Genome-wide genotyping elucidates the geographical diversification and dispersal of the polyploid and clonally propagated yam (*Dioscorea alata* L.)

Bilal Muhammad Sharif[1,2,3], Concetta Burgarella[1,2,4,#], Fabien Cormier[2,5,#], Pierre Mournet[1,2], Sandrine Causse[1,2], Kien Nguyen Van[6], Juliane Kaoh[7], Mamy Tiana Rajaonah[8], Senanayake Ravinda Lakshan[9], Jeffrey Waki[10], Ranjana Bhattacharjee[11], Gueye Badara[11], Babil Pachakkil[12], Gemma Arnau[2,5], and Hana Chaïr[1,2 *]

[1]CIRAD, UMR AGAP, F34398-Montpellier, France; [2]AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France; [3]University of Vienna, Department of Evolutionary Anthropology, 1090 Vienna, Austria; [4]Uppsala University, Department of Organismal Biology, Uppsala, Sweden; [5]CIRAD, UMR AGAP, F-97170, Petit Bourg, Guadeloupe, France; [6]Plant Resources Center (PRC), An Khanh, Hoai Duc, Hanoi, Vietnam; [7]Vanuatu Agricultural Research and Technical Centre (VARTC), Espiritu Santo PB 231, Vanuatu; [8] Kew Madagascar Conservation Centre, Antananarivo 101, Madagascar,

[9]*Field Crops Research and Development Institute (FCRDI), 50270, Mahailluppallama, Anuradhapura, Sri Lanka;* [10]*National Agricultural Research Institute (NARI), P.O. Box 1639, Lae, Morobe Province, Papua New Guinea;* [11]*International Institute of Tropical Agriculture (IITA), PMB 5320, Ibadan, Oyo State, Nigeria;* [12]*Tokyo University of Agriculture (TUA), Sakuragaoka 1-1-1, Setagaya-ku, Tokyo 156-0054, Japan.*

**For correspondence. E-mail hana.chair@cirad.fr*

# contributed equally to the paper.

- **Background and Aims** Inferring the diffusion history of many human-dispersed species is still not straightforward due to unsolved past human migrations. The centre of diversification and routes of migration of the autopolyploid and clonally propagated greater yam, *Dioscorea alata*, one of the oldest edible tubers is still unsolved. Here, we address yam demographic and dispersal history using a worldwide sample.
- **Methods** We characterized genome-wide patterns of genetic variation by genotyping by sequencing 643 greater yam accessions spanning four continents. First, we disentangled the polyploid and clonal components of yam diversity using alleles frequency distribution and identity by descent approaches. Then, we addressed yam geographical origin and diffusion history with a model-based coalescent inferential approach.
- **Key Results** Diploid genotypes were more frequent than triploids and tetraploids in all the continents. Genetic diversity was generally low and clonality appeared to be a main factor of diversification. The most likely evolutionary scenario supported an early divergence of mainland Southeast Asian and Pacific gene pools with continuous migration between them. Triploids and tetraploids genetic make-up suggests that they have originated from these two regions before westward yam migration. The Indian

Peninsula gene pool gave origin to the African gene pool, which was later introduced in the Caribbean region.

- **Conclusions** Our results are congruent with the hypothesis of independent domestication origins of the two main Asian and Pacific gene pools. The low genetic diversity and high clonality observed suggest a strong domestication bottleneck followed by thousands of years of widespread vegetative propagation and polyploidisation. Both processes reduced the extent of diversity available for breeding, which most likely threaten future adaptation.

Key words: clonal propagation, demography, geographical distribution, polyploidy, population genomics, yam, *Dioscorea alata*.

INTRODUCTION

The present geographical distribution of flora and fauna is often the results of a long process of human-mediated dispersal, which had profound effects on the genetic diversity and the demographic histories of plant and animal populations (Almathena *et al.* 2016; Boivin *et al*. 2012). Inferring the origin of domesticated species and their dispersal constitute a long-standing interest for evolutionary biologists because it informs on species ability to adapt to environmental changes and opens a window on cultural transitions and spatial expansions in human history. From a practical point of view, addressing the diffusion and diversification of domesticates also gives insight on relevant diversity for breeding programs and metrics for biodiversity conservation.

After domestication in their centres of origin, most of crop species have gone through different waves of translocation through human migrations and therefore have colonised different continents. Archaeological and linguistic data allowed inferring the origin and the dispersal path of several crops (Beaujard 2011; Fuller *et al.* 2011; Boivin *et al*. 2012). Concomitantly, the use of molecular markers has led to explore the assumptions withdrawn by such studies through the lens of species genetic make-up and extend the sampling to larger geographical areas (Perrier *et al.* 2011; Diez *et al*. 2015). A significant bound in the comprehension of crop domestication and evolution has been taken in recent years thanks to the development of whole genomic approaches and the availability of genome sequences for many species e.g. cassava (Bredeson *et al*. 2016); potato (Hardigan *et al*. 2017); maize (Kistler *et al*. 2018); and rice (Choi and Purugganan, 2018).

However, reconstructing the routes of crop translocation is not straightforward due to unsolved geographical migrations and in some cases to the biological specificities of the species itself. For Asian crops, two domestications centres were identified, one in Southeast

Asia such as for rice (Fuller *et al.*, 2009) and common millet (Lu 83 *et al.*, 2009) and one in Papua New Guinea as for banana (Perrier *et al.*, 2011). Later these species have spread in different continents through several corridors. The Indian Ocean has been the focus of attention as it is considered one of the most important corridor of wild and domestic plant and animal translocation from Asia to Africa and vice versa, playing an important role in fauna and flora globalisation (Boivin *et al*. 2013; Boivin *et al*. 2014). The Wallace's lane, separating Sunda (Eurasian) and Sahul (mainland Australia, Tasmania, and New Guinea) plates, which converged during the late Miocene and Pliocene resulting in the dispersal of diverse Sunda fauna and flora into the newly formed lowland areas of New Guinea, is another corridor involved in the still unresolved geographical domestication and translocation of many species between Asia and Oceania (Richardson *et al*. 2012 Morley 2018).

Greater yam, *Dioscorea alata* L., belongs to the Dioscoreaceae family. Unlike other edible yam species, its pantropical distribution could mainly be explained by its ease of cultivation and broad tolerance to different environments (Orkwor and Asadu, 1998). *Dioscorea alata* is cultivated, for its starchy tubers, in upland parts of Asia (ZhiGang *et al*. 2014), tropical America (Siqueira *et al*. 2014), Africa (Egesi *et al.* 2003; Girma *et al.* 2014) and in the Pacific (Lebot *et al*. 1998). In the two latter regions, it is of utmost importance for food security while also having a considerable social and cultural status. In traditional agrosystems, yams are cultivated exclusively through clonal propagation and selection of somaclonal mutants (Vandenbroucke *et al.* 2016). Flowering is erratic or absent in many cultivars, but new combinations can still be obtained via residual sexual reproduction (Abraham *et al*. 2013). *Dioscorea alata* is a dioecious, autopolyploid species (2n = 2X= 40, 3X=60 and 4x = 80) (Arnau *et al.* 2009), triploids and tetraploids are the result of unreduced gametes (Nemorin *et al.*, 2013). The species is not found in the wild and no wild relatives have been clearly identified (Lebot 2009). It has been suggested that its closest relative is *D.*

hamiltonii Hook. f., based on taxonomical relatedness (Coursey 1967), or *D. nummularia* Lam. based on AFLP markers (Malapa *et al*. 2005). These hypotheses were later dismissed by the studies on chloroplastic markers showing that *D. nummularia* is not the closest relative of *D. alata* (Wilkin *et al*. 2005; Chaïr *et al*. 2016). Therefore, the close relative remains unknown.

The Enantiophyllum clade, which includes *D. alata*, has a Laurasian origin placed in East Asia in the Late Oligocene with *D. alata* speciation dated, using plastid markers, at the late Miocene around seven million years ago (Viruel *et al*. 2016). The domestication of the species, its introduction in Africa and its worldwide dispersal are still matter of debate due to the scarcity of genomic data and archaeological remains. The most ancient archaeological evidence of the use of greater yam was obtained through an analysis of starch grains extracted from stone artefacts from Papua New Guinea dating back more than 44000 years (Summerhayes *et al.* 2010). The greater yam was dispersed from New Guinea by the first Lapita settlers who migrated eastwards, from the Bismarck Archipelago more than 3000 years ago (Kirch 2000; Bedford *et al*. 2006). Paleobotanical scholars linked its introduction in Africa to the joint diffusion of the vegecultural trio "taro (*Colocasia esculenta* (L.) Schott), banana (*Musa sp*.) and greater yam", during the westward expansion of Austronesian-speaking seafarers while colonizing Madagascar through a central Indian Ocean corridor (Fuller *et al*. 2011). Different *Dioscorea* species are currently found in America including Asian and African species. The introduction of greater yam in America was more likely concomitant to the introduction of other African crops, in the 16 - 17th centuries, following the maritime expansion of Iberians (Carney 2001).

Few studies have tackled the issue of *D. alata* worldwide genetic diversity as most works were conducted at country level, such as in Brazil (Siqueira *et al*. 2014), Jamaica (Asemota *et al*. 1996), China (ZhiGang *et al*. 2014), Vanuatu (Vandenbroucke *et al*. 2016;

Malapa *et al*. 2005) and Nigeria (Obidiegwu *et al*. 2009). The first worldwide study showed a complex pattern of zymotypes shared between continents (Lebot *et al*. 1998). A recent work analysed a sample from the Caribbean, West Africa, Vanuatu and India with microsatellite markers. In Vanuatu yam is known as an important crop and farmers are maintaining large diversity (Lebot *et al*. 1998; Vandenbroucke *et al*. 2016). The highest diversity was encountered in India and Vanuatu, which leaded to the hypothesis of two centres of diversification (Arnau *et al*. 2017).

Despite the rising importance of greater yam for food supply, little genomic resources have been developed to understand its domestication and dispersal dynamics. Based on a worldwide sampling spanning four continents and a genome-wide approach, this study aims at exploring the demographic history and dispersal of greater yam as well as the contribution of vegetative reproduction and polyploidy to the diversity of this species. Our findings enabled us to reconstruct the genetic origin and dispersal of greater yam through Indian and Atlantic oceans.

MATERIALS AND METHODS

Sample Collection and Genotype Calling

We collected 600 accessions from the species main growing areas in Asia: India, Japan, Sri Lanka and Vietnam [222], Africa: Benin, Burkina Faso, Congo, Côte d'Ivoire, Equatorial Guinea, Ghana, Madagascar, Nigeria, Sierra Leone and Togo [141], the Caribbean: Cuba, Dominican Republic, Guadeloupe, Haiti, Jamaica, Martinique, Puerto Rico, Saint Lucia and Saint Vincent [157] and the Pacific New Caledonia, Papua New Guinea and Vanuatu [123] (Supplementary Data Fig. S1). Since we had just three accessions and one accession from French Guyana and Brazil respectively, we pooled them with the Caribbean ones to simplify

the analysis (Supplementary Data Fig. S1). To set a threshold for clonal lineage identification, the sampling included two progenies of 15 accessions each with one shared parent, that were previously used to build up the *D. alata* genetic map (Cormier *et al*. 2019) along with an additional 25 known hybrids and replicates. DNA extraction and Genotyping by Sequencing (GBS) were performed as described in Cormier *et al*. 2019. GBS libraries were constructed as described by Elshire *et al*. 2011 using PstI-MseI restriction enzymes (New England Biolabs, Hitchin, UK) on 200 ng genomic DNA for ech sample, followed by ligation with a barcode adapter and a common Illumina sequencing adapter. Amplified multiplexed libraries were purified and verified to ensure that most of the DNA fragments were between 150–300 bp. Sequencing was conducted on Illumina HiSeq 3000 system (150 bp, single-end reads) at the GeT-PlaGe platform in Toulouse, France.

Fastq files were demultiplexed with GBSX, then Illumina adaptators were removed from the sequences using Cutadapt v1.9 (Martin, 2011). A total of 5,925,921,969 raw reads were obtained from the 643 accessions, available in the NCBI SRA (Sequence Read Archive), under the BioProject number PRJNA576311. The SNP calling was performed using VcfHunter package (Garsmeur *et al*. 2018) consisting of mapping step on *D. rotundata* reference genome (Tamiru *et al*. 2018) using "process_reseq" pipeline, followed by site pre-filtering using the VcfPreFilter.1.0.py program with the default parameters. SNP calling was done separately for each chromosome and 21 variant call format (VCF) files were generated and then concatenated to produce a single final VCF file.

SNPs were filtered using VCFtools v0.1.14 (Danecek *et al.*, 2011) based on the following criteria: minimum read depth 8, minimum allele count 3, minimum allele frequency at 0.05, biallelic loci only and maximum missing rate of 10%, leading to genotyping matrix of 643 genotypes and 6017 SNPs. To build up the Site Frequency Spectrum for demographic analysis, the same parameters were used on 72 non-clonally

related diploid genotypes, from Asia, Africa and Pacific, identified subsequently without applying the minimum allele frequency filtering, leading to 384469 SNPs.

Ploidy level inference from the GBS data

The ploidy level of the accessions received from the different sources was mostly unknown. It was thus crucial to first assess it. We designed a method based on the allelic frequency distribution at heterozygotic loci (PolynomPloidy.R) (Supplementary Data Method S1). First, the allelic frequency per accession and per site was calculated by dividing the number of allele reads by the total number of reads. Then the distribution of these observed allelic frequencies was plotted as a barplot (Supplementary Data Fig. S2) on which a polynomial function was fitted with no prior on its mode (*e.g.* unimodal, bimodal or trimodal). Finally, the ploidy level was inferred from the number of local maximum of the fitted polynomial function. Indeed, for a diploid individual, a normal distribution centred at mean = 0.5 was expected. While for a triploid or tetraploid the expected distributions were bimodal (⅓, ⅔) and trimodal (¼, ½, and ¾), respectively. Accessions for which the fitted polynomial function did not contain any local maximum or more than three were set as unknown ploidy.

Different parameters/thresholds (*e.g.* number of reads per site, minor allele frequency) were tested to filter the GBS data and barplots were drawn using 100, 250 or 500 classes (Supplementary Data Table S1). All combinations of parameters values, thresholds and barplot classes were tested and the combination that maximized the accuracy on a training dataset was then used on the whole dataset. Indeed, a training phase was first conducted using 33 accessions consisting of 15 diploids, 8 triploids and 10 tetraploids, as determined by flow

cytometry on which the combinations of parameters values that maximized accuracy well predicted 78% of ploidy levels.

Relatedness between diploid and polyploids

The genotyping matrix (643 genotypes x 6017 SNPs) was converted into a 012 incidence matrix with 0, 1 and 2 corresponding to homozygote for the reference allele, heterozygote and homozygote for the alternate allele, respectively. Three matrices were produced, the first with the whole dataset, the second after removal of accessions with unknown ploidy (487 genotypes), and the third included the 93 independent diploid accessions (ie after removing multi-locus genotypes, see below) as well as the triploid and tetraploid accessions. Missing data were imputed using the mean allele frequencies. Kinship between samples was calculated as 1 – the pairwise genetic distance derived from the Manhattan function and thus assessed as a percentage of common alleles. Significant kinship between samples was detected fitting a normal law on the kinship distribution between independent diploids (n=93) using the fitdist function of the fitdistrplus R package (Delignette-Muller and Dutang 2015). The significance threshold was set at a *p-value* of 0.05. These analyses were conducted under the R CRAN 3.4.0 environment (R Core Team). Finally, a network visualization based on significant kinship was drawn up using the prefuse force directed layout implemented in the Cytoscape 3.4.0 software (Shannon *et al.* 2003).

Duplicate identification

To identify multi-locus lineages (MLLs) represented by distinct multi-locus genotypes (MLG), kinship coefficient, π, and identity-by-descent (IBD) probabilities for all possible pairs among the 347 diploid accessions were calculated using PLINK software (Purcell *et al*. 2007), as described in grape (Myles *et al*. 2011). Information from the two offspring, full and half siblings, cohorts was used to set the threshold for clones, *ie* the valley between the non-related accessions and the close accessions because probably likely parent-offsprings (Supplementary Data Fig. S3 A) and the non-related accessions but genetically close as result of clonal multiplication (Supplementary Data Fig. S3B). After removing progenies, hybrids and replicates, an index of genotypic richness (*R*) was calculated as described in Dorken and Eckert 2001 along with the fixation index *Fis*. Then, we generated a sample with unique diploid genotypes and randomly selected one MLG representative from each MLL.

Genetic population analyses

The 93 unique diploid accessions were used to investigate the population structure. Diversity statistics, nucleotide diversity, Tajima's *D* (Tajima 1989), and *Fst* (Weir and Cockerham 1984), were calculated and averaged on 100-kb genomic bins containing at least three SNPs using VCFtools v0.1.14 (Danecek *et al.* 2011). Kruskal-Wallis rank sum and pairwise Wilcoxon tests were performed to assess their significance. The Principal Component Analysis (PCA) was carried out using the *FactoMiner R* package (Lê *et al.* 2008). Population structure was calculated using ADMIXTURE (Alexander *et al*. 2009), testing replicates of five at $K = 2$ to 6. After selection of the optimal number of *K* reflecting the most probable number of groups, an 80% ancestry threshold was set to assign each

individual genome to a group. Accessions with membership probabilities under 80% were considered to be of admixed origin.

Demography modelling

To more thoroughly investigate the genetic history of the identified gene pools and their evolutionary relationships, coalescent simulations were performed to compare different demographic scenarios using FASTSIMCOAL v 2.6 (Excoffier *et al*. 2013). To build the Site Frequency Spectrum (SFS), 384,469 SNPs were used after filtering for quality parameters without removing the rare variants. To account for missing data, we estimated the allele frequency by downsizing the sample size of each population to the value of the smallest population in our data set, and we built 2-Dimensional SFSs, adapting the customised R-script used in Burgarella *et al*. 2018. The likelihood of each model and the parameters estimations were optimized using 100 independent runs, each using 1 million simulations and 100 cycles of the conditional maximization algorithm (ECM). For each model, we used the point estimate of parameters resulting in the best likelihood value from the best run of the first round of simulations to refine the likelihood estimation. Likelihoods were re-estimated for each model using 100 additional runs of 1 million simulations. Then, Akaike Information Criteria (AIC) were calculated for model comparison .

First, three simple demographic models were tested, featuring Mainland Southeast Asia (MSEA), Indian Peninsula, and Pacific gene pools. Then migration was added between the Indian Peninsula and the Pacific in the two best topologies obtained from previous analyses. After analysing the relationship between these three populations, Africa was included in the best model obtained previously and three further models were tested. The Caribbean sample was not included in this analysis as yam was introduced in Americas relatively recently,

primarily from Africa via slave trade (Carney 2001). Each model is described in Supplementary Data Method S2.

RESULTS

Distribution of ploidy levels across continents

The Genotyping-By-Sequencing (GBS) of all samples produced more than five billions raw reads that were mapped to the *D. rotundata* V1 reference genome (Tamiru *et al*. 2017), resulting in the identification of 15,048,820 SNPs. Then, filters of different stringency were applied depending on the requirements of each performed analysis (*Material and Methods*).

To define individual ploidy level and sample clonal structure, we used 6017 high-quality bi-allelic SNPs well distributed across the 20 *D. alata* linkage groups (Cormier *et al*. 2019) (Supplementary Data Fig. S4). We computed the distribution of allele frequencies per accession at heterozygous loci, and were able to infer the ploidy of 487 (75.7%) accessions. We determined that 352, 100 and 34 accessions were diploid, triploid and tetraploid, respectively (Table 1). Out of the 352 diploids, 302 accessions were landraces and 50 were replicates and hybrids (Supplementary Data Table S2). No significant difference was observed in the distribution of diploid accessions among continents. In contrast, the number of triploids in Asia and the Caribbean was higher than in Africa and the Pacific. For 162 accessions (25.19 %), the model was unable to define the ploidy level due to the sequencing depth. Indeed, a stringent threshold of minimum depth per accession and per site (30X) was defined as an optimal parameter during the model training (Supplementary Data Table S1).

The genetic distance-based network computed to assess the geographical and genetic origin of polyploids showed that the accessions with unfitted ploidy level were spread evenly over the network in agreement with the lower depth sequencing rather than the geographical or phylogenetic relatedness and were thus removed (Supplementary Data Fig. S5). Diploids

belonging to the same geographical origin were clustered (Fig. 1). Most triploid accessions clustered in two groups. One was close to diploids from Asia (mostly accessions from Vietnam). Their genetic similarity, suggests that there were very few clonal lineages. This group included 60 out of the 100 triploid accessions from Asia (34 out of 41), Africa (6 out of 14) and the Caribbean (17 out of 29). The second triploid group (17 out of 100) was close to accessions from the Pacific, with more genetically distant genotypes, suggesting a higher rate of different polyploidisation events. The tetraploid genotypes were close to these two main triploid groups. The genetic proximity between diploids, triploids and tetraploids from the same geographical origin suggested that the polyploids were derived from unreduced gametes from the same diploid gene pool. The few remaining triploid and tetraploid accessions were close to the African and Caribbean diploid gene pools which may be explained by tuber dispersals via human migrations.

Populations of mixed ploidy cannot be assessed using current statistical approaches (Meirmans *et al*. 2018). As reported in previous studies and confirmed here, the natural frequency of diploid lineages outweighs the contribution of polyploids to diversity (Abraham *et al*. 2013). Moreover, genetic diversity within the diploid gene pool well represented the overall diversity of the species as polyploids arise from diploids through unreduced gametes formation (Nemorin *et al.* 2013). We thus focused the subsequent analyses on the 347 diploid accessions.

Clonal relationship between diploid accessions

After calculating the Identity by Descent (IBD) density distribution of all possible pairs of the 352 diploid accessions and setting the clonality threshold, only 53 landraces were identified as unique genotypes (UG). The other 249 landraces had at least one clone and were spread over 40 MLLs, containing between 2 and 19 accessions (Fig. 2a). The 302 diploid landraces were thus resumed into only 93 independent diploid genotypes (40 MLL and 53 UG), which highlighted high clonality within the species. The high extent of clonality was confirmed by the low genotypic richness (R = 17.98 to 39.99) and by the negative inbreeding coefficients (excess of heterozygosity) within continents (Table 1).

Among the 40 MLLs, only one MLL had clones shared between the four continents while 13 had an intercontinental distribution (Fig. 2b). Most of the MLLs (68%) had an intracontinental distribution with 19 exclusive to a single country. The Caribbean had the highest number of MLLs shared across continents (12/13), while Asia had the lowest (4/13). The non-homogeneous geographical span of some of the clonal lineages could be the consequence of their adaptation to a particular environment versus a wide range of environments, although sampling bias cannot be excluded.

Population structure and nucleotide diversity

Within each MLL, a single genotype representative of the clonal lineage was randomly chosen and added to the 53 unique genotypes leading to a set of 93 diploid genotypes that were used for the following analyses. We first assessed the population structure through a principal component analysis. PC1 separated the Asian accessions from the rest while PC2 distinguished the Pacific samples (8.48 % and 5.23% of variance explained, respectively) (Fig 3a). This genetic structure was confirmed by the ADMIXTURE analysis (Fig. 3b). We set an 80% ancestry threshold for the assignment of individuals to a genetic group, which resulted in four clusters (Supplementary Data Fig. S6). In the first cluster, 90% of the accessions were from Mainland Southeast Asia (mainly Vietnam). The second cluster included two accessions from India and Sri Lanka, four from Caribbean and one from Africa and Pacific. The third cluster was mainly from the Pacific region (77%) and the fourth cluster was mainly from Africa (66.6%) and Caribbean (25%). The 42 remaining accessions were of admixed ancestry between the three clusters, suggesting a complex history among these geographical areas. Since there was a good correspondence between genetic structure and geographical origin, and we sought to determine the origin and dispersal of yam among continents, we subdivided our sample in five gene pools for the subsequent analyses: Mainland Southeast Asia (MSEA), Indian Peninsula (InP), Pacific (Pac), African (Afr) and Caribbean gene pools (Supplementary Data Table S3). Nucleotide diversity was low in all gene pools, with the highest values obtained for Pacific and MSEA (π= 1.29e^{-5} and 1.26e^{-5} respectively) and the lowest for Caribbean and Africa (π=0.96e^{-5} and 1.10e^{-5}, respectively). A negative Tajima's *D* was obtained for all gene pools, indicating an excess of low frequency alleles, relative to intermediate frequencies, a potential signal, which could reflect demographic expansion as well as clonal reproduction. The highest negative values were obtained for Caribbean and Indian Peninsula gene pools (*D* = -0.57 and -0.43, respectively)

with lower values obtained for MSEA and Pacific (*D*=-0.07 and -0.12, respectively. The *Fst* values were very low (*Fst* = 0.001 to 0.055), indicating a weak but generally significant differentiation between the gene pools. The lowest *Fst* was obtained between Africa and Caribbean and the highest between MSEA and Indian Peninsula gene pools.

Inference of *demographic history*

First, we assessed the most likely tree topology for MSEA, Indian Peninsula and Pacific gene pools in the presence and absence of gene flow. First, three simple demographic models were tested, featuring Mainland Southeast Asia (MSEA), Indian Peninsula, and Pacific gene pools. Then migration was added between the Indian Peninsula and the Pacific in the two best topologies obtained from previous analyses. After analysing the relationship between these three populations, an AIC-based model comparison supported a scenario in which the Pacific lineages split early from an Asian ancestral population. Asian lineages split later into MSEA and IndP, and continuous migration occurred between Indian and Pacific gene pools (Supplementary Data Fig. S7a-e). We then used this inferred scenario to determine the origin of the African gene pool by comparing three scenarios whereby Africa originated from Pacific or IndP or both (Fig. 4a-c). We did not test an African origin from MSEA because previous studies provided evidence of the colonisation of Madagascar and Africa from Southeast Asia, mainly through the Austronesians (Boivin *et al*. 2013; Crowther *et al.* 2016) and our ADMIXTURE analysis did not support a common ancestry between the MSEA and African accessions. Indeed, no African accessions were assigned to the MSEA gene pool and only one was assigned to the Pacific one (Fig. 3, Supplementary Data Table S3). This analysis yielded evidence of the introduction of the African gene pool from the Indian one, and excluded major contributions from the Pacific gene pool (Fig. 4a).

DISCUSSION

Investigating the evolutionary history of a crop such as *Dioscorea alata* is a challenging task for multiple reasons. Greater yam is unknown in the wild and its wild relatives have yet to be identified. Greater yam domestication origin and dispersal are closely tied to three main historical events that are still under investigation, i.e. the Sunda or Sahul (mainland Australia, Tasmania, and New Guinea) domestication origin of many crops, human migrations from Asia to Africa via the Indian Ocean and the more recent Colombian exchange. We advanced our understanding of *D. alata* diversity and evolution by analyzing the widest and most comprehensive sample of greater yam to date, spanning the four main continents where the species is cultivated.

Clonality as factor of yam diversification

We showed that greater yam world-wide diversity is characterized by a high extent of clonality, corroborating findings from our previous studies (Arnau *et al*. 2017). Indeed, the 302 diploid accessions in our sampling were actually derived from only 93 independent genotypes (70%), which is in line with the fact that greater yam has long been cultivated by vegetative propagation in farming systems so diversification has mainly occurred by somaclonal mutation. Our results were therefore the direct consequences of this somaclonal selection and allows quantifying it. Farmers actually collect tubers in fallows, evaluate and then select them during cultivation by vegetative propagation. This is a common practice in root and tuber crop farming (McKey *et al*. 2010), and has been described in African yam (Chaïr *et al*. 2010) and cassava (Duputié *et al*. 2009).

Genetic and geographical origin of yam polyploids

We found that the most common forms were diploids, followed by triploids and tetraploids, in agreement with previous studies (Abraham and Nair 1991; Arnau *et al*. 2009). Indeed, triploid genotypes are created from a rare single unreduced gamete, while two unreduced gametes are required for tetraploid formation which is even less frequent (Nemorin *et al*. 2013). The reduced number of triploids and tetraploids relative to diploids could thus be explained by the rare occurrence of autopolyploidisation added to the erratic flowering of diploids. As most polyploid accessions were genetically close to diploids from the Asian or Pacific gene pools, independently of their geographical origin, we could reasonably hypothesize that polyploidisation occurred several times independently in these two regions before the migration of greater yam from Asia and the Pacific to Africa and the Caribbean. Interestingly, most of the African and Caribbean triploid accessions appeared to be close to diploid Asian accessions, suggesting that the Asian gene pool contributed more than the Pacific one to westward diffusion of greater yam. The high genetic similarity found within the Asian triploid accessions suggested that they arose from a few autopolyploidisation events, which have been multiplied by vegetative propagation. In the Pacific region, where yam is an important staple food crop, unlike in Asia where rice is predominant, the diversity observed in polyploids might be the result of preservation of greater diversity by farmers, or recent polyploidisation events such as for cassava in Vanuatu where recruitment of spontaneous triploid accessions by farmers is still on-going (Sardos *et al*. 2009).

Genetic evidence of yam early diffusion

Dioscorea alata harboured very low nucleotide diversity in comparison to potato diploid landraces (Hardigan *et al*. 2017) or cultivated and wild cassava (Ramu *et al.* 2017). It is also generally recognised that the reduction in sexual fitness is a domestication trait in asexually propagated crops (Meyer and Purugganan 2013; Denham *et al.*, 2020), as recently shown in potato (Hardigan *et al*. 2017). Consequently, despite the absence of identified wild relatives, based on the narrow nucleotide diversity, the high number of clonal lineages and the scarcity of flowering observed in the field, we could reasonably deduce that greater yam has undergone a strong domestication bottleneck. Most previous studies were focused on the Indian and Pacific regions, which generated incomplete and contrasting pictures of the domestication origin of *D. alata*. By expanding the sampling beyond these areas, we revealed an early split between Mainland Southeast Asia and Pacific gene pools and later the split between MSEA and Indian Peninsula. Our demographic inference findings indicated continuous migration between the Pacific and Indian Peninsula regions, which might also explain the divergence in previous hypotheses on the origin of *D. alata* (Arnau *et al*. 2017). We could not date the split between the MSEA and Pacific gene pools. However, considering that the peopling of Sahul dates back to at least 50,000 years ago (Bird *et al*. 2019), and that hunter-gatherer societies exploited endemic yams in New Guinea 49000 to 36000 years ago (Summerhayes *et al*. 2010), it is likely that yam had reached the Sahul in wild or pre-domesticated form and was domesticated later in both in MSEA and the Pacific regions. Similarly, multiple geographical domestication regions have been suggested for different crops such as maize (Kistler *et al*. 2018) and barley (Dai *et al*. 2014). Greater yam would subsequently have been dispersed eastwards by the first Lapita settlers (Kirch 2000; Bedford *et al*. 2006).

African gene pool originating from Indian peninsula

The demographic analysis indicated an Indian Peninsula origin of the African yams. None of the African accessions was assigned to the MSEA genetic pool and only one was assigned to Pacific one. The same applies to accessions from Madagascar, which is assumed to be one of the gates by which yam reached Africa (Beaujard 2011), as all but one were assigned to the Indian Peninsula gene pool. Moreover, only two polyploids were found to be genetically close to the Pacific gene pool while all African triploid and tetraploid accessions were close to the Asian gene pool. This is in accordance with an introduction of yam in Africa via the Indian Ocean. African accessions had the lowest nucleotide diversity levels, suggesting a strong founder effect. While we obtained genetic evidence of the Indian Peninsula origin of greater yam in Africa, the routes of its introduction could not be traced from this study. Two main routes are often discussed, one involves an introduction through Madagascar followed by the colonisation of Africa, while the second one assumes an entry from East Africa, mainly via the Swahili coast (Boivin *et al*. 2013). The importance of greater yam cultivation in Madagascar is well documented, even though this crop has gradually been substituted by rice (Beaujard 2011; Crowther *et al*. 2016). How yam, banana and taro all reached West Africa remains unclear but likely occurred much later than the first introductions in East Africa.

From Africa to the Carribean, slave trade as a mean of yam introduction in Americas

We gathered different lines of evidence supporting that the Caribbean gene pool is mainly originated from Africa. Genetic differentiation was almost null between African and the Caribbean diploid gene pools. Most of the accessions were assigned to the Indian Peninsula and African gene pools. Moreover, most of the clonal lineages were shared with Africa. Greater yam was most probably introduced in America during the Colombian exchange. The slave trade is known to have been a factor of introduction of African crops to tropical Americas during Colombian exchange (Boivin *et al*. 2012). Actually, yams had provisioned the slave ships that traversed the Middle Passage of Atlantic slavery for some 350 years (Carney 2001). Reports indicated the introduction of African yams without providing any taxonomic information on the species involved. The greater yam could have been introduced with Guinea yam (*D. rotundata*). While nowadays African rice (*Oryza glaberrima*), most probably introduced concomitantly to yams, was supplanted by Asian rice (*O. sativa*) (van Andel *et al*. 2016), the most cultivated species in tropical Americas is greater yam (*D. alata*).

CONCLUSIONS

In the present study, our genetic analysis and demographic inference supported an early divergence of greater yam between Mainland Southeast Asia and Pacific, followed most probably by two independent domestication events. Then the species would have reached the Indian Peninsula, subsequently Africa and from there the Caribbean. We also revealed high clonality and low nucleotide diversity, which are indicators of a strong domestication bottleneck and a diversification process achieved mainly via somaclonal accumulation. The narrow diversity raises concerns about the scope for genetic improvement of traits of interest. Future research efforts should have the double aim of exploring the adaptation of worldwide-distributed clonal lineages under different cropping systems and environmental constraints, and identify useful alleles within the untapped diversity of greater yam wild relatives.

ACKNOWLEDGEMENTS

We are grateful to Centre de Ressources Biologiques (CRP-PT) for providing us with the yam accessions. We thank Xavier Perrier for comments on the draft manuscript, Guillaume Martin for valuable advices on SNP calling, Denis Cornet for assistance with maps and Denis Filloux for providing us dried leaves of the BGPI yam accessions. Finally, we are grateful to Vincent Lebot for valuable comments and critical discussions on this paper. Part of this work was carried out using SouthGreen bioinformatics platform at Cirad (Montpellier, France). KeyGene N.V. owns patents and patent applications protecting its sequence based genotyping technologies.

FUNDING

This work was funded by the CGIAR Research Program on Roots, Tubers and Bananas (CRP-RTB) and Agropolis Foundation (N° 1403-023).

LITERATURE CITED

Abraham K, Nair PG. 1991. Polyploidy and sterility in relation to sex in *Dioscorea alata* L. (Dioscoreaceae). *Genetica* **83:** 93-97

Abraham K, Nemorin A, Lebot V, Arnau G. 2013. Meiosis and sexual fertility of autotetraploid clones of greater yam *Dioscorea alata* L. *Genetic Resources and Crop Evolution* **60:** 819-823.

Almathena F, Charruauc P, Mohandesanc E, Mwacharob JM, *et al.* 2016 Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proceedings of the National Academy of Sciences* **113:** 6707–6712.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19:** 1655-1664.

Arnau G, Bhattacharjee R, Sheela MN, *et al.* 2017. Understanding the genetic diversity and population structure of yam (*Dioscorea alata* L) using microsatellite markers. *PLoS ONE* **12.**

Arnau G, Nemorin A, Maledon E, Abraham K. 2009. Revision of ploidy status of *Dioscorea alata* L (Dioscoreaceae) by cytogenetic and microsatellite segregation analysis. *Theoretical and Applied Genetics* **118:** 1239-1249.

Asemota HN, Ramser J, LopezPeralta C, Weising K, Kahl, G. 1996. Genetic variation and cultivar identification of Jamaican yam germplasm by random amplified polymorphic DNA analysis. *Euphytica* **92:** 341-351

Beaujard P. 2011. The first migrants to Madagascar and their introduction of plants: linguistic and ethnological evidence. *Azania* **46:** 169-189

Bedford S, Spriggs M, Regenvanu R. 2006. The Teouma Lapita site and the early human settlement of the Pacific Islands. *Antiquity* **80**: 812-828

Bird MI, Condie SA, O'Connor S, *et al.* 2019. Early human settlement of Sahul was not an accident. *Scientific reports* **9:** 8220

Boivin N, Crowther A, Helm R, Fuller DQ. 2013. East Africa and Madagascar in the Indian Ocean world Journal of World. *Prehistory* **26**: 213-281.

Boivin N, Crowther A, Prendergast M, Fuller DQ. 2014. Indian Ocean Food Globalisation and Africa African. *Archaeological Review* **31:** 547-581.

Boivin N, Fuller DQ, Crowther A. 2012. Old World globalization and the Columbian exchange: comparison and contrast. *World Archaeology* **44:** 452-469.

Burgarella C, Cubry P, Kane NA, *et al.* 2018. A Western Sahara centre of domestication inferred from pearl millet genomes. *Nature Ecology and Evolution* **2:** 1377-1380.

Carney JA. 2001. African rice in the Columbian Exchange. *Journal of African History* **42:** 377-396.

Chaïr H, Cornet D, Deu M, *et al.* 2010. Impact of farmer selection on yam genetic diversity. *Conservation Genetics* **11:** 2255-2265.

Chaïr H, Sardos J, Supply A, Mournet P, Malapa R, Lebot V. 2016. Plastid phylogenetics of Oceania yams (*Dioscorea spp*, Dioscoreaceae) reveals natural interspecific hybridization of the greater yam (*D alata*). *Botanical Journal of the Linnean Society* **180:** 319-333.

Choi JY, Purugganan MD 2018. Multiple Origin but Single Domestication Led to *Oryza sativa*. *G3-Genes Genomes Genetics* **8:** 797-803.

Cormier F, Lawac F, Maledon E, *et al.* 2019. A reference high-density genetic map of greater yam (*Dioscorea alata* L.). *Theoretical and Applied Genetics* **132:** 1733-1744.

Coursey DG. 1967. Yams An account of the Nature, Origins, Cultivation and Utilisation of the Useful Members of the Dioscoreaceae. Longmans, Green and Co Ltd, Londres, UK.

Crowther A, Lucas L, Helm R, *et al.* 2016. Ancient crops provide first archaeological signature of the westward Austronesian expansion. *Proceedings of the National Academy of Sciences* **113:** 6635-6640.

Dai F, Chen Z-H, Wang X, *et al.* 2014. Transcriptome profiling reveals mosaic genomic origins of modern cultivated barley. *Proceedings of the National Academy of Sciences* **111:** 13403-13408.

Danecek P, Auton A, Abecasis G, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156-2158.

Denham T, Barton H, Castillo C, et al. 2020. The domestication syndrome in vegetatively propagated field crops. *Annals of Botany* 1–17.

Delignette-Muller ML, Dutang C. 2015. fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software* **64:** 34.

Diez CM, Trujillo I, Martinez-Urdiroz N, *et al.* 2015. Olive domestication and diversification in the Mediterranean Basin. *New Phytologist* **206:** 436-44

Dorken ME, Eckert CG. 2001. Severely reduced sexual reproduction in northern populations of a clonal plant, *Decodonverticillatus* (Lythraceae). *Journal of Ecology* **89:** 339-350.

Duputié A, Masol F, David P, Haxaire C, McKey D. 2009. Traditional Amerindian cultivators combine directional and ideotypic selection for sustainable management of cassava genetic diversity. *Journal of Evolutionary Biology* **22:** 1317-1325.

Egesi CN, Asiedu R, Egunjobi JK, Bokanga M. 2003. Genetic diversity of organoleptic properties in water yam (*Dioscorea alata* L). *Journal of the Science of Food and Agriculture* **83:** 858-865.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, *et al.* 2011. A Robust Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* **9:** e1003905.

Fuller DQ, Qin L, Zheng YF, *et al.* 2009. The Domestication Process and Domestication Rate in Rice: Spikelet Bases from the Lower Yangtze. *Science* **323**: 1607-1610.

Fuller DQ, Boivin N, Hoogervorst T, Allaby R. 2011. Across the Indian Ocean: the prehistoric movement of plants and animals. *Antiquity* **85:** 544-558.

Garsmeur O, Droc G, Antonise R, *et al.* 2018. A mosaic monoploid reference sequence for the highly complex genome of sugarcane *Nature Communications* **9:** 2638.

Girma, G, Hyma, KE, Asiedu, R, Mitchell, SE, Gedil, M, Spillane, C, 2014 Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. *Theoretical and Applied Genetics* 127, 1783-1794.

Hahn SK, Osiru DSO, Akoroda MO, Otoo JA. 1987. Yam production and its future prospects. *Outlook on Agriculture* **16:** 105-110.

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. *Proceedings of the National Academy of Sciences* **114:** E9999-E10008.

Kirch PV. 2000. Hanamiai: Prehistoric colonization and cultural change in the Marquesas Islands (East Polynesia). *Journal of Anthropological Research* **56:** 256-257.

Kistler L, Maezumi SY, Gregorio de Souza J, *et al.* 2018. Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* **362:** 1309-1313.

Lê S, Josse J, Husson F. 2008. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software* **25:** 1-18.

Lebot V. 1999. Biomolecular evidence for plant domestication in Sahul. *Genetic Resources and Crop Evolution* **46:** 619-628.

Lebot V. 2009. Tropical root and tuber crops: cassava, sweet potato, yams, aroids CABI.

Lebot V, Trilles B, Noyer JL, Modesto J. 1998. Genetic relationships between *Dioscorea alata* L. cultivars. *Genetic Resources and Crop Evolution* **45**, 499-509.

Lu HY, Zhang JP, Liu KB, *et al*. 2009. Earliest domestication of common millet (*Panicum miliaceum*) in East Asia extended to 10000 years ago. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 7367-7372.

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. *Genetic resources and crop evolution* **7:** 919-929.

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

McKey, D, Elias M, Pujol B, Duputie A. 2010. The evolutionary ecology of clonally propagated domesticated plants. *New Phytologist* **186:** 318-332.

Meirmans PG, Liu S, Van Tienderen PH. 2018. The analysis of polyploid genetic data. *Journal of Heredity* 1-14.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* **14:** 840.

Morley RJ. 2018. Assembly and division of the South and South-East Asian flora in relation to tectonics and climate change. *Journal of Tropical Ecology* **34:** 209-234.

Myles S, Boyko AR, Owens CL, *et al.* 2011. Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 3530-3535.

Nemorin A, David J, Maledon E, Nudol E, Dalon J, Arnau G. 2013. Microsatellite and flow cytometry analysis to help understand the origin of *Dioscorea alata* polyploids *Annals of Botany* **112:** 811-819.

Obidiegwu JE, Asiedu R, Ene-Obong EE, Muoneke CO, Kolesnikova-Allen M. 2009. Genetic characterization of some water yam (*Dioscorea alata* L) accessions in West Africa with simple sequence repeats. *Journal of food agriculture & environment* **7:** 634-638.

Orkwor GC, Asadu CLA 1998 Agronomy. In: Orkwor GC, Asiedu R, Ekanayake IJ (eds) Food yams: advances in research. NRCRI and IITA Ibadan, Nigeria.

Perrier X, De Langhe E, Donohue M, *et al.* 2011. Multidisciplinary perspectives on banana (*Musa spp*.) domestication. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 11311-11318.

Purcell S, Neale B, Todd-Brown K, *et al.* 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81:** 559-575.

Ramu P, Esuma W, Kawuki R, *et al.* 2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation *Nature Genetics* **49:** 959.

Richardson J, Costion C, Muellner A. 2012. The Malesian floristic interchange: plant migration patterns across Wallace's Line. In: D. Gower, K.J., J. Richardson, B. Rosen, L. Rüber, & S. Williams (Eds) (Ed.), Biotic Evolution and Environmental Change in Southeast Asia (Systematics Association Special Volume Series). Cambridge University Press, Cambridge, pp. 138-163.

Sardos J, Rodier-Goud M, Dambier D, Malapa R, Noyer J-L, Lebot. V. 2009. Evidence for spontaneous polyploidization in cassava *Manihot esculenta* Crantz. *Plant Systematics and Evolution* **283:** 203-209.

Shannon P, Markiel A, Ozier O, *et al.* 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* **13:** 2498-2504.

Siqueira M, Bonatelli M, Gunther T, *et al.* 2014. Water yam (*Dioscorea alata* L) diversity pattern in Brazil: an analysis with SSR and morphological markers. *Genetic Resources and Crop Evolution* **61:** 611-624.

Summerhayes GR, Leavesley M, Fairbairn A, *et al.* 2010. Human Adaptation and Plant Use in Highland New Guinea 49,000 to 44,000 Years Ago. *Science* **330:** 78-81.

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

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 Biology* 15.

van Andel TR, Meyer RS, Aflitos SA, *et al.* 2016. Tracing ancestor rice of Suriname Maroons back to its African origin. *Nature Plants* **2:** 16149.

Vandenbroucke H, Mournet P, Vignes H, *et al.* 2016. Somaclonal variants of taro (*Colocasia esculenta* Schott) and yam (*Dioscorea alata* L) are incorporated into farmers' varietal portfolios in Vanuatu. *Genetic Resources and Crop Evolution* **63:** 495-511.

Viruel J, Segarra-Moragues JG, Raz L, *et al.* 2016. Late Cretaceous-Early Eocene origin of yams (*Dioscorea*, Dioscoreaceae) in the Laurasian Palaearctic and their subsequent Oligocene-Miocene diversification. *Journal of biogeography* **43:** 750-762.

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

Wilkin P, Schols P, Chase MW, *et al.* 2005. A plastid gene phylogeny of the yam genus, *Dioscorea*: Roots, fruits and Madagascar. *Systematic Botany* **30:** 736-749.

ZhiGang W, XiaoXia L, XinChun L, *et al.* 2014. Genetic diversity analysis of yams (*Dioscorea spp*) cultivated in China using ISSR and SRAP markers *Genetic Resources and Crop Evolution* **61:** 639-650.

Table 1. The number of accessions per continent according to the ploidy level inferred from GBS data and the genetic parameters for diploid landraces.

N_t, Sample size; N_{2x} Sample size for diploids (after removing full and half-siblings, and replicates); UG, Number of unique genotypes; MLL, number of multi-locus lineages, between brackets shared MLLs with another continent; G, number of independent genotypes; *R*, genotypic richness index; *Fis*, fixation index per continent.

Continent		Ploidy in the whole dataset (%)				Genetic parameters for diploid landraces					
	N_t	2X	3X	4X	NA	N_{2x}	UG	MLL	G	R	*Fis*
Africa	141	44.68	9.92	7.09	38.29	63	6	13(7)	19	17.98	-0.16
Asia	222	57.20	18.47	3.15	21.17	118	20	21(4)	41	39.99	-0.15
Caribbean	157	66.87	15.92	3.18	14.01	65	18	13(12)	31	29.98	-0.07
Pacific	123	46.34	13.00	9.75	30.89	56	9	12(8)	21	19.98	-0.14
Total	643	54.74	14.93	5.28	25.03	302	53	-	112	-	-

Table 2. Population genetic statistics for the 93 diploid genotypes per geographical origin.

ΠΙ: nucleotide diversity, *D* Tajima's D, IQR : Inter-Quartile Range.

Continent	*Π* (IQR)	*D* (IQR)	*Fst* Africa	*Fst* MSEA	*Fst* Indian Peninsula	*Fst* Caribbean
All	$0.84e^{-5}$ ($0.92e^{-5}$)	-1.06 (0.98)	-	-	-	-
Africa	1.10e-5 (0.87e-5)	-0.14 (1.4)	-	-	-	-
MSEA	1.26e-5 (0.97e-5)	-0.08 (1.35)	0.03	-	-	-
Indian Peninsula	1.20e-5 (0.94e-5)	-0.43 (1.14)	0.026	0.055	-	-
Caribbean	0.96e-5 (0.81e-5)	-0.57 (1.19)	0.001	0.023	0.011	-
Pacific	1.29e-5 (0.97e-5)	-0.12 (1.4)	0.019	0.037	0.03	0.015

Figure legends

Fig. 1. The network shows the genetic relationships between diploid, triploid and tetraploid accessions. Combination of colours and shape represents ploidy levels and geographical origin.

Fig. 2. a- Clonal relationships within the 40 multi-locus lineages (MLL) identified among the 302 *Dioscorea alata* diploid accessions. Each cluster represents one MLL. Node colours correspond to the geographical origin of the clones. **b** - Geographical distribution of MLLs shared between continents. Total number of MLLs within each continent is reported: Africa 13 (shared 7); Asia (shared 4); Caribbean 13 (shared 12) and Pacific 12 (shared 8). (See Table 1)

Fig. 3. Visualization of genetic relationships between the 93 diploid *Dioscorea alata* accessions from Africa, Asia, Caribbean and Pacific after removing the clones. **a-** Principal Component Analysis depicting the 93 accessions. Square, Africa; circle; Asia; triangle, Caribbean; and diamond, Pacific. Accessions are coloured according to their assignment to the four genetic clusters after Admixture analysis. The threshold to assign a genotype to a cluster is set at 80%. **b-** Admixture barplot showing the distribution of the *K=4* genetic clusters. Within each continent, accessions are ordered according to cluster assignment proportions. **c**- Map showing the geographical distribution of the 93 accessions in each continent according to their genetic clustering.

Fig. 4. Demographic scenarios of domestication simulated with FASTSIMCOAL2.6

Split between Mainland South East Asia (MSEA) and Pacific (Pac) followed by split of Indian Peninsula (InP) populations with constant migration between InP and Pac. Scenario (a) origin of Africa (Afr) from the Indian Peninsula, (b) origin of Africa from the Pacific, (c) origin of Africa from the Indian Peninsula with gene flow between the Pacific and Africa.
N_{ANC}, ancestral population size. TDIV, time of divergence; TDI1<TIV2<TDIV3. GF, gene flow. 0.005 is the migration rate between two populations.

Figure 1

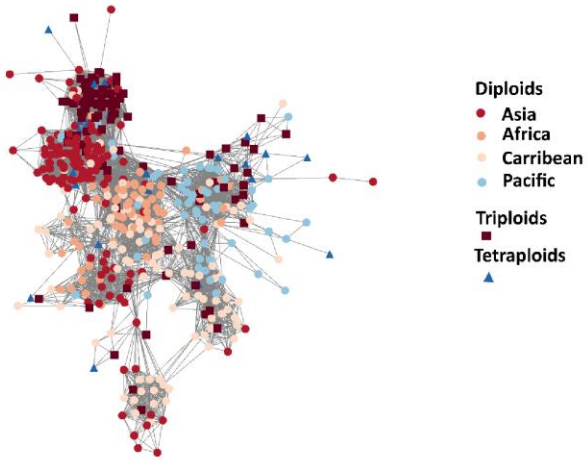

Figure 2

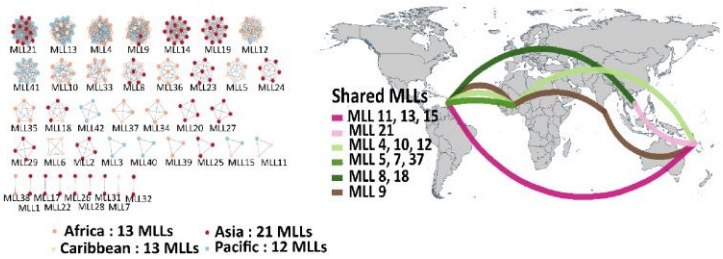

Figure 3

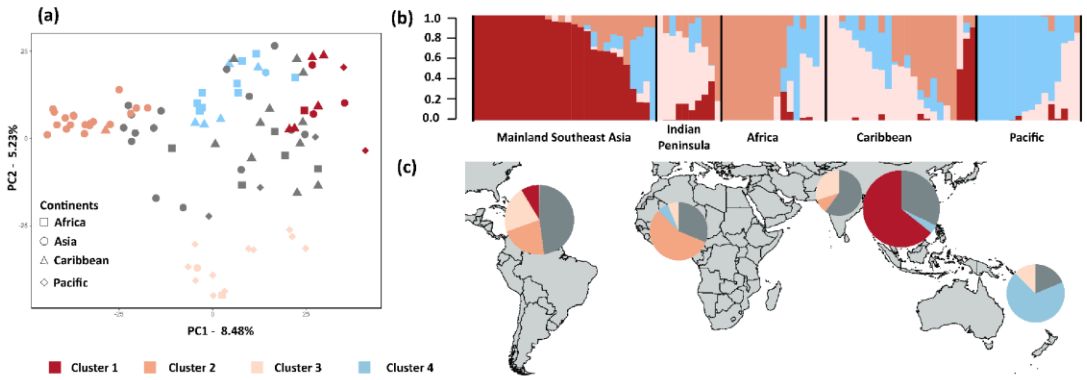

Figure 4