



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

Roland Akakpo, Nora Scarcelli, Hana Chair, Alexandre Dansi, Gustave Djedatin, Anne-Céline Thuillet, Benedicte Rhone, Olivier François, Karine Alix, Yves Y. Vigouroux

► To cite this version:

Roland Akakpo, Nora Scarcelli, Hana Chair, Alexandre Dansi, Gustave Djedatin, et al.. Molecular basis of African yam domestication: analyses of selection point to root development, starch biosynthesis, and photosynthesis related genes. BMC Genomics, 2017, 18 (1), 10.1186/s12864-017-4143-2 . hal-01814687

HAL Id: hal-01814687

<https://agroparistech.hal.science/hal-01814687>

Submitted on 26 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

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 Chair⁴, Alexandre Dansi³, Gustave Djedatin³, Anne-Céline Thuillet¹, Bénédicte Rhoné^{1,5}, Olivier François⁶, Karine Alix² 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 domesticate 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

¹Institut 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/demuladapt>). 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. praeheensis* 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. praeheensis* (Additional file 2: Table S3 a and b). Secondly, the cultivated diversity θ was 28% and 44% lower than *D. abyssinica* and *D. praeheensis* 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

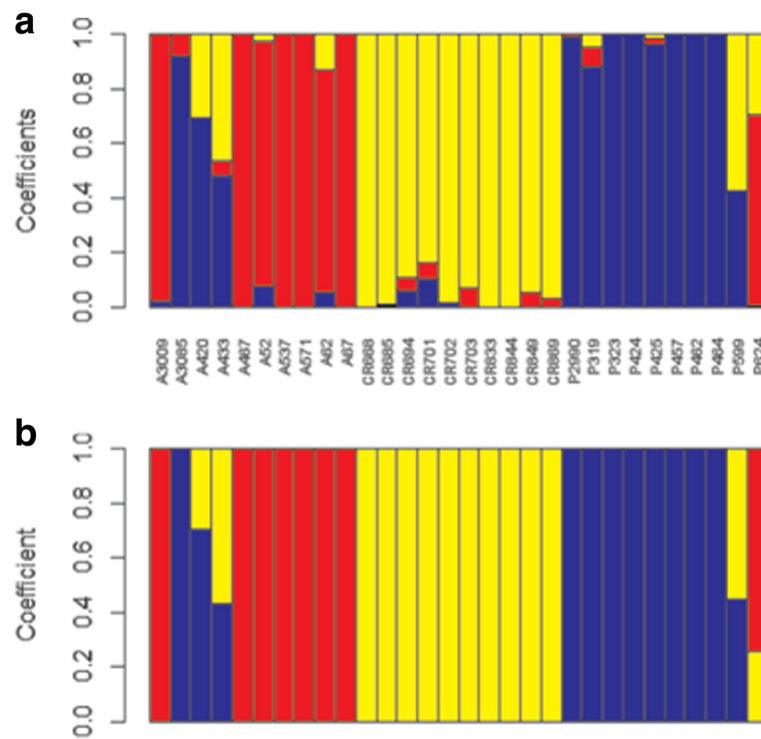


Fig. 1 Structure analysis using sNMF(a) and NgsAdmix (b). Each color represents one population. The length of each segment in each vertical bar represents the proportion of ancestry in each population

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% ($-\log_{10}(\pi_c/\pi_w) > 1.34$), 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. praezensilis* 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. praezensilis* 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

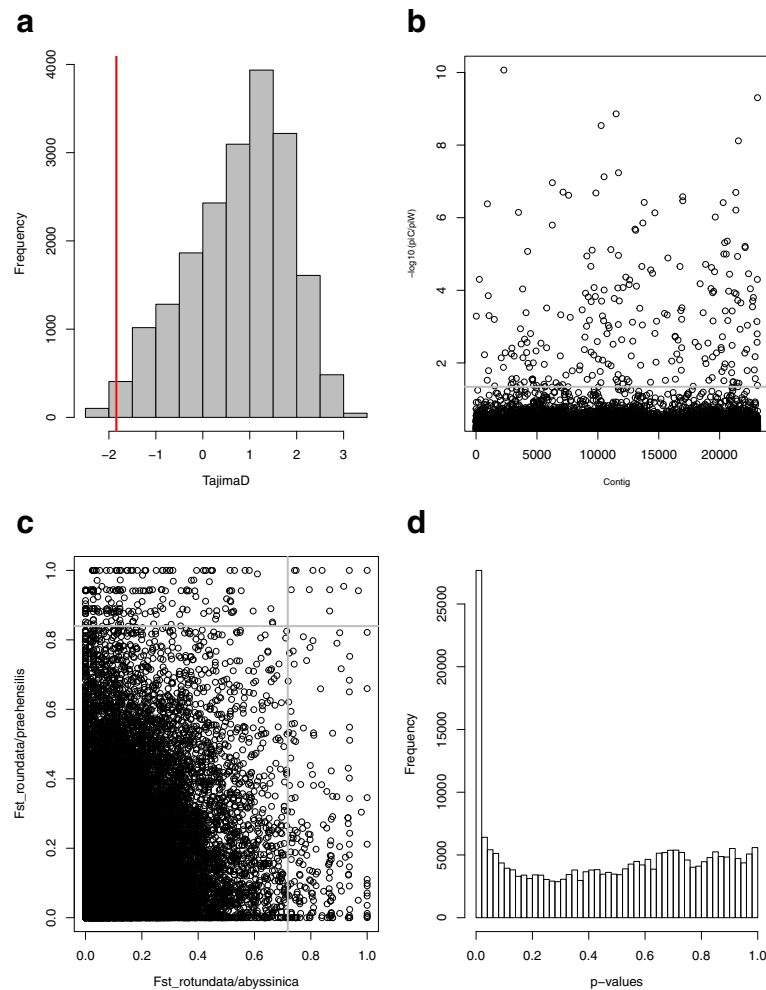


Fig. 2 Summary of the different tests used to identify outlier contigs. In the distribution of Tajima's D value of the cultivated species (a), the red line indicates the 1% threshold used to consider contigs as candidates. In the of reduction of nucleotide diversity π (b), the $-\log_{10}(\pi_c/\pi_w)$ 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

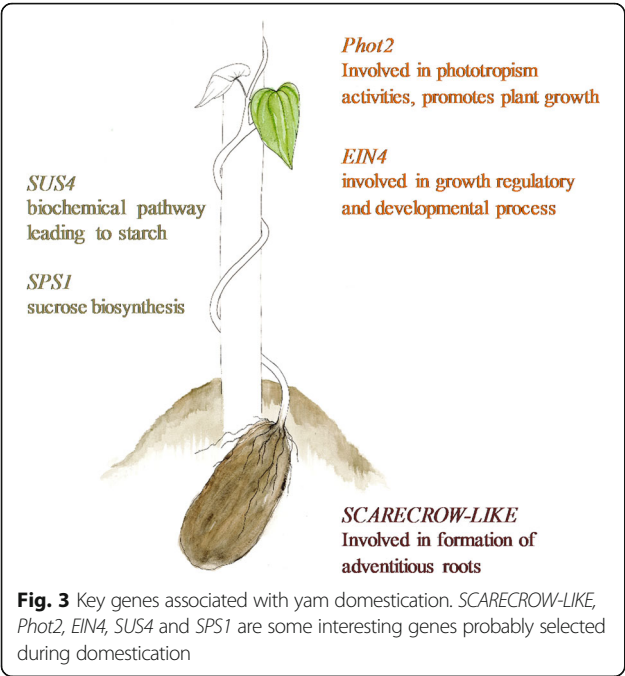
(Additional file 2: Table S6). When we restricted our analysis to the 81 candidate contigs detected by at least two methods, we obtained nine significant GO terms (Additional file 2: Table S7). The most significant GO terms were identical whether we considered all the candidate contigs or only the 81 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. praeensis*. 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.



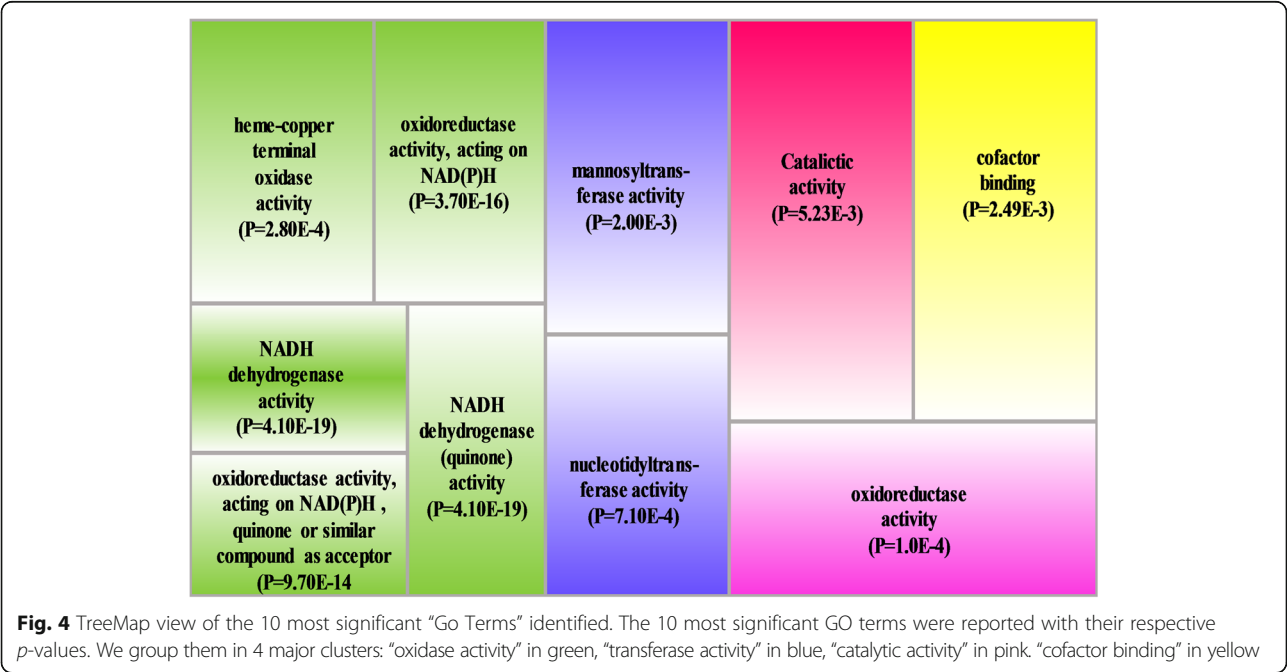
Linkage disequilibrium (LD) also decreased rapidly, like in other outcrossing crops. This LD decay is more similar to that observed in maize [53–55] than to that reported in self-pollinating crops such as rice [56]. However, our estimation of LD is based on a small sample and we might overestimate the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



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: *SUS4* 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 *NADPH DH* complex is an important set of enzymes for chlororespiration [66]. The *NADPH 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, *NADPH 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 *NADPH 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 PCA. **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 Dainou (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

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Institut 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. ²Unité Mixte de Recherche Génétique Quantitative et Evolutive – Le Moulon, INRA – Univ. Paris-Sud – CNRS – AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France. ³Faculté 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. ⁴Centre International de la Recherche Agronomique pour le Développement, UMR AGAP, F-34398 Montpellier, France. ⁵Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Lyon, France. ⁶Université de Grenoble, Grenoble, France.

Received: 4 April 2017 Accepted: 2 October 2017

Published online: 12 October 2017

References

- Diamond J. Evolution, consequences and future of plant and animal domestication. *Nature*. 2002;418:700–7.
- Fuller DQ. Contrasting patterns in crop domestication and domestication rates: recent Archaeobotanical insights from the old world. *Ann Bot*. 2007; 100:903–24.
- Purugganan MD, Fuller DQ. The nature of selection during plant domestication. *Nature*. 2009;457:843–8.
- Harris DR. Foraging and Farming: The Evolution of Plant Exploitation. eds Harris, D. R. & Hillman, G. C. 1989. p. 11–26.
- Purugganan MD, Fuller DQ. Archaeological data reveal slow rates of evolution during plant domestication. *Evolution*. 2011;65:171–83.
- Meyer RS, Purugganan MD. Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet*. 2013;14:840–52.
- McKey D, Elias M, Pujol B, Duputié A. The evolutionary ecology of clonally propagated domesticated plants. *New Phytol*. 2010;186:318–32.
- Whitt SR, Wilson LM, Tenaillon MJ, Gaut BS, Buckler ES. Genetic diversity and selection in the maize starch pathway. *Proc Natl Acad Sci*. 2002;99: 12959–62.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Hamon P, Brizard J-P, Zoundjihékon 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.
- Scarcelli N, Dainou 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.
- 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.
- Hamon P, Dumont R, Zoundjihékon 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, Kernytzky 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, Molteke 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. *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. Cloutault 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. Zoundjikepon 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. Garriss 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.