.1$ after 100 bp (Additional file 2: Figure S2).

The combination of selection tests identified a large set of candidate contigs

Contigs were searched for selection signatures using four different methods: Tajima's D, marked reduction in the diversity in the cultivated samples, differentiation between wild and cultivated species, and principal component analysis. Using the four methods, a total of 998 candidate contigs were identified (Additional file 2: Table S4), among which 81 were detected by at least two methods (Additional file 2: Figure S3).

(i) Tajima's D in the cultivated yam showed a skewed distribution to positive values (Fig. 2-a), with a mean of 0.77. The distribution reflected an excess of contigs with low diversity (Fig. 2-a). The distribution of Tajima's values in the two wild species is centered on zero and consequently reflects a more global equilibrium between SNP occurrence and their frequencies (Additional file 2: Figure S4). Using a 1% threshold (Tajima D < −1.84), a

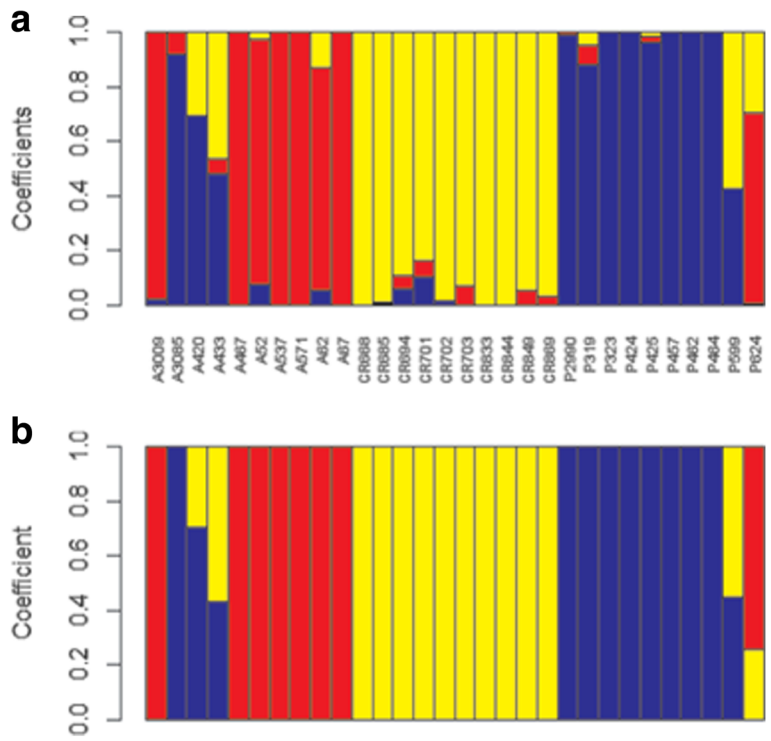


Fig. 1 Structure analysis using sNMF(**a**) and NgsAdmix (**b**). Each color represents one population. The length of each segment in each vertical bar represents the proportion of ancestry in each population

total of 187 contigs were identified as potential candidates under selection in the cultivated sample.

(ii) The reduction of nucleotide diversity and the nucleotide polymorphism were highly correlated (r = 0.997, p < 0.001, (Additional file 2: Figure S5). Consequently, we only used the reduction of nucleotide diversity (π_c/π_w) for further analysis. Using a threshold of 1% (–log10 (π_c/π_w) > 1.34), a total of 232 contigs were identified as having an extremely low diversity in the cultivated sample compared to their wild relatives, and were therefore considered as candidates. (Fig. 2-b).

(iii) The average differentiation between *D. rotundata* and *D. praehensilis* was higher than between *D. rotundata* and *D. abyssinica,* (F_{ST} = 0.21 and 0.16, respectively, p-value <0.001). Using a 1% threshold (F_{ST} > 0.73 and 0.84 for *D. rotundata* with *D. praehensilis* and *D. abyssinica* respectively), 422 contigs were identified with extremely high F_{ST} values with one or the other wild species. Among them, 12 showed extreme values with the two wild species simultaneously (Fig. 2-c).

(iv) Last, we used a SNP-based approach. The two first principal components were used to perform the genome scan for selection using Pcadapt V.2.2 (Additional file 2: Figure S6a). The Mahalanobis statistic distance fitted a normal distribution (Additional file 2: Figure S6b). The histogram of p-values showed an excess of small p-values, indicating the presence of outliers (Fig. 2d). Using a 5% threshold, we identified 2502 SNPs in 1602 candidate contigs with extremely low p-values. A total of 238 contigs that showed at least two SNPs putatively under selection were retained as candidates.

Root development, starch biosynthesis, phototropism and photosynthesis candidate genes were selected

We compared the candidate contigs with the available annotation of the yam transcriptome reference [24]. Thus, we retrieved some genes corresponding to putative targets for selection during yam domestication. In particular, among the genes annotated for the candidate genes, we identified five candidate contigs that were relevant in the light of yam domestication (Fig. 3 and Additional file 2: Table S5). These five candidate contigs showed strong diversity loss in the cultivated group compared to the wild species (Additional file 2: Figure S7). A candidate contig was a putative *SCARECROW-LIKE* gene involved in root development [42, 43]. Two other genes were associated with the earliest stages of starch biosynthesis and storage i.e., genes coding for the sucrose synthase 4 [44] and the sucrose-phosphate synthase 1 [45]. We also identified two genes associated with growth and phototropism, respectively*: Ethylene Insensitive 4* genes *(EIN4)* [46] and *Phototropin 2* gene (*Phot2,* [47]. The 998 candidate contigs were significantly enriched for a total of 21 significant GO terms

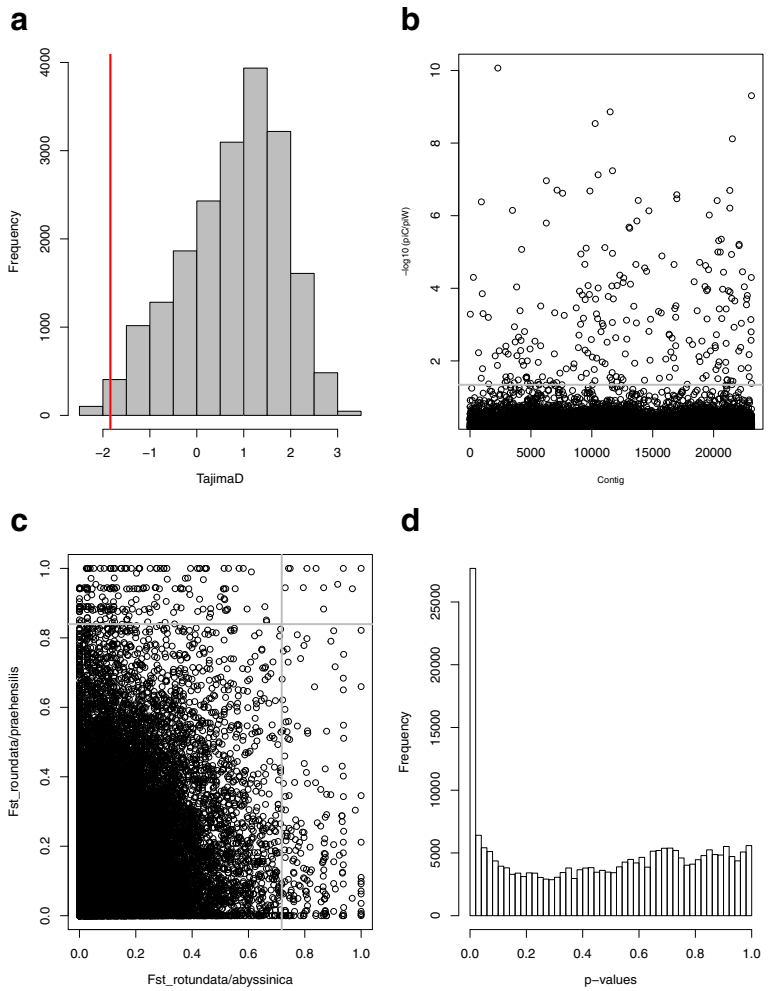


Fig. 2 Summary of the different tests used to identify outlier contigs. In the distribution of Tajima's D value of the cultivated species (**a**), the red line indicates the 1% threshold used to consider contigs as candidates. In the of reduction of nucleotide diversity π (**b**), the -log10 (π_c/π_w) for each contig is represented by one dot. The gray line corresponds to the 1% threshold used to consider contigs as candidates. In the comparison of F_{ST} between the cultivated and the two-wild species (**c**), each dot represents a contigs. The grey lines indicate the 1% threshold used to consider contigs as candidates. Finally, in the histogram of *p*-value (**d**), the peak of SNP close to zero indicates the presence of outliers. Here, the SNPs were considered as candidates using an FDR of 0.05

(Additional file 2: Table S6). When we restricted our analysis to the 81 candidate contigs detected by at least two methods, we obtained nine significant GO terms (Additional file 2: Table S7). The most significant GO terms were identical whether we considered all the candidate contigs or only the 81 candidate contigs. The set of GO terms found across these two enrichment tests was associated with dehydrogenase and oxidoreductase (*NADH DH*) activities (Fig. 4).

Discussion

The domestication diversity loss observed in yam is comparable to an outcrossing crop

Today, the *D. rotundata* yam species is vegetatively propagated. However, the nucleotide diversity loss associated with domestication is relatively modest: the cultivated sample had 26% and 36% diversity loss respectively relative to *D. abyssinica* and *D. praehensilis*. In out-crossing species like pearl millet and maize, diversity losses of 32% [48] and 35% [49] were reported. In self-pollinating species, the diversity loss can be much higher, for example, 62% in barley [50], and 70% in wheat [51]. The loss of diversity observed in our study is more similar to outcrossing crops. We do not know when the transition from an outcrossing crop to a preferentially vegetative crop occurred. It is likely that during the first step of domestication, the crop reproduced mainly through seed. Even today, the reproduction system of *D. rotundata* is not purely vegetative [13, 52], and some cultivated varieties were found to have been recently obtained by cross-pollination. So, this modest loss of diversity is not surprising.

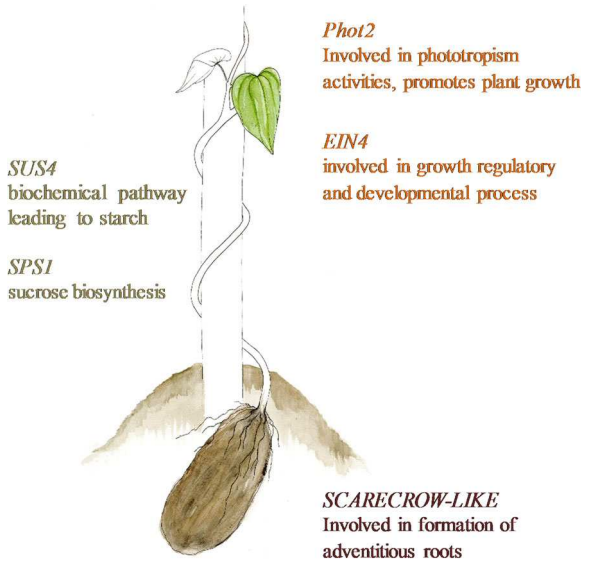


Fig. 3 Key genes associated with yam domestication. *SCARECROW-LIKE, Phot2, EIN4, SUS4* and *SPS1* are some interesting genes probably selected during domestication

Linkage disequilibrium (LD) also decreased rapidly, like in other outcrossing crops. This LD decay is more similar to that observed in maize [53–55] than to that reported in self-pollinating crops such as rice [56]. However, our estimation of LD is based on a small sample and we might overestimate the rapidity of its decrease.

Overall, despite the mode of reproduction of the cultivated yam, both the diversity loss and the LD decay observed were similar to those in outcrossing crops.

Identifying selected genes during domestication

We found 2% of yam genome classified as candidates for selected genes during domestication. A very similar rate of genome under selection was previously observed in maize, ranging from 2 to 5% [49, 57, 58]. Among the contigs we identified, roughly 10% of the candidate contigs were commonly identified by a least two different methods used for detecting signatures of selection.

Depending of the strength and the timing of selection, its resulting impact on diversity could differ. Consequently, each test has different strength and power to detect these specific signatures of selection. For example, when strongly selected, alleles could be fixed. These specific genes showing strong selection could be detected by differentiation F_{ST} based test, but not by Tajima's D test because of their fixed polymorphism [31]. So, the specificity of each test could lead to the discovery of only a small set of the same contigs by all different methods. However, each method could also identify false positives [59]. These false positives could be specific of a test. In conclusion, both false positives and different impacts of selection on diversity resulted in roughly 10% of genes being simultaneously identified by all the methods performed. Furthermore, signature of selection on two contigs could be associated with a single selection events one of them. Even if we found that linkage disequilibrium decreased fast, our list of selected

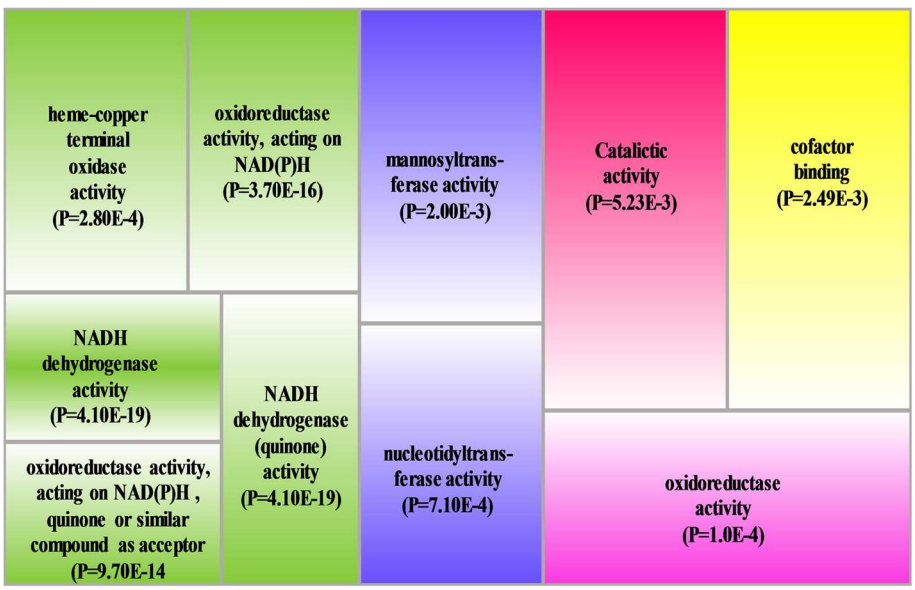


Fig. 4 TreeMap view of the 10 most significant "Go Terms" identified. The 10 most significant GO terms were reported with their respective *p*-values. We group them in 4 major clusters: "oxidase activity" in green, "transferase activity" in blue, "catalytic activity" in pink. "cofactor binding" in yellow

genes might represent fewer selection events than their actual numbers.

Domestication is associated with selection of root development, sugar metabolism, and phototropism genes

Cultivated yams are known to have less ramified and larger roots than wild yams. Remarkably, we found a contig homologous to a gene coding for a *SCARECROW-LIKE* protein. As demonstrated in *Arabidopsis*, this gene is a key player in root development [42, 43] and consequently may have been mobilized during yam domestication. We also pinpointed a contig homologous to an *EIN4* gene. *EIN4* is a receptor of ethylene [46] involved in growth regulation and many developmental processes including seed germination, leaf and flower senescence [60]. At this stage, we do not know if this gene may affect root development itself or its above ground development.

Domestication of root and cereal crops is notably associated with the increase of starch production. Several studies on cereals suggest that starch biosynthesis and storage were important targets for selection [61]. In our study, we observed the selection of two genes involved in the production of sugar: *SUS*4 and *SPS1*. *SUS* catalysis is the first step leading to starch formation [44] by converting sucrose to fructose and UDP-glucose. In wheat, selection for increased starch content was associated with selection of *SUS* genes [62], and enhancing *SUS* activities also resulted in increasing starch content in maize [63]. The *SPS* gene has also been reported to play a major role in sucrose biosynthesis under osmotic stress conditions [45]. In conclusion, similar set of genes were selected during cereal, root and tuber crops.

Beyond starch production, cultivated yam underwent a major change in its living environment during domestication. Yams are now grown in open fields, whereas its wild relatives grow as vines in the shade of tutor trees. This environmental change during domestication certainly required adaptation due to such changes in light and heat. We observed strong signatures of selection in genes associated with physiological processes of regulation of photosynthesis for light tracking and for plant growth. Indeed, one of our candidate contigs is homologous to the *Phototropin 2* gene (*Phot2*). In higher plants, Phot2 enables perception of blue light and consequently optimization of photosynthetic performance and growth [47].

Adaptation to high intensity light was selected during yam domestication

Beyond specific genes associated with the change from shade to light environment, we also found a significant enrichment of interesting gene ontology terms. The most significant GO terms observed were and oxidoreductase activities associated with *NADPH DH* complex genes [64, 65]. Whatever the strategy of enrichment test used, the results were robust for these functions. The *NADH DH* complex is an important set of enzymes for chlororespiration [66]. The *NADH DH* complex is involved in photosynthesis [67], more specifically in the photosystems I (PSI) and II (PSII). It plays a role in protection against photo-oxidative stresses associated with the formation of reactive oxygen species (ROS) [68]. High light and heat could favour the production of ROS [69, 70]. In oats, NADH DH is over-expressed with increasing light [67]. Consequently, it has been postulated that this type of complex plays a role in mitigating ROS stress associated with increasing intensity of light or heat. In Brassica plants, the same *NADH DH* complex has also been reported to be associated with the domestication process [71]. The wild species of *Brassica* showed higher tolerance to high light and heat intensity than the cultivated species [71]. In this specific case, domestication was associated with a decrease in photosynthetic parameters under stress conditions in the cultivated species [71]. The two wild species of yam are vines that grow in partial shade. The cultivated species *D. rotundata* grows under full sunlight in the field. We hypothesize that adaptation of the cultivated yam led to the selection of genes that enable efficient photosynthesis with increasing light and heat intensity. Optimizing photosynthesis is also an important way to enhance production of carbohydrate, later stored as starch in the tuber.

Conclusions

Selection in the early step of sugar biosynthesis is detected in yam, and previously detected in cereal. This result suggests that key step in starch biosynthesis were necessary both in cereal as well as in root and tuber crops. More interestingly, drastic changes in habitat associated with domestication is certainly retraced in selection in phototropism genes. Selection on dehydrogenase and oxidoreductase activities associated with *NADPH DH* complex genes, was certainly the consequence of adaptation to optimize photosynthesis in full light. If some convergence is observed at the molecular level, very specific adaptations were necessary for the domestication of African yam. Beyond domestication, this study highlight the molecular mechanism associated with changes from shade-tolerant plant to a full light environment.

Additional files

Additional file 1: We assess if the mapping of genomic DNA reads on a transcriptome reference could impact SNP calling in our special case.

Table S1. Summary of mapping and SNP calling using simulated data. (DOCX 15 kb)

Additional file 2: Molecular basis of African yam domestication: analyses of selection point to starch biosynthesis, root development and photosynthesis related genes. **Table S1.** Passport data of plant material collected from Benin. **Table S2.** Metric information of data filtering and mapping. **Table S3.** Mean Nucleotide diversity (π) and polymorphism (θ). **Table S4.** List of the contigs detected as selected by at least one method. **Table S5.** Remarkable candidate genes showing selection signature. **Table S6.** Gene Ontology (GO) terms significantly enriched (p-value ≤ 0.05) among the 998 candidate contigs. **Table S7.** Gene Ontology (GO) terms significantly enriched (p-value ≤ 0.05) among the 81 candidates contigs detected by a least two methods. **Figure S1.** Cross-entropy calculated using sNMF (Frichot et al., 2014) for K = 1 to 6. Ten repetitions of the run were done. **Figure S2.** Intra-contigs linkage disequilibrium (LD) as a function of physical distance between SNPs pairs from 1% of all contigs. **Figure S3.** Venn Diagram comparing the candidate contigs obtained using the 4 methods. **Figure S4.** Distribution of Tajima's D value calculated for *D. abyssinica* (a) and *D. praehensilis* (b). **Figure S5.** Comparison of diversity lost. **Figure S6.** Variance explained by PCA axis (a) and distribution of Mahalanobis distance (b) from PCAdapt. **Figure S7.** Nucleotide diversity within five candidate contigs for cultivated and the wild species (XLSX 45 kb)

Abbreviations
BWA: Burrows-wheeler aligner; GATK: Genome analysis tool kit; GeT: Genome and transcriptome; LD: Linkage disequilibrium; LDcorSV: Linkage disequilibrium corrected by the structure and the relatedness; SOAP: Short oligonucleotide analysis package; SPS: Sucrose-phosphate synthase; SUS: Sucrose synthase

Acknowledgments
We thank the GeT-genotoul platform in Toulouse for DNA sequencing. Samples were previously obtained from a collaboration between Serge Tostain (IRD), Clément Agbangla (Université d'Abomey-Calavi, Cotonou, Benin), Ougbi Daïnou (Université d'Abomey-Calavi, Cotonou, Benin). We thank Marie Couderc and Cédric Mariac for advices during genomic bank preparation and sequencing. We thank Cécile Berthouly-Salazar and Philippe Cubry for their advices in carrying out data analysis.

Funding
This work was supported by a PhD grant to RA by the BID. This work was supported by the *Agence Nationale de la Recherche* with a grant to YV: ANR-13-BSV7–0017.

Availability of data and materials
Raw data (fastq) files are available from SRA (SRX3035965-SRX3035994). Code as a Additional file 1: Table S1.

Authors' contributions
RA, NS, HC, AD, GD, OF, KA, YV designed the study; NS generated the data; BR and OF contributed to analytic tools; RA performed the population genetic analyses; RA, NS, HC, AD, OF, KA, YV interpreted the results; ACT designed Fig. 3, RA, NS, KA and YV wrote the draft and the different authors contribute to its corrections. All authors read and approved the final manuscript.

Ethics approval and consent to participate
All samples were collected according to international rules. An agreement was signed between IRD and Université d'Abomey-Calavi (Benin) and sampling was performed together with local researchers. Plants were identified by Serge Tostain (yam specialist, IRD), Nora Scarcelli (yam specialist, IRD) and local yam farmers.

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Institut de Recherche pour le Développement, Université de Montpellier, Unité Mixte de Recherche Diversité Adaptation et Développement des Plantes (UMR DIADE), 911, avenue Agropolis, 34394 Montpellier, France. [2]Unité Mixte de Recherche Génétique Quantitative et Evolutive – Le Moulon, INRA – Univ. Paris-Sud – CNRS – AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France. [3]Faculté des Sciences et Techniques de Dassa, Laboratoire de Biotechnologie, Ressources Génétiques et Amélioration des Espèces Animales et Végétales (BIORAVE), Université d'Abomey, Dassa-Zoumè, Benin. [4]Centre International de la Recherche Agronomique pour le Développement, UMR AGAP, F-34398 Montpellier, France. [5]Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Lyon, France. [6]Université de Grenoble, Grenoble, France.

Received: 4 April 2017 Accepted: 2 October 2017
Published online: 12 October 2017

References
1. Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002;418:700–7.
2. Fuller DQ. Contrasting patterns in crop domestication and domestication rates: recent Archaeobotanical insights from the old world. Ann Bot. 2007; 100:903–24.
3. Purugganan MD, Fuller DQ. The nature of selection during plant domestication. Nature. 2009;457:843–8.
4. Harris DR. Foraging and Farming: The Evolution of Plant Exploitation. eds Harris, D. R. & Hillman, G. C. 1989. p. 11–26.
5. Purugganan MD, Fuller DQ. Archaeological data reveal slow rates of evolution during plant domestication. Evolution. 2011;65:171–83.
6. Meyer RS, Purugganan MD. Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet. 2013;14:840–52.
7. McKey D, Elias M, Pujol B, Duputié A. The evolutionary ecology of clonally propagated domesticated plants. New Phytol. 2010;186:318–32.
8. Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES. Genetic diversity and selection in the maize starch pathway. Proc Natl Acad Sci. 2002;99: 12959–62.
9. Sosso D, Luo D, Li Q-B, Sasse J, Yang J, Gendrot G, et al. Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. Nat Genet. 2015;47:1489–93.
10. Mignouna HD, Dansi A. Yam (Dioscorea Ssp.) domestication by the Nago and Fon ethnic groups in Benin. Genet Resour Crop Evol. 2003;50:519–28.
11. Hamon P. Structure, origine génétique des ignames cultivées du complexe Dioscorea cayenensis-rotundata et domestication des ignames en Afrique de l'Ouest. Paris: ORSTOM; 1987 p. 223. (Travaux et Documents Microédités; 47). Th.: Sci. Nat., Paris 11: Orsay. 1987/09/22. ISBN 2-7099-0923-5.
12. Terauchi R, Chikaleke VA, Thottappilly G, Hahn SK. Origin and phylogeny of Guinea yams as revealed by RFLP analysis of chloroplast DNA and nuclear ribosomal DNA. TAG Theor Appl Genet Theor Angew Genet. 1992;83:743–51.
13. Scarcelli N, Tostain S, Vigouroux Y, Agbangla C, Dainou O, Pham J-L. Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. Mol Ecol. 2006;15:2421–31.
14. Girma G, Hyma KE, Asiedu R, Mitchell SE, Gedil M, Spillane C. Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theor Appl Genet. 2014;127:1783–94.
15. Hamon P, Brizard J-P, Zoundjihékpon J, Duperray C, Borgel A. Étude des index d'ADN de huit espèces d'ignames (Dioscorea sp.) par cytométrie en flux. Can J Bot. 1992;70:996–1000.
16. Scarcelli N, Daïnou O, Agbangla C, Tostain S, Pham J-L. Segregation patterns of isozyme loci and microsatellite markers show the diploidy of African yam Dioscorea Rotundata (2n = 40). TAG Theor Appl Genet Theor Angew Genet. 2005;111:226–32.
17. Scarcelli N, Couderc M, Baco MN, Egah J, Vigouroux Y. Clonal diversity and estimation of relative clone age: application to agrobiodiversity of yam (Dioscorea Rotundata). BMC Plant Biol. 2013;13:178.
18. Hamon P, Dumont R, Zoundjihèkpon J, Tio-Touré B, Hamon S. Les ignames sauvages d'Afrique de l'ouest : caractéristiques morphologiques = Wild yams in West Africa : morphological characteristics - 010004065.pdf. 1995. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-05/010004065.pdf. Accessed 25 Jul 2016.

19. Shiwachi H, Ayankanmi T, Asiedu R. Effect of photoperiod on the development of inflorescences in white Guinea yam (Dioscorea Rotundata). Trop Sci. 2005;45:126–30.
20. Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A, et al. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour. 2014;14:1103–13.
21. Scarcelli N, Mariac C, Couvreur TLP, Faye A, Richard D, Sabot F, et al. Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? Mol Ecol Resour. 2015;16:434–45.
22. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10–2.
23. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinforma Oxf Engl. 2009;25:1754–60.
24. Sarah G, Homa F, Pointet S, Contreras S, Sabot F, Nabholz B, et al. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Mol Ecol Resour. 2016;17:565–580.
25. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19:1124–32.
26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
27. Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. Fast and efficient estimation of individual ancestry coefficients. Genetics. 2014;196:973–83.
28. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013;195:693–702.
29. Nei M. Molecular evolutionary genetics. New York: Columbia University Press; 1987.
30. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7:256–76.
31. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics. 2013;14:289.
32. Hill WG, Robertson A. Linkage disequilibrium in finite populations. TAG Theor Appl Genet Theor Angew Genet. 1968;38:226–31.
33. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. Bioinformatics. 2012;28:3326–3332.
34. Desrousseaux D, Sandron F, Siberchicot A, Cierco-Ayrolles C, Mangin B. LDcorSV: Linkage disequilibrium corrected by the structure and the relatedness. 2013. https://CRAN.R-project.org/package=LDcorSV.
35. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
36. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132:583–9.
37. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. ArXiv150404543 Q-Bio. 2015. http://arxiv.org/abs/1504.04543. Accessed 27 Nov 2015.
38. Mahalanobis PC. On the generalized distance in statistics. In: Proceedings National Institute of Science, India. 1936;2:49–55.
39. Dabney A, Storey JD. Qvalue: Q-value estimation for false discovery rate control. 2010. R package version 2.8.0. http://github.com/jdstorey/qvalue.
40. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6:e21800.
41. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinforma Oxf Engl. 2006;22:1600–7.
42. Sánchez C, Vielba JM, Ferro E, Covelo G, Solé A, Abarca D, et al. Two SCARECROW-LIKE genes are induced in response to exogenous auxin in rooting-competent cuttings of distantly related forest species. Tree Physiol. 2007;27:1459–70.
43. Heo J-O, Chang KS, Kim IA, Lee M-H, Lee SA, Song S-K, et al. Funneling of gibberellin signaling by the GRAS transcription regulator SCARECROW-LIKE 3 in the Arabidopsis root. Proc Natl Acad Sci. 2011;108:2166–71.
44. Baroja-Fernández E, Muñoz FJ, Li J, Bahaji A, Almagro G, Montero M, et al. Sucrose synthase activity in the sus1/sus2/sus3/sus4 Arabidopsis mutant is sufficient to support normal cellulose and starch production. Proc Natl Acad Sci. 2012;109:321–6.
45. Huber SC, Huber JL. Role and regulation of sucrose-phosphate synthase in higher plants. Annu Rev Plant Physiol Plant Mol Biol. 1996;47:431–44.
46. Hua J, Sakai H, Nourizadeh S, Chen QG, Bleecker AB, Ecker JR, et al. EIN4 and ERS2 are members of the putative ethylene receptor gene family in Arabidopsis. Plant Cell. 1998;10:1321–32.
47. Takemiya A, Inoue S, Doi M, Kinoshita T, Shimazaki K. Phototropins promote plant growth in response to blue light in low light environments. Plant Cell. 2005;17:1120–7.
48. Clotault J, Thuillet A-C, Buiron M, De Mita S, Couderc M, Haussmann BIG, et al. Evolutionary history of pearl millet (Pennisetum Glaucum [L.] R. Br.) and selection on flowering genes since its domestication. Mol Biol Evol. 2012;29:1199–212.
49. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, et al. The effects of artificial selection on the maize genome. Science. 2005;308:1310–4.
50. Kilian B, Ozkan H, Kohl J, von Haeseler A, Barale F, Deusch O, et al. Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. Mol Genet Genomics MGG. 2006;276:230–41.
51. Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, et al. Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol Biol Evol. 2007;24:1506–17.
52. Zoundjihekpon J, Hamon S, Tio-Touré B, Hamon P. First controlled progenies checked by isozymic markers in cultivated yams Dioscorea Cayenensis-Rotundata. Theor Appl Genet. 1994;88:1011–6.
53. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A. 2001;98:11479–84.
54. Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea Mays Ssp. Mays L.). Proc Natl Acad Sci U S A. 2001;98:9161–6.
55. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012;44:803–7.
56. Garris AJ, McCouch SR, Kresovich S. Population structure and its effect on Haplotype diversity and linkage disequilibrium surrounding the xa5 locus of Rice (Oryza Sativa L.). Genetics. 2003;165:759–69.
57. Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, Smith JSC, et al. An analysis of genetic diversity across the maize genome using microsatellites. Genetics. 2005;169:1617–30.
58. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44:808–11.
59. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philos Trans R Soc Lond Ser B Biol Sci. 2010;365:185–205.
60. Davies PJ. Ethylene in plant biology. Cell. 1993;72:11–2.
61. Campbell BC, Gilding EK, Mace ES, Tai S, Tao Y, Prentis PJ, et al. Domestication and the storage starch biosynthesis pathway: signatures of selection from a whole sorghum genome sequencing strategy. Plant Biotechnol J. 2016;14:2240–2253.
62. Hou J, Jiang Q, Hao C, Wang Y, Zhang H, Zhang X. Global selection on sucrose synthase haplotypes during a century of wheat breeding. Plant Physiol. 2014;164:1918–29.
63. Li J, Baroja-Fernández E, Bahaji A, Muñoz FJ, Ovecka M, Montero M, et al. Enhancing sucrose synthase activity results in increased levels of starch and ADP-glucose in maize (Zea Mays L.) seed endosperms. Plant Cell Physiol. 2013;54:282–94.
64. Quiles MJ. Regulation of the expression of chloroplast ndh genes by light intensity applied during oat plant growth. Plant Sci. 2005;168:1561–9.
65. Rumeau D, Bécuwe-Linka N, Beyly A, Louwagie M, Garin J, Peltier G. New subunits NDH-M, –N, and -O, encoded by nuclear genes, are essential for plastid Ndh complex functioning in higher plants. Plant Cell. 2005;17:219–32.
66. Quiles MJ, Cuello J. Association of ferredoxin-NADP oxidoreductase with the chloroplastic pyridine nucleotide dehydrogenase complex in barley leaves. Plant Physiol. 1998;117:235–44.
67. Quiles MJ. Stimulation of chlororespiration by heat and high light intensity in oat plants. Plant Cell Environ. 2006;29:1463–70.
68. Quiles MJ, López NI. Photoinhibition of photosystems I and II induced by exposure to high light intensity during oat plant growth: effects on the chloroplast NADH dehydrogenase complex. Plant Sci. 2004;166:815–23.
69. Miller G, Schlauch K, Tam R, Cortes D, Torres MA, Shulaev V, et al. The plant NADPH oxidase RBOHD mediates rapid systemic signaling in response to diverse stimuli. Sci Signal. 2009;2:ra45.
70. Baxter A, Mittler R, Suzuki N. ROS as key players in plant stress signalling. J Exp Bot. 2012;28:3326–3328.
71. Díaz M, de Haro V, Muñoz R, Quiles MJ. Chlororespiration is involved in the adaptation of Brassica plants to heat and high light intensity. Plant Cell Environ. 2007;30:1578–85.

Supporting information: Molecular basis of African yam domestication: analyses of selection point to starch biosynthesis, root development and photosynthesis related genes

Table

Table S1. Passport data of plant material collected from Benin.

Table S2. Metric information of data filtering and mapping

Table S3. Mean Nucleotide diversity (π) and polymorphism ($\ominus$)

Table S4. List of the contigs detected as selected by at least one method

Table S5. Remarkable candidate genes showing selection signature

Table S6. Gene Ontology (GO) terms significantly enriched (p-value≤0.05) among the 998 candidate contigs.

Table S7. Gene Ontology (GO) terms significantly enriched (p-value≤0.05) among the 81 candidates contigs detected by a least two methods.

Figure

Fig. S1. Cross-entropy calculated using sNMF (Frichot *et al.*, 2014) for K=1 to 6. Ten repetitions of the run were done.

Fig. S2. Intra-contigs linkage disequilibrium (LD) as a function of physical distance between SNPs pairs from 1% of all contigs.

Fig. S3. Venn Diagram comparing the candidate contigs obtained using the 4 methods.

Fig. S4. Distribution of Tajima's D value calculated for *D. abyssinica* (a) and *D. praehensilis* (b).

Fig. S5. Comparison of diversity lost.

Fig. S6. Variance explained by PCA axis (a) and distribution of Mahalanobis distance (b) from PCAdapt.

Fig. S7. Nucleotide diversity within five candidate contigs for cultivated and the wild species.

Table S1. Passport data of plant material collected in Benin.

The table lists the 30 individuals used in this study. The code A; P and C correspond respectively to the species *D. abyssinica*, *D. praehensilis* and *D. rotundata*. Each species was represented by 10 individuals. For each sample, its collected code, its local name, the species identity in the field, latitude and longitude, the village of origin are given.

Code	Name	Species	Latitude	Longitude	Location
A52	1461	*D. abyssinica*	10.48	2.5	Yarra
A433	1638		7.43	1.75	Gounoukouin
A571	1930		10.22	2.31	Gorobani
A62	1000		8.46	2.5	Idawa-Attata

A67	1419		8.37	1.98	Assaba
A420	1624		7.52	1.8	Lanhougbon-Amakpa
A467	1826		10.23	2.32	Gorobani
A537	1896		10.18	2.31	Gorobani
A3009	1484		11.05	1.51	Parc Pendjari
A3085	2017		8.37	2.02	Tchaguessé
P319	1655	*D. praehensilis*	8.37	2.02	Tchaguessé-Assaba's
P323	1660		8.37	2.02	Tchaguessé-Assaba's
P424	1629		7.5	1.82	Ammazoumé
P425	1630		7.5	1.82	Ammazoumé
P457	1816		6.39	2.16	Ahozon
P462	1821		6.39	2.16	Ahozon
P464	1823		6.39	2.16	Ahozon
P599	1959		7.5	1.82	Langbon
P624	1985		7.53	1.8	Langbon
P2990	1441		8.43	2.03	Djagballo
CR694	Ourou Yessingué CR1	*D. rotundata*	10.24	2.42	Wari
CR701	Anago		7.53	1.8	Langbon
CR702	Akpakodje		7.53	1.8	Langbon
CR703	Kodjewe		7.53	1.8	Langbon
CR668	Soagana		10.2	2.3	Gorobani
CR685	Morokorou 2		10.2	2.3	Gorobani
CR833	Odor 43		8.37	1.97	Assaba
CR844	Aklatchi Go38		7.43	1.75	Gounoukouin
CR849	Tabane Gu2		10.4	2.27	GuessouBani
CR869	Kpakara Y6		10.48	2.52	Yarra

Table S2. Metric information of data filtering and mapping

For each sample, the table provide information related to the total sequencing raw reads (Raw_Reads); the total recovered after filtering (Reads_Passing_Filter); the total reads that mapped on the transcriptome (Mapped_reads) and the percentage of reads that mapped (Mapped_Reads (%))

Sample	Raw_Reads	Reads_Passing_Filter	Mapped_reads	Mapped_Reads (%)
A52	19482684	17221766	2962684	17,20
A62	28003444	24434104	3183689	13,03
A67	88650184	78389652	14313408	18,26
A420	69939054	60659952	8602228	14,18

Sample	Raw_Reads	Reads_Passing_Filter	Mapped_reads	Mapped_Reads (%)
P319	9179494	8000486	996137	12,45
P424	41786178	36065282	4490937	12,45
P425	19259482	16935350	1947025	11,50
P457	51119260	43876208	4891141	11,15
P462	52009008	45061196	4893580	10,86
P323	50813916	43923142	5173532	11,78
P464	62352432	55000220	6410330	11,66
A467	66626892	59446908	6986083	11,75
A537	74226998	64749708	8728937	13,48
P599	77565138	68938948	7580072	11,00
P624	50259584	44771024	5343884	11,94
CR668	63344054	56618236	6690835	11,82
CR685	23473574	20346262	3007977	14,78
A433	9142726	7992442	1079547	13,51
CR833	34441020	30266518	3956029	13,07
CR844	50212974	43867422	5240980	11,95
CR694	45359084	40416586	2772969	6,86
CR849	59911430	52769044	6869498	13,02
CR869	77831534	69482998	6970085	10,03
P2990	33803026	29920070	4511755	15,08
A3009	39775674	34661188	5661700	16,33
A3085	36251032	31685990	3548737	11,20
CR701	9894214	8763076	1171514	13,37
CR702	24088444	21238732	2936548	13,83
CR703	37133718	32665614	4247848	13,00
A571	117595320	117595320	17797393	15,13

37

Table S3. Mean nucleotide diversity (π) and polymorphism ($\ominus$).
Diversity was calculated for 26 individuals (a), using groups without hybrids defined based on structure results, and for field identification using all 30 individuals (b). *** = p<0.001; ** = p<0.01; * = p<0.5; NS = p>0.5.

(a)

	π	$\ominus$
D. abyssinica (C)	3.97×10^{-3}	6.72×10^{-3}
D. praehensilis (P)	4.64×10^{-3}	8.55×10^{-3}
D. rotundata (A)	2.94×10^{-3}	4.79×10^{-3}
C/A ratio	0.74 **	0.72 ***
C/P ratio	0.64 ***	0.56 ***
A/P ratio	0.86 *	0.79 ***

(b)

	π	$\ominus$
D. abyssinica (C)	4.97×10^{-3}	1.06×10^{-2}
D. praehensilis (P)	4.60×10^{-3}	9.30×10^{-3}
D. rotundata (A)	2.94×10^{-3}	4.79×10^{-3}
C/A ratio	0.59 ***	0.45 ***
C/P ratio	0.64 ***	0.51 ***
A/P ratio	1.08 NS	1.14 ***

Table S4. List of the 998 selected Contigs

Provided in an excel file

The table lists the 998 selected contigs. The five contigs putted forward were in the top of the list. For each contigs the methods of selection were provided. The annotation was also provided if available for 467 contigs was also provided.

Table S5. Remarkable candidate genes showing selection signature

We present the five most interesting candidates genes involved in yam domestication. The genes were grouped in 4 categories according to their functions.

Contig	Genes names	Description and references
	Root development	
singlet___33548	*SCARCEROW-LIKE*	Earliest stages of adventitious root formation (Sanchez et al. 2007)
	Starch formation and storage	
Contig12423	Sucrose Synthase 4 (*SUS4*)	First step of starch formation: conversion of sucrose to fructose and UDP-glucose (Fu et al. 1995; Hou et al. 2003; Baroja-Fernánde et al. 2011; Li et al. 2012)
singlet___203418	Sucrose-Phosphate Synthase 1 (SPS1)	Sucrose biosynthesis and phosphate regulation (Huber et al. 1996)
	Phototropism activity	
Contig12174	Phototropin 2 (*Phot2*)	Promotes plant growth by controlling and integrating a variety of responses that optimize photosynthetic performance (Takemiya et al. 2005.
	Developement	
Contig11327	Ethylene Insensitive 4 (*EIN4*)	Ethylene receptor (Hua *et al.*, 1998). Ethylene is an endogenous growth regulator involved in many developmental processes, including seed germination, leaf and flower senescence, and fruit ripening (Abeles, 1992).

Table S6. Gene Ontology (GO) terms significantly enriched (p-value≤0.05) among the 998 candidate contigs.

GO.ID	Term	Annotated	Significant	Expected	P-value
GO:0003954	NADH dehydrogenase activity	32	19	1.29	4.1e-19
GO:0050136	NADH dehydrogenase (quinone) activity	32	19	1.29	4.1e-19
GO:0016655	oxidoreductase activity, acting on NADPH, quinone or similar compound as acceptor	42	19	1.69	3.7e-16
GO:0016651	oxidoreductase activity, acting on NADPH	54	19	2.17	9.7e-14
GO:0016491	oxidoreductase activity	667	47	26.85	1.0e-04
GO:0015002	heme-copper terminal oxidase activity	9	4	0.36	0.00028
GO:0016779	nucleotidyltransferase activity	101	12	4.07	0.00071
GO:0000030	mannosyltransferase activity	7	3	0.28	0.00200
GO:0048037	cofactor binding	87	10	3.5	0.00249
GO:0003824	catalytic activity	4550	207	183.13	0.00523
GO:0004540	ribonuclease activity	5	2	0.2	0.01490
GO:0015077	monovalent inorganic cation transmembran...	51	6	2.05	0.01593
GO:0034061	DNA polymerase activity	25	4	1.01	0.01670
GO:0016407	acetyltransferase activity	38	5	1.53	0.01729
GO:0051536	iron-sulfur cluster binding	53	6	2.13	0.01899
GO:0019829	cation-transporting ATPase activity	26	4	1.05	0.01913
GO:0016863	intramolecular oxidoreductase activity	6	2	0.24	0.02176
GO:0015078	hydrogen ion transmembrane transporter	29	4	1.17	0.02767
GO:0016860	intramolecular oxidoreductase activity	29	4	1.17	0.02767
GO:0016787	hydrolase activity	1477	73	59.45	0.03127
GO:0051540	metal cluster binding	62	6	2.5	0.03775

Table S7. Gene Ontology (GO) terms significantly enriched (p-value≤0.05) among the 81 candidate contigs detected by at least two methods

GO.ID	Term	Annotated	Significant	Expected	p-value
GO:0003954	NADH dehydrogenase activity	32	6	0.07	5.7e-11
GO:0050136	NADH dehydrogenase (quinone) activity	32	6	0.07	5.7e-11
GO:0016655	oxidoreductase activity, acting on NADH, quinone or similar compound as acceptor	42	6	0.1	3.2e-10
GO:0016651	oxidoreductase activity, acting on NADPH	54	6	0.12	1.6e-09
GO:0016491	oxidoreductase activity	667	7	1.53	0.00048
GO:0005198	structural molecule activity	234	4	0.54	0.00176
GO:0000030	mannosyltransferase activity	7	1	0.02	0.01599
GO:0048037	cofactor binding	87	2	0.2	0.01673
GO:0003824	catalytic activity	4550	15	10.46	0.03324

Figure S1. Cross-entropy calculated using sNMF (Frichot *et al.*, 2014) for K=1 to 6. Ten repetitions of the run were done.

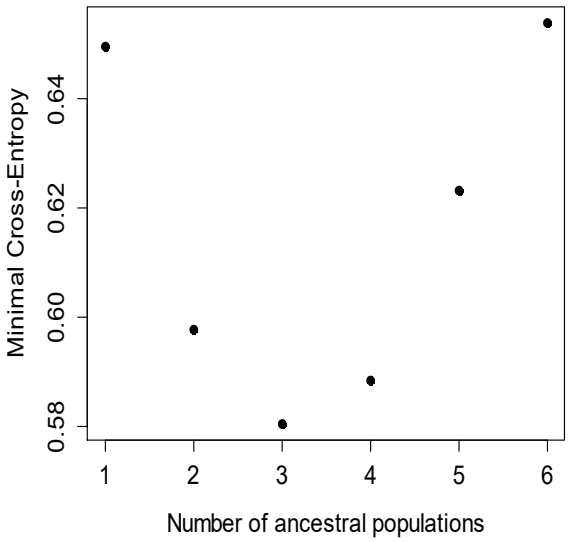


Figure S2. Intra-contigs linkage disequilibrium (LD) as a function of the physical distance between SNPs pairs from 1% of all contigs.

The grey line indicates that the LD decreased rapidly over a very short distance less than $r^2<0.1$ at 100 bp.

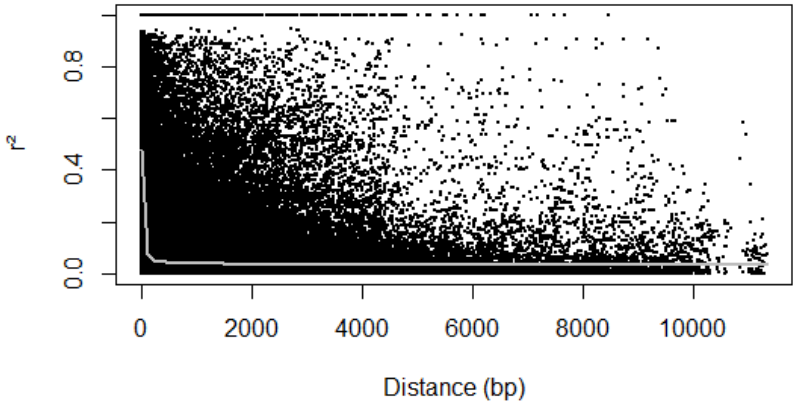


Figure S3. Venn Diagram comparing the candidate contigs obtained using the 4 methods.

Each oval corresponds to the contigs detected by one method. Blue = Tajima D's method; pink = PCAdapt method; purple = Diversity method; orange = F_{ST} method. Numbers where the ovals overlap correspond to the contigs detected by all related methods.

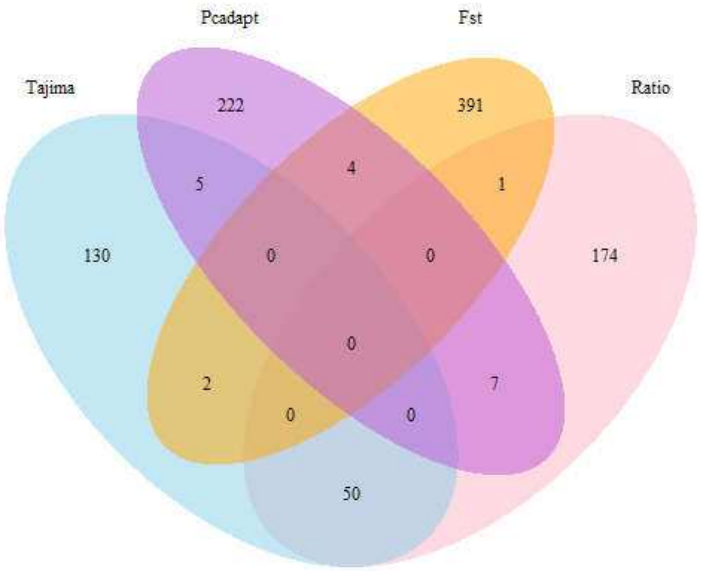


Figure S4. Distribution of Tajima's D value calculated for *D. abyssinica* (a) and *D. praehensilis* (b).

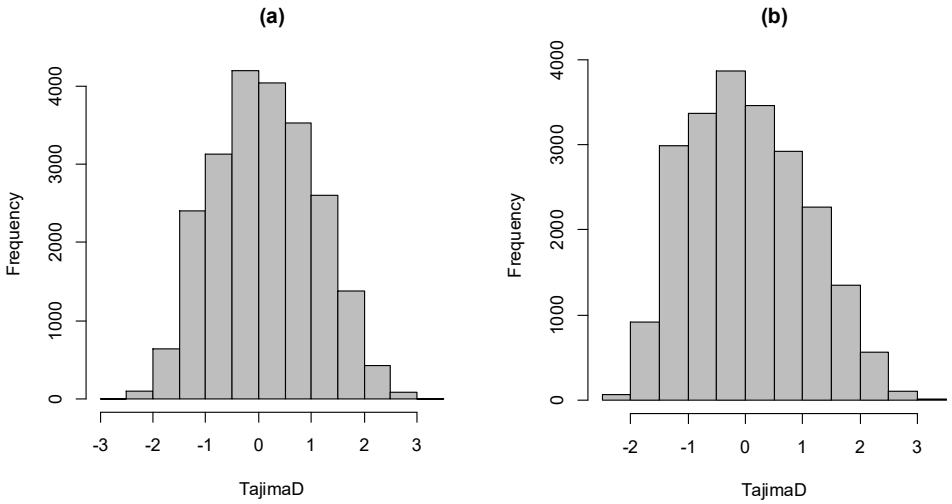


Figure S5. Correlation between the reduction of nucleotide diversity and the reduction of nucleotide polymorphism: r = 0.997, p < 0.001.

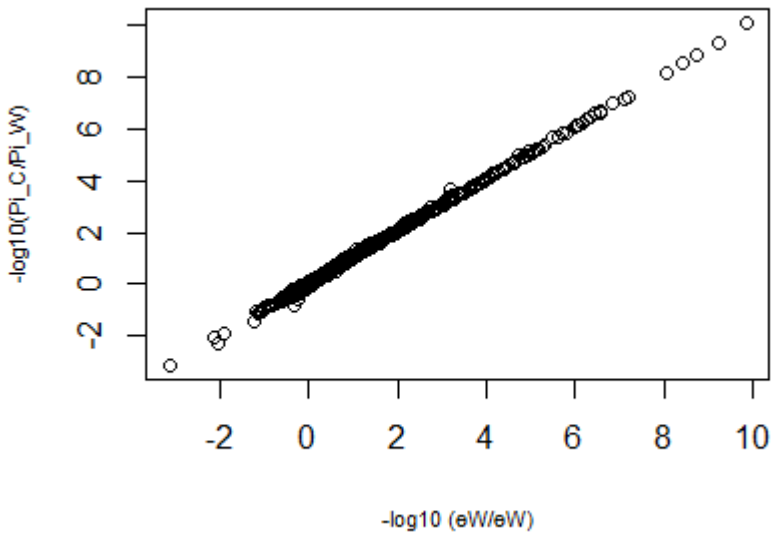


Figure S6. Variance explained by PCA axis (a) and distribution of Mahalanobis distance (b) from PCAdapt.

The two first axis were used (a). The distribution of Mahalanobis distance (b) fitted the normal distribution (b).

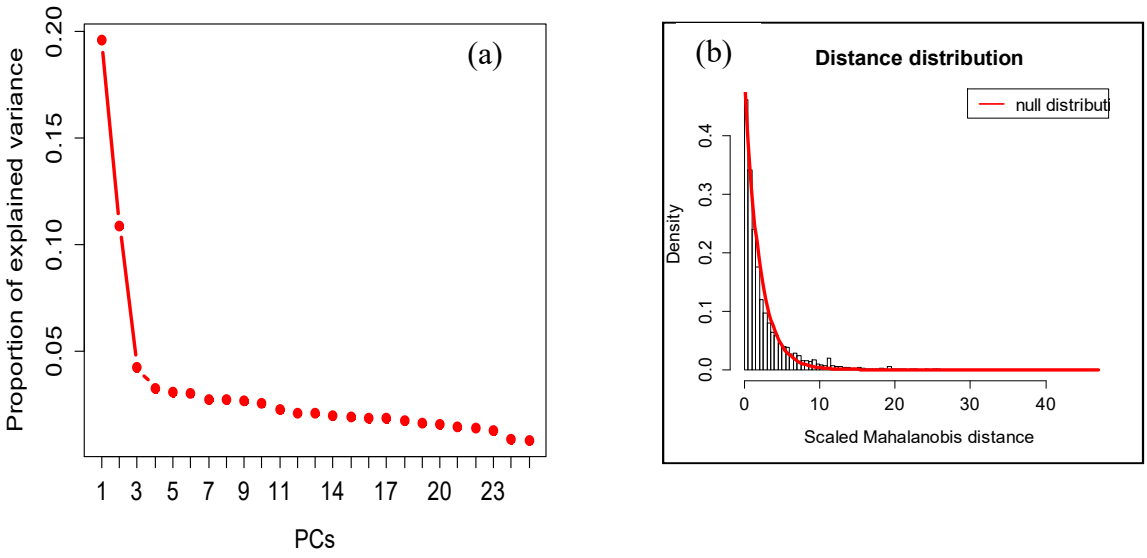
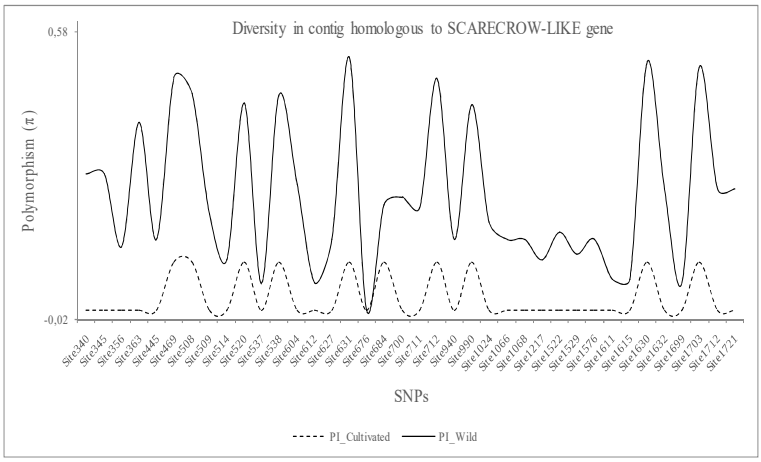
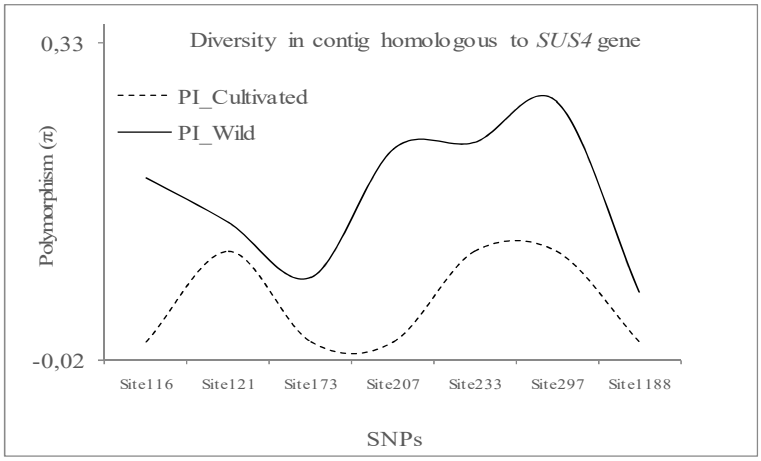


Figure S7. Comparison of the nucleotide diversity of the cultivated species to the mean nucleotide diversity of the two wild species for five interesting candidate contigs

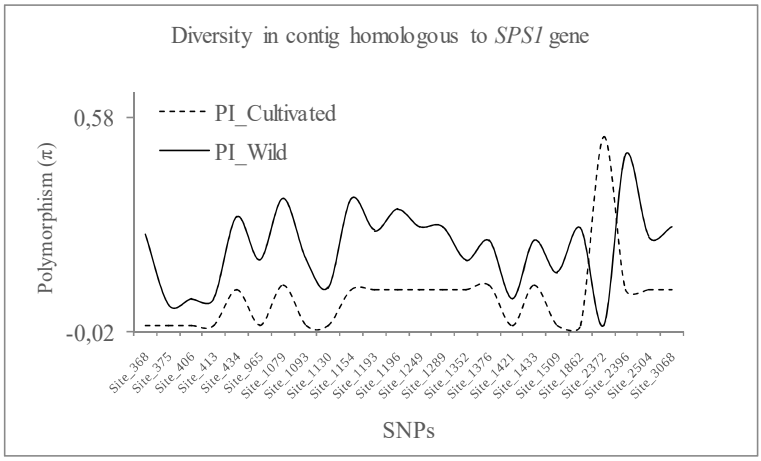
We plotted nucleotide diversity in 5 key candidate contigs homologous to *SCARECROW-LIKE* (a); *SUS4* (b); *SPS1* (c); *EIN4* (d) and *Phot2* (e) genes.



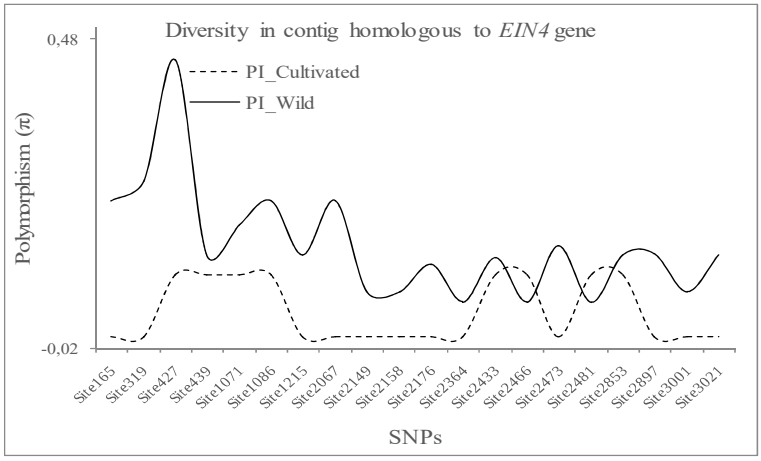
(a)



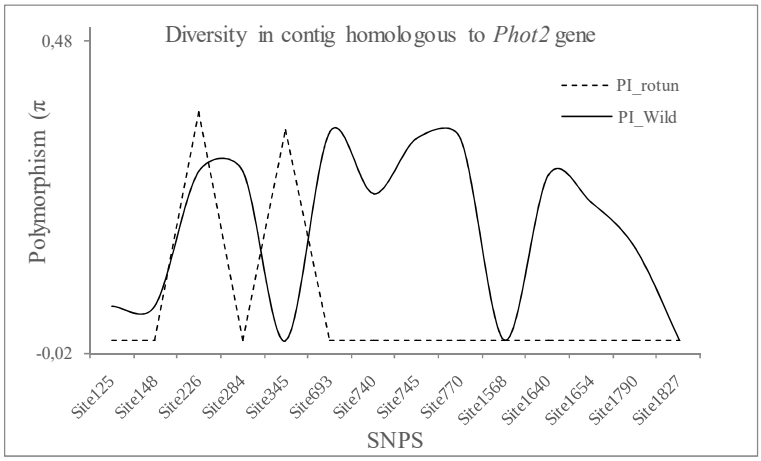
(b)



(c)


Diversity in contig homologous to *EIN4* gene
0,48
-0,02
Polymorphism (π)
PI_Cultivated
PI_Wild
Site165
Site319
Site427
Site439
Site1071
Site1086
Site1215
Site2067
Site2149
Site2158
Site2176
Site2364
Site2433
Site2466
Site2473
Site2481
Site2853
Site2897
Site3001
Site3021
SNPs

(d)


Diversity in contig homologous to *Phot2* gene
0,48
-0,02
Polymorphism (π
PI_rotun
PI_Wild
Site125
Site148
Site226
Site284
Site345
Site693
Site740
Site745
Site770
Site1568
Site1640
Site1654
Site1790
Site1827
SNPS

(e)

Supplementary file

We assess if the mapping of genomic DNA reads on a transcriptome reference could impact SNP calling in our special case.

Material and Methods.

We simulated a 100kb reference genome. From this genome, using BedTools (Quinlan and Hall, 2010) we deleted sequence in intervals of 400bp corresponding to introns, to simulate a transcriptome reference. The 400bp deletions were closed to what is observed rice (http://rice.plantbiology.msu.edu/analyses_facts.shtml).

Using wgsim (https://github.com/lh3/wgsim), we simulated synthetic paired-end reads based on the normal 100kb reference reads with an expected coverage of 30x. Wgsim was set for reads length=150, number of reads=10,000, the base quality was set at 30 and we used the option –e0.001 for allowing sequencing error. Finally, we simulated the mapping of genomic DNA reads on a transcriptom reference to validate our approach. We used default option of bwa aln-sampe to map our reads on the two reference genomes i.e. the normal reference and the truncated one. In each case, we used GATK HaplotypeCaller (Citation) to call SNPs variant.

Simulating mapping and SNP calling using synthetic data

The truncated reference lead to less read mapped as expected: roughly 35% of the read mapped (Table 1). But, whatever the reference used, i.e. the normal genome of reference and the truncated one, we did not call more variant in the truncated version or the normal version (Table 1). So we do have major issue of our mapping strategy. We might however, perhaps detected some effect if sequencing depth is low.

Table 1. Summary of mapping and SNP calling using simulated data

		Number of mapped reads		Number of SNPs	
ID_name	Total_Reads	Normal_Ref	Truncated_Ref	Normal_Ref	Truncated_Ref
ID1	20000	20000	6836	0	0
ID2	20000	20000	7062	0	0

Reference:

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297-303.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841-842.

Chapitre 3 : Inférence de l'histoire de la domestication de l'igname Africaine : *D. praehensilis* est à l'origine de *D. rotundata*

1. Contexte et objectifs

L'hypothèse actuelle est que l'origine de l'agriculture en Afrique serait située dans les régions de savane (Harlan 1976). Suivant cette hypothèse l'igname cultivée *D. rotundata* dériverait de l'espèce sauvage d'origine savanicole *D. abyssinica.* Des auteurs ont aussi postulé l'origine de l'igname à partir de l'espèce *D. praehensilis* inféodée aux régions de forêt (Coursey 1976). Nous avons testé ces hypothèses sur la base d'inférence démographique à l'aide de données génomiques.

2. Méthodes

Nous avons utilisé des modèles de génétique des populations basés sur la coalescence pour inférer les événements historiques. Nous avons testé l'origine à partir d'une ou des deux espèces sauvages au moyen de la méthode d'inférence du logiciel Fastsimcoal (Excoffier *et al.* 2013). Une étude de structuration de la diversité génétique des ignames d'Afrique a révélé 4 sous populations : *D. abyssinica* d'Afrique de l'Ouest ; *D. praehensilis* d'Afrique de l'Ouest ; toutes les ignames cultivées *D. rotundata* et un dernier groupe sauvage du Cameroun (Scarcelli et all, en Préparation). Sur la base de cette structuration, nous avons testé l'origine de l'igname basée sur ces quatre populations. Un total de 10 individus a été utilisé par groupe. Le spectre de fréquences (SFS), utilisant 100.000 SNPs pris au hasard, a été utilisé pour les analyses.

3. Principaux résultats

Le modèle donnant une divergence récente entre *D. rotundata* et *D. praehensilis* d'Afrique de l'ouest est le mieux supporté par les données. Cette divergence aurait lieu il y a environ 6500 générations. Ainsi, notre étude indique que l'igname aurait été domestiquée en zone de forêt, à partir de l'espèce sauvage *D. praehensilis*. Etant donné que *D. rotundata* pousse aujourd'hui aussi en zones de savane, il y a certainement eu des adaptations spécifiques à ce milieu après sa domestication.

Cette étude est présentée sous forme de chapitre intitulé « **Origin forest or savannah for African domesticated crop: a study case for the African cultivated yam *Dioscorea rotundata*** ». Elle fera l'objet d'un article plus conséquent (en préparation), par intégration de mes résultats à ceux obtenus par Nora Scarcelli, Philippe Cubry et Anne-Céline Thuillet, et abordera de manière plus exhaustive l'histoire de la domestication de l'igname africaine.

Origin forest or savannah for African domesticated crops: a study case for the African cultivated yam *Dioscorea rotundata*

Abstract

Several selection pressures have reshaped population dynamics of species in space and time, leading to their present genetic structure. New genetic technologies now offer unprecedented opportunities to study the past state of population dynamics based on their current genetic and genomic information. We undertake here, for the first time, the coalescent genetic modelling of the origin of the African cultivated yam *D. rotundata,* based on whole genome sequencing data. We showed that *D. rotundata and D. praehensilis* shared a recent ancestral population, based on a simple divergence scenario. Regarding the adaptation of *D. praehensilis* to the forested areas of West Africa allows invalidating the overall hypothesis that major African domesticated crops originated only from savannah.

Introduction

The origin of major African domesticated crops is hypothesized to have occurred in savannah, namely in the dry areas of Sahel (Harlan 1976). After domestication, agriculture and crops then later colonized forested areas. In support of this hypothesis, the wild forms of major African cereals, i.e. pearl millet (*Pennisetum glaucum*) and sorghum (*Sorghum bicolour*), are native to the savannah areas of Africa (Harlan 1976; Portères 1976; Wetterstrom 1998; Oumar *et al.* 2008). Moreover, the oldest archeological remains of domesticated crops were also found in the savannah. The oldest archeobotanical evidence of the pearl millet cultivation was found in northern Mali and dated at 4,500 BP, before present (Manning *et al.* 2011). Recently, an archaeological evidence of sorghum was discovered near Kassala in Eastern Sudan, dating from 4,000 BP to 4,500 (Winchell *et al.* 2017). However, for other domesticated crops like yam, the savannah hypothesis for its origin is still unproved even if yam is considered to fit this hypothesis (Vernier *et al.* 2003; Baco *et al.* 2004; Scarcelli *et al.* 2006b; Dumont *et al.* 2010). One difficulty for yam is the absence of any archaeological remains, and it is considered that yam domestication occurred ~5000 years ago (Coursey 1976).

African yam (*Dioscorea rotundata*) provides an interesting model to test the savannah hypothesis as the main origin of African crops. *D. rotundata* has two closed wild relatives: *D. abyssinica* and *D. praehensilis* (Scarcelli *et al.* 2017). Interestingly, these two wild species have distinct ecological distributions: *D. abyssinica* is found in the wooded savannah areas while *D. praehensilis* is found in tropical forested areas (Hamon *et al.* 1995). To address the question of

the savannah origin of yam, we could address the question: from which one of *D. abyssinica* and *D. Praehensilis* did the African cultivated yam share a common ancestor with?

We used genomic dataset and model based inference to answer this question. We inferred the most likely evolutionary scenario of the origin of the domestication of *D. rotundata*: (1) Did the domestication of *D. rotundata* originate from the savannah species *D. abyssinica*? (2) Did the domestication of *D. rotundata* originate from the African forest species *D. praehensilis*? or (3) Was *D. rotundata* an inter-specific hybridization product of the two species?

Materials and methods

Plant material

We sampled our dataset from a data constituted in 3,570,940 polarized SNP loci of 167 yam individual (Scarcelli *et al.* unpublished). This dataset presented a mean missing data of 7% and a mean sequencing depth of 7. We discarded SNP with more than 80% heterozygosity across all individuals. For the site frequency spectrum (SFS) calculation, we randomly selected 100,000 SNPs and 11 individuals per per subpopulation without considering hybrid individuals.

SFS estimation

Previous structure analysis (Akakpo *et al.* unpublished) revealed four ancestral clusters, three are formed by wild species: (1) Western Africa *D. abyssinica* including accessions *from* Ghana, Benin and Nigeria*;* (2) Western Africa *D. praehensilis* including acessions from Ghana, Benin and Nigeria and (3) wild yams from Central Africa. The fourth cluster included all the cultivated *D. rotundata* genotypes sampled from the four countries above. SFS calculation was performed for these four groups by randomly sampling 11 individuals. For each group, allelic frequencies were estimated using the 11 individual per subpopulation individuals. To reduce bias potentially due to the 7% of missing data, the estimated allelic frequencies were multiplied by 0.93 times (100% of data – 7% of missing data) the number of genotypes. We estimated multidimensional SFS between the different groups.

Evolutionary scenarios tested

We first tested topologies on the three wild groups. We built three alternative scenarios (Figure 1), with one wild group being the first to diverge from the others. The scenarios were kept simple with constant population size, and resized only occurring at a diverging event. When the best wild topology was found, we tested the best scenario for the cultivated groups considering a unique origin (Figure 2) and admixture models (Figure 3).

Estimation procedure and validation criterion

Fastsimcoal v2.5.2.21 (Excoffier *et al.* 2013) was used to estimate the demographic parameters from the SFS. We used multidimensional SFS in order to perform model comparisons using an Akaike Information Criterion (AIC). For each tested scenario, we performed up to 100 runs of estimation procedure. We used at least 250,000 and at more 1,000,000 simulations, and between 40 and 150 loops of ECM algorithm (with a criterion of convergence of 1%). We kept the parameters inferred for the best run, and as suggested by the author (Excoffier *et al.* 2013), we performed 100 new runs of 1,000,000 simulations that were used to better estimate the model likelihood. Using this likelihood (L), we estimated the AIC with the formula: *AIC=2k-2log10[L]* where k is the number of estimated parameters. We then calculated the Akaike's weight of evidence in favour of each model (Excoffier *et al.* 2013). Finally, confidence intervals for each parameter were estimated for the selected model using parametric bootstraps based on 20 predicted SFSs (Excoffier *et al.* 2013).

Results

The first test revealed which scenario of divergence was the best for the wild relatives (Figure 1). The scenario (model M2, Figure 1) with an early divergence of *D. abyssinica* was favoured (*p=1*, Table 1). Adding the unique cultivated group to the tree wild groups of this model, we found that the most likely demographic scenario (model M5, Figure 2) suggested a recent common ancestor for *D. praehensilis and D. rotundata (p=1,* Table 2*)*. On the contrary, all the tests performed for early admixture between the two wild species that might lead to the formation of *D. rotundata* domestication, gave less likely scenarios (Figure 3, Table 2).

Discussion

Our results revealed that *D. rotundata* and *D. praehensilis* shared a recent ancestral population. The best scenario we obtained corresponds to a simple divergence. Consequently, *D. rotundata* could not be depicted as the result of interspecific hybridization between the two wild species, as previously proposed (Coursey 1976; Hamon 1987; Terauchi *et al.* 1992).

Interestingly, *D. praehensilis* is adapted to the forested areas of West Africa. Therefore, our results invalidate the overall hypothesis that major domesticated crops originate from savannah (Harlan 1976; Marshall and Hildebrand 2002) and only diffuse later to forest zones. Yam, which is a major root and tuber crop in Africa, was demonstrated here to show an unexpected domestication process: originating first from the forest area, and then colonizing progressively the savannah area. An interesting and new question that have thus emerged from our study, is to understand how adaptation to the more pronounced dry season of African savannah was

acquired by cultivated yam during its evolution. Evidence has been brought for gene flow that might occur between the cultivated species *D. rotundata* and its wild relative from savannah areas *D. abyssinica* (Scarcelli *et al.* 2006a, b). To what extend such a gene flow would have favoured the transfer of adaptive alleles from the wild into the cultivated genetic background, allowing adaptation of yam to savannah, still requires further analysis.

References

Baco MN, Tostain S, Mongbo RL, et al (2004) Bulletin de Ressources Phytogntiques - Gestion dynamique de la diversité variétale des ignames cultivées (Dioscorea cayenensis–D. rotundata) dans la commune de Sinendé au nord Bénin. http://www.bioversityinternational.org/fileadmin/PGR/article-issue_139-art_4-lang_fr.html. Accessed 23 Jan 2018

Coursey DG (1976) The Origins and Domestication of Yams in Africa. In: Origins of African Plant Domestication, Reprint 2011. De Gruyter Mouton, Berlin, Boston

Dumont R, Zoundjihekpon J, Vernier P (2010) Origin and diversity of Dioscorea rotundata Poir yams. How African peasants' knowledge makes it possible for them to use wild biodiversity in farming. Cah Agric 19:255–261. doi: 10.1684/agr.2010.0411

Excoffier L, Dupanloup I, Huerta-Sánchez E, et al (2013) Robust Demographic Inference from Genomic and SNP Data. PLOS Genet 9:e1003905. doi: 10.1371/journal.pgen.1003905

Hamon P (1987) Structure, origine génétique des ignames cultivées du complexe Dioscorea cayenensis-rotundata et domestication des ignames en Afrique de l'Ouest.

Hamon P, Dumont R, Zoundjihèkpon J, et al (1995) Les ignames sauvages d'Afrique de l'ouest : caractéristiques morphologiques = Wild yams in West Africa : morphological characteristics - 010004065.pdf. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-05/010004065.pdf. Accessed 25 Jul 2016

Harlan JR (1976) Origins of African Plant Domestication, Reprint 2011. De Gruyter Mouton, Berlin, Boston

Manning K, Pelling R, Higham T, et al (2011) 4500-Year old domesticated pearl millet (Pennisetum glaucum) from the Tilemsi Valley, Mali: new insights into an alternative cereal domestication pathway. J Archaeol Sci 38:312–322. doi: 10.1016/j.jas.2010.09.007

Marshall F, Hildebrand E (2002) Cattle Before Crops: The Beginnings of Food Production in Africa. J World Prehistory 16:99–143. doi: 10.1023/A:1019954903395

Oumar I, Mariac C, Pham J-L, Vigouroux Y (2008) Phylogeny and origin of pearl millet (Pennisetum glaucum [L.] R. Br) as revealed by microsatellite loci. Theor Appl Genet 117:489–497. doi: 10.1007/s00122-008-0793-4

Portères R (1976) Origins of African Plant Domestication. Walter de Gruyter

Scarcelli N, Chaïr H, Causse S, et al (2017) Crop wild relative conservation: Wild yams are not that wild. Biol Conserv 210:325–333. doi: 10.1016/j.biocon.2017.05.001

Scarcelli N, Tostain S, Mariac C, et al (2006a) Genetic nature of yams (Dioscorea sp.) domesticated by farmers in Benin (West Africa). Genet Resour Crop Evol 53:121–130. doi: 10.1007/s10722-004-1950-5

Scarcelli N, Tostain S, Vigouroux Y, et al (2006b) Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. Mol Ecol 15:2421–2431. doi: 10.1111/j.1365-294X.2006.02958.x

Terauchi R, Chikaleke VA, Thottappilly G, Hahn SK (1992) Origin and phylogeny of Guinea yams as revealed by RFLP analysis of chloroplast DNA and nuclear ribosomal DNA. TAG Theor Appl Genet Theor Angew Genet 83:743–751. doi: 10.1007/BF00226693

Vernier P, Orkwor GC, Dossou AR (2003) Studies on Yam Domestication and Farmers' Practices in Benin and Nigeria. Outlook Agric 32:35–41. doi: 10.5367/000000003101294244

Wetterstrom W (1998) The origins of agriculture in Africa: With particular reference to Sorghum and Pearl Millet. https://www.academia.edu/5811773/The_origins_of_agriculture_in_Africa_With_particular_reference_to_Sorghum_and_Pearl_Millet. Accessed 15 Dec 2017

Winchell F, Stevens CJ, Murphy C, et al (2017) Evidence for Sorghum Domestication in Fourth Millennium BC Eastern Sudan: Spikelet Morphology from Ceramic Impressions of the Butana Group. Curr Anthropol 58:673–683. doi: 10.1086/693898

Title: Origin forest or savannah for African domesticated crops: a study case for the African cultivated yam *Dioscorea rotundata*

Tables

Table 1: Statistical parameters of the wild typology

Model	Max_likelihood	AIC	ExpDelta	Weight
M1	-441195.9	2031796	7.39E-49	7.39E-49
M2	-441147.8	2031575	1.00E+00	1.00E+00
M3	-441372.7	2032610	1.38E-225	1.38E-225

The ancestral population includes the three wild groups and the unique cultivated group. Max_likelihood is the maximum likelihood calculated across simulation; AIC is the Akaike Information Criterion computed for the model; ExpDelta corresponds to the exp(-0.5*DELTAi) where i is the ith model and DELTAi = AICi – AICmin; Weight is the Akaike's weight of evidence in favour of ith model.

Table 2: Statistical parameters of the four ancestral populations

Model	Max_likelihood	AIC	ExpDelta	Weight
M4	-515435	2373683	0	0
M5	-513062	2362754	1	1
M6	-515171	2372466	0	0
M7	-516083	2376674	0	0
M8	-513436	2364484	0	0
M9	-516899	2380431	0	0

The ancestral population includes the three wild groups and the unique cultivated group. Max_likelihood is the maximum likelihood calculated across simulation; AIC is the Akaike Information Criterion computed for the model; ExpDelta corresponds to the exp(-0.5*DELTAi)

where i is the ith model and DELTAi = AICi – AICmin; Weight is the Akaike's weight of evidence in favour of ith model.

Figures

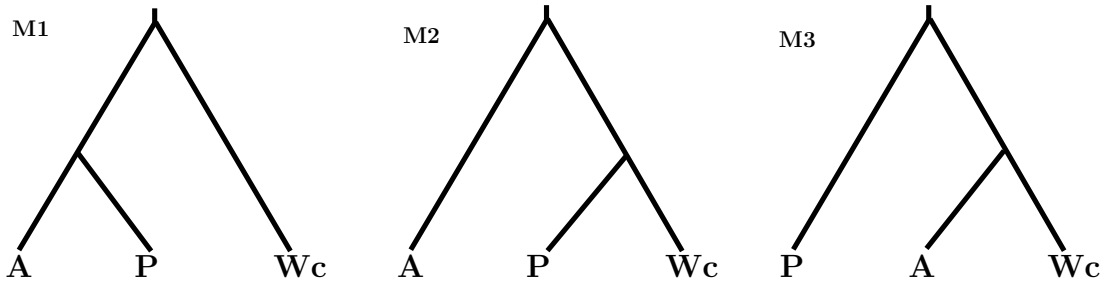


Figure 1: The three models tested to identify the best structure of the wild groups. M1 is the model considering an earlier divergence of Wild Cameroon; M2: earliest divergence of D. abyssinica and in M3, D. praehensilis first diverge. A = *D. abyssinica*; P = *D. Praehensilis* west; Wc = Wild *D. praehensilis* from Cameroun

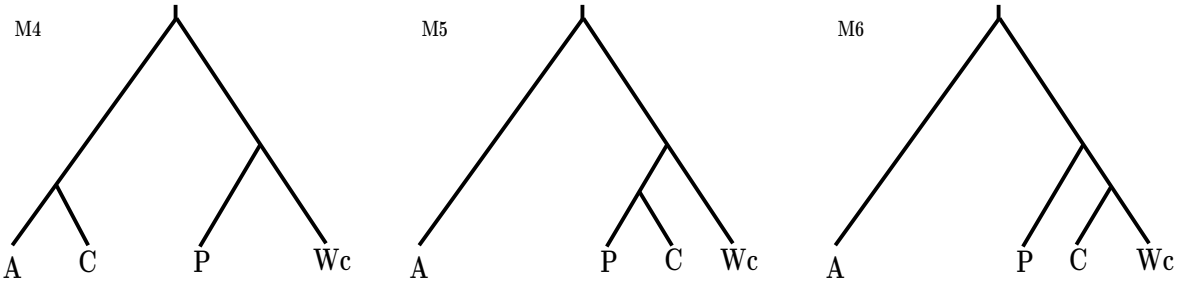


Figure 2: Models considering the unique cultivated group as diverged from each of the three wild group. M4: the cultivated group derived from *D. abyssinica*; M5: the cultivated group derived from *D. praehensilis* and M6: the cultivated group derived from wild Cameroon. A = *D. abyssinica*; P = *D. praehensilis* west; Wc = Wild *D. praehensilis* from Cameroun

C = *D. rotundata*; N = Resizing

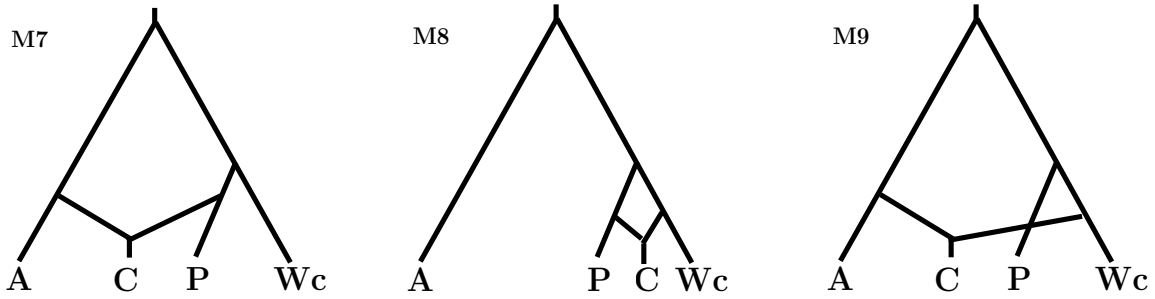


Figure 3: Models with admixture considering two origin of the cultivated group. M7: the cultivated group derived from an admixture of *D. abyssinica* and *D. praehensilis*; M8: the cultivated group derived from an admixture of *D. praehensilis* and Wild Cameroon and M9: *D. abyssinica* and wild Cameroon led to the cultivated *D. rotundata*. A = *D. abyssinica*; P = *D. Praehensilis* west; Wc = Wild *D. praehensilis* from Cameroun; C = *D. rotundata*.

Chapitre 4 : Analyse de la variabilité génomique en éléments transposables et autres éléments répétés et recherche de corrélations à différents environnements, chez l'igname Africaine.

1. Contexte et objectifs

Les génomes des eucaryotes possèdent une fraction d'éléments répétés (ER) dont font partie les éléments transposables (ET). Ces derniers, du fait de leur activité de multiplication et de transposition insertionnelle dans les génomes, peuvent contribuer à l'évolution du génome (Vicient et Casacuberta, 2017). L'insertion des ETs peut avoir des conséquences fonctionnelles et être associée à des adaptations : chez le maïs, l'insertion d'ETs est associée à l'adaptation des maïs aux milieux tempérés (Yang *et al.* 2013) ou la tolérance à différents stress abiotiques (Makarevitch *et al.* 2015). Nous avons estimé et caractérisé la fraction répétée du génome de l'igname Africaine. Puis, nous avons comparé la diversité des ER et en particulier des ETs entre l'espèce cultivée *D. rotundata* et ses deux espèces sauvages apparentées *D. abyssinica* et *D. praehensilis*.

2. Méthodes

Nous avons procédé à la construction *de novo* d'une base de données annotées en ERs, en exploitant des données de séquençage d'un génome représentatif de chaque espèce. La base de données générée est le bilan des résultats de prédiction obtenus à l'aide des logiciels Repdenovo (Chu *et al.* 2016) et Repark (Koch *et al.* 2014). Les données de séquençage des génomes d'un total de 180 individus ont ensuite été utilisées pour l'estimation de l'abondance des éléments prédits, au niveau de chacune des espèces étudiées. Des analyses de corrélation ont permis de comparer la diversité des ERs entre les trois espèces. Enfin, l'analyse de corrélation entre la composition de la fraction en ERs pour chacun des 180 individus et des données climatiques obtenues par localisation géographique (WorldClim database http://worldclim.org/bioclim) a été réalisée.

3. Principaux résultats

Ce travail a permis de générer une base de données des ERs de l'igname constituée d'un total de 5 047 ERs prédits. Cette base de données comprend 1 159 ETs annotés dont 70% des éléments consensus uniques sont des éléments de classe I (LTR, LINE, SINE, TRIM) avec 50% de rétrotransposons à LTR. Les génomes des trois espèces présentent des abondances relatives en chacun des 1 159 ETs annotés relativement stables. Cependant, 11 ETs (dont 10 de classe I)

ont été identifiés comme différentiellement représentés entre les trois espèces. La structure de la diversité génétique des 180 individus évaluée selon leur profil d'abondance en ETs présente une corrélation significative avec celle définie par les variants SNPs. Enfin, 46% des ERs prédits, comprenant 593 ETs annotés dont 8 des 11 ETs qui varient en abondance entre les trois espèces, présentent une fréquence corrélée de manière significative à certaines des variables environnementales intégrées à notre étude.

4. Conclusion

Cette étude a contribué de manière originale à la caractérisation génomique de l'igname Africaine en s'intéressant spécifiquement à la fraction répétée de son génome. Elle a permis de démontrer l'intérêt des données de séquençage haut débit dans l'identification des ERs d'un génome, même sans assemblage. Nous avons ainsi identifié une contribution significative des ERs à la structuration de la diversité de l'igname Africaine en lien avec sa répartition géographique et les données climatiques.

L'ensemble de cette étude a donné lieu à la rédaction de l'article avec un objectif de soumission à *Frontiers in Plant Science* : « **Evaluating the relationships between transposable element content, genetic diversity and geographical distribution in African yam (Dioscorea rotundata) and wild relatives** ».
Roland Akakpo, Florian Maumus, Nora Scarcelli, Hana Chaïr, Christine Tranchant, Alexandre Dansi, Gustave Djedatin, Yves Vigouroux and Karine Alix.

1

2

Evaluating the relationships between transposable element content, genetic diversity and geographical distribution in African yam (*Dioscorea rotundata* and wild relatives)

Running title: **Diversity of *Dioscorea* transposable elements**

Roland Akakpo[1,2], Florian Maumus[4], Nora Scarcelli[2], Hana Chaïr[5], Christine Tranchant[2], Alexandre Dansi[3], Gustave Djedatin[3], Yves Vigouroux[2], Karine Alix[1]

[1]GQE – Le Moulon, INRA, Univ Paris Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France ;

[2]Institut de Recherche pour le Développement, Université de Montpellier, Unité Mixte de Recherche Diversité Adaptation et Développement des Plantes (UMR DIADE), Montpellier, France ;

[3]Université d'Abomey, Faculté des Sciences et Techniques de Dassa, Laboratoire de Biotechnologie, Ressources Génétiques et Amélioration des Espèces Animales et Végétales (BIORAVE), Dassa-Zoumè, Benin ;

[4]Unité de Recherche Génomique-Info (URGI) – INRA, Route de Saint-Cyr RD 10 78026 Versailles cedex

[5]Centre International de la Recherche Agronomique pour le Développement, UMR AGAP, F-34398 Montpellier, France ;

Authors' emails

1. Roland AKAKPO: roland.akakpo@ird.fr
2. Florian MAUMUS : florian.maumus@inra.fr
3. Nora SCARCELLI : nora.scarcelli@ird.fr
4. Hana CHAIR: hana.chair@cirad.fr
5. Christine TRANCHANT : christine.tranchant@ird.fr
6. Alexandre DANSI : adansi2001@gmail.com
7. Gustave DJEDATIN : djedatingustave@yahoo.fr
8. Yves VIGOUROUX: yves.vigouroux@ird.fr

Corresponding author:

Karine ALIX

GQE – Le Moulon, INRA, Univ Paris Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

Phone: 33(0)1 69 33 23 72 ; email: karine.alix@inra.fr

Abstract

Repetitive sequences are the most abundant component of plant genomes and can dramatically affect genome evolution and genetic variation. Understanding their evolutionary dynamics might thus contribute to understand plant genome evolution and function. The availability of new high-throughput sequencing techniques facilitates the implementation of low-cost analyses to uncover this knowledge, even in orphan crop such as African yam, without need of any reference genome.

Using paired-end sequence data and *de novo* repeat inferring strategies, we identified and annotated the most repetitive elements in the African yam genome, based on the survey of three *Dioscorea* species: the cultivated yam *D. rotundata* and its two wild relatives *D. praehensilis* and *D. abyssinica*. We identified a total of 5,047 unique consensus repeat elements (REs), which represent ~40% of the yam genome, including a total of 1,159 different transposable elements (TEs). These genomic resources were used to characterize and compare TE contents between wild and cultivated yams, in order to evaluate the global contribution of repeat elements to the structure of genetic diversity and genome shaping during evolution and adaptation of African yam.

The qualitative composition of TE fractions was very similar between the three *Dioscorea* species and only slight differences in TE frequencies were detected when comparing the three species. Nevertheless, relative abundances of TEs allowed partitioning the 180 *Dioscorea* accessions under survey, in accordance with the structure of genetic diversity we determined using SNP variants. A few retrotransposons differentiated one species to another because of highly significant variation in their copy numbers. Correlation between relative abundances of each predicted RE and various bioclimatic data indicated that the whole fraction of REs was strongly associated with environmental factors. Our findings have thus highlighted the contribution of genome size, via RE content, and retrotransposons in the dynamic process of adaptation in African yam.

Keywords: *Dioscorea spp.*, population genomics, NGS, repetitive sequences; transposable element, bioclimatic variable, adaptation.

Introduction

Repetitive sequences represent large portions of all plant genomes. These repeats elements include tandemly-repeated sequences such as telomeric sequences, micro- or mini-satellites and satellite DNAs but also ribosomal genes – reviewed by Mehrotra and Goyal, (2014) and repetitive sequences dispersed throughout the genome. They mainly correspond to transposable elements (transposons) and are well documented as representing the repeat fraction of the

plant genome. Of particular interest are transposable elements (TEs) because of their capacity to replicate and reach very high copy numbers in the genome, as well as to move across the host genome (Bennetzen, 2000). Thanks to such dynamics specificities, TEs can indeed be considered as the main players in plant genome size variation (Bennetzen and Wang, 2014) in addition to polyploidy (Alix *et al.*, 2017) as well as in chromosome and karyotype evolution (Li *et al.*, 2017). Their major contribution to plant genome diversity and gene expression variation has also been recurrently reported (for review, Vicient and Casacuberta, 2017).

TEs are subdivided into two major classes characterized by their mechanism of transposition and the nature of their transposition intermediate (Finnegan, 1989; Wicker *et al.*, 2007). The class I elements (confined in eukaryotes) are retrotransposons that transpose via a so-called ''copy-and-paste'' mechanism via a RNA intermediate obtained after transcription of the element. This class includes different TE orders from which we can cite Long terminal repeat retrotransposons (LTR-RTs), Long interspersed nuclear elements (LINEs) and Short interspersed nuclear elements (SINEs) (Xiong and Eickbush, 1990). Terminal-repeat retrotransposons in miniature (TRIM) represent particular class I non-autonomous retrotransposons (Witte *et al.*, 2001). The class II elements are characterized by the ''cut-and-paste'' mechanism that allows the element to transpose via a DNA intermediate cut out from the host DNA prior to insertion (Wang *et al.*, 2012). Class II comprises the main TIR order (transposons characterized by terminal inverted repeats) that is then divided in several superfamilies (e.g. *Tc1-Mariner*, *hAT*, *CACTA*, *Mutator, PIF-Harbinger*) according to the similarity of transposases, the element-encoded protein that catalyzes transposition and integration (Flavell, 1995). The Helitrons that mobilize single-stranded DNA through a rolling-circle mechanism for their replication (Kapitonov and Jurka, 2001) also belong to the class II like the Maverick elements (Pritham *et al.*, 2007). Finally, Miniature inverted-repeat transposable elements (MITEs) are considered as a special type of class II non-autonomous TIR elements, with specificities in size (a few hundred base pairs), copy number (several thousand copies in Poaceae genomes) and location in the genome as they are usually associated with the gene space of the host genome (Sarilar *et al.*, 2011; Stelmach *et al.*, 2017). Preponderance of TEs in all plant genomes and preferential association of specific TE families with protein-coding genes have thus led to the hypothesis of a major role of TEs in genome shaping as well as gene function or regulation of gene expression.

Correlation between plant genome size and genome TE content has been observed since decades (Pearce *et al.*, 1996; Sanmiguel and Bennetzen, 1998). Trends for accumulation of specific TE orders / super families have been demonstrated as quite similar in closely related lineages: in plant genomes, class I retrotransposons and particularly LTR-RTs are the most represented TEs (reviewed by Vitte *et al.*, 2014). Interestingly, LTR-RTs were also shown as

being the most abundant source of genome size variation in all the plant genomes investigated. For example, the genome of sorghum (730 Mb) contains a total of 62% TEs with 55% retrotransposons that contributed mainly to the expansion of the sorghum genome in comparison to rice (430 Mb and 26% LTR-RTs) while the number of gene families in sorghum and rice are the same (Paterson *et al.*, 2009). For additional comparison within the Poaceae, TE content was estimated as 85% of the maize genome (2,3 Mb) including a total of 75% LTR-RTs (Schnable *et al.*, 2009). Another particularity of TEs is their ability to move across the genome that may generate insertional mutations and a large variety of structural modifications at the origin of the huge plant genome diversity. Structural variants like copy number variations (CNV) and presence/absence variations (PAV) may find their origins in TEs (Saxena *et al.*, 2014) due to the role of TEs in mediating unequal and illegitimate recombination (Lisch, 2013). Movements of TEs can also affect function of neighboring genes, usually resulting in detrimental mutations but sometimes in highly beneficial modifications that are then selected (domesticated) along evolution (Sinzelle *et al.*, 2009). Such cases have been reported in response to selection, notably during plant domestication and breeding (Vitte *et al.*, 2014) and more generally in a context of plant adaptation (Lisch, 2013). TEs have been demonstrated to be reactivated under stress conditions in all eukaryotes (Horváth *et al.*, 2017) and notably in plants in response to highly diverse stresses (Grandbastien, 1998; Negi *et al.*, 2016; Parisod *et al.*, 2010) that may indicate a specific role of TEs in adaptation to constraints. To support this hypothesis, several studies have provided evidence that TEs/repeat elements and the molecular changes they trigger contribute to phenotypic variation (Huang *et al.*, 2018, 9; Jian *et al.*, 2017; Yang *et al.*, 2013) favoring adaptation to environmental and climatic patterns (Diez *et al.*, 2014; Makarevitch *et al.*, 2015; Song and Cao, 2017). Accordingly, it appears now necessary to conduct population dynamics studies dedicated to TEs (Petrov *et al.*, 2011) to add to our knowledge of their contribution to the plant genome evolution and function. Only few studies have indeed focused on population genomics of TEs in plants (Ågren *et al.*, 2016; Bardil *et al.*, 2015; Domb *et al.*, 2017) while such studies should contribute to evaluate the role of TEs in shaping population structure along evolution, and particularly their role in modeling the genome of plant populations adapted to different environments.

The present study focused on the characterization of the repetitive fraction of the genome of the cultivated African yam *Dioscorea rotundata* in comparison with its two closest wild relatives *D. abyssinica* and *D. praehensilis*. These three species are all diploids (2n = 40) (Girma *et al.*, 2014; Hamon *et al.*, 1992) and the size of the *D. rotundata* genome was recently estimated of 594 Mb (Tamiru *et al.*, 2017). Yam is the second most widely grown tuber crop in Africa after cassava, and *D. rotundata* is the most widely cultivated species providing staple food for over 100 million people (Mignouna and Dansi, 2003). However, compared to other major cultivated plants, many genomic features of yam are still poorly characterized and

notably, the repetitive fraction of its genome has been only marginally investigated so far. The recent analysis of the complete genome sequence of *D. rotundata* quickly mentioned that TEs may represent ~46% of the genome (Tamiru *et al.*, 2017). In particular, domestication of yam in Africa is still unclear and the relationships between the African cultivated yam *D. rotundata* and its two wild relatives have not been elucidated yet. Only few studies analyzed the structure of genetic diversity in African yam (Girma *et al.*, 2014; Loko *et al.*, 2016; Ngwe *et al.*, 2015; Scarcelli *et al.*, 2017) or have addressed the question of the molecular basis of domestication (Akakpo *et al.*, 2017; Scarcelli *et al.*, 2013), but nothing has been performed concerning the characterization of the repetitive elements of the yam genome. The contribution of REs and notably of TEs to the genetic diversity of African yam remains to be addressed.

We initiated the present study to provide new insights into African yam genome evolution and the structure of its genetic diversity. We exploited high-throughput genome sequencing data generated from 180 yam accessions and used appropriate bioinformatic tools to predict the repeat fraction of the yam genome and create *de novo* a *Dioscorea* TE database. Our main objectives were: (i) to characterize and compare TE contents between the three *Dioscorea* species surveyed; (ii) to evaluate the relationship between variation in TE contents and population structure assessed using SNPs; and (iii) establish correlations between TE contents and bioclimatic data that type various geographical locations. The contribution of TEs to the genetic diversity of African yam and the role of REs and particularly of TEs in shaping the genome of cultivated and wild yams in different environments are discussed.

Materials and Methods

Plant material and genome sequencing data

For *de novo* creation of the repeat element (RE) database, we used whole genome sequencing data from one individual of the three *Dioscorea* species (Akakpo *et al.* 2017): the cultivated yam *D. rotundata* and its two closest wild relatives, *D. abyssinica* and *D. praehensilis.* The average coverage was 30× for each of these three individuals (accession codes: CR869 for *D. rotundata*; A571 for *D. abyssinica;* and P464 for *D. praehensilis*); the sequence data were Illumina 2× 100bp paired-reads corresponding to a total of 220 million reads.

To characterize the TE content of each of the three yam species surveyed, we analysed the re-sequencing data of a collection of 180 single plants, including 85 individuals from *D. rotundata*, 38 from *D. abyssinica*, and 57 from *D. praehensilis*. Plant material and sequencing data are fully described in Scarcelli *et al.* (submitted) and available in GenBank (SRA XXXX-XXXX).

De novo creation of a *Dioscorea* transposable element database

We used Repdenovo (Chu *et al.* 2016) and Repark (Koch *et al.* 2014) to infer *de novo* repeat motifs from each of the three *Dioscorea* species independently (represented by one accession each). Both programs were performed separately to predict repeat elements (REs) using their default settings, and predictions were fused into one single file per species. Each file of predicted REs was then cleaned up and filtered as follows: we filtered out i) all repeat sequences that were less than 150 bp-long as recommended by Koch *et al.* (2014) (this filter was particularly relevant for RepARK consensuses that were enriched in short REs, i.e. 50-149 bp-long, due to extensive fragmentation of the libraries); ii) putative chloroplast and mitochondrial sequences. For chloroplast sequences, the repeats were mapped onto the yam chloroplast reference genome (NC_024170.1), using BLASTN with 80% of homology. For the mitochondrial sequences, the repeats were mapped onto the mitochondrial reference genomes of rice (BA000029.3) and wheat (NC_007579.1) using BLASTN with 80% of homology. The three filtered RE libraries created from the three *Dioscorea* species were then combined by removing redundancy (95% coverage, 98% identity): a unique yam RE file was obtained. From the unique consensus RE file, as proposed by Flutre *et al.* (2011), annotation and classification of transposable elements and other repeats were performed using the program PASTEClassifier (Hoede *et al.*, 2014).

Comparative analysis of TE contents among the three *Dioscorea* species

To estimate the full contents in repeats (all predicted REs or only annotated TEs) for each of the 180 *Dioscorea* individual genomes surveyed, we had to choose the best tool that allowed to optimally align short genome sequence reads to the *de novo* yam RE library. The two mappers BWA and Stampy were previously highlighted for their high efficiency (Tian *et al.*, 2016). We tested these two tools with their default settings, by mapping the sets of short reads generated for the three individuals used for the creation of the yam RE library, against the *de novo* yam RE library itself. The "idxstat" option of Samtools (Li *et al.*, 2009) was used to report the total number of RE hits per RE family for each individual, and comparison of RE content estimates obtained with each mapper was performed.

Estimated raw TE contents were obtained for the collection of 180 yam individuals by mapping their short genome sequence reads to the *de novo* yam TE database using Stampy-1.0.31 (Lunter and Goodson, 2011). To allow comparisons, the estimated raw TE contents were corrected for differences in the total number of reads per library and the length of the reference TE sequence, by calculating reads per kilobase per million reads (RPKM). The RPKM was calculated following Tenaillon *et al.* (2011) formula: $RPKM = H_i / (L_i \times M \times 10^{-6})$, where H_i is the number of reads mapped onto the TE sequence i; L_i is the length in kilobases (kb) of the

TE sequence i and M is the total number of reads mapped onto the whole TE fraction. RPKM values represent normalized TE content estimates and were used for further comparative analysis.

To compare repeat contents between the three *Dioscorea* species, we first used RPKM values estimated for the elements of the annotated TE database only. To characterize the variability in TE content among the 180 yam individuals, we performed Pearson correlation tests on the log scale of RPKM between species. Then, to observe potential difference between species in relative TE abundances, we performed a Wilcoxon paired test using RPKM for all the TEs. We also calculated TE based difference between pairs of species and we used a threshold of 5,000 RPKM units to identify TEs that might be overrepresented in one species compared to another. Presence of differentiation might allow understanding the contribution of TEs in the genomic structure of population. All these different analyses were performed using R scripts.

Variability of TE contents in relation to the structure of genetic diversity

We analysed the association of TE content variation with population structure. First, a principal component analysis (PCA) based on the RKPM data was performed to assess the differences in relative TE abundances between the 180 individuals. Then, population structure was determined on the basis of SNP data identified using previously described methods (Akakpo *et al.*, 2017). Briefly, best quality sequencing reads generated from the 180 accessions were mapped onto the *D. rotundata* transcriptome reference (Sarah *et al.*, 2016), using default options in BWA-mem V0.7.5a–r405 (Li and Durbin, 2009). SNP calling was then performed using default options and the "-rf BadCigar" options in GATK-UnifiedGenotyper (McKenna *et al.*, 2010). SNPs were filtered for low missing rate <4%, minor allele frequencies MAF > 5% and a mean depth ≥ 2. Based on SNP data, a genome-wide genetic diversity analysis was performed using default options in the program ADMIXTURE (Alexander *et al.*, 2009). For K sub-group ranging between 1 and 10, we identified the most likely number of K as the one minimizing the cross-entropy criterion. Consequently, relationships between TE contents and structure of population were tested for the more reliable K groups by performing a regression analysis through the linear model: $Y_i = x_0 + {}_1q_1 + {}_2q_2 + \ldots\ldots + {}_nqn$ (where Y_i represents the accession coordinates related to the ith PC from the Principal Component Analysis performed above based on TE contents, and qi corresponds to the ancestries from the structure analysis in population i). To determine if TE contents may fit the structure of genetic diversity, we compared the genetic groups obtained using ANOVA with those defined on the basis of genomic TE contents generated by PCA (i.e. the coordinates of accessions from PCA).

Relationships between total repeat element contents and bioclimatic data

From the WorldClim database (http://worldclim.org/bioclim), geographical coordinates were used to retrieve data for ninety bioclimatic variables for each of the 180 yam accessions (Diez *et al.*, 2014; Hijmans *et al.*, 2005). The ninety bioclimatic variables correspond to different measures of variation of temperature and precipitation (Supplementary Table S1).

To assess the relationships between RE contents (represented by RPKM) and the different bioclimatic variables, we performed Pearson's correlation test using R scripts. All the classified and unclassified repeat elements (representing 5,047 REs) were analyzed. The ninety climatic variables characterizing each geographical position of the 180 yam accessions were summarized by performing a Principal Component Analysis (PCA). The two first axes of PCA were respectively correlated with RE contents; the p-values associated to the correlation tests were recovered. The R package qvalue (Dabney and Storey, 2010) was used to compute corresponding q-values. A false discovery rate (FDR) of 5% was used to identify all REs putatively associated with environmental variables.

Results

De novo creation of a transposable element database for *Dioscorea*

The two repeat prediction softwares allowed inferring a total of 31,342 repeat sequences for the three yam species surveyed (Supplementary Table S2). Our filtering strategies and the merging of the two batches of repeat sequences (i.e. from the two softwares) allowed identifying a total of 7,275 REs. Then, the filtered RE libraries created for the three *Dioscorea* species were merged by removing redundancy: a total of 5,047 unique repeat elements were recovered, providing a list of 1,159 unique (consensus) transposable elements (TEs) that constitute the *de novo Dioscorea* TE database. In detail, we annotated a total of 527 TEs in *D. abyssinica*, 588 in *D. praehensilis* and 552 in *D. rotundata*, with lengths ranging from 150 bp to more than 21.9 kb (Table 1). The composition of the TE fraction, in terms of TE categories, is identical between the three *Dioscorea* species: 70% of the consensus TEs is class I retrotransposons with 50% of unique LTR-retrotransposons, and 30% of consensus TEs are class II elements with a majority of TIR transposons (25%). The composition of the TE fraction predicted for the three *Dioscorea* species is provided in Figure 1. It can be noticed that more than half of the predicted REs were annotated as other repeats than transposable elements, with a majority of REs that were not classified (noCat in Table 1).

Global TE contents were relatively stable among the three *Dioscorea* species

To compare TE content variability within the three *Dioscorea* species, we used a set of 180 yam accessions and we first tested and optimised the two mapping tools BWA-MEM and Stampy. The two mapping tools were used to search for repeat contents within the genome of the three species *D. abyssinica, D. praehensilis* and *D. rotundata* (using the total sequencing reads from the three individuals used for the creation of the yam RE library). The mean percentage of reads that mapped onto all the predicted repeat elements was 1.49-fold higher using Stampy (42.9%) than BWA-MEM (28.7%). Similar results were obtained when we restricted the analysis to the 1,159 TE families: 1.36-fold more reads were aligned using Stampy (7.6%) than BWA-MEM (5.6%). We also observed high correlation between the mapping rate of BWA-MEM and Stampy for the three species. For *D. abyssinica*, correlation between the rate of mapped reads from BWA-MEM and Stampy was r=0.88 (p-*value*$<3.10^{-16}$); for *D. praehensilis* it was r=0.78 (p-*value*$<3.10^{-16}$) and for *D. rotundata* the correlation was r=0.71 (p-*value*$<23.10^{-16}$). We finally chose to use Stampy to map reads onto the TE library and define TE contents for all the 180 yam accessions under survey.

Using Stampy, we mapped the genome sequencing reads from each of the 180 yam accessions onto the complete predicted *Dioscorea* RE library. Percentages of mapped reads were very similar between species, with an average of 39.3%, 40.4% and 39.8% of mapped reads for *D. abyssinica, D. Praehensilis* and *D. rotundata,* respectively (Figure 2-a). The average rate of mapped reads onto the sequences contained in the TE library with 1,159 annotated TEs was 7.2%, 7.1% and 7.2% for *D. abyssinica, D. praehensilis* and *D. rotundata,* respectively (Figure 2-b). Plots of the RPKM (Reads per Kilobase per Million mapped, see Methods) values for individual TE subfamilies yielded similar distributions within the three species. The TE families had a unimodal distribution (Figure 3). Among the 1,159 TEs, we found that the most represented order of TEs among the three species was the LTR-retrotransposons. Notably, abundance in LTR-retrotransposons represented on average 4.6% of the *D. abyssinica* genome, corresponding to 63.9% of the total predicted TE content; 4.2% of the *D. praehensilis* genome (59.19% out of total TE contents) and 4.3% of the *D. rotundata* genome (59.7% out of total TE contents). Pairwise correlations of RKPM values between the three species, based on the 180 accessions and estimated for the 1,159 annotated TE families, were also very high: correlations (Pearson r^2) between *D. rotundata* and *D. praehensilis* were $r^2 = 0.96$ (p-value $<$ 0.001), $r^2 = 0.93$ (p-value < 0.001) between and *D. rotundata* and *D. abyssinica*, and $r^2 = 0.88$ (p-value < 0.001) between *D. praehensilis* and *D. abyssinica* (Figure 4). The non parametric Wilcoxon paired tests revealed significant differences between pairs of species ($P.value <$ 0.001). Slight but significant differences in RPKM between pairs of species also revealed that relative TE abundances were on average the highest in *D. rotundata* and the lowest in *D.*

praehensilis, being intermediate in *D. abyssinica* (as shown by the blue line in Figure 4). In addition, we identified eleven TEs that showed extreme difference in their relative abundances between at least two of the three *Dioscorea* species (Figure 5). These outliers comprised ten class I elements (including 4 TRIM, 2 SINEs and 3 LTR-RTs) and only one MITE belonging to class II (Supplementary Table S3). Our results indicate that the global TE content and the outlier TEs might contribute, at least partly, to the structure of the population of *Dioscorea* we studied.

Correlation between the structure of genetic diversity based on SNPs and the relative abundances of TEs

To compare differential TE contents among species with the structure of genetic diversity determined using SNP data, we first performed a genome-wide structure analysis. We identified 218,266 high quality SNPs corresponding to 25,921 contigs, as previously reported (Akakpo et al; 2017). Based on this SNP data set, the program ADMIXTURE partitioned the 180 yam accessions into four genetic groups corresponding to *D. rotundata*, *D. abyssinica*, *D. praehensilis* and a wild Cameroonian accession group (i.e. including accessions from the two wild species *D. abyssinica* and *D. Praehensilis* collected in Cameroon (Supplementary figure S1 and S2). Performing a PCA on the RKPM values estimated for each of the 1,159 unique TEs of the TE library allowed highlighting a significant difference between species on the basis of their TE content. Notably, the second axis divided the population of 180 yam individuals into four groups (but not as well defined as those based on SNP data), corresponding globally to *D. abyssinica, D. rotundata, D. praehensilis* and wild accessions (*D. abyssinica* and *D. praehensilis)* from Cameroon. An ANOVA test on the PCA coefficient also confirmed such difference (p-value <0.001, Supplementary Figure S3). To compare population structure and genomic structure, we performed a Linear Model regression between each K matrix of ancestry (from K=3 to K=10) and the PCA coefficient following the second principal component from the PCA (Figure 6). The results showed that the successive R-squared from LM regression for each K group were on average 0.42 following the PC2 (p-value < 0.001, Supplementary Table S4).

Abundance estimates of specific repeat elements are correlated to environmental variables

As many bioclimatic variables are highly correlated (e.g. precipitation of driest month BIO14 and precipitation of driest quarter BIO17), we first performed a PCA on the 19 climatic variables used in the present study. The two first axes explained 78% of the variance (Supplementary Figure S4). We found that precipitation variables (positive values – BIO 14, 17 and 18) and temperature variables in warmest conditions (negative values – BIO 4, 5, 7 and

10) contributed to axis 1, and that other temperature variables contributed to axis 2 (negative values – BIO 1, 6, 8, 9, and 11). We then correlated these two PCA axes with the RPKM variations estimated for each of the repeat element of the total RE library we created. Using a False Discovery Rate of 5%, we detected a total of 2,436 candidate elements, corresponding to 48% of the repeat fraction of the yam genome, that were associated with the axis 1 of the PCA obtained on the 19 bioclimatic variables. These candidate REs included 908 TEs from the 1,159 annotated TE database. We also identified a total of 2,322 candidate REs associated with the environmental variables of the axis 2 (FDR of 5%). These candidate REs corresponded to 46% of the repeat fraction of the yam genome including 593 TEs from the yam TE database. We observed that associations between the candidate REs and the climatic variables harboured a bimodal distribution, whether following axis 1 or axis 2 (Supplementary Figure S5). Interestingly, the two distributions were widely *s*kewed to negative values on the two first PCs, indicating that differential amplifications of repeats were mainly correlated with variations in temperature (Supplementary Figure S5).

Discussion

In the present study, we used the most important cultivated African yam species *D. rotundata* and its two closest wild relatives *D. abyssinica* and *D. praehensilis* (Akakpo *et al.*, 2017) to *de novo* predict and assemble the repeated components of the *Dioscorea* genome. According to the comparative analysis of RE and TE contents we performed between 180 yam accessions from various geographical locations, we provide new insights into the role of repeated sequences in the genome evolution and adaptation of African yam.

A repeat element library for yam that includes more than 1,000 annotated TEs

We identified a total of 7,275 repeat elements for the three *Dioscorea* species surveyed representing 5,047 unique consensus elements, among which 1,159 elements were assembled and annotated as class I or class II transposable elements (as shown in Figure 1). Based on mapping results, predicted REs represent ~40% of the yam genome corresponding to the estimate of 46% of interspersed sequences obtained from the analysis of the whole *D. rotundata* genome (Tamiru *et al.* 2017). For the dataset of annotated TEs we constructed, the mapping results indicate that it does not exceed 10% of the yam genome. The annotated TE content we obtained for yam is thus quite low compared to the 32% fraction of the recently released genome of *D. rotundata* annotated as TEs (Tamiru *et al.* 2017). The relatively small fraction of TEs we estimated might result from an incomplete annotation of TEs, as supported by the high proportion of unclassified repeat elements we obtained (listed in Table 1 as the category 'noCat'). Difficulties in accessing full length consensuses for TEs present at low copy numbers in the genome using the Illumina sequencing technology (generating short reads) necessarily

reduce the number of clearly annotated TEs when structure-based methods for identification are used (Bergman and Quesneville, 2007). Part of the unclassified REs should also correspond to TEs specific to *Dioscorea*, which are then too divergent from the TE sequences registered in Repbase (Bao *et al.*, 2015) to be retrieved. Repeat elements and notably TEs are involved in genome divergence process at the origins of population structure and speciation (reviewed by Belyayev 2014; Domb *et al.*, 2017). Taking into consideration the distant phylogenetic position of *Dioscorea* compared to other monocotyledons with complete genome sequences (APG IV 2016), it is expected that the TE fraction of yam shows relatively low level of identity to already registered TE sequences. The fact that we obtained a low representation of the LTR-retrotransposons within the genome of *Dioscorea* (approximately 4.4%) also supports this hypothesis. In plant genomes, LTR-RTs are indeed the most abundant class of TEs representing a significant part of the total nuclear genome (Bennetzen *et al.*, 2012): for instance, considering plant genomes of similar sizes, LTR-RTs are estimated as representing 14,32% of the *Vitis vinifera* genome (Jaillon *et al.*, 2007), up to 22.13% of the *Brassica oleracea* genome and at least 9.43% of the *B. rapa* genome (Liu *et al.*, 2014). Further work might be thus needed to deliver a more complete annotated fraction of the assembled repetitive fraction that remains challenging for the wild *Dioscorea* species. In the present study, in order to provide a global view of the contribution of REs and TEs in shaping the genome of yam, we considered separately annotated TEs and the whole fraction of predicted REs including the unannotated ones. We obtained consistent results with the different approaches, indicating that the annotated TE database we created is a good representative of the whole TE fraction of the yam genome.

Repeat element contents show slight but significant variation between the three African yam species

The comparisons made for genome contents in total REs and annotated TEs between the three African yam species showed a high conservation of the fraction of repeat elements. The estimated total abundances for both the REs and the fully annotated TEs were similar across the three species. This is in accordance with the usually observed correlation between plant genome size and genome TE content, considering a total of 594 Mb for the cultivated yam genome (Tamiru *et al.* 2017) very closed to the ~530 Mb for the two wild yam genomes (calculated with the 0.63pg/1C values for the two wild species taking into account 0.71 pg/1C for the cultivated one; C-values from Bennett and Leitch, 2012). Among the wild species, *D. praehensilis* showed a greatest variability in repeat content than *D. abyssinica*. Interestingly, the cultivated species *D. rotundata* harboured the highest intra-species variation for repeat contents with several genotypes displaying values for abundance in REs outside the main distribution (as illustrated by the outliers observed in figure 2). Such variation may illustrate

the still ongoing process of domestication of yam associated to farmer's crop management practices (i.e. ennoblement), resulting in relatively high genome diversity. Thanks to such cultivation practices (based on the recurrent introduction of genetic variability to yam that is clonally propagated; Scarcelli *et al.* 2013), it was demonstrated that the genetic diversity of cultivated yam had a geographical component (Tostain *et al.* 2007). It is thus not surprising to observe here the highest genetic diversity for the repetitive fraction of the cultivated genome, which was accessed by the survey of a set of *D. rotundata* accessions covering a large area of cultivation (i.e. 85 accessions collected from Ghana to Cameroon).

The composition of the TE fraction was remarkably stable within the three species. Approximately 70% of the annotated TE fraction corresponded to class I transposable elements with 50% of LTR-RTs, in accordance with what is commonly observed for the composition of the TE fraction in plants (Vitte *et al.* 2014). We also evaluated congruence in the relative abundance of TEs between the three species by comparing RPKM values across the 1,159 annotated TEs. We obtained strong correlations in RPKM for the three inter-species comparisons, with the highest one being between *D. rotundata* and *D. praehensilis* ($r = 0.96$) and the lowest between *D. rotundata* and *D. abyssinica* ($r = 0.88$). Here, we clearly showed that the relative abundances of TEs were only slightly variable between species: our results thus indicate that genome divergence between the three *Dioscorea* species was not accompanied by any global amplification / elimination of individual TEs. Our results in *Dioscorea* echo the results from Tenaillon *et al.* (2011) who demonstrated similar relative abundances for ~1,500 individual TE families between maize and one of its wild distant relatives *Zea luxurians* with no outlier, even if their genomes differ in size by 50%. Interestingly, while the genomes of the three *Dioscorea* species are quite similar in size, we could identify eleven outliers that corresponded to TEs with highly significant differences in their relative abundances between species (Fig.5). Such outliers suggest the involvement of TEs in the genomic differentiation between the three species. In order to consolidate the role of TEs in the differentiation between species, we performed a PCA on the relative TE abundances (i.e. RPKM values) to highlight differences in TE contents between individuals. The results showed that RPKM values allowed partitioning the 180 yam accessions we surveyed, and clearly discriminated between the three species. Compared to the structure of genetic diversity we determined using SNP variants, the variation in relative TE abundances displayed high correlations with the population structure obtained. Our study has thus clearly demonstrated the effective role of the evolutionary dynamics of TEs in shaping the wild and cultivated yam genomes.

The repeat fraction of the yam genome contributed to adaptation of *Dioscorea* to environment

We found strong relationships between the total repeat fraction of the yam genome and the bioclimatic variables: whatever the bioclimatic variable, around 50% of the total RE content was correlated with the given bioclimatic variable. Genome size and genome content in repeat elements are known to be highly correlated (reviewed by Kejnovsky *et al.*, 2012; Lyu *et al.* 2018). Moreover, previous studies in plants also reported significant relationships between genome size and climate (e.g. in maize Díez *et al.*, 2013; Jian *et al.*, 2017 and in Lilium, Du *et al.*, 2017;), indicating that plant genome size might be under both neutral evolution and adaptive selection. Accordingly, through the analysis of the whole RE fraction of the yam genome, we might have access to the global relationship between genome size of yam and climate. In the present study, the question arising now is thus to know if the observed correlation between the content in repeat elements and bioclimatic variables might be driven by an adaptive selection pattern (at least partly as genetic drift might also be involved). As it is unlikely to imagine specific selection, if any, on each single RE family to explain that half of the predicted REs is associated to variation in bioclimatic data, we can make the confident conclusion that if selection is acting, it might be on the overall genome size in yam.

As reported earlier, relative abundances of TEs correlated with the genetic structure of the yam collection we surveyed, and we thus might expect having specific profiles of diffusion/colonization for specific climatic area in association with changes in the frequency of TEs. The correlations characterized between climate and RE content would then mainly illustrate the activation of specific repeat elements during colonization of new environments. In more detail, the majority (10 out of 11) of the differentially represented TEs between species (i.e. the outliers depicted above) were class I elements characterized by a replicative transposition process. Five of them corresponded to LTR-RTs (1) and TRIMs (4) (i.e. non-autonomous LTR-RTs, Witte *et al.* 2001) that are known to be particularly reactive to genome instability and environmental stresses (reviewed by Granbastien 2015). It can be thus hypothesized that such TEs may have contributed to the genomic differentiation of the three *Dioscorea* species in response to global environmental constraints in a dynamic process of adaptation. This hypothesis is in accordance with the particular case of the wild species from Cameroon that constitute a separate group from the three others (with one for each species): Cameroon presents specific climatic conditions with respect to its geographical location and it was revealed that the individuals from the two wild *Dioscorea* species located in this area harbor specific TE frequencies. Interestingly, among the 10 class I outliers, 8 of them including the 5 LTR-RTs were significantly correlated with bioclimatic variables: 7 with PC1 and 4 with PC2 (data not shown). The *Dioscorea* species are adapted to specific ecological niches with

contrasting environments (Scarcelli *et al.* 2017). The coding regions under selection (related to domestication of African yam) were previously identified, and genes associated with photosynthesis and phototropism were among the main targets for adaptive selection in relation with environment (Akakpo *et al.* 2017). Here, our analysis focusing on the non-coding regions of the yam genome indicates a complementary role of TEs, and notably of retrotransposons, in adaptation.

To go further in the analysis of the implication of REs/TEs in adaptation, developing markers anchored in the TE candidates (outliers) identified here could add to our knowledge of the concrete significance of TEs in the genetic determinism of adaptation in yam. For instance, GWAS performed on polymorphic TE-derived markers would allow accessing the non-coding regions of the genome, and then deciphering the whole genome contribution to adaptation in yam. These findings would be highly valuable for the valorization of yam genetic resources in Africa.

Acknowledgments

This work was supported by a PhD grant to RA by the Islamic Development Bank. This work was supported by the *Agence Nationale de la Recherche – France* with a grant to YV (ANR-13-BSV7-0017). We thank the GeT platform in Toulouse (http://get.genotoul.fr) for DNA sequencing. We thank Marie Couderc and Cédric Mariac for advices during genomic bank preparation and sequencing. We thank Ndomassi Tando for advices in carrying out data processing.

Author contributions

RA, FM, NS, HC, AD, GD, YV, and KA, designed the study; NS generated the whole genome sequencing data; CT and FM contributed to analytic tools; RA predicted the REs, assembled the TE database and performed all the analyses on REs / TEs; RA and KA analyzed data and RA prepared figures and tables; RA, YV and KA interpreted the results; RA, YV and KA wrote the manuscript, and the different authors contribute to its corrections.

References

Ågren, J. A., Huang, H.-R., and Wright, S. I. (2016). Transposable element evolution in the allotetraploid Capsella bursa-pastoris. *Am. J. Bot.* 103, 1197–1202. doi:10.3732/ajb.1600103.

Akakpo, R., Scarcelli, N., Chaïr, H., Dansi, A., Djedatin, G., Thuillet, A.-C., *et al.* (2017). Molecular basis of African yam domestication: analyses of selection point to root

development, starch biosynthesis, and photosynthesis related genes. *BMC Genomics* 18, 782. doi:10.1186/s12864-017-4143-2.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi:10.1101/gr.094052.109.

Alix, K., Gérard, P. R., Schwarzacher, T., and Heslop-Harrison, J. S. (Pat) (2017). Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Ann. Bot.* 120, 183–194. doi:10.1093/aob/mcx079.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6, 11. doi:10.1186/s13100-015-0041-9.

Bardil, A., Tayalé, A., and Parisod, C. (2015). Evolutionary dynamics of retrotransposons following autopolyploidy in the Buckler Mustard species complex. *Plant J. Cell Mol. Biol.* 82, 621–631. doi:10.1111/tpj.12837.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42, 251–269.

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., *et al.* (2012). Reference genome sequence of the model plant Setaria. *Nat. Biotechnol.* 30, 555. doi:10.1038/nbt.2196.

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65, 505–530. doi:10.1146/annurev-arplant-050213-035811.

Dabney, A., and Storey, J. D. (2010). qvalue: Q-value estimation for false discovery rate control.

Díez, C. M., Gaut, B. S., Meca, E., Scheinvar, E., Montes-Hernandez, S., Eguiarte, L. E., *et al.* (2013). Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytol.* 199, 264–276. doi:10.1111/nph.12247.

Diez, C. M., Meca, E., Tenaillon, M. I., and Gaut, B. S. (2014). Three Groups of Transposable Elements with Contrasting Copy Number Dynamics and Host Responses in the Maize (Zea mays ssp. mays) Genome. *PLOS Genet.* 10, e1004298. doi:10.1371/journal.pgen.1004298.

Domb, K., Keidar, D., Yaakov, B., Khasdan, V., and Kashkush, K. (2017). Transposable elements generate population-specific insertional patterns and allelic variation in genes of wild emmer wheat (Triticum turgidum ssp. dicoccoides). *BMC Plant Biol.* 17, 175. doi:10.1186/s12870-017-1134-z.

Du, Y., Bi, Y., Zhang, M., Yang, F., Jia, G., and Zhang, X. (2017). Genome Size Diversity in Lilium (Liliaceae) Is Correlated with Karyotype and Environmental Traits. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.01303.

564 Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends Genet.*
565 *TIG* 5, 103–107.
566 Flavell, A. J. (1995). Retroelements, reverse transcriptase and evolution. *Comp. Biochem.*
567 *Physiol. B Biochem. Mol. Biol.* 110, 3–15.
568 Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering Transposable
569 Element Diversification in De Novo Annotation Approaches. *PLOS ONE* 6, e16526.
570 doi:10.1371/journal.pone.0016526.
571 Girma, G., Hyma, K. E., Asiedu, R., Mitchell, S. E., Gedil, M., and Spillane, C. (2014). Next-
572 generation sequencing based genotyping, cytometry and phenotyping for understanding
573 diversity and evolution of guinea yams. *Theor. Appl. Genet.* 127, 1783–1794.
574 doi:10.1007/s00122-014-2339-2.
575 Grandbastien, M.-A. (1998). Activation of plant retrotransposons under stress conditions.
576 *Trends Plant Sci.* 3, 181–187. doi:10.1016/S1360-1385(98)01232-1.
577 Hamon, P., Brizard, J.-P., Zoundjihékpon, J., Duperray, C., and Borgel, A. (1992). Étude des
578 index d'ADN de huit espèces d'ignames (Dioscorea sp.) par cytométrie en flux. *Can. J.*
579 *Bot.* 70, 996–1000. doi:10.1139/b92-123.
580 Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high
581 resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–
582 1978. doi:10.1002/joc.1276.
583 Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., *et al.* (2014).
584 PASTEC: An Automatic Transposable Element Classification Tool. *PLOS ONE* 9,
585 e91929. doi:10.1371/journal.pone.0091929.
586 Horváth, V., Merenciano, M., and González, J. (2017). Revisiting the Relationship between
587 Transposable Elements and the Eukaryotic Stress Response. *Trends Genet. TIG* 33,
588 832–841. doi:10.1016/j.tig.2017.08.007.
589 Huang, C., Sun, H., Xu, D., Chen, Q., Liang, Y., Wang, X., *et al.* (2018). ZmCCT9 enhances
590 maize adaptation to higher latitudes. *Proc. Natl. Acad. Sci. U. S. A.* 115, E334–E341.
591 doi:10.1073/pnas.1718058115.
592 Jian, Y., Xu, C., Guo, Z., Wang, S., Xu, Y., and Zou, C. (2017). Maize (Zea mays L.) genome
593 size indicated by 180-bp knob abundance is associated with flowering time. *Sci. Rep.* 7,
594 5954. doi:10.1038/s41598-017-06153-8.
595 Kapitonov, V. V., and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proc. Natl.*
596 *Acad. Sci. U. S. A.* 98, 8714–8719. doi:10.1073/pnas.151269298.
597 Kejnovsky, E., Hawkins, J. S., and Feschotte, C. (2012). "Plant Transposable Elements:
598 Biology and Evolution," in *Plant Genome Diversity Volume 1* (Springer, Vienna), 17–
599 34. doi:10.1007/978-3-7091-1130-7_2.
600 Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
601 transform. *Bioinforma. Oxf. Engl.* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

602 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009). The
603 Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* 25, 2078–2079.
604 doi:10.1093/bioinformatics/btp352.
605 Li, S.-F., Su, T., Cheng, G.-Q., Wang, B.-X., Li, X., Deng, C.-L., *et al.* (2017). Chromosome
606 Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes* 8.
607 doi:10.3390/genes8100290.
608 Lisch, D. (2013). How important are transposons for plant evolution? *Nat. Rev. Genet.* 14, 49.
609 doi:10.1038/nrg3374.
610 Loko, Y. L., Bhattacharjee, R., Agre, A. P., Dossou-Aminon, I., Orobiyi, A., Djedatin, G. L., *et*
611 *al.* (2016). Genetic diversity and relationship of Guinea yam (Dioscorea cayenensis
612 Lam.–D. rotundata Poir. complex) germplasm in Benin (West Africa) using
613 microsatellite markers. *Genet. Resour. Crop Evol.*, 1–15. doi:10.1007/s10722-016-0430-z.
614 Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast
615 mapping of Illumina sequence reads. *Genome Res.* 21, 936–939.
616 doi:10.1101/gr.111120.110.
617 Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J., *et al.*
618 (2015). Transposable Elements Contribute to Activation of Maize Genes in Response to
619 Abiotic Stress. *PLOS Genet.* 11, e1004915. doi:10.1371/journal.pgen.1004915.
620 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., *et al.*
621 (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-
622 generation DNA sequencing data. *Genome Res.* 20, 1297–1303.
623 doi:10.1101/gr.107524.110.
624 Mehrotra, S., and Goyal, V. (2014). Repetitive Sequences in Plant Nuclear DNA: Types,
625 Distribution, Evolution and Function. *Genomics Proteomics Bioinformatics* 12, 164–
626 171. doi:10.1016/j.gpb.2014.07.003.
627 Mignouna, H. D., and Dansi, A. (2003). Yam (Dioscorea ssp.) domestication by the Nago and
628 Fon ethnic groups in Benin. *Genet. Resour. Crop Evol.* 50, 519–528.
629 doi:10.1023/A:1023990618128.
630 Negi, P., Rai, A. N., and Suprasanna, P. (2016). Moving through the Stressed Genome:
631 Emerging Regulatory Roles for Transposons in Plant Stress Response. *Front. Plant Sci.*
632 7. doi:10.3389/fpls.2016.01448.
633 Ngwe, M. F. S. N., Omokolo, D. N., and Joly, S. (2015). Evolution and Phylogenetic Diversity
634 of Yam Species (Dioscorea spp.): Implication for Conservation and Agricultural
635 Practices. *PLOS ONE* 10, e0145364. doi:10.1371/journal.pone.0145364.
636 Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., *et al.* (2010). Impact of
637 transposable elements on the organization and function of allopolyploid genomes. *New*
638 *Phytol.* 186, 37–45. doi:10.1111/j.1469-8137.2009.03096.x.

639 Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., *et*
640 *al.* (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature* 457,
641 551–556. doi:10.1038/nature07723.
642 Pearce, S. R., Harrison, G., Li, D., Heslop-Harrison, J., Kumar, A., and Flavell, A. J. (1996).
643 The Ty1-copia group retrotransposons in Vicia species: copy number, sequence
644 heterogeneity and chromosomal localisation. *Mol. Gen. Genet. MGG* 250, 305–315.
645 Petrov, D. A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K., and González, J. (2011).
646 Population Genomics of Transposable Elements in Drosophila melanogaster. *Mol. Biol.*
647 *Evol.* 28, 1633–1644. doi:10.1093/molbev/msq337.
648 Pritham, E. J., Putliwala, T., and Feschotte, C. (2007). Mavericks, a novel class of giant
649 transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390,
650 3–17. doi:10.1016/j.gene.2006.08.008.
651 Sanmiguel, P., and Bennetzen, J. L. (1998). Evidence that a Recent Increase in Maize Genome
652 Size was Caused by the Massive Amplification of Intergene Retrotransposons. *Ann.*
653 *Bot.* 82, 37–44. doi:10.1006/anbo.1998.0746.
654 Sarah, G., Homa, F., Pointet, S., Contreras, S., Sabot, F., Nabholz, B., *et al.* (2016). A large
655 set of 26 new reference transcriptomes dedicated to comparative population genomics in
656 crops and wild relatives. *Mol. Ecol. Resour.* doi:10.1111/1755-0998.12587.
657 Sarilar, V., Marmagne, A., Brabant, P., Joets, J., and Alix, K. (2011). BraSto, a Stowaway
658 MITE from Brassica: recently active copies preferentially accumulate in the gene space.
659 *Plant Mol. Biol.* 77, 59–75. doi:10.1007/s11103-011-9794-9.
660 Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant
661 genomes. *Brief. Funct. Genomics* 13, 296–307. doi:10.1093/bfgp/elu016.
662 Scarcelli, N., Chaïr, H., Causse, S., Vesta, R., Couvreur, T. L. P., and Vigouroux, Y. (2017).
663 Crop wild relative conservation: Wild yams are not that wild. *Biol. Conserv.* 210, 325–
664 333. doi:10.1016/j.biocon.2017.05.001.
665 Scarcelli, N., Couderc, M., Baco, M. N., Egah, J., and Vigouroux, Y. (2013). Clonal diversity
666 and estimation of relative clone age: application to agrobiodiversity of yam (Dioscorea
667 rotundata). *BMC Plant Biol.* 13, 178. doi:10.1186/1471-2229-13-178.
668 Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., *et al.* (2009). The
669 B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115.
670 doi:10.1126/science.1178534.
671 Sinzelle, L., Izsvák, Z., and Ivics, Z. (2009). Molecular domestication of transposable elements:
672 from detrimental parasites to useful host genes. *Cell. Mol. Life Sci. CMLS* 66, 1073–
673 1093. doi:10.1007/s00018-009-8376-3.
674 Song, X., and Cao, X. (2017). Transposon-mediated epigenetic regulation contributes to
675 phenotypic diversity and environmental adaptation in rice. *Curr. Opin. Plant Biol.* 36,
676 111–118. doi:10.1016/j.pbi.2017.02.004.

677 Stelmach, K., Macko-Podgórni, A., Machaj, G., and Grzebelus, D. (2017). Miniature Inverted
678 Repeat Transposable Element Insertions Provide a Source of Intron Length
679 Polymorphism Markers in the Carrot (Daucus carota L.). *Front. Plant Sci.* 8, 725.
680 doi:10.3389/fpls.2017.00725.
681 Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., *et al.* (2017).
682 Genome sequencing of the staple food crop white Guinea yam enables the development
683 of a molecular marker for sex determination. *BMC Biol.* 15. doi:10.1186/s12915-017-
684 0419-x.
685 Tenaillon, M. I., Hufford, M. B., Gaut, B. S., and Ross-Ibarra, J. (2011). Genome size and
686 transposable element content as determined by high-throughput sequencing in maize
687 and Zea luxurians. *Genome Biol. Evol.* 3, 219–229. doi:10.1093/gbe/evr008.
688 Tian, S., Yan, H., Neuhauser, C., and Slager, S. L. (2016). An analytical workflow for accurate
689 variant discovery in highly divergent regions. *BMC Genomics* 17. doi:10.1186/s12864-
690 016-3045-z.
691 Vicient, C. M., and Casacuberta, J. M. (2017). Impact of transposable elements on polyploid
692 plant genomes. *Ann. Bot.* 120, 195–207. doi:10.1093/aob/mcx078.
693 Vitte, C., Fustier, M.-A., Alix, K., and Tenaillon, M. I. (2014). The bright side of transposons
694 in crop evolution. *Brief. Funct. Genomics* 13, 276–295. doi:10.1093/bfgp/elu002.
695 Wang, X., Yang, Y., Moore, D. R., Nimmo, S. L., Lightfoot, S. A., and Huycke, M. M. (2012).
696 4-Hydroxy-2-Nonenal Mediates Genotoxicity and Bystander Effects Caused by
697 Enterococcus faecalis–Infected Macrophages. *Gastroenterology* 142, 543-551.e7.
698 doi:10.1053/j.gastro.2011.11.020.
699 Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., *et al.* (2007). A
700 unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8,
701 nrg2165. doi:10.1038/nrg2165.
702 Witte, C.-P., Le, Q. H., Bureau, T., and Kumar, A. (2001). Terminal-repeat retrotransposons
703 in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad.*
704 *Sci. U. S. A.* 98, 13778–13783. doi:10.1073/pnas.241341898.
705 Xiong, Y., and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their
706 reverse transcriptase sequences. *EMBO J.* 9, 3353–3362.
707 Yang, Q., Li, Z., Li, W., Ku, L., Wang, C., Ye, J., *et al.* (2013). CACTA-like transposable
708 element in ZmCCT attenuated photoperiod sensitivity and accelerated the
709 postdomestication spread of maize. *Proc. Natl. Acad. Sci.* 110, 16969–16974.
710 doi:10.1073/pnas.1310949110.
711

Title: Evaluating the relationships between transposable element content, genetic diversity and geographical distribution in yam (*Dioscorea*)

Tables

Table 1: Summary of the 5,047 unique repeat elements (REs) predicted and annotated within the three *Dioscorea* genomes

Figures

Figure 1: Composition of the TE fraction predicted from *D. rotundata* (accession CR806) illustrating the diversity of TEs in *Dioscorea*. Class I TEs (LTR-RTs, LINE, SINE, TRIM) and class II TEs (TIR, MITE, Helitron, Maverick) represent ~70% and ~30% of the TE categories, respectively. The LTR-retrotransposons are the most diverse TEs (~50% of the TE consensuses) in the *Dioscorea* genome.

Figure 2: Rate of mapped reads against (a) the full RE library and (b) the TE database for the three *Dioscorea* species. The percentage of mapped reads were relatively stable between *D. abyssinica, D. praehensilis* and *D. rotundata* (p-value<0.001 based on Pearson's correlation test). The boxes indicate the first quartile (bottom line), the median (central line) and the third quantile (top line). The whiskers represent the standard deviation.

Figure 3: Distribution of TE abundance in the three species

Figure 4: Correlations between the relative abundances of the annotated TEs for the three *Dioscorea* species, in pairwise comparisons. The coefficients of correlation are (a) $r = 0.98$ between *D. rotundata* and *D. praehensilis*, (b) r = 0.96 between *D. rotundata* and *D. abyssinica*, (c) $r = 0.94$ between *D. praehensilis* and *D. abyssinica*.

Figure 5: Plot of difference in TEs abundance between species showing outliers represented TEs. (a) Difference of abundance between *D. abyssinica* and *D. rotundata*; (b) difference between *D. praehensilis* and *D. rotundata;* (c) Difference between *D. praehensilis* and *D. abyssinica*

Figure 6: Principal component analysis of TE copy numbers (as RPKM) for all the 180 yam accessions surveyed. (**a**) The PCA depicts a large diversity in TE copy numbers for the two species *D. abyssinica* and *D. rotundata* but does not support any clear species clustering. (**b**)

741 More specifically, the PCA does not allow clustering wild species from Cameroon separately, as
742 observed with the analysis performed on SNP data.

Tables

Table 1: Summary of the 5,047 unique repeat elements (REs) predicted and annotated within the three *Dioscorea* genomes

		D. abyssinica				***D. praehensilis***				***D. rotundata***			
		Number	*Min Length*	*Median*	*Max length*	*Number*	*Min Length*	*Median*	*Max length*	*Number*	*Min Length*	*Median*	*Max length*
Transposable elements													
Order	*Class*												
LTR	I	266	150	376	19,287	289	116	353	19,290	279	112	349	3,492
LINE	I	19	184	2,388	19,550	24	159	2,110	16,670	19	210	736	3,461
SINE	I	42	150	251	518	49	109	236	664	47	147	270	699
TRIM	I	0	0	0	0	2	348	398	449	4	265	297	616
noCat class I		45	150	192	21,909	48	105	205	21,909	43	113	185	1585
TIR	II	124	152	277	1,982	142	150	456	4,581	122	150	277	2,310
MITE	II	12	155	184	448	15	163	198	365	18	151	191	414
Helitron	II	9	152	260	541	1	284	284	284	4	219	249	359
Maverick	II	1	4,376	4,376	4,376	2	1,070	3,080	5,087	2	1,269	2,603	3,937
noCat class II		9	182	360	569	16	155	264	659	14	157	342	735
Total number:		**527**				**588**				**552**			
Others													
PotentialHostGene		19	200	4,469	16,967	19	186	6,617	16,970	25	203	1,581	14,850
SRR		76	104	176	467	105	102	185	439	106	100	175	477
noCat		1,397	100	202	21,909	1,870	100	202	21,910	1,991	101	215	10,310
Total number:		**1,492**				**1,994**				**2,122**			

Filtered predicted REs from RepARK and Repdenovo were merged by removing redundancy and annotated using PASTEC. Classification of the repetitive elements in the nomenclature from PASTEC, and TE classification is from Wicker *et al.* 2007; noCat=no classification at this level.

Figures

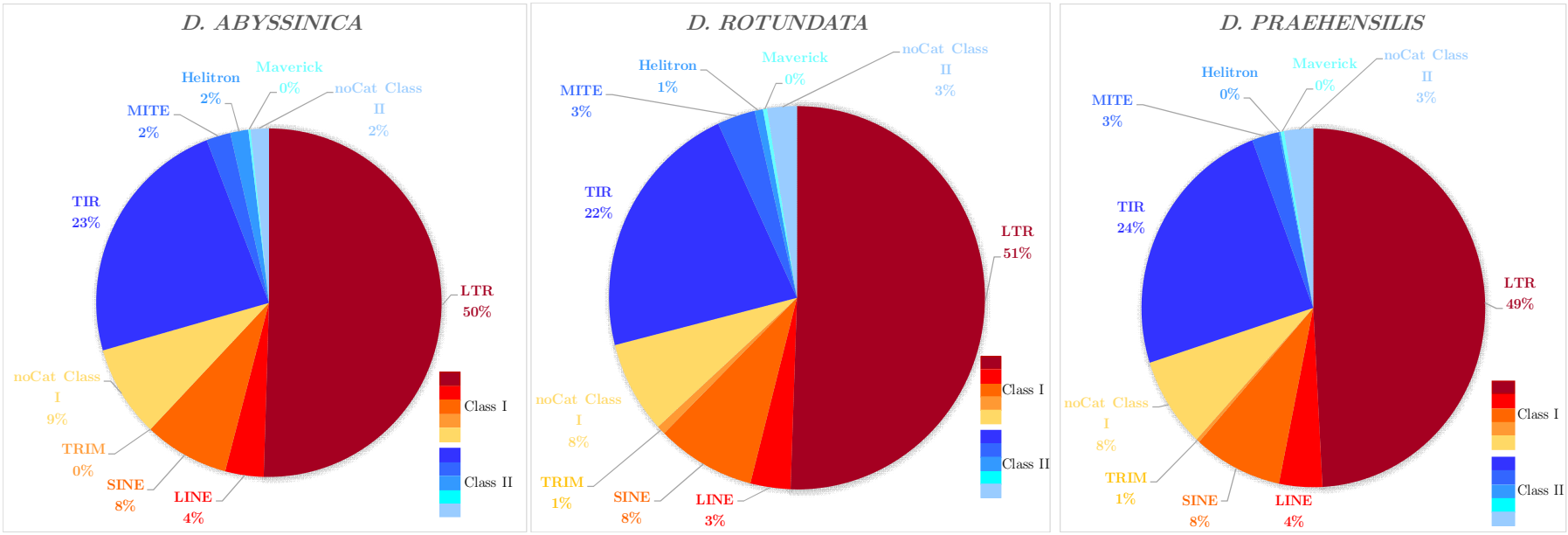
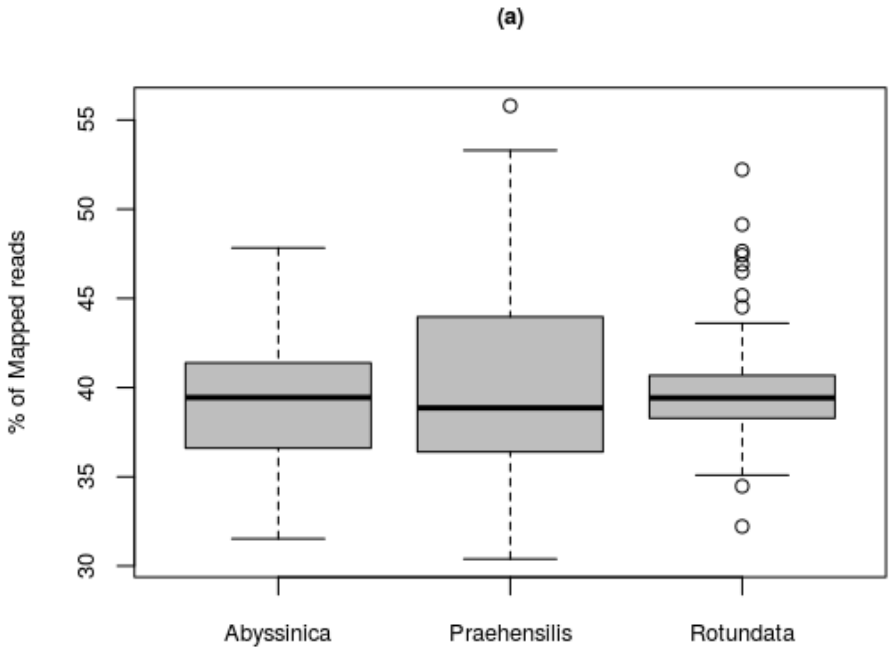
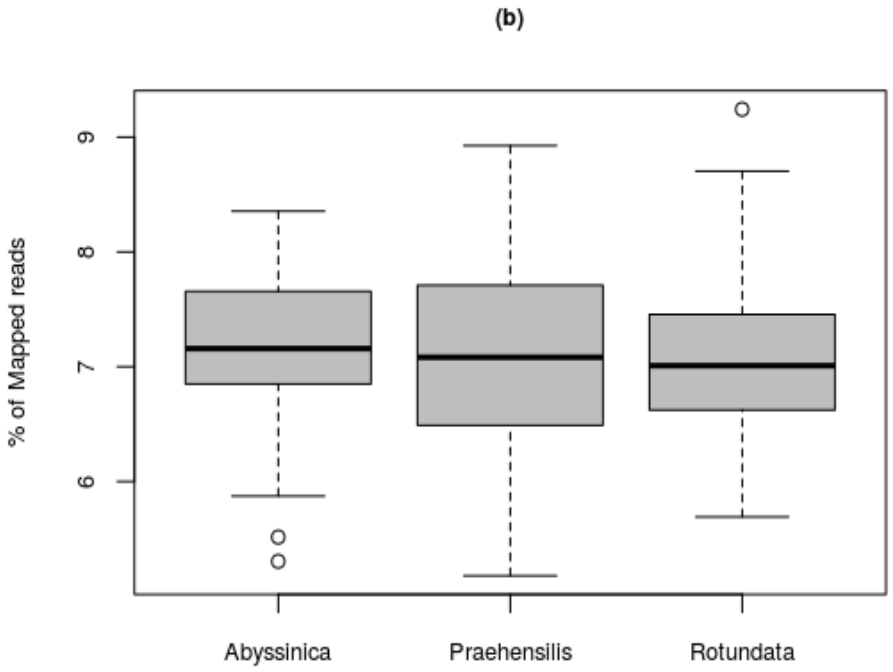


Figure 1: Composition of the TE fraction predicted from *D. abyssinica* (accession A571); *D. rotundata* (accession CR806) and *D. praehensilis* (accession P464) illustrating the diversity of TEs in *Dioscorea*.

753

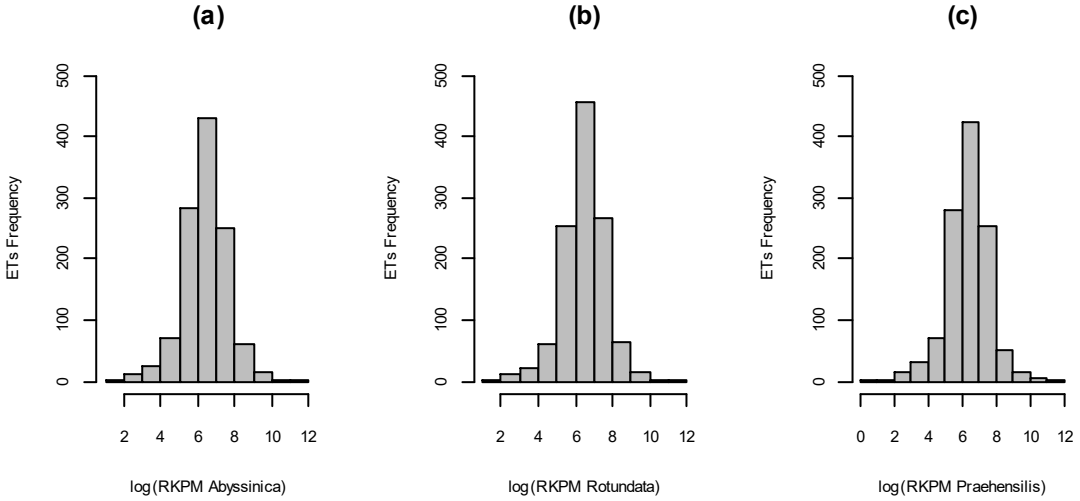


754



755

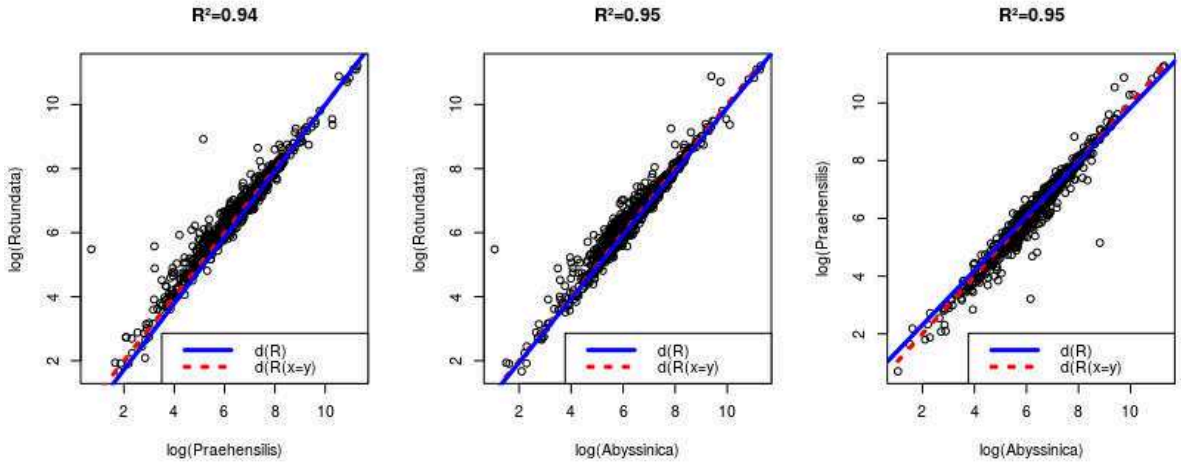
756 **Figure 2:** Rate of mapped reads onto (a) the full RE library and (b) the TE database for the
757 three *Dioscorea* species.



758

759 **Figure 3:** Distribution of TE abundance in the three species: (a) *D. abyssinica*, (b) *D.*
760 *rotundata* and (c) *D. praehensilis*.

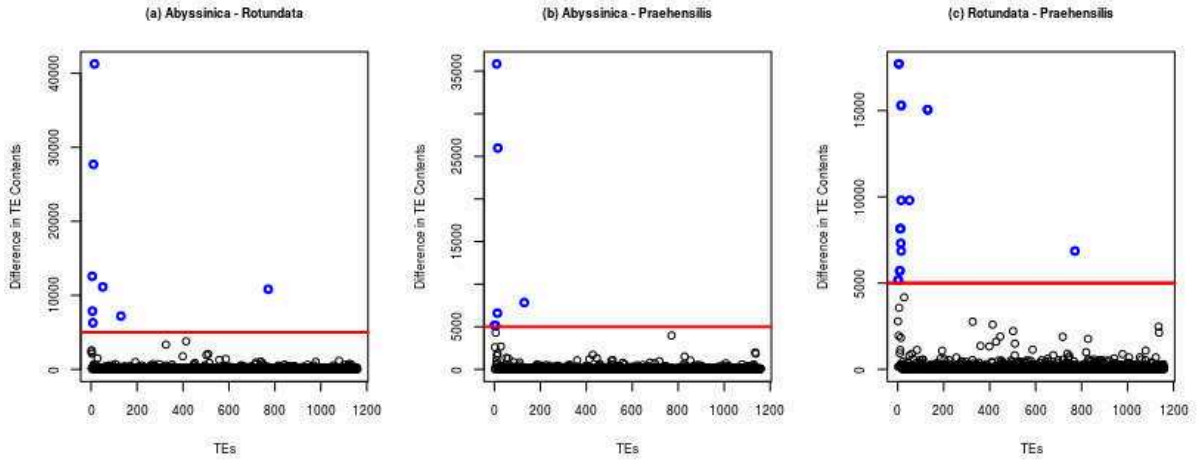
761



762

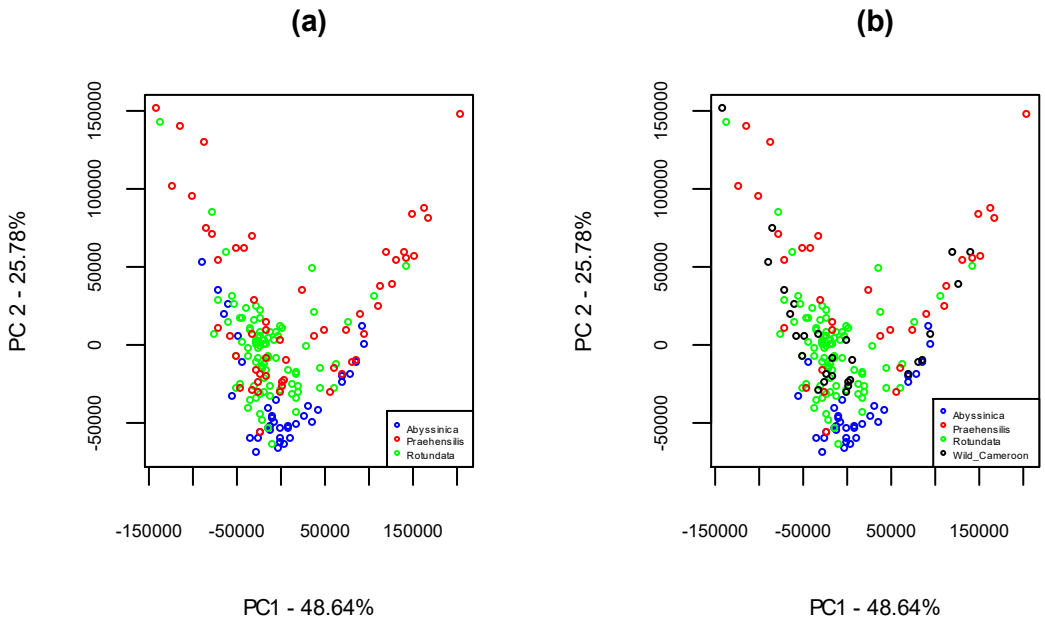
763 **Figure 4:** Correlations between the relative abundances of the annotated TEs for the three
764 *Dioscorea* species, in pairwise comparisons.
765

766



767

768 **Figure 5:** Plot of the differences in TE content between species showing outliers that
769 correspond to the TEs with the most differential copy numbers between two *Dioscorea* species;
770 (a) between *D. abyssinica* and *D. rotundata*, (b) between *D. abyssinica* and *D. praehensilis*, (c)
771 between *D. rotundata* and *D. praehensilis*.



772

773 **Figure 6:** Principal component analysis of TE copy numbers (as RPKM) for all the 180 yam
774 accessions surveyed.
775

776 **Supplementary Information**

777

778 **Supplementary Table S1:** List of the bioclimatic variables used in the present study

Variables	Description
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

779

780 **Supplementary Table S2:** Metrics of the raw repeat elements (REs) predicted by RepARK
781 and Repdenovo

Accession (species)	*Minimum length*		*Mean length*		*3rd Quantile*		*Maximum length*		*Total number of REs*	
	RPAK	RDNV	RPAK	RDNV	RPAK	RDNV	RPAK	RDNV	RPAK	RDNV
A571 (*D. abyssinica*)	57	100	198	1,092	142	450	21,910	25,490	8,898	143
P464 (*D. praehensilis*)	57	100	182	715	140	393	21,910	25,490	11,211	352
CR806 (*D. rotundata*)	57	100	165.9	516	150	496	17,200	16700	10,344	394

782 RPAK=RepARK; RDNV= Repdenovo

783 **Supplementary Table S3:** List of the eleven most discriminant TE families in terms of copy number between at least two of the three *Dioscorea*
784 species

All outliers	Class	Order	Abyssinica	Rotundata	Praehensilis
NEW_CONTIG_MERGE_8	I	TRIM	11965	53231	37935
NEW_CONTIG_MERGE_248	I	TRIM	17088	44780	52937
NEW_CONTIG_MERGE_140	I	SINE	24330	11764	29466
NODE_108_length_192_cov_2.026042_30_450_3830	I	noCat	76952	65813	75609
NODE_501_length_294_cov_2.241497_30_613_30675	I	noCat	61591	50767	57624
NEW_CONTIG_MERGE_151	I	SINE	2544	10409	6844
NODE_126_length_307_cov_2.009772_60_450_14265	II	MITE	21170	13972	29019
NEW_CONTIG_MERGE_180	I	LTR	79898	73604	79313
NEW_CONTIG_MERGE_477	I	TRIM	6770	7483	173
NEW_CONTIG_MERGE_11	I	TRIM	9594	11990	14772
NEW_CONTIG_MERGE_13	I	SINE	8838	6266	11436

785 **Supplementary Table S4:** Relationships between TE content and SNP diversity structure

	PC1		PC2		PC3	
	R-squared	*p*-value	R-squared	*p*-value	R-squared	*p*-value
K2	0.01	0.29	0.00	0.82	0	0.89
K3	0.02	0.19	0.31	**<0.001***	0.03	0.05*
K4	0.04	0.08	0.46	**<0.001***	0.17	<0.001*
K5	0.05	0.07	0.46	**<0.001***	0.18	<0.001*
K6	0.05	0.11	0.47	**<0.001***	0.18	<0.001*
K7	0.1	**0.01***	0.48	**<0.001***	0.18	<0.001*
K8	0.11	**0.01***	0.48	**<0.001***	0.19	<0.001*
K9	0.16	**<0.001***	0.55	**<0.001***	0.19	<0.001*
K10	0.17	**<0.001***	0.55	**<0.001***	0.21	<0.001*

786 PC1, 2 and 3 = Principal Component 1, 2 and 3. K 1 to 10 are the K subgroups from structure
787 analysis

788

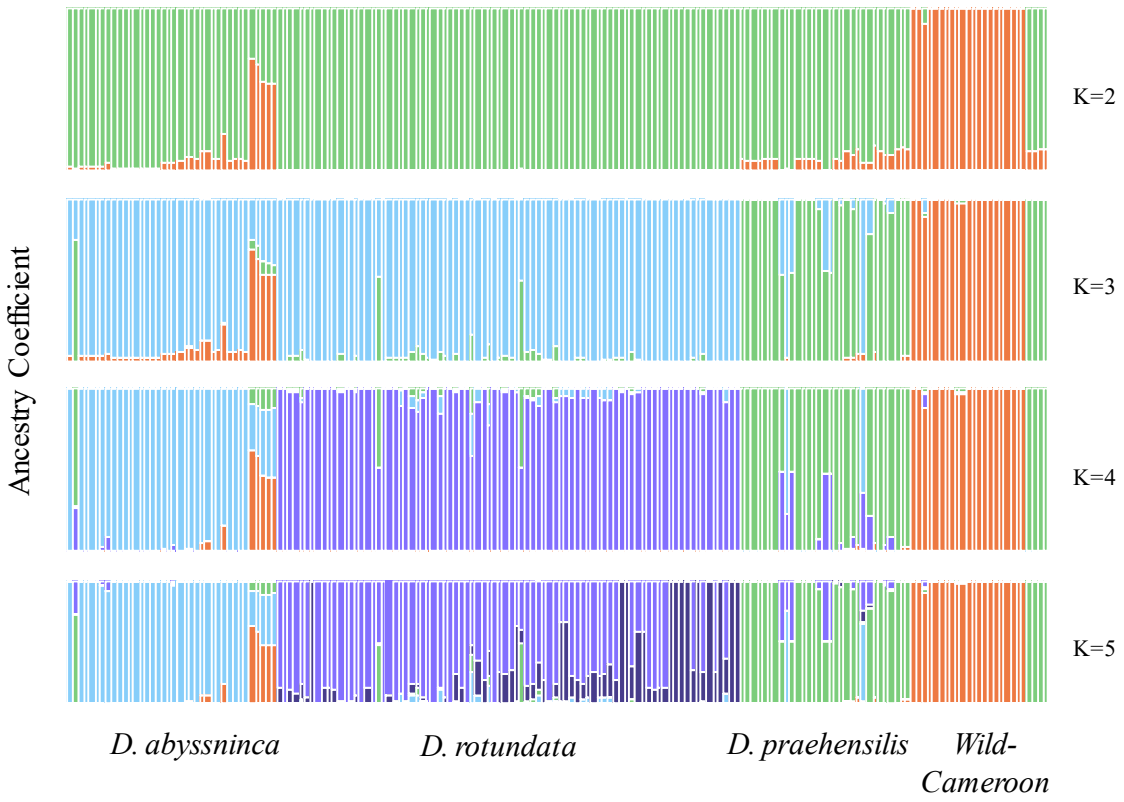
789

790

791

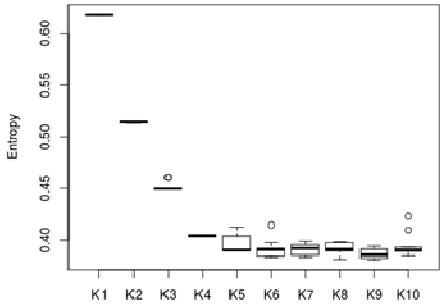
792

793

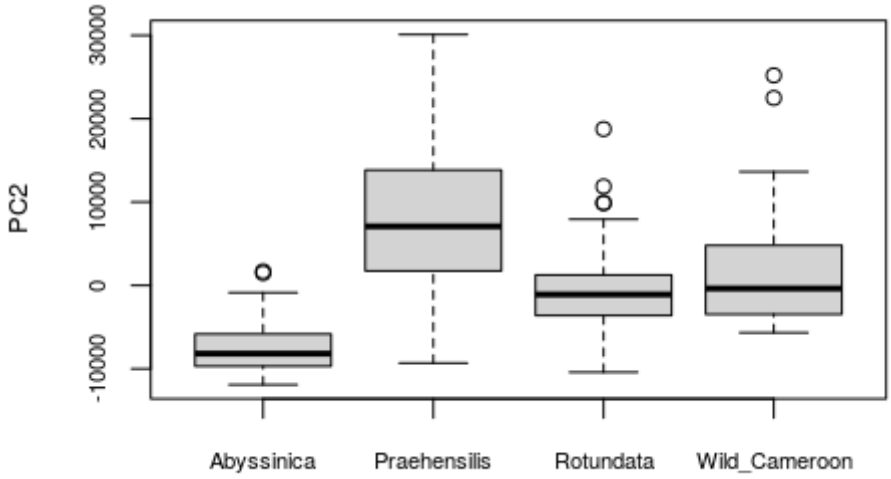


794

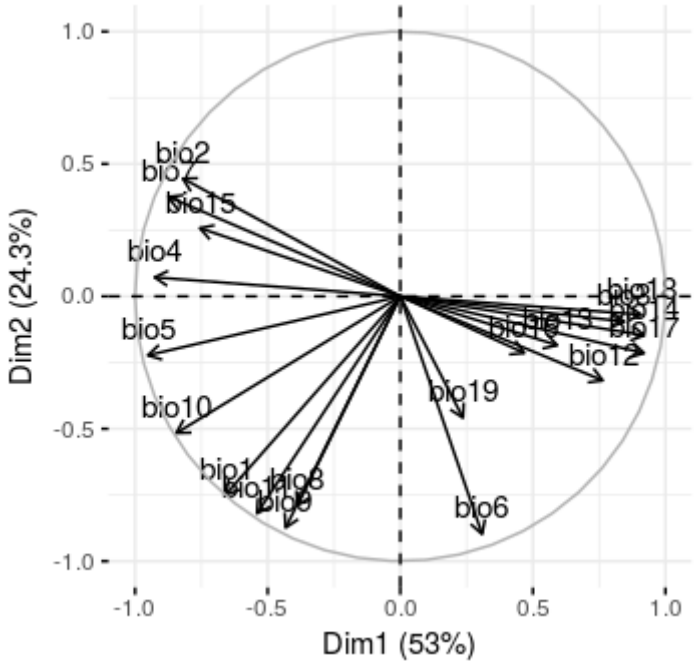
795 **Supplementary Figure S1:** (a) Population structure plot of 180 yam accessions inferred
796 using Admixture. K is the number of ancestral populations considered. Each colour represents
797 one population and each vertical bar represents an individual. The proportion contributed by
798 ancestral populations to the individual corresponds to the length of the segment.



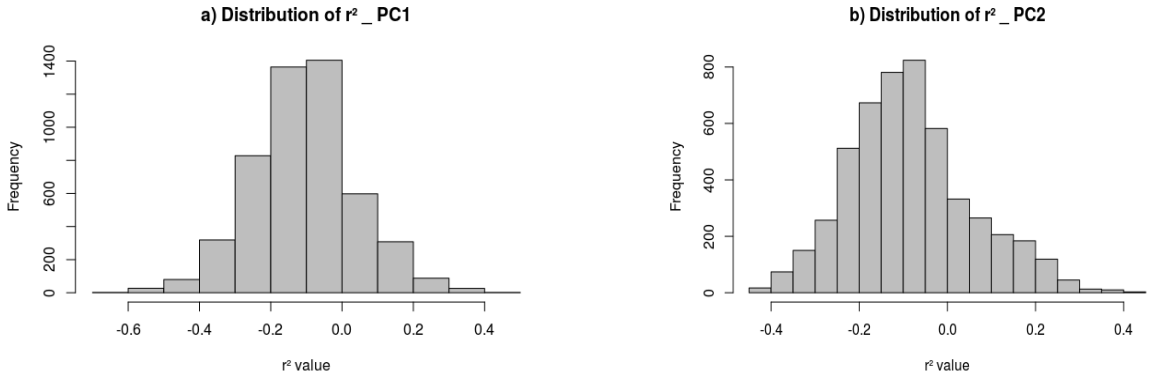
Supplementary Figure S2: Cross-entropy information for K=1 to 10. Ten repetitions of the run were done. The box plot of the cross-entropy information for K=2 to K=10 across the 10 runs revealed that the cross-entropy less varied for K=4 indicating that K=4 most reflected the structure of our yam accessions. K=4 led to 4 genetics groups corresponding to *D. rotundata*; *D. abyssinica*; *D. praehensilis* and Cameroonian wilds groups.



Supplementary Figure S3: PCA show significant difference in TEs contents between the four species *D. abyssinica, D. praehensilis, D. rotundata* and the Wilds species from Cameroon (p-value<0.001 based on ANOVA). The boxes indicate the first quartile (bottom line), the median (central line) and the third quantile (top line). The whiskers represent the standard deviation.



Supplementary Figure S4: Graph of the 19 variables used to correlate environment with TE contents. Axis 1 explained rainfall variables (Bio 14, 17, and 18) whereas axis 2 explained the heat variable (BIO1, 6, 8; 9 and 11).



Supplementary Figure S5: Distribution of the correlation between candidate TEs associated with bioclimatic variables

Chapitre 5 : Base génétique de l'adaptation de l'igname Africaine à différents climats

1. Contexte et objectifs

L'adaptation des plantes aux conditions d'agriculture et aux variabilités environnementales s'accompagne de modifications majeures dans la morphologie et le développement des plantes (Bijlsma et Loeschcke 2005). L'igname africaine *D. rotundata* est cultivée dans une gamme variée de climats, ce qui suppose qu'elle s'est adaptée aux environnements de savane ou de forêt. Dans cette étude, nous avons utilisé des données de variants génétiques de type SNP pour rechercher les signatures de telles adaptations.

2. Approches d'étude

Au total, 153 accessions d'ignames des trois espèces ont été testées. Nous avons utilisé les approches de génétique d'association implémentées dans les logiciels LFMM et EMMA qui prennent en compte la structure génétique des populations. Les variants SNPs qui présentent une association significative au seuil FDR de 5% ont été retenus, et les gènes associés ont été recueillis. Ces gènes ont, ensuite, été testés pour leur potentiel enrichissement en fonction spécifique.

3. Principaux résultats

Des gènes du complexe NADH-DH associés aux activités de déshydrogénase et d'oxydoréduction ont été détectés dans cette étude. Ce même complexe a été identifié au cours de la détection de sélection associée à la domestication (Chapitre 1). On se demande alors si, par cette analyse d'association gènes environnement, on ne regarde pas plutôt de la domestication, en lieu et place de l'adaptation aux différents climats. Des analyses supplémentaires sont nécessaires.

Cette étude fera l'objet d'un article scientifique. Nous présentons ici les résultats ainsi que leurs conclusions au stade actuel. **« Genetic variation associated with climate variability in the African yam ».**

Roland Akakpo, Philippe Cubry, Nora Scarcelli, Anne-Céline Thuillet, Karine Alix, Olivier François, Yves Vigouroux.

1

Genetic variation associated with climate variability in the African yam domestication

Roland Akakpo, Philippe Cubry, Nora Scarcelli, Anne-Céline Thuillet, Karine Alix, Olivier François, Yves Vigouroux

Introduction

The development of agricultural societies is tightly associated with the domestication of crops (Diamond 2002; Ross-Ibarra *et al.* 2007). After domestication, agriculture spread to new environments, and adaptation occured (Tenaillon and Charcosset 2011; Swarts *et al.* 2017). The adaptation of plants to cultivation and to a variety of environments was accompanied by major changes in plant morphology and development (Bijlsma and Loeschcke 2005). Abiotic stresses determine species distribution across different types of environments (Díaz *et al.* 2007). Understanding the genetic basis of plant adaptation associated with the transition from natural environment to field conditions could allow understanding responses of plants to stress during domestication and the course of their evolution.

Reshaping chlororespiration process is one of plant strategies to adapt to environmental stress (Quiles 2006; Li *et al.* 2016; Parades and Quiles, 2017 and references therein). Chlororespiration is a "respiratory electron-transport chain in the thylakoid membrane of chloroplasts, which interacts with photosynthetic electron transport, involving both the non-photochemical reduction and oxidation of plastoquinones with the consumption of oxygen" (Bennoun 1982; Bennoun 1994). Chlororespiration might be stimulated by heat/chilling and high light intensity (Quiles 2006; Paredes and Quiles 2017). Heat and high light intensity could also stimulate chlororespiration in oat plants (Quiles 2006). The plastid-encoded NADH DH complex is one of the more important enzyme in chlororespiration (Jose Quiles M and Cuello 1998; Rumeau *et al.* 2005; Desplats *et al.* 2009). The NADH DH complex is involved in the photosystems I (PSI) and II (PSII), and it plays a key role in protection against photo-oxidative stresses associated with the formation of reactive oxygen species (ROS) (Quiles and López 2004; Peng *et al.* 2008). ROS production is also stimulated by stress signalling (Miller *et al.* 2009; Baxter *et al.* 2013).

We have recently found that adaptation to high intensity light was selected during yam (*Dioscorea spp.*) domestication (Akakpo *et al.* 2017). The African wild yams are wine living in forest (*D. praehensilis*) and savannah (*D. abyssinica*) (Hamon *et al.* 1995), that grow in partial shade of higher trees. On the contrary, the cultivated species *D. rotundata* grows under full sunlight in the field. One would then expect to identify genes of domestication as genes

involved in adaptation to full light, associated with environmental stress such as climatic variability.

In the present study, we performed genome-wide environmental association studies in an African yam population. We used re-sequencing data of both cultivated and wild accessions to identify loci exhibiting association with environmental variables.

Materials and methods

We used a whole-genome resequencing dataset consisting of 167 individuals (including 33 *D. abyssinica*, 50 *D. praehensilis* and 84 *D. rotundata*) sampled from Ghana, Benin, Nigeria and Cameroon (Supporting Information Table S1). This dataset allows detecting 3,570,940 SNP loci (Scarcelli *et al.* in prep). The dataset presented a mean missing data of 7% and a mean sequencing depth of 7. We further filtered out SNP having a Minor Allele Frequency (MAF) lower than or equal to the 5%. Missing SNP data were imputed using R package LEA (Frichot and François 2015).

Population structure analysis of 167 *Dioscorea* individuals

In order to discard eventual hybrid lineage for the association test, an unsupervised clustering program ADMIXTURE (Alexander *et al.* 2009) was used to perform maximum likelihood clustering analysis. The number of K sub-group varied from 2 to 10. Individual admixture coefficients were confirmed using the clustering approach of the program sNMF (Frichot *et al.* 2014). A supervised structure analysis based on the most likely K cluster was finally performed, using the Principal Component Analysis (PCA) approach implemented in the R package SNPrelate (Zheng *et al.* 2012).

Genome-wide environmental association studies

Based on individual admixture coefficients, we excluded all accessions with hybrid profiles (value of ancestry 20 - 80%). We used the new dataset to screen genomes for signatures of environmental genetic adaptation. Our analysis aims to detect loci exhibiting association with environmental variables with controlling population structure. We retrieved 19 bioclimatic environmental variables from the WorldClim database (http://worldclim.org/bioclim) based on sampling geographical coordinate of each individual (Table 1). These variables are often highly correlated. We first performed a principal component analysis (PCA) in order to obtain synthetic variables (axes of the PCA) decorrelated from each other. The two first axes of the PCA were then used for association. For genome-wide gene-environment studies, sampling

independent SNP allowed increasing the power of detection of SNP/climate association (Winham and Biernacka 2013). We then pruned a subset of the SNPs, for SNPs being in linkage disequilibrium with each other. We used the option –indep in Plink v1.90b4 (Purcell *et al.* 2007) to prune SNPs, using a window size of 50 SNPs, a step size of 5 (variant count used to shift the window at the end of each step). We used a variance inflation factor (VIF) threshold of 2 (*VIF=1/(1-R^2)* where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously. We performed two methods for gene-environment association tests: (1) a regression model defined by a combination of fixed and latent effects, implemented in the Latent Factor Mixed Models (LFMM) R package (Frichot and François 2015); (2) the mixed model which corrects for population structure and genetic relatedness implemented in the Efficient Mixed-Model Association (EMMA) R package (Kang *et al.* 2008). A False Discovered Rate of 0.05 was used for determining significance. Putative genes were recovered by performing the screening of annotation file for overlaps SNP position: we used bedtools-intersect option in the program bedtools-2.26.0 (Quinlan and Hall 2010).

Enrichment test for annotated candidate regions

A total of 26,198 genes were predicted in the *D. rotundata* reference genome (Tamiru *et al.* 2017). Based on this gene annotation, our candidate regions were tested for enrichment of gene ontology (GO) molecular function terms. We used a standard Fisher's exact test implemented in the R package TopGO (Alexa *et al.* 2006). We restricted our analysis to a minimum of five annotated genes per term in order to limit statistical artifacts of GO terms with less annotated genes.

Results

Genetic structure of population reveals hybrids and four ancestral populations

The filtered whole-genome sequencing dataset contains a total of 341,050 SNPs. These markers supported a genetic structuration of the 167 yam individuals, consisting in four subpopulations (Supplementary information Figure S1; Fig. 1). Three groups corresponded, respectively, to 1) *D. abyssinica* and 2) *D. praehensilis* from West-Africa and 3) all *D. rotundata* accessions. The fourth subpopulation included the accessions from the wild species *D. abyssinica* and *D. praehensilis* originating from Cameroon (Fig.1). This structuration is in perfect agreement with previous analysis (Akakpo *et al.* in prep). We found that the ancestry coefficients were almost identical between the two programs sNMF and Admixture with a correlation coefficient of ancestry of r^2=0.99 ($p.value$<0.001). PCA also divided the 167 yams in four subgroups

following the likely K=4 clusters. We identified fourteen accessions showing hybrid patterns (ancestry ranged from 15% to 78%) whatever the number of subpopulations taken into account. These individuals were discarded from the dataset for further analysis.

Nineteen candidates genes were associated with climatic variability

We performed association tests using a total of 153 (=167-14) yam accessions. PCA analysis on the 19 bioclimatic variables captured 76.8% of the variance on the two first principal components (PC). The two first axes explained 78% of the variance (Supplementary Figure S6). Precipitation variables (positive values – BIO 14, 17 and 18) and temperature variables in warmest conditions (negative values – BIO 4, 5 and 10) contributed to axis 1, and that other temperature variables contributed to axis 2 (negative values – BIO 1, 6, 8, 9, and 11) (Supplementary information Figure S2). Together, the LFMM and EMMA methods for association tests allowed detecting 100 candidate SNPs at 5% of significance, as associated with the PC1 (Fig. 3-a) and PC2 (Fig. 3-b). The 100 SNPs position overlapped with 19 19 different putative genes region. Among these 19 genes, 16 were annotated (Supplementary Table S2), and were thus used to test for gene ontology (GO) enrichment.

Annotated candidates genes were enriched in function involved in chlororespiration

Among the 26,198 predicted genes for the reference genome, 13,923 were annotated with sufficiently supported GO terms belonging to the biological process (BP) category. Our 16 annotated candidate genes were enriched in twenty significant GO terms (Supplementary information Table S1). Particularly, we found enrichment of genes associated with oxidoreductase (GO:0016651 ; GO:0050664) and dehydrogenase (NADH DH) activities (GO:0008137 ; GO:0050136 ; GO:0003954) (Fig. 4).

Conclusion

The two programs used to structure the genetic diversity in yam revealed that the wild yam from Cameroun were genetically different from the other west-African one. This structuration is in perfect agreement with previous analysis (Scarcelli *et al.* in prep). Additionally, we observe hybrids lines individuals that were not relevant for association tests. These individuals were then discarded. The association test between SNP and environment variability led to detect nineteen significant genes under selection among which sixteen were annotated according to the reference genome. Surprisingly, these sixteen annotated genes were significantly enriched in function associated with the NADPH complex (Miller *et al.* 2009). NADPH is a gene complex involved in chlororespiration activity. Such genes complex was relevant to understand

yam adaptation, since the domestication of yam led to spread from shading environment (savanna/forest) to full sunlight in field: the photosynthesis activity might be reshaped. We previously detected the NADPH complex as selected during yam domestication (Akakpo *et al.* 2017).

Litterature cited

Akakpo R, Scarcelli N, Chaïr H, Dansi A, Djedatin G, Thuillet A-C, Rhoné B, François O, Alix K, Vigouroux Y. 2017. Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes. BMC Genomics 18:782.

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinforma. Oxf. Engl. 22:1600–1607.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. [Internet]. Available from: http://genome.cshlp.org/content/early/2009/07/31/gr.094052.109

Baxter A, Mittler R, Suzuki N. 2013. ROS as key players in plant stress signalling. J. Exp. Bot.:ert375.

Bennoun P. 1982. Evidence for a respiratory chain in the chloroplast. Proc. Natl. Acad. Sci. U. S. A. 79:4352–4356.

Bennoun P. 1994. Chlororespiration revisited: Mitochondrial-plastid interactions in Chlamydomonas. Biochim. Biophys. Acta BBA - Bioenerg. 1186:59–66.

Bijlsma R, Loeschcke V. 2005. Environmental stress, adaptation and evolution: an overview. J. Evol. Biol. 18:744–749.

Desplats C, Mus F, Cuiné S, Billon E, Cournac L, Peltier G. 2009. Characterization of Nda2, a Plastoquinone-reducing Type II NAD(P)H Dehydrogenase in Chlamydomonas Chloroplasts. J. Biol. Chem. 284:4148–4157.

Diamond J. 2002. Evolution, consequences and future of plant and animal domestication. Nature 418:700–707.

Díaz M, de Haro V, Muñoz R, Quiles MJ. 2007. Chlororespiration is involved in the adaptation of Brassica plants to heat and high light intensity. Plant Cell Environ. 30:1578–1585.

Frichot E, François O. 2015. LEA: An R package for landscape and ecological association studies. Methods Ecol. Evol. 6:925–929.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics 196:973–983.

Hamon P, Dumont R, Zoundjihèkpon J, Tio-Touré B, Hamon S. 1995. Les ignames sauvages d'Afrique de l'ouest : caractéristiques morphologiques = Wild yams in West Africa :

morphological characteristics - 010004065.pdf. Available from: http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-05/010004065.pdf

Jose Quiles M null, Cuello null. 1998. Association of ferredoxin-NADP oxidoreductase with the chloroplastic pyridine nucleotide dehydrogenase complex in barley leaves. Plant Physiol. 117:235–244.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723.

Miller G, Schlauch K, Tam R, Cortes D, Torres MA, Shulaev V, Dangl JL, Mittler R. 2009. The plant NADPH oxidase RBOHD mediates rapid systemic signaling in response to diverse stimuli. Sci. Signal. 2:ra45.

Paredes M, Quiles MJ. 2017. Chilling stress and hydrogen peroxide accumulation in Chrysanthemum morifolium and Spathiphyllum lanceifolium. Involvement of chlororespiration. J. Plant Physiol. 211:36–41.

Peng L, Shimizu H, Shikanai T. 2008. The chloroplast NAD(P)H dehydrogenase complex interacts with photosystem I in Arabidopsis. J. Biol. Chem. 283:34873–34879.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Quiles MJ. 2006. Stimulation of chlororespiration by heat and high light intensity in oat plants. Plant Cell Environ. 29:1463–1470.

Quiles MJ, López NI. 2004. Photoinhibition of photosystems I and II induced by exposure to high light intensity during oat plant growth: Effects on the chloroplast NADH dehydrogenase complex. Plant Sci. 166:815–823.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc. Natl. Acad. Sci. 104:8641–8648.

Rumeau D, Bécuwe-Linka N, Beyly A, Louwagie M, Garin J, Peltier G. 2005. New Subunits NDH-M, -N, and -O, Encoded by Nuclear Genes, Are Essential for Plastid Ndh Complex Functioning in Higher Plants. Plant Cell 17:219–232.

Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, Kruse-Peeples M, Lepak N, Prim L, Romay MC, *et al.* 2017. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science 357:512–515.

Tamiru M, Natsume S, Takagi Hiroki, White B, Yaegashi H, Shimizu M, Yoshida K, Uemura A, Oikawa K, Abe A, *et al.* 2017. Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination.

204 BMC Biol. [Internet] 15. Available from:
205 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5604175/
206 Tenaillon MI, Charcosset A. 2011. A European perspective on maize history. C. R. Biol.
207 334:221–228.
208 Winham SJ, Biernacka JM. 2013. Gene-environment interactions in genome-wide association
209 studies: current approaches and new directions. J. Child Psychol. Psychiatry 54:1120–
210 1134.
211 Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A High-performance
212 Computing Toolset for Relatedness and Principal Component Analysis of SNP Data.
213 Bioinformatics:bts606.
214

215 # Title: Genetic variation associated with climate variability in the African yam domestication

216 ## Tables

217 **Table 1. Gene Ontology terms significantly enriched the 16 annotated candidate genes**

N°	GO.ID	Term	Annotated	Significant	Expected	classicFisher
1	GO:0016651	oxidoreductase activity, acting on NAD(P...	79	2	0.1	0.0044
2	GO:0043167	ion binding	4361	11	5.55	0.0047
3	GO:0000287	magnesium ion binding	100	2	0.13	0.007
4	GO:0050664	oxidoreductase activity, acting on NAD(P...	6	1	0.01	0.0076
5	GO:0016408	C-acyltransferase activity	9	1	0.01	0.0114
6	GO:0004743	pyruvate kinase activity	13	1	0.02	0.0164
7	GO:0030955	potassium ion binding	13	1	0.02	0.0164
8	GO:0031420	alkali metal ion binding	13	1	0.02	0.0164
9	GO:0015416	ATPase-coupled organic phosphonate trans...	18	1	0.02	0.0227
10	GO:0015604	organic phosphonate transmembrane transp...	18	1	0.02	0.0227
11	GO:0010333	terpene synthase activity	22	1	0.03	0.0277
12	GO:0043225	ATPase-coupled anion transmembrane trans...	23	1	0.03	0.0289
13	GO:0016838	carbon-oxygen lyase activity, acting on ...	25	1	0.03	0.0314
14	GO:0015605	organophosphate ester transmembrane tran...	27	1	0.03	0.0338
15	GO:0035091	phosphatidylinositol binding	28	1	0.04	0.0351
16	GO:0048038	quinone binding	30	1	0.04	0.0375

17	GO:0051539	4 iron, 4 sulfur cluster binding	32	1	0.04	0.04
18	GO:0008137	NADH dehydrogenase (ubiquinone) activity	34	1	0.04	0.0424
19	GO:0050136	NADH dehydrogenase (quinone) activity	34	1	0.04	0.0424
20	GO:0003954	NADH dehydrogenase activity	37	1	0.05	0.0461
21	GO:0016655	oxidoreductase activity, acting on NAD(P...	47	1	0.06	0.0582
22	GO:0016407	acetyltransferase activity	52	1	0.07	0.0642

Figures

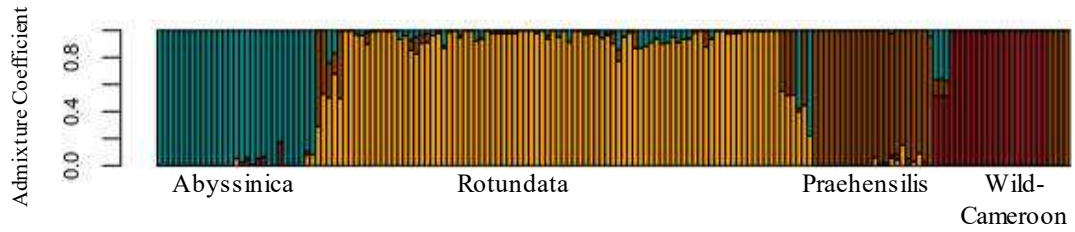


Fig 1. Structure analysis using Admixture. Each colour represents one subpopulation. The length of each segment in each vertical bar represents the proportion of ancestry in each population.

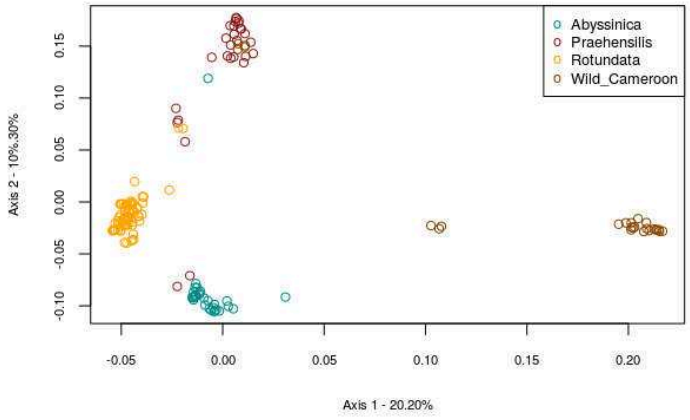
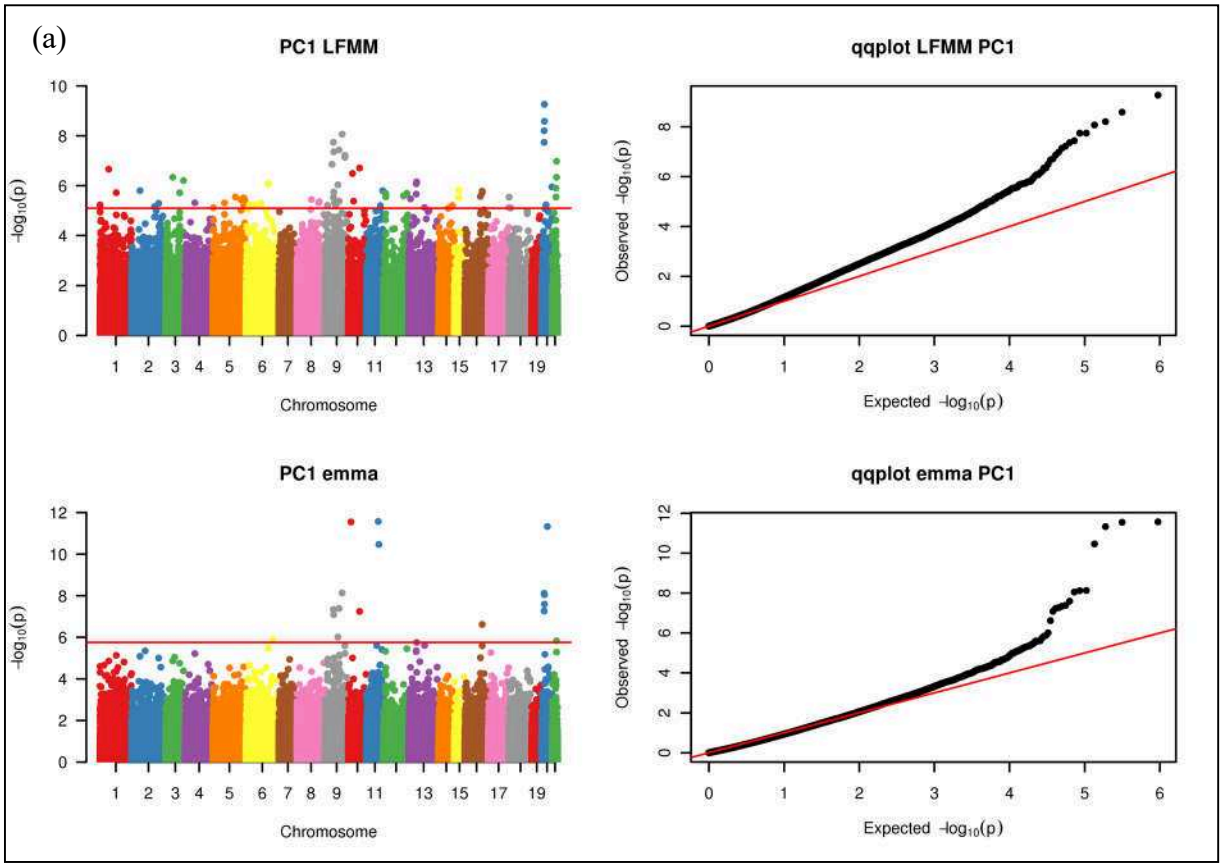


Fig 2. Genetic structure of populations using PCA approach. Each colour represents one subpopulation and each dot represents one individual.


(a)
PC1 LFMM
qqplot LFMM PC1
PC1 emma
qqplot emma PC1
−log10(p)
Observed −log10(p)
Expected −log10(p)
Chromosome
1 2 3 4 5 6 7 8 9 11 13 15 17 19

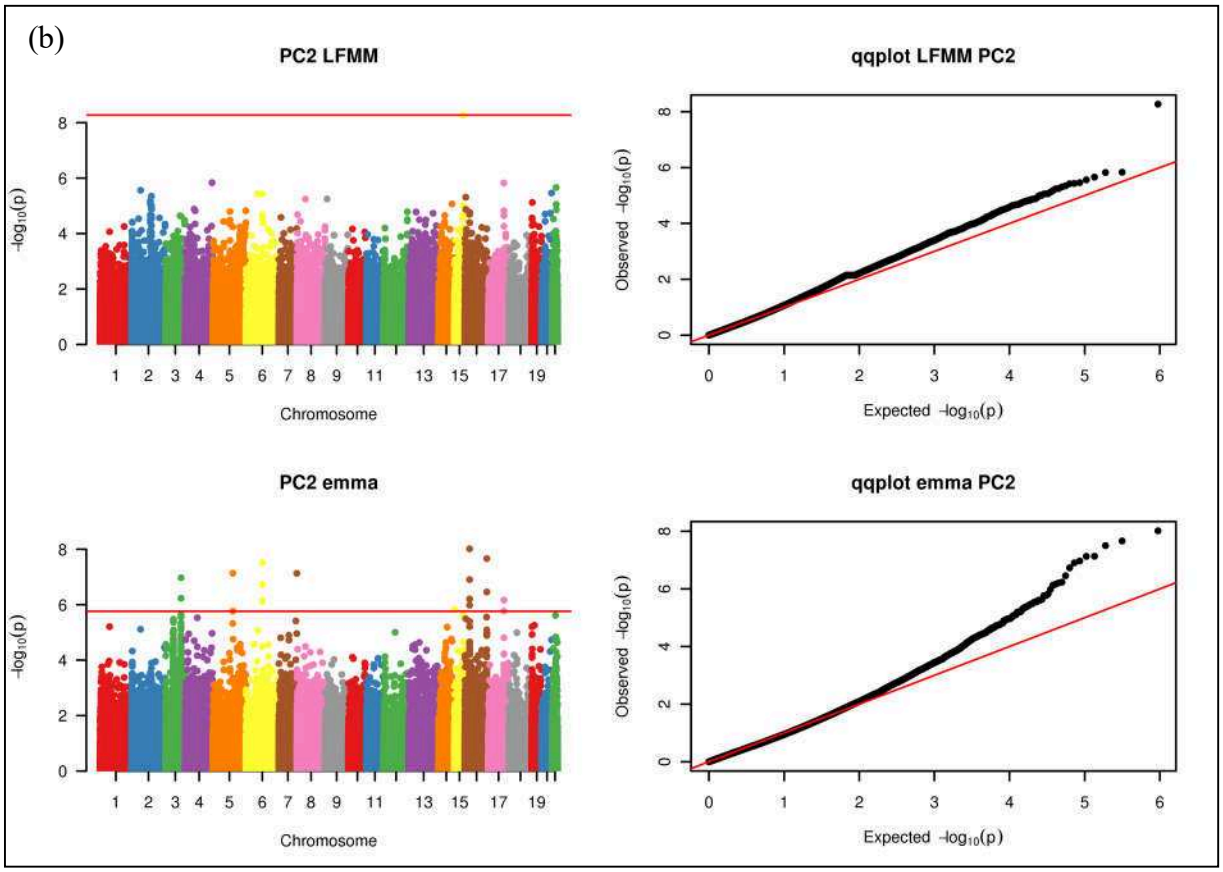

(b)
PC2 LFMM
qqplot LFMM PC2
PC2 emma
qqplot emma PC2
−log10(p)
Observed −log10(p)
Expected −log10(p)
Chromosome
1 2 3 4 5 6 7 8 9 11 13 15 17 19

Fig 3. Visualization of association test between: (a) SNPs and warmest temperature / rainfall variables (PC1); and (b) SNPs and other temperature variables (PC2). On the left are the Manhattan plots from LFMM (on top) and EMMA (on the bottom). The X axis is the genomic position of the SNPs in the genome, and the Y axis is the negative log base 10 of the p-values. Each colour represents one chromosome. Each dot represents one SNP. The red line designates the threshold of 5% significance. The SNPs at the top of the red line showed stronger associations. At the right are the corresponding qqplot.

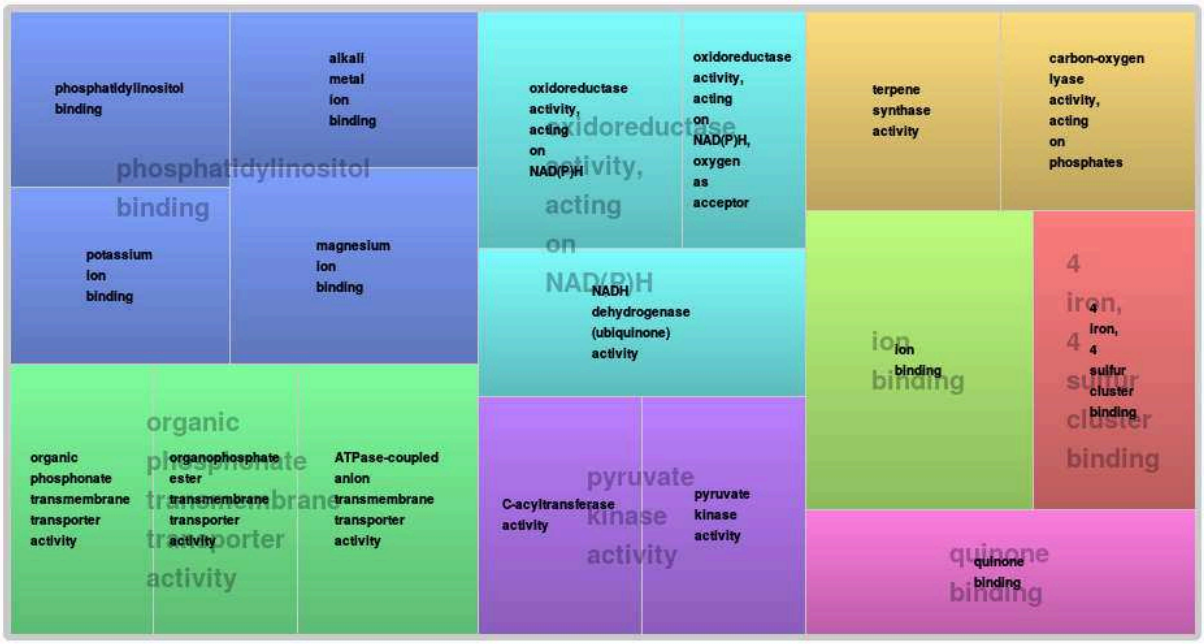


Figure 4. Gene Ontology treemap of the 19 most significant GO terms with their *p.value*

Title: Genetic variation associated with climate variability in the African yam domestication

Supplementary Information

Table S1. List of used bioclimatic variables

Variables	Description
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

Table S2. List and description of the candidate genes for association with environmental variables

Gene_ID	Chromosome	Description
Dr11523	chr10	NAC domain
Dr11547	chr10	
Dr21691	chr10	VHS|GAT|ENTH/VHS|VHS subgroup
Dr06501	chr11	ABC transporter-like|ABC-2 type transporter|Plant PDR ABC transporter associated|P-loop containing nucleoside triphosphate hydrolase|AAA+ ATPase domain
Dr19926	chr12	Protein kinase domain|Leucine-rich repeat|Leucine-rich repeat-containing N-terminal, type 2|Leucine rich repeat 4|Protein kinase-like domain|Concanavalin A-like lectin/glucanase, subgroup|Leucine-rich repeat, typical subtype|Protein kinase, ATP binding site
Dr11226	chr14	Protein of unknown function DUF247, plant|Zinc finger, CCHC-type
Dr10367	chr16	NB-ARC|P-loop containing nucleoside triphosphate hydrolase
Dr01482	chr17	Glutamyl/glutaminyl-tRNA synthetase|Glutamyl/glutaminyl-tRNA synthetase, class Ib, catalytic domain|Glutamyl/glutaminyl-tRNA synthetase, class Ib, anti-codon binding domain|Glutaminyl-tRNA synthetase, class Ib, non-specific RNA-binding domain 2|Glutaminyl-tRNA synthetase, class Ib, non-specific RNA-binding domain, N-terminal|Ribosomal protein L25/Gln-tRNA synthetase, anti-codon-binding domain|Glutamyl/glutaminyl-tRNA synthetase, class Ib, alpha-bundle domain|Ribosomal protein L25/Gln-tRNA synthetase, beta-barrel domain|Rossmann-like alpha/beta/alpha sandwich fold|Glutamine-tRNA synthetase|Aminoacyl-tRNA synthetase, class I, conserved site
Dr21242	chr18	Transcription factor IIA, alpha/beta subunit|Transcription factor IIA, helical|Transcription factor IIA, beta-barrel|Transcription factor IIA, alpha subunit, N-terminal
Dr10845	chr18	ATPase, AAA-2|P-loop containing nucleoside triphosphate hydrolase|Double Clp-N motif
Dr09277	chr20	Pyruvate kinase|Pyruvate kinase, barrel|Pyruvate kinase, C-terminal|Pyruvate kinase-like, insert domain|Pyruvate/Phosphoenolpyruvate kinase-like domain|Pyruvate kinase, beta-barrel insert domain|Pyruvate kinase, alpha/beta|Pyruvate kinase, active site

Dr00778	chr21	Terpene synthase, N-terminal domain|Terpene synthase, metal-binding domain|Terpenoid cyclases/protein prenyltransferase alpha-alpha toroid|Terpenoid synthase
Dr17208	chr4	Ferric reductase transmembrane component-like domain|FAD-binding 8|Ferric reductase, NAD binding|NADPH oxidase Respiratory burst|Cytochrome b245, heavy chain|Riboflavin synthase-like beta-barrel|EF-hand domain pair|EF-hand domain|Ferredoxin reductase-type FAD-binding domain|EF-Hand 1, calcium-binding site
Dr09852	chr5	F-box domain|Galactose oxidase/kelch, beta-propeller|Kelch-type beta propeller
Dr07452	chr5	Thiolase|Thiolase, N-terminal|Thiolase, C-terminal|Alpha-ketoglutarate-dependent dioxygenase AlkB-like|Thiolase-like|Thiolase-like, subgroup|Thiolase, active site|Thiolase, conserved site
Dr07505	chr5	NADH:ubiquinone oxidoreductase-like, 20kDa subunit
Dr11933	chr7	Phox homologous domain|Vps5 C-terminal
Dr20537	chr9	
Dr20545	chr9	

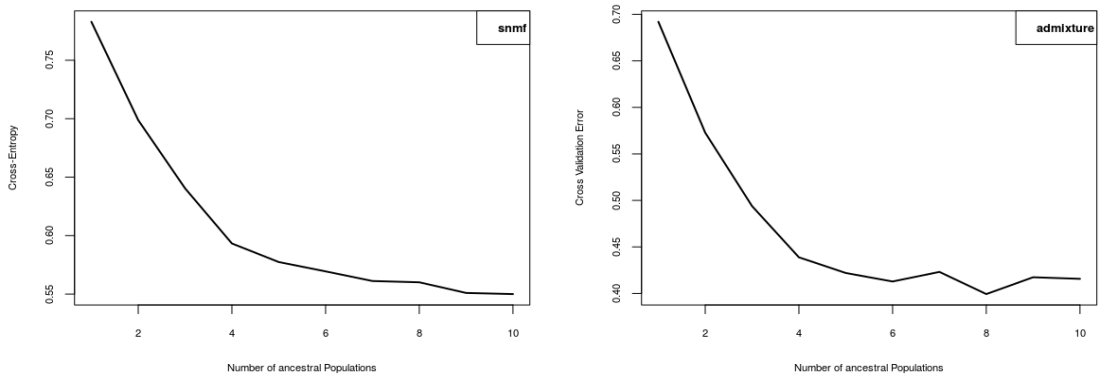


Figure S1. Cross-entropy plot for the number of cluster K = 1 - 10. At left, the cross entropy plot from sNMF, at right the cross validation error plot from Admixture. The retained value of K is K = 4.

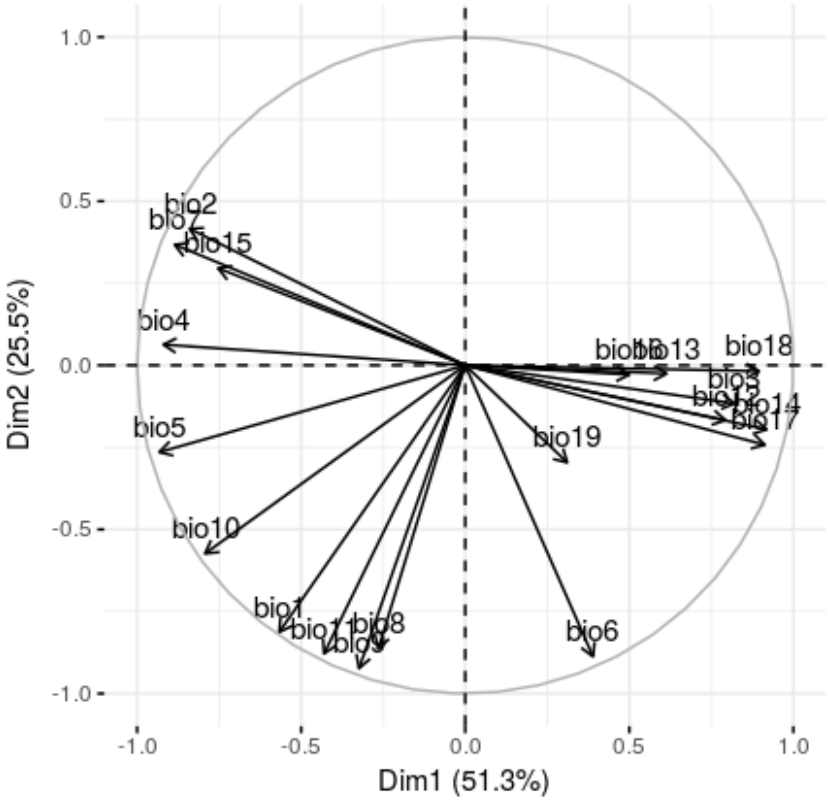


Figure S2. Graph of the 19 variables used to correlate environment with genetic variation. Axis 1 explained warmest temperature (left – Bio 4, 5, 7, 10) and rainfall variables (right – Bio 14, 17, 18) whereas axis 2 explained other temperature variables (Bio 6, 8, 9, 11).

Discussion Générale et Perspectives

« Le but de la discussion ne doit pas être la victoire, mais l'amélioration », Joseph Joubert

Les travaux de ma thèse de doctorat ont apporté des contributions originales sur la domestication de l'igname Africaine, par l'analyse et la caractérisation des gènes sélectionnés durant cette domestication, et de la structuration spatiale de la composition en éléments répétés.

Parmi les résultats importants, nous avons observé chez l'igname des sélections durant sa domestication qui rappellent celles observées chez les céréales. Des gènes impliqués dans la biosynthèse et le stockage de l'amidon : le gène *SUS4* (Baroja-Fernández *et al.* 2012) et le gène *SPS1* (Huber and Huber 1996) ont été notamment identifiés comme sélectionnés. De façon similaire, ces deux gènes ont été identifiés comme impliqués dans la domestication chez le blé (Hou *et al.* 2014) et le maïs (Li *et al.* 2013). Cette convergence dans la sélection des mêmes gènes entre les céréales et l'igname est assez remarquable. Pourquoi la sélection humaine favorisant les stockages importants d'amidon touche –t-elle ces mêmes gènes ? Bien sûr, les voix de synthèses de ces composés sont partagées entre ces espèces végétales. Mais cela souligne clairement que ces deux gènes sont « bridés » dans les populations naturelles, où la sélection semble favoriser le développement de l'embryon et non du tissu nourricier (Salamini *et al.* 2002) et que la sélection humaine favorise la sélection de variants maximisant le stockage dans les organes de réserve.

Nous avons aussi observé des sélections spécifiques à l'igname, notamment sur des gènes impliqués dans le développement racinaire et dans le phototropisme. Les gènes *SCARECROW-LIKE* et *Phot2* avaient été caractérisé pour leur implication respective dans la production des racines adventices (Sánchez *et al.* 2007) et dans le phototropisme chez les plantes supérieures (Takemiya *et al.* 2005). Les caractères relatifs à la morphologie et au développement des tubercules (notamment la taille et le nombre de tubercules) sont différents chez les plantes cultivées, et peut-être que ces gènes expliquent une partie des différences avec les populations sauvages. Nous avons enfin observé que des gènes du complexe NADH-DH, associés aux activités de chlororespiration, ont été sélectionnés lors de la domestication. Nous avons alors fait l'hypothèse que cette sélection serait potentiellement associée à une adaptation de l'igname

à la forte luminosité de plein champ. A ce stade, cela reste une hypothèse qui nécessiterait d'être validée par des expérimentations.

Le même complexe de gènes a été détecté lors de l'étude menée sur l'identification des gènes associés aux variations environnementales. Il semblerait ainsi que notre étude menée sur un gradient environnemental ait permis aussi de retrouver les gènes différenciant forme sauvage et forme cultivée. Par conséquent, une nouvelle hypothèse de travail intéressante à tester est de voir si l'étude de l'adaptation de l'igname aux variations climatiques n'est pas un moyen d'étudier la différence cultivé/sauvages et d'identifier ainsi de nouvelles cibles de la domestication.

L'étude que nous avons menée sur la fraction répétée du génome a par ailleurs révélé que la variabilité génomique serait fonction de la variabilité climatique. La question que pose ce résultat original est de savoir i) si l'association observée est reliée à l'histoire des populations ou ii) si la partie répétée du génome varie effectivement du fait d'une sélection plus globale, notamment sur la taille du génome. Il existe une corrélation entre la fraction répétée du génome et la taille du génome (Bennetzen 2000). Le fait qu'un grand nombre d'éléments répétés soit corrélé aux variables bioclimatiques pourrait favoriser une hypothèse globale de sélection sur la taille du génome complet (d'ailleurs démontrée sur d'autres complexes d'espèces, e.g. Du *et al.* 2017). Mais à ce stade, pour avancer sur ces questions, il serait nécessaire d'avoir accès à la variabilité de la taille du génome en fonction de l'environnement. Nous n'avons pas ces données actuellement, mais nos résultats suggèrent qu'elles seraient pertinentes pour la suite de nos études concernant les bases génomiques de l'adaptation chez l'igname. Nous avons aussi observé que la variation pour les abondances relatives des éléments répétés était corrélée à la structuration de la diversité génétique révélée par étude de variants SNP. Certains éléments transposables sont ainsi apparus différentiellement amplifiés au sein des sous-populations d'ignames, même si globalement l'abondance relative en éléments transposables (pour les 1 159 étudiés) était fortement corrélée entre les trois espèces d'igname étudiées.

Enfin, nos travaux ont permis de proposer une hypothèse d'origine de l'igname à partir de l'espèce de forêt *D. prahensilis*. L'agriculture en Afrique a été définie de manière récurrente comme étant originaire des savanes (Harlan 1976). Nos résultats indiqueraient plutôt une mosaïque complexe d'origines de l'agriculture des milieux de savane mais aussi de forêt ? Pour l'igname, Coursey (1976) fut un des premiers à avancer une hypothèse d'origine de forêt à partir de *D. praehensilis*, sur la base des similarités morphologiques entre l'espèce cultivée et cette espèce sauvage.

Nos résultats actuels soulèvent des questions et des voies de recherche nouvelles.

1. Introgression adaptative

Si l'igname cultivée trouve son origine en forêt à partir de l'espèce sauvage *D. praehensilis*, est-ce que son adaptation au milieu de savane ne se serait pas réalisée par des phénomènes de flux de gènes, à partir des populations sauvages de l'espèce *D abyssinica*, déjà adaptée à cet environnement ? Est-ce que des introgression adaptative récurrentes permettraient d'envisager l'adaptation de l'igname cultivée aux différentes zones climatiques ?

2. Identification de l'origine géographique

Au stade actuel de nos investigations, on ne peut identifier précisément la localisation géographique de l'événement de domestication de l'igname Africaine. Nous avons pu démontrer sa forte proximité génétique avec l'espèce forestière *D. praehensilis*, mais sans caractériser où s'est produit cette domestication en Afrique. On peut envisager aujourd'hui d'utiliser des modèles de génétique des populations permettant d'inférer le lieu géographique le plus probable de cette domestication (Tournebize *et al.* 2017). Ces modélisations démo-génétiques spatialisées seront très utiles pour une telle inférence (Ray *et al.* 2010). Cette ligne de recherche est poursuivie au laboratoire de l'équipe DynaDiv – DIADE-IRD, notamment par N. Scarcelli.

3. Etude fine du rôle des éléments transposables dans la domestication

Nous avons procédé à une annotation partielle des d'éléments transposables (ET) de l'igname, sur la base de données de séquençage en lectures courtes. Une partie des ETs de l'igname sont aussi potentiellement spécifiques de l'espèce, et il serait intéressant d'envisager une étude plus approfondie de chacun d'eux. Le génome de *D. rotundata* étant désormais disponible (Tamiru *et al.* 2017), il serait intéressant de compléter en premier lieu la base de données en éléments répétés que nous avons développée pour affiner l'identification de la fraction répétée de l'igname.

Notre approche a été une étude globale, mais nous pourrions aussi mener une étude ciblée des variations structurelles associées aux ETs au sein du génome. Il serait ensuite intéressant de savoir si ces insertions d'ETs pourraient être associées à des sélections que nous avons identifiées dans la présente étude, ou éventuellement à de nouvelles sélections, spécifiques du compartiment non codant du génome. Une analyse des insertions voisines des régions géniques permettrait de caractériser des polymorphismes de présence/absence. Par exemple, des marqueurs de type S-SAP (*Sequence-specific amplification polymorphism,* e.g. Senerchia *et al.*

2016) ancrés dans les ETs d'intérêt et développés sur le panel d'accessions d'igname étudié, en association avec du séquençage Illumina qui permettrait un alignement sur le génome de référence, amènerait à identifier des marqueurs potentiellement sous sélection. Nous pourrions ensuite étudier spécifiquement ces polymorphismes en relation avec la sélection durant la domestication, ou en relation avec les variations environnementales. Ces approches d'analyse des ETs sont bien maîtrisées au sein de l'équipe DyGAP de GQE - Le Moulon (e.g. Sarilar *et al.* 2013) et pourraient être développées en collaboration.

L'étude menée sur le complexe d'espèces de l'igname Africaine constitué des espèces sauvages *D. praehensilis* et *D. abyssinica* et de l'espèce cultivée *D. rotundata* a permis de faire des avancées importantes sur la compréhension du processus de domestication chez cette espèce cultivée, du point de vue de ses bases génomiques. Elle a mis en évidence la continuité du processus de sélection que connaît cette espèce, depuis son passage du compartiment sauvage au compartiment cultivé jusqu'à sa diffusion dans toute la *yam belt* et son adaptation à des milieux environnementaux variés. Elle a permis d'aboutir à l'hypothèse que le phénomène d'introgression adaptative, basé sur des événements d'hybridation interspécifique, pourrait jouer un rôle majeur dans l'adaptation de l'igname cultivée, espèce où la multiplication végétative est pourtant majoritaire. Des études visant à mieux comprendre ces processus adaptatifs apparaissent dès lors comme des questions particulièrement importantes. En effet, cela permettrait d'envisager une meilleure valorisation des ressources génétiques pour l'amélioration de l'igname en termes de capacités adaptatives, ce qui représente un enjeu majeur pour le développement de la culture de l'igname en Afrique de l'Ouest.

Références bibliographiques

Aldrich PR, Doebley J (1992) Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild Sorghum bicolor. TAG Theor Appl Genet Theor Angew Genet 85:293–302. doi: 10.1007/BF00222873

Ayensu ES, Coursey DG (1972) Guinea yams the botany, ethnobotany, use and possible future of yams in West Africa. Econ Bot 26:301–318. doi: 10.1007/BF02860700

Baco MN (2007) Gestion locale de la diversité cultivée au Nord Bénin : éléments pour une politique publique de conservation de l'agrobiodiversité de l'igname (Dioscorea spp.). Orléans

Baroja-Fernández E, Muñoz FJ, Li J, et al (2012) Sucrose synthase activity in the sus1/sus2/sus3/sus4 Arabidopsis mutant is sufficient to support normal cellulose and starch production. Proc Natl Acad Sci 109:321–326. doi: 10.1073/pnas.1117099109

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42:251–269

Bradshaw JE (ed) (2010) Root and Tuber Crops. Springer New York, New York, NY

Bricas N, Attaie H (1998) La consommation alimentaire des ignames. Synthèse des connaissances et enjeux par la recherche. In: L'igname, plante séculaire et culture d'avenir. Cirad, Inra, Orstom, Coraf, Coll Colloques, pp 21–30

Briggs WR, Christie JM (2002) Phototropins 1 and 2: versatile plant blue-light receptors. Trends Plant Sci 7:204–210. doi: 10.1016/S1360-1385(02)02245-8

Britten RJ (2010) Transposable element insertions have strongly affected human evolution. Proc Natl Acad Sci U S A 107:19945–19948. doi: 10.1073/pnas.1014330107

Campbell BC, Gilding EK, Mace ES, et al (2016) Domestication and the storage starch biosynthesis pathway: signatures of selection from a whole sorghum genome sequencing strategy. Plant Biotechnol J 14:2240–2253. doi: 10.1111/pbi.12578

Candolle A de (1886) Origine des plantes cultivées. F. Alcan, Paris

Chandrasekara A, Kumar TJ (2016) Roots and Tuber Crops as Functional Foods: A Review on Phytochemical Constituents and Their Potential Health Benefits. Int J Food Sci 2016:. doi: 10.1155/2016/3631647

Cooperation TC for A and R (1996) Les ignames sauvages d'Afrique de l'Ouest. Spore

Coursey DG (1976) The Origins and Domestication of Yams in Africa. In: Origins of African Plant Domestication, Reprint 2011. De Gruyter Mouton, Berlin, Boston

Dansi A, Mignouna HD, Zoundjihékpon J, et al (1999) Morphological diversity, cultivar groups and possible descent in the cultivated yams (Dioscorea cayenensis/D. rotundata) complex in Benin Republic. Genet Resour Crop Evol 46:371–388. doi: 10.1023/A:1008698123887

Dansi A, Mignouna HD, Zoundjihekpon J, et al (2000) Identification of some Benin Republic's guinea yam (Dioscorea cayenensis Dioscorea rotundata complex) cultivars using randomly amplified polymorphic DNA. Genet Resour Crop Evol 47:619–625. doi: 10.1023/A:1026589702426

Darwin C (1859) On the Origin of Species. http://www.gutenberg.org/files/1228/1228-h/1228-h.htm#link2H_4_0003. Accessed 13 Oct 2017

Darwin C, Gray A (1868) The variation of animals and plants under domestication. New York : Orange Judd & Co.

De Mita S, Thuillet A-C, Gay L, et al (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol Ecol 22:1383–1399. doi: 10.1111/mec.12182

de Wet JMJ (1978) Systematics and Evolution of Sorghum Sect. Sorghum (Gramineae). Am J Bot 65:477–484. doi: 10.2307/2442706

Demol J (2002) Amélioration des plantes: application aux principales espèces cultivées en régions tropicales. Presses Agronomiques de Gembloux

Dhont M (2010) History of oral contraception. Eur J Contracept Reprod Health Care 15:S12–S18. doi: 10.3109/13625187.2010.513071

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418:700–707. doi: 10.1038/nature01019

Díez CM, Gaut BS, Meca E, et al (2013) Genome size variation in wild and cultivated maize along altitudinal gradients. New Phytol 199:264–276. doi: 10.1111/nph.12247

Doebley J (1992) Mapping the genes that made maize. Trends Genet TIG 8:302–307

Doebley J (1989) Isozymic Evidence and the Evolution of Crop Plants. In: Isozymes in Plant Biology. Springer, Dordrecht, pp 165–191

Doebley J, Stec A, Gustus C (1995) Teosinte Branched1 and the Origin of Maize: Evidence for Epistasis and the Evolution of Dominance. Genetics 141:333–346

Doebley J, Stec A, Hubbard L (1997) The evolution of apical dominance in maize. Nature 386:485–488. doi: 10.1038/386485a0

Doebley JF, Gaut BS, Smith BD (2006) The Molecular Genetics of Crop Domestication. Cell 127:1309–1321. doi: 10.1016/j.cell.2006.12.006

Driscoll CA, Macdonald DW, O'brien SJ (2009) From Wild Animals to Domestic Pets, an Evolutionary View of Domestication. National Academies Press (US)

Dumont R, Dansi A, Vernier P, Zoundjihèkpon J (2005) Biodiversité I et domestication des ignames en Afrique de l'Ouest. 136

Dumont R, Zoundjihekpon J, Vernier P (2010) Origin and diversity of Dioscorea rotundata Poir yams. How African peasants' knowledge makes it possible for them to use wild biodiversity in farming. Cah Agric 19:255–261. doi: 10.1684/agr.2010.0411

Elias M, Panaud O, Robert T (2000) Assessment of genetic variability in a traditional cassava (Manihot esculenta Crantz) farming system, using AFLP markers. Heredity 85 Pt 3:219–230

Everett TH (1981) The New York Botanical Garden Illustrated Encyclopedia of Horticulture. Courier Corporation

Eyre-Walker A, Gaut RL, Hilton H, et al (1998) Investigation of the bottleneck leading to the domestication of maize. Proc Natl Acad Sci U S A 95:4441–4446

Fernie AR, Willmitzer L (2001) Molecular and Biochemical Triggers of Potato Tuber Development. Plant Physiol 127:1459–1465. doi: 10.1104/pp.010764

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405. doi: 10.1038/nrg2337

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Mol Biol Evol 30:1687–1699. doi: 10.1093/molbev/mst063

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Fuller DQ, Hildebrand E (2013) Domesticating Plants in Africa. doi: 10.1093/oxfordhb/9780199569885.013.0035

Fustier M-A (2016) Adaptation locale des téosintes Zea mays ssp. parviglumis et Zea mays ssp. mexicana le long de gradients altitudinaux. Université Paris sud

Fustier M-A, Brandenburg J-T, Boitard S, et al (2017) Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. Mol Ecol 26:2738–2756. doi: 10.1111/mec.14082

Germonpré M, Lázničková-Galetová M, Sablin MV (2012) Palaeolithic dog skulls at the Gravettian Předmostí site, the Czech Republic. J Archaeol Sci 39:184–202. doi: 10.1016/j.jas.2011.09.022

Girma G, Hyma KE, Asiedu R, et al (2014) Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of Guinea yams. TAG Theor Appl Genet Theor Angew Genet 127:1783–1794. doi: 10.1007/s00122-014-2339-2

Gladieux P, Zhang X-G, Róldan-Ruiz I, et al (2010) Evolution of the population structure of Venturia inaequalis, the apple scab fungus, associated with the domestication of its host. Mol Ecol 19:658–674. doi: 10.1111/j.1365-294X.2009.04498.x

González J, Karasov TL, Messer PW, Petrov DA (2010) Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila. PLOS Genet 6:e1000905. doi: 10.1371/journal.pgen.1000905

Govaerts R, Wilkin P, Saunders RMK (2007) World Checklist of Dioscoreales: Yams and their Allies | Kew Gardens Gift Shop. http://shop.kew.org/world-checklist-of-dioscoreales-yams-and-their-allies. Accessed 19 Aug 2015

Gracia MD (2014) Génomique évolutive de l'agent pathogène de la tavelure du pommier, Venturia inaequalis, dans le cadre de la domestication de son hôte. Phdthesis, Université d'Angers

Graur D, Li W-H (2000) Fundamentals of Molecular Evolution. Sinauer

Gray MM, Sutter NB, Ostrander EA, Wayne RK (2010) The IGF1 small dog haplotype is derived from Middle Eastern grey wolves. BMC Biol 8:16. doi: 10.1186/1741-7007-8-16

Griffith F (1928) The Significance of Pneumococcal Types. J Hyg (Lond) 27:113–159

Grosseteste R (1997) On the Six Days of Creation: A Translation of the Hexaëmeron. Oxford University Press, Oxford

Haldane JBS (1965) Data needed for a blueprint of the first organism. In: The origins of prebiological systems and of their molecular matrices. Ny : Academic Press

Hammer K (1984) The domestication syndrome. Kult 32:11–34. doi: 10.1007/BF02098682

Hamon P (1987) Structure, origine génétique des ignames cultivées du complexe Dioscorea cayenensis-rotundata et domestication des ignames en Afrique de l'Ouest.

Hamon P, Dumont R, Zoundjihèkpon J, et al (1995) Les ignames sauvages d'Afrique de l'ouest : caractéristiques morphologiques = Wild yams in West Africa : morphological characteristics - 010004065.pdf. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-05/010004065.pdf. Accessed 25 Jul 2016

Hancock AM, Brachi B, Faure N, et al (2011) Adaptation to Climate Across the Arabidopsis thaliana Genome. Science 334:83–86. doi: 10.1126/science.1209244

Hardigan MA, Laimbeer FPE, Newton L, et al (2017) Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proc Natl Acad Sci 114:E9999–E10008. doi: 10.1073/pnas.1714380114

Harlan JR (1971) Agricultural origins: centers and noncenters. Science 174:468–474. doi: 10.1126/science.174.4008.468

Harlan JR (1976) Origins of African Plant Domestication, Reprint 2011. De Gruyter Mouton, Berlin, Boston

Harlan JR (1992) Crops and Man. https://www.goodreads.com/work/best_book/2266751-crops-and-man. Accessed 20 Feb 2018

Haudry A, Cenci A, Ravel C, et al (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol Biol Evol 24:1506–1517. doi: 10.1093/molbev/msm077

He Z, Zhai W, Wen H, et al (2011) Two Evolutionary Histories in the Genome of Rice: the Roles of Domestication Genes. PLOS Genet 7:e1002100. doi: 10.1371/journal.pgen.1002100

Hermann D (2011) Caractérisation d'éléments transposables de type mariner chez les microalgues marines. Le Mans

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet 10:639–650. doi: 10.1038/nrg2611

Horváth V, Merenciano M, González J (2017) Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. Trends Genet TIG 33:832–841. doi: 10.1016/j.tig.2017.08.007

Hou J, Jiang Q, Hao C, et al (2014) Global selection on sucrose synthase haplotypes during a century of wheat breeding. Plant Physiol 164:1918–1929. doi: 10.1104/pp.113.232454

Huber SC, Huber JL (1996) ROLE AND REGULATION OF SUCROSE-PHOSPHATE SYNTHASE IN HIGHER PLANTS. Annu Rev Plant Physiol Plant Mol Biol 47:431–444. doi: 10.1146/annurev.arplant.47.1.431

Hufford MB, Martínez-Meyer E, Gaut BS, et al (2012) Inferences from the Historical Distribution of Wild and Domesticated Maize Provide Ecological and Evolutionary Insight. PLOS ONE 7:e47659. doi: 10.1371/journal.pone.0047659

Hurst LD (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet TIG 18:486

Idumah FO, T OP, Ighodaro UB (2014) Economics of Yam Production under Agroforestry System in Sapoba Forest Area, Edo State, Nigeria. Int J Agric For 4:440–445

Iltis HH, Doebley JF (1980) Taxonomy of Zea (Gramineae). II. Subspecific Categories in the Zea Mays Complex and a Generic Synopsis. Am J Bot 67:994–1004. doi: 10.2307/2442442

Int J Toxicol (2004) Final report of the amended safety assessment of Dioscorea Villosa (Wild Yam) root extract. Int J Toxicol 23 Suppl 2:49–54. doi: 10.1080/10915810490499055

Jia G, Huang X, Zhi H, et al (2013) A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (Setaria italica). Nat Genet 45:957–961. doi: 10.1038/ng.2673

Jian Y, Xu C, Guo Z, et al (2017) Maize (Zea mays L.) genome size indicated by 180-bp knob abundance is associated with flowering time. Sci Rep 7:5954. doi: 10.1038/s41598-017-06153-8

Jin J, Huang W, Gao J-P, et al (2008) Genetic control of rice plant architecture under domestication. Nat Genet 40:1365–1369. doi: 10.1038/ng.247

Kumar S, Das G, Shin H-S, Patra JK (2017) Dioscorea spp. (A Wild Edible Tuber): A Study on Its Ethnopharmacological Potential and Traditional Use by the Local People of Similipal Biosphere Reserve, India. Front Pharmacol 8:. doi: 10.3389/fphar.2017.00052

Kwon Y-K, Jie EY, Sartie A, et al (2015) Rapid metabolic discrimination and prediction of dioscin content from African yam tubers using Fourier transform-infrared spectroscopy

combined with multivariate analysis. Food Chem 166:389–396. doi: 10.1016/j.foodchem.2014.06.035

Lam H-M, Xu X, Liu X, et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059. doi: 10.1038/ng.715

Lebot V (2008) Section III. Yams: origin and history. In: Atherton J, Rees A (eds) Tropical root and tuber crops: cassava, sweet potato, yams and aroids. CABI, Wallingford, pp 183–190

Leimu R, Fischer M (2008) A Meta-Analysis of Local Adaptation in Plants. PLOS ONE 3:e4010. doi: 10.1371/journal.pone.0004010

Li J, Baroja-Fernández E, Bahaji A, et al (2013) Enhancing sucrose synthase activity results in increased levels of starch and ADP-glucose in maize (Zea mays L.) seed endosperms. Plant Cell Physiol 54:282–294. doi: 10.1093/pcp/pcs180

Li X, Jian Y, Xie C, et al (2017) Fast diffusion of domesticated maize to temperate zones. Sci Rep 7:2077. doi: 10.1038/s41598-017-02125-0

Luu K, Bazin E, Blum MGB (2017) pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour 17:67–77. doi: 10.1111/1755-0998.12592

Macko-Podgórni A, Machaj G, Stelmach K, et al (2017) Characterization of a Genomic Region under Selection in Cultivated Carrot (Daucus carota subsp. sativus) Reveals a Candidate Domestication Gene. Front Plant Sci 8:. doi: 10.3389/fpls.2017.00012

Makarevitch I, Waters AJ, West PT, et al (2015) Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. PLOS Genet 11:e1004915. doi: 10.1371/journal.pgen.1004915

Malapa R, Arnau G, Noyer JL, Lebot V (2005) Genetic Diversity of the Greater Yam (Dioscorea alata L.) and Relatedness to D. nummularia Lam. and D. transversa Br. as Revealed with AFLP Markers. Genet Resour Crop Evol 52:919–929. doi: 10.1007/s10722-003-6122-5

Manning K, Pelling R, Higham T, et al (2011) 4500-Year old domesticated pearl millet (Pennisetum glaucum) from the Tilemsi Valley, Mali: new insights into an alternative cereal domestication pathway. J Archaeol Sci 38:312–322. doi: 10.1016/j.jas.2010.09.007

Manu-Aduening JA, Lamboll RI, Dankyi AA, Gibson RW (2005) Cassava diversity in Ghanaian farming systems. Euphytica 144:331–340. doi: 10.1007/s10681-005-8004-8

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402. doi: 10.1146/annurev.genom.9.081307.164359

Mariac C, Scarcelli N, Pouzadou J, et al (2014) Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour 14:1103–1113. doi: 10.1111/1755-0998.12258

Marshall F, Hildebrand E (2002) Cattle Before Crops: The Beginnings of Food Production in Africa. J World Prehistory 16:99–143. doi: 10.1023/A:1019954903395

Matsuoka Y, Vigouroux Y, Goodman MM, et al (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci 99:6080–6084. doi: 10.1073/pnas.052125199

Mazoyer M, Roudart L (2002) Histoire des agricultures du monde. Du néolithique à la crise contemporaine. Le Seuil

McClintock B (1984) The significance of responses of the genome to challenge. Science 226:792–801

Merilä J, Hendry AP (2014) Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. Evol Appl 7:1–14. doi: 10.1111/eva.12137

Métailié J-P (2016) Les plantes cultivées : la contribution précolombienne à l'agriculture mondiale. 751–755

Meyer RS, DuVal AE, Jensen HR (2012) Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol 196:29–48. doi: 10.1111/j.1469-8137.2012.04253.x

Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 14:840–852. doi: 10.1038/nrg3605

Mignouna HD, Abang MM, Wanyera NW, et al (2005) PCR marker-based analysis of wild and cultivated yams (Dioscorea spp.) in Nigeria: Genetic relationships and implications for ex situ conservation. Genet Resour Crop Evol 52:755–763. doi: 10.1007/s10722-004-6128-7

Mignouna HD, Dansi A (2003) Yam (Dioscorea ssp.) domestication by the Nago and Fon ethnic groups in Benin. Genet Resour Crop Evol 50:519–528. doi: 10.1023/A:1023990618128

Mignouna HD, Dansi A, Zok S (2002) Morphological and isozymic diversity of the cultivated yams (Dioscorea cayenensis/Dioscorea rotundata complex) of Cameroon. Genet Resour Crop Evol 49:21–29. doi: 10.1023/A:1013805813522

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218. doi: 10.1146/annurev.genet.39.073003.112420

Olsen KM (2012) One gene's shattering effects. Nat Genet 44:616. doi: 10.1038/ng.2289

Oumar I, Mariac C, Pham J-L, Vigouroux Y (2008) Phylogeny and origin of pearl millet (Pennisetum glaucum [L.] R. Br) as revealed by microsatellite loci. Theor Appl Genet 117:489–497. doi: 10.1007/s00122-008-0793-4

Pyhäjärvi T, Hufford MB, Mezmouk S, Ross-Ibarra J (2013) Complex Patterns of Local Adaptation in Teosinte. Genome Biol Evol 5:1594–1609. doi: 10.1093/gbe/evt109

Quiles MJ, López NI (2004) Photoinhibition of photosystems I and II induced by exposure to high light intensity during oat plant growth: Effects on the chloroplast NADH dehydrogenase complex. Plant Sci 166:815–823. doi: 10.1016/j.plantsci.2003.11.025

Quiros CF, Ortega R, Raamsdonk L van, et al (1992) Increase of potato genetic resources in their center of diversity: the role of natural outcrossing and selection by the Andean farmer. Genet Resour Crop Evol 39:107–113. doi: 10.1007/BF00051229

Ray N, Currat M, Foll M, Excoffier L (2010) SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26:2993–2994. doi: 10.1093/bioinformatics/btq579

Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46:21–42. doi: 10.1146/annurev-genet-110711-155621

Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical Divergence and Gene Flow in the Genus Zea. Genetics 181:1399–1413. doi: 10.1534/genetics.108.097238

Sahore A (2011) Propriétés physico-chimiques et fonctionnelles des tubercules et des amidons d'igname (Dioscorea). Editions Publibook

Salamini F, Özkan H, Brandolini A, et al (2002) Genetics and geography of wild cereal domestication in the near east. Nat Rev Genet 3:429–441. doi: 10.1038/nrg817

Sánchez C, Vielba JM, Ferro E, et al (2007) Two SCARECROW-LIKE genes are induced in response to exogenous auxin in rooting-competent cuttings of distantly related forest species. Tree Physiol 27:1459–1470

Sang T, Li J (2013) Molecular Genetic Basis of the Domestication Syndrome in Cereals. In: Cereal Genomics II. Springer, Dordrecht, pp 319–340

Sarah G, Homa F, Pointet S, et al (2016) A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Mol Ecol Resour. doi: 10.1111/1755-0998.12587

Sautour M, Mitaine-Offer A-C, Lacaille-Dubois M-A (2007) The Dioscorea genus: a review of bioactive steroid saponins. J Nat Med 61:91–101. doi: 10.1007/s11418-006-0126-3

Saxena RK, Edwards D, Varshney RK (2014) Structural variations in plant genomes. Brief Funct Genomics 13:296–307. doi: 10.1093/bfgp/elu016

Scarcelli N (2005) Structure et dynamique de la diversité d'une plante cultivée à multiplication végétative : le cas des ignames au Bénin (Dioscorea sp.). https://tel.archives-ouvertes.fr/tel-00482798. Accessed 17 Feb 2015

Scarcelli N, Chaïr H, Causse S, et al (2017) Crop wild relative conservation: Wild yams are not that wild. Biol Conserv 210:325–333. doi: 10.1016/j.biocon.2017.05.001

Scarcelli N, Couderc M, Baco MN, et al (2013) Clonal diversity and estimation of relative clone age: application to agrobiodiversity of yam (Dioscorea rotundata). BMC Plant Biol 13:178. doi: 10.1186/1471-2229-13-178

Scarcelli N, Tostain S, Mariac C, et al (2006a) Genetic nature of yams (Dioscorea sp.) domesticated by farmers in Benin (West Africa). Genet Resour Crop Evol 53:121–130. doi: 10.1007/s10722-004-1950-5

Scarcelli N, Tostain S, Vigouroux Y, et al (2006b) Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. Mol Ecol 15:2421–2431. doi: 10.1111/j.1365-294X.2006.02958.x

Scarcelli N, Tostain S, Vigouroux Y, et al (2006c) Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. Mol Ecol 15:2421–2431. doi: 10.1111/j.1365-294X.2006.02958.x

Simons KJ, Fellers JP, Trick HN, et al (2006) Molecular Characterization of the Major Wheat Domestication Gene Q. Genetics 172:547–555. doi: 10.1534/genetics.105.044727

Sinzelle L, Izsvák Z, Ivics Z (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell Mol Life Sci CMLS 66:1073–1093. doi: 10.1007/s00018-009-8376-3

Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet 43:1160–1163. doi: 10.1038/ng.942

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tajima F (1983) Evolutionary Relationship of DNA Sequences in Finite Populations. Genetics 105:437–460

Takemiya A, Inoue S, Doi M, et al (2005) Phototropins Promote Plant Growth in Response to Blue Light in Low Light Environments. Plant Cell 17:1120–1127. doi: 10.1105/tpc.104.030049

Takuno S, Ralph P, Swarts K, et al (2015) Independent Molecular Basis of Convergent Highland Adaptation in Maize. Genetics 200:1297–1312. doi: 10.1534/genetics.115.178327

Tamiru M, Natsume S, Takagi H, et al (2017) Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. BMC Biol 15:. doi: 10.1186/s12915-017-0419-x

Tenaillon MI, Charcosset A (2011) A European perspective on maize history. C R Biol 334:221–228. doi: 10.1016/j.crvi.2010.12.015

Tenaillon MI, Hollister JD, Gaut BS (2010) A triptych of the evolution of plant transposable elements. Trends Plant Sci 15:471–478. doi: 10.1016/j.tplants.2010.05.003

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. Genome Biol Evol 3:219–229. doi: 10.1093/gbe/evr008

Tostain S, Agbangla C, Daïnou O (2002) Les ignames Dioscorea abyssinica et D. praehensilis en Afrique de l'Ouest : diversité génétique estimée par les marqueurs AFLP. Ann Sci Agron BéninSpécial Colloq 3:1–20

Tostain S, Agbangla C, Scarcelli N, et al (2007) Genetic diversity analysis of yam cultivars (Dioscorea rotundata Poir.) in Benin using simple sequence repeat (SSR) markers. Plant Genet Resour Charact Util 5:71–81. doi: 10.1017/S1479262107672323

Tournebize R, Manel S, Vigouroux Y, et al (2017) Two disjunct Pleistocene populations and anisotropic postglacial expansion shaped the current genetic structure of the relict plant Amborella trichopoda. PLOS ONE 12:e0183412. doi: 10.1371/journal.pone.0183412

Trut L, Oskina I, Kharlamova A (2009) Animal evolution during domestication: the domesticated fox as a model. BioEssays News Rev Mol Cell Dev Biol 31:349–360. doi: 10.1002/bies.200800070

van Heerwaarden J, Doebley J, Briggs WH, et al (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc Natl Acad Sci U S A 108:1088–1092. doi: 10.1073/pnas.1013011108

Vavilov NI, Dorofeev VF (1992) Origin and Geography of Cultivated Plants. Cambridge University Press

Vernier P, Orkwor GC, Dossou AR (2003) Studies on Yam Domestication and Farmers' Practices in Benin and Nigeria. Outlook Agric 32:35–41. doi: 10.5367/000000003101294244

Vicient CM, Casacuberta JM (2017) Impact of transposable elements on polyploid plant genomes. Ann Bot 120:195–207. doi: 10.1093/aob/mcx078

Vigouroux Y, Jaqueth JS, Matsuoka Y, et al (2002) Rate and Pattern of Mutation at Microsatellite Loci in Maize. Mol Biol Evol 19:1251–1260. doi: 10.1093/oxfordjournals.molbev.a004186

Vitte C, Fustier M-A, Alix K, Tenaillon MI (2014) The bright side of transposons in crop evolution. Brief Funct Genomics 13:276–295. doi: 10.1093/bfgp/elu002

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. Annu Rev Genet 47:97–120. doi: 10.1146/annurev-genet-111212-133526

Wallace B (1975) HARD AND SOFT SELECTION REVISITED. Evol Int J Org Evol 29:465–473. doi: 10.1111/j.1558-5646.1975.tb00836.x

Wang R-L, Stec A, Hey J, et al (1999) The limits of selection during maize domestication. Nature 398:236–239. doi: 10.1038/18435

Wang W, Feng B, Xiao J, et al (2014) Cassava genome from a wild ancestor to cultivated varieties. Nat Commun 5:5110. doi: 10.1038/ncomms6110

Whitt SR, Wilson LM, Tenaillon MI, et al (2002) Genetic diversity and selection in the maize starch pathway. Proc Natl Acad Sci 99:12959–12962. doi: 10.1073/pnas.202476999

Wicker T, Sabot F, Hua-Van A, et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:nrg2165. doi: 10.1038/nrg2165

Wilkes HG (1977) Hybridization of maize and teosinte, in mexico and guatemala and the improvement of maize. Econ Bot 31:254–293. doi: 10.1007/BF02866877

Winchell F, Stevens CJ, Murphy C, et al (2017) Evidence for Sorghum Domestication in Fourth Millennium BC Eastern Sudan: Spikelet Morphology from Ceramic Impressions of the Butana Group. Curr Anthropol 58:673–683. doi: 10.1086/693898

Wright SI, Bi IV, Schroeder SG, et al (2005) The effects of artificial selection on the maize genome. Science 308:1310–1314. doi: 10.1126/science.1107891

Yamasaki M, Wright SI, McMullen MD (2007) Genomic Screening for Artificial Selection during Domestication and Improvement in Maize. Ann Bot 100:967–973. doi: 10.1093/aob/mcm173

Yang MH, Yoon KD, Chin YW, Kim JW (2009) Phytochemical and pharmacological profiles of Dioscorea species in Korea, China and Japan. Korean J Pharmacogn

Yang Q, Li Z, Li W, et al (2013) CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci 110:16969–16974. doi: 10.1073/pnas.1310949110

Yeh C-T, White F, Bai G, et al (2012) Parallel domestication of the *Shattering1* genes in cereals. Nat Genet 44:ng.2281. doi: 10.1038/ng.2281

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160. doi: 10.1016/j.copbio.2006.02.003

Zoundjihekpon J, Hamon S, Tio-Touré B, Hamon P (1994) First controlled progenies checked by isozymic markers in cultivated yams Dioscorea cayenensis-rotundata. Theor Appl Genet 88:1011–1016. doi: 10.1007/BF00220809

Annexes

Annexes 1. Formations (Au total 22 heures enrégistrées par l'école doctorale)

1. Communication et médiation scientifique

Ateliers de Formations à destination de doctorants de l'IRD IRD Délégation Nord – Bondy

Formation à la Recherche et à la gestion de l'information scientifique Centre de documentation IRD Montpellier

2. Formations scientifiques pluridisciplinaires proposées par une école doctorale

Cours doctoral de génomique environnementale AgroParisTech/Paris

Forum Biotechno 2016 Réseau Biotechno

Software and Statistical Methods for Population Genetics 11-15 sept 2017 Laboratoire TIMC - IMAG, Aussois

3. Formations générales en langues et ouverture culturelle

Préparation au TOEIC - Niveau avancé, Université d'Evry val d'Essonne

Annexe 2. Valorisations des résultats

1. Congrès

European Plant Science Retreat (EPSR), Barcelonne 2016 EPSR (Poster)
Plant and Animal Genome XXV 2017, San Diego, CA, USA (Communication)
PhdDays 2015 ED Science du Végétal
PhdDays 2016 Ecole Doctorale Science du Végétal (Poster)
PhdDays 2017 Ecole Doctorale Science du Végétal (Poster)
Réunion Dynamique des génomes de végétaux (DynaGeV) 2015 (Poster)
Réunion Dynamique des génomes de végétaux (DynaGeV) 2016 (Communication)

2. Publications

Publiée :

Roland Akakpo, Nora Scarcelli, Hana Chaïr, Alexandre Dansi, Gustave Djedatin, Anne-Céline Thuillet, Bénédicte Rhoné, Olivier François, Karine Alix and Yves Vigouroux. 2017. Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes. BMC Genomics (IF : 3.729)

Soumis :

Roland Akakpo, Florian Maumus, Nora Scarcelli, Hana Chaïr, Christine Tranchant, Alexandre Dansi, Gustave Djedatin, Yves VIGOUROUX, Karine Alix. Evaluating the relationships between transposable element content, genetic diversity and geographical distribution in African yam (Dioscorea rotundata) and wild relatives. Frontier in Plants Sciences (IF: 4.298).

En preparation:

Nora Scarcelli, Hana Chaïr, Philippe Cubry, **Roland Akakpo**, Anne-Celine Thuillet, Cedric Mariac, Marie Couderc, Olivier François, Jude Obidiegwu, Mohammed Nasser Baco, Emmanuel Otoo, Bonaventure Sonké, Karine Alix, Yves Vigouroux. Yam genomic sequences reshape our understanding of domesticaiton in Africa Sequencing of 167 yam genomes.

Roland Akakpo, Philippe Cubry, Nora Scarcelli, Anne-Céline Thuillet, Karine Alix, Olivier François, Yves Vigouroux. « Genetic variation associated with climate variability in the African yam ».

Annexe 3. Liste des accessions utilisées

N°	Code Acession	Code Séquençage	Pays ce Collecte	Latitude	Longitude
1	A3009	RunHi1c-TAG-53	Benin	11,05215	1,50501
2	PA3085	RunHi1c-TAG-54	Benin	8,3702	2,01572
3	A420	RunHi1c-TAG-12	Benin	7,52029	1,80386
4	A467	RunHi1c-TAG-33	Benin	10,22723	2,31591
5	A5045	RunHi4c-TAG-59	Ghana	10,51775	-0,849216667
6	A5047	RunHi4c-TAG-60	Ghana	10,5179333	-0,8492
7	A5048	RunHi4c-TAG-61	Ghana	10,51795	-0,849216667
8	A5059	RunHi4c-TAG-62	Ghana	10,51865	-0,850283333
9	A5061	RunHi4c-TAG-95	Ghana	10,5186333	-0,850216667
10	A5063	RunHi4c-TAG-96	Ghana	10,5186	-0,850133333
11	A5066	RunHi4c-TAG-98	Ghana	10,5184167	-0,849733333
12	A5067	RunHi4c-TAG-99	Ghana	10,5183833	-0,849783333
13	A5068	RunHi4c-TAG-100	Ghana	10,5182	-0,849533333
14	A52	RunHi1c-TAG-1	Benin	10,47672	2,49758
15	A537	RunHi1c-TAG-34	Benin	10,18262	2,30669
16	A5495	RunHi7c-TAG-48	Cameroun	7,87281	13,59043
17	A5496	RunHi7c-TAG-49	Cameroun	7,87281	13,59043
18	A5497	RunHi7c-TAG-50	Cameroun	9,95327	13,67433
19	A5498	RunHi7c-TAG-51	Cameroun	9,95331	13,67423
20	A5499	RunHi7c-TAG-52	Cameroun	9,95331	13,67423
21	A5689	RunHi7c-TAG-70	Nigeria	9,7999444	7,457055556
22	A5690	RunHi7c-TAG-71	Nigeria	9,3864722	7,314472222
23	A5691	RunHi7c-TAG-72	Nigeria	9,2669444	7,166388889
24	A5692	RunHi7c-TAG-73	Nigeria	9,2721389	7,097861111
25	A5693	RunHi7c-TAG-74	Nigeria	10,7280278	7,511833333
26	A5694	RunHi7c-TAG-75	Nigeria	10,9325	7,644694444
27	A5695	RunHi7c-TAG-76	Nigeria	10,7516944	7,516222222
28	A5696	RunHi7c-TAG-77	Nigeria	10,9963889	7,656194444
29	A5697	RunHi7c-TAG-78	Nigeria	10,4395833	7,55175
30	A5699	RunHi7c-TAG-80	Nigeria	9,4406667	7,971638889
31	A5700	RunHi7c-TAG-81	Nigeria	9,8289444	7,967166667
32	A5701	RunHi7c-TAG-82	Nigeria	8,9079167	7,883222222
33	A5702	RunHi7c-TAG-83	Nigeria	9,1433611	7,948055556
34	A5703	RunHi7c-TAG-84	Nigeria	10,3385	7,709611111
35	A5704	RunHi7c-TAG-85	Nigeria	9,9737222	7,986861111
36	A5705	RunHi7c-TAG-86	Nigeria	8,8712222	8,432388889
37	A571	RunHi1c-TAG-9	Benin	10,21508	2,30734
38	A62	RunHi1c-TAG-10	Benin	8,45607	2,50219
39	A67	RunHi1c-TAG-11	Benin	8,37443	1,97862
40	CR3586	RunHi8c-TAG-48	Benin	10,2675	2,307222
41	CR3725	RunHi8c-TAG-49	Benin	10,229722	2,308056

42	CR4003	RunHi8c-TAG-50	Benin	10,2675	2,307222
43	CR4229	RunHi8c-TAG-51	Benin	10,275556	2,313889
44	CR4275	RunHi8c-TAG-52	Benin	10,256944	2,310556
45	CR4583	RunHi8c-TAG-53	Benin	10,2675	2,307222
46	CR4818	RunHi8c-TAG-54	Benin	10,243333	2,328611
47	CR4923	RunHi8c-TAG-7	Ghana	6,5427	-0,370483333
48	CR4941	RunHi8c-TAG-8	Ghana	6,94265	-1,258683333
49	CR4950	RunHi8c-TAG-9	Ghana	7,5408	-1,3263
50	CR4952	RunHi4c-TAG-48	Ghana	7,54115	-1,326233333
51	CR4963	RunHi8c-TAG-10	Ghana	7,7899833	-0,965733333
52	CR4965	RunHi8c-TAG-11	Ghana	7,8069667	-0,974283333
53	CR4967	RunHi4c-TAG-49	Ghana	7,8194167	-0,96355
54	CR4969	RunHi8c-TAG-12	Ghana	7,8193667	-0,963583333
55	CR4980	RunHi4c-TAG-50	Ghana	8,1473833	-1,837066667
56	CR4987	RunHi4c-TAG-51	Ghana	8,1480667	-1,837383333
57	CR4990	RunHi8c-TAG-13	Ghana	8,14805	-1,837433333
58	CR4991	RunHi4c-TAG-52	Ghana	8,14705	-1,835966667
59	CR5003	RunHi4c-TAG-54	Ghana	8,1294833	-1,7818
60	CR5005	RunHi8c-TAG-14	Ghana	8,1297	-1,782433333
61	CR5014	RunHi8c-TAG-15	Ghana	9,4227333	-0,337166667
62	CR5018	RunHi8c-TAG-16	Ghana	9,4187	-0,3299
63	CR5029	RunHi4c-TAG-56	Ghana	9,4311333	-0,186183333
64	CR5031	RunHi4c-TAG-57	Ghana	9,4311833	-0,185683333
65	CR5036	RunHi4c-TAG-58	Ghana	9,4998	-0,025583333
66	CR5097	RunHi8c-TAG-17	Nigeria	7,383333333	5,05
67	CR5099	RunHi8c-TAG-18	Nigeria	8,2	9,5
68	CR5111	RunHi8c-TAG-19	Nigeria	9,42	8,3
69	CR5118	RunHi8c-TAG-22	Benin	9,4200001	1,53
70	CR5120	RunHi8c-TAG-23	Nigeria	9,8100093	8,8551156
71	CR5345	RunHi8c-TAG-24	Cameroun	4,5781944	13,76961111
72	CR5346	RunHi7c-TAG-12	Cameroun	4,5781944	13,76961111
73	CR5348	RunHi7c-TAG-13	Cameroun	4,5781944	13,76961111
74	CR5353	RunHi7c-TAG-16	Cameroun	4,5651944	13,43747222
75	CR5387	RunHi7c-TAG-21	Cameroun	4,4568056	11,56594444
76	CR5390	RunHi8c-TAG-25	Cameroun	4,4826111	11,48544444
77	CR5392	RunHi7c-TAG-22	Cameroun	4,5546389	11,44775
78	CR5395	RunHi7c-TAG-23	Cameroun	4,5513889	11,44097222
79	CR5456	RunHi7c-TAG-37	Cameroun	4,3302222	9,437527778
80	CR5457	RunHi8c-TAG-26	Cameroun	4,3302222	9,437527778
81	CR5458	RunHi8c-TAG-27	Cameroun	4,32525	9,440444444
82	CR5461	RunHi7c-TAG-38	Cameroun	4,3728056	9,432638889
83	CR5462	RunHi7c-TAG-39	Cameroun	4,3728056	9,432638889
84	CR5465	RunHi7c-TAG-40	Cameroun	4,2311667	9,608305556
85	PCR5467	RunHi7c-TAG-41	Cameroun	4,3330278	9,685694444
86	PCR5468	RunHi8c-TAG-28	Cameroun	4,3330278	9,685694444

87	CR5477	RunHi8c-TAG-29	Cameroun	4,3821389	9,534
88	CR5478	RunHi7c-TAG-43	Cameroun	4,3821389	9,534
89	CR5489	RunHi7c-TAG-47	Cameroun	4,3665833	9,554583333
90	CR5517	RunHi7c-TAG-55	Nigeria	5,8844722	6,890555556
91	CR5522	RunHi7c-TAG-56	Nigeria	6,4708056	5,5025
92	CR5523	RunHi7c-TAG-57	Nigeria	6,4145	5,770555556
93	CR5525	RunHi8c-TAG-31	Nigeria	6,4128056	5,768972222
94	CR5526	RunHi8c-TAG-32	Nigeria	7,1538333	4,907166667
95	CR5527	RunHi8c-TAG-33	Nigeria	7,2218611	5,546861111
96	CR5533	RunHi8c-TAG-34	Nigeria	8,0751389	3,539027778
97	CR5537	RunHi8c-TAG-35	Nigeria	8,7605278	3,628305556
98	CR5540	RunHi8c-TAG-36	Nigeria	8,0797222	3,541388889
99	CR5543	RunHi7c-TAG-59	Nigeria	8,0797222	3,541388889
100	CR5546	RunHi7c-TAG-60	Nigeria	7,8465833	4,926972222
101	CR5553	RunHi7c-TAG-61	Nigeria	7,8474167	6,755777778
102	CR5555	RunHi8c-TAG-38	Nigeria	7,7277778	6,484333333
103	CR5558	RunHi8c-TAG-39	Nigeria	7,7685278	5,739
104	CR5565	RunHi8c-TAG-40	Nigeria	7,8038333	5,419027778
105	CR5567	RunHi8c-TAG-41	Nigeria	7,8253611	6,046583333
106	CR5573	RunHi8c-TAG-42	Nigeria	9,2583889	7,054527778
107	CR5574	RunHi7c-TAG-62	Nigeria	9,2583889	7,054527778
108	CR5583	RunHi8c-TAG-43	Nigeria	8,9085833	7,879888889
109	CR5589	RunHi7c-TAG-63	Nigeria	9,9063056	7,950777778
110	CR5591	RunHi8c-TAG-44	Nigeria	10,3301944	7,709638889
111	CR5599	RunHi8c-TAG-45	Nigeria	8,3443889	8,57475
112	CR5602	RunHi7c-TAG-64	Nigeria	8,8999444	8,411388889
113	CR5615	RunHi7c-TAG-65	Nigeria	7,2938056	8,815861111
114	CR5639	RunHi7c-TAG-66	Nigeria	6,0283889	7,489805556
115	CR5663	RunHi7c-TAG-68	Nigeria	7,158	7,783666667
116	CR5682	RunHi7c-TAG-69	Nigeria	7,5356944	9,7155
117	CR629	RunHi8c-TAG-46	Benin	10,2	2,3
118	CR668	RunHi1c-TAG-38	Benin	10,2	2,3
119	CR685	RunHi1c-TAG-39	Benin	10,2	2,3
120	CR694	RunHi1c-TAG-5	Benin	10,24	2,42
121	CR702	RunHi1c-TAG-7	Benin	7,52921	1,79751
122	CR703	RunHi1c-TAG-8	Benin	7,52921	1,79751
123	CR837	RunHi8c-TAG-47	Benin	7,24	1,74
124	CR844	RunHi1c-TAG-47	Benin	7,43	1,75
125	CR849	RunHi1c-TAG-50	Benin	10,4	2,27
126	CR869	RunHi1c-TAG-51	Benin	10,48	2,52
127	P2990	RunHi1c-TAG-52	Benin	8,4282	2,03437
128	P323	RunHi1c-TAG-3	Benin	8,36629	2,01634
129	P424	RunHi1c-TAG-22	Benin	7,49582	1,8156
130	P425	RunHi1c-TAG-23	Benin	7,49582	1,8156
131	P457	RunHi1c-TAG-24	Benin	6,38818	2,15634

132	P462	RunHi1c-TAG-25	Benin	6,38787	2,15628
133	P464	RunHi1c-TAG-32	Benin	6,38784	2,15641
134	P4917	RunHi4c-TAG-39	Ghana	5,6951167	-0,615683333
135	P4918	RunHi4c-TAG-40	Ghana	5,71195	-0,664666667
136	P4919	RunHi4c-TAG-41	Ghana	5,7119167	-0,6647
137	P4920	RunHi4c-TAG-42	Ghana	5,712	-0,664816667
138	P4921	RunHi4c-TAG-43	Ghana	5,7118167	-0,66495
139	P4928	RunHi4c-TAG-44	Ghana	6,5385667	-0,394383333
140	P4936	RunHi4c-TAG-46	Ghana	6,4729333	-1,073083333
141	P4937	RunHi4c-TAG-47	Ghana	6,4726833	-1,07325
142	P5307	RunHi7c-TAG-7	Cameroun	3,8940556	12,089
143	P5318	RunHi7c-TAG-8	Cameroun	3,8207222	13,33413889
144	P5331	RunHi7c-TAG-9	Cameroun	3,2435278	13,60616667
145	P5344	RunHi7c-TAG-11	Cameroun	4,5748333	13,777
146	P5350	RunHi7c-TAG-15	Cameroun	4,5640278	13,45838889
147	P5358	RunHi7c-TAG-17	Cameroun	4,5994722	13,16844444
148	P5369	RunHi7c-TAG-18	Cameroun	4,6546667	12,43386111
149	P5378	RunHi7c-TAG-19	Cameroun	4,4386111	11,87844444
150	P5381	RunHi7c-TAG-20	Cameroun	4,2959167	11,37836111
151	P5398	RunHi7c-TAG-24	Cameroun	4,3116389	11,28877778
152	P5404	RunHi7c-TAG-25	Cameroun	3,9729444	11,45416667
153	P5413	RunHi7c-TAG-26	Cameroun	3,6323889	11,49591667
154	P5417	RunHi7c-TAG-27	Cameroun	3,4791389	11,70922222
155	P5420	RunHi7c-TAG-28	Cameroun	3,2093056	11,79241667
156	P5424	RunHi7c-TAG-29	Cameroun	2,8124722	12,13127778
157	P5427	RunHi7c-TAG-30	Cameroun	2,82825	12,28416667
158	P5430	RunHi7c-TAG-31	Cameroun	2,95975	11,90480556
159	P5434	RunHi7c-TAG-32	Cameroun	3,0235278	11,66608333
160	P5438	RunHi7c-TAG-33	Cameroun	2,9306944	11,39494444
161	P5441	RunHi7c-TAG-34	Cameroun	2,9706944	10,97936111
162	P5448	RunHi7c-TAG-35	Cameroun	3,2964167	11,45877778
163	P5472	RunHi7c-TAG-42	Cameroun	4,3442778	9,591027778
164	P5483	RunHi7c-TAG-45	Cameroun	4,5044167	9,566694444
165	P5708	RunHi7c-TAG-87	Nigeria	7,8571944	3,680861111
166	P5709	RunHi7c-TAG-88	Nigeria	7,3898056	5,702305556
167	P5710	RunHi7c-TAG-89	Nigeria	7,8465833	4,926972222
168	P5713	RunHi7c-TAG-90	Nigeria	7,5911111	4,351027778
169	P5716	RunHi7c-TAG-91	Nigeria	8,0686111	6,786555556
170	P5717	RunHi7c-TAG-92	Nigeria	8,4476389	6,98575
171	P5720	RunHi7c-TAG-93	Nigeria	7,82525	6,031888889
172	P5723	RunHi7c-TAG-94	Nigeria	5,8426944	7,4745
173	P5725	RunHi7c-TAG-95	Nigeria	5,32125	7,630222222
174	P5728	RunHi7c-TAG-96	Nigeria	7,158	7,783666667
175	P5729	RunHi7c-TAG-97	Nigeria	6,6533889	7,377472222
176	P5731	RunHi7c-TAG-98	Nigeria	6,3769167	7,496194444

177	P5744	RunHi7c-TAG-101	Nigeria	6,4748889	8,09325
178	P5746	RunHi7c-TAG-102	Nigeria	7,2014167	5,032694444
179	P599	RunHi1c-TAG-35	Benin	7,49699	1,81651
180	P624	RunHi1c-TAG-36	Benin	7,52927	1,79755

Titre : Etude de la domestication et de l'adaptation de l'igname (*Dioscorea spp*) en Afrique par des approches génomiques

Mots clés : Adaptation – *Dioscorea spp* – Domestication – Eléments transposables – Génomique des populations – NGS

Résumé : L'igname (*Dioscorea spp*) est un aliment de base de plus de 100 millions de personnes en Afrique. L'objectif de cette thèse était d'étudier la diversité génomique de l'igname, comprendre les bases génétiques de sa domestication, et d'étudier son adaptation à différentes zones climatiques. L'étude du processus de domestication de l'igname a été menée par une approche de génomique comparée entre l'espèce cultivée *D. rotundata* et deux espèces sauvages apparentées *D. praehensilis* et *D. abyssinica*, en utilisant des données de séquençage NGS génomique. Nous avons mis en évidence des sélections fortes de gènes de la voie de biosynthèse de l'amidon. Des gènes impliqués dans la morphologie des tubercules ou l'aptitude au phototropisme, ainsi que des gènes du complexe NADH deshydrogenase ont également été identifiés comme sélectionnés durant la domestication. Ce même complexe NADH-DH a également été identifié lors de la recherche de gènes associés à la distribution d'une collection d'ignames selon la variabilité climatique. Nous avons aussi créé la première banque de novo d'éléments transposables (ET) de l'igname L'étude que nous avons menée sur les éléments répétés (ER) du génome de l'igname nous a permis d'identifier une forte corrélation entre la variabilité des abondances relatives d'un grand nombre d'ERs et la variabilité climatique. Enfin, nous avons pu proposer une hypothèse quant à l'origine de l'igname cultivée *D. rotundata.* La domestication de l'igname dériverait de l'espèce inféodée au milieu forestier, *D. praehensilis*. Ces résultats remettent en cause l'hypothèse d'une origine stricte en zone de savane pour les espèces cultivées et l'agriculture en Afrique de l'Ouest.

Title : Study of the domestication and adaptation of yams (*Dioscorea spp*) in Africa using genomic approaches

Key words : Adaptation – *Dioscorea spp* – Domestication - NGS – Population genomics – Transposable elements

Abstract : Yam (*Dioscorea spp*) is a major staple for more than 100 million people in Africa. The main objectives of the present PhD project were to study yam genomic diversity, its domestication, and to characterize the genomic determinism of its adaptation to different climatic zones. We investigated the genetic basis of yam domestication in a comparative genomic approach between the cultivated species *D. rotundata* and two wild close relatives *D. praehensilis* and *D. abyssinica*, by exploiting NGS sequencing data. We demonstrated that genes from the starch biosynthesis were selected during yam domestication. Genes related to tuber morphology or phototropism ability, as well as genes of the NADH dehydrogenase complex were also under selection. The same NADH-DH complex was also identified when assessing adaptation to climate variability. We also created the first de novo database of yam transposable elements (TEs). The study we performed on these repeat elements (REs) highlighted a strong correlation between the variability in relative abundances of numerous REs and climatic variability. Finally, we were able to propose an hypothesis on the origin of the cultivated yam *D. rotundata*. Our hypothesis identifies the origin of yam in the forest areas, with the species *D. praehensilis* as the putative progenitor. Our results question the generally admitted hypothesis of savannah origins for crops and agriculture in Africa.