The Discovery of Wild Date Palms in Oman Reveals a Complex Domestication History Involving Centers in the Middle East and Africa

M. Gros-Balthazard, M. Galimberti, A. Kousathanas, C. Newton, Sarah Ivorra, L. Paradis, Y. Vigouroux, R. Carter, M. Tengberg, V. Battesti, et al.

Report

Current Biology

The Discovery of Wild Date Palms in Oman Reveals a Complex Domestication History Involving Centers in the Middle East and Africa

Highlights

- We discovered wild populations of the date palm *Phoenix dactylifera* in remote Oman
- Wild date palms differ from modern cultivars both morphologically and genetically
- The domestication of date palms involved sources from the Middle East and Africa

Authors

Muriel Gros-Balthazard, Marco Galimberti, Athanasios Kousathanas, ..., Jean-Christophe Pintaud, Jean-Frédéric Terral, Daniel Wegmann

Correspondence

muriel.grosb@gmail.com (M.G.-B.), daniel.wegmann@unif.ch (D.W.)

In Brief

Gros-Balthazard et al. report the discovery of wild date palms, the ancestral species of one of the oldest cultivated fruit trees and the cornerstone of the oasis agricultural system for thousands of years. Comparing the genomes of wild and modern date palms reveals a secondary domestication event in Africa, but only weak artificial selection.

Gros-Balthazard et al., 2017, Current Biology *27*, 2211–2218
July 24, 2017
http://dx.doi.org/10.1016/j.cub.2017.06.045

CellPress

Current Biology
Report

The Discovery of Wild Date Palms in Oman Reveals a Complex Domestication History Involving Centers in the Middle East and Africa

Muriel Gros-Balthazard,[1,2,3,4,*] Marco Galimberti,[3,4] Athanasios Kousathanas,[3,4,5] Claire Newton,[1,6] Sarah Ivorra,[1] Laure Paradis,[1] Yves Vigouroux,[2] Robert Carter,[7] Margareta Tengberg,[8] Vincent Battesti,[9] Sylvain Santoni,[10] Laurent Falquet,[3,4] Jean-Christophe Pintaud,[2,11] Jean-Frédéric Terral,[1,12] and Daniel Wegmann[3,4,12,13,*]

[1]Institut des Sciences de l'Evolution, Université de Montpellier, UMR 5554 CNRS / Université de Montpellier / IRD / EPHE, CC065, Equipe Dynamique de la Biodiversité, Anthropo-écologie, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
[2]Institut de Recherche pour le Développement, Université de Montpellier, UMR DIADE, 911 Avenue Agropolis, 34394 Montpellier Cedex 5, France
[3]Department of Biology, University of Fribourg, Chemin du Musée 10, 1700 Fribourg, Switzerland
[4]Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland
[5]Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France
[6]Laboratoire d'Archéologie et de Patrimoine, Université du Québec à Rimouski, 300 Allée des Ursulines, Rimouski, QC G5L 3AI, Canada
[7]UCL Qatar, University College London, Doha, Qatar
[8]UMR 7209 Archéozoologie et Archéobotanique, CNRS and Muséum National d'Histoire Naturelle, 75005 Paris, France
[9]UMR 7206 Éco-anthropologie et Ethnobiologie, CNRS and Muséum National d'Histoire Naturelle (Musée de l'Homme), 75016 Paris, France
[10]UMR Genetic Improvement and Adaptation of Mediterranean and Tropical Plants, INRA Montpellier, 34398 Montpellier Cedex 5, France
[11]Deceased
[12]These authors contributed equally
[13]Lead Contact
*Correspondence: muriel.grosb@gmail.com (M.G.-B.), daniel.wegmann@unif.ch (D.W.)
http://dx.doi.org/10.1016/j.cub.2017.06.045

SUMMARY

For many crops, wild relatives constitute an extraordinary resource for cultivar improvement [1, 2] and also help to better understand the history of their domestication [3]. However, the wild ancestor species of several perennial crops have not yet been identified. Perennial crops generally present a weak domestication syndrome allowing cultivated individuals to establish feral populations difficult to distinguish from truly wild populations, and there is frequently ongoing gene flow between wild relatives and the crop that might erode most genetic differences [4]. Here we report the discovery of populations of the wild ancestor species of the date palm (*Phoenix dactylifera* L.), one of the oldest and most important cultivated fruit plants in hot and arid regions of the Old World. We discovered these wild individuals in remote and isolated mountainous locations of Oman. They are genetically more diverse than and distinct from a representative sample of Middle Eastern cultivated date palms and exhibit rounded seed shapes resembling those of a close sister species and archeological samples, but not modern cultivars. Whole-genome sequencing of several wild and cultivated individuals revealed a complex domestication history involving the contribution of at least two wild sources to African cultivated date palms. The discovery of wild date palms offers a unique chance to further elucidate the history of this iconic crop that has constituted the cornerstone of traditional oasis polyculture systems for several thousand years [5].

RESULTS AND DISCUSSION

The Identification of Wild Date Palms

Archeological evidence suggests that date palms have been used for millennia in North Africa, the Middle East, and as far as northwestern India [5], where they are still of huge social and economic importance [6]. Yet, their domestication history remains poorly understood, with recent genomic studies hinting at a contribution of multiple wild populations as evidenced by the surprisingly large genetic differentiation between cultivated individuals from Africa and the Middle East [7–10].

Although no wild populations have been described to date [11], uncultivated date palms occur across the entire distribution area [12, 13]. However, whether they are feral (derived from cultivated individuals but not tended) or truly wild is not known. Recently, we discovered uncultivated populations in remote, mountainous locations in Oman that exhibit unusually rounded seeds resembling those of the sister species *Phoenix sylvestris* [14, 15].

Here we present a systematic screening of 102 individuals sampled from nine such candidate wild populations (Table 1; Table S1), corroborating their outlier status. We first compared the shape of 763 seeds from 39 of these individuals to 5,353 seeds from 271 cultivated date palms from the entire distribution area and 760 seeds from 38 *Phoenix sylvestris* individuals (Table 1; Table S1). Normal mixture modeling of seed shapes

Table 1. *Phoenix* spp. Accessions Analyzed

	Total	Seed Morphology	Microsatellites	Whole-Genome
African/South European cultivated *P. dactylifera*	275	161 (3,210)	231	3
Middle Eastern/Indian/Pakistan cultivated *P. dactylifera*	173	110 (2,143)	141	13
wild *P. dactylifera*	102	39 (763)	102	3
P. atlantica	37	0	37	1
P. sylvestris	74	38 (760)	58	1
Archeological material from *P .dactylifera*	4	4 (4)	0	0
Total	665	352 (6,880)	569	21

For seed morphology, the number of seeds is given in parentheses. See also Table S1.

captured by elliptic Fourier transforms [16] clustered most of the wild candidates with the sister species *P. sylvestris* (Figure 1A), and not with cultivated individuals. This was also the case when assigning individuals to three clusters (Figure S1). Additionally, the rounded seeds of the putatively wild individuals matched the shape of four archeological seeds from Kuwait (Table S1; Figure S1) that date back to the assumed onset of cultivation in the region about 5000 BCE according to archaeological evidence [5].

We next compared the genetic diversity and structure of all 102 putatively wild individuals to 372 cultivated date palms and 58 *P. sylvestris* individuals using 17 autosomal microsatellites (Table 1; Table S1; Data S1A). The putatively wild individuals had a significantly larger diversity (allelic richness [AR] 6.74; private allelic richness [PAR] 1.25; Table S2) than cultivated individuals from the Middle East (AR 6.21; PAR 0.46; Table S2). Interestingly, the limited genetic data used in this screen was sufficient to identify the putatively wild individuals as a unique cluster, both in an admixture analysis (Figure 1B; Figure S2) and in a principal component analysis (PCA) (Figure 1C; Figure S2). Finally, a population tree shows the putatively wild individuals at the base of the Middle Eastern clade (Figure 1D).

A Secondary Domestication Event in Africa

Although these results support the status of many of the Omani uncultivated individuals studied here as representatives of wild date palms sister to modern cultivars from the Middle East, our data also suggest a more complex domestication history. In particular, African accessions showed the highest diversity among all groups (AR 6.94; PAR 1.53; Table S2) and appeared to be twice as differentiated from Middle Eastern cultivated date palms as the latter are from the putatively wild individuals (F_{ST} of 8.64% and 3.95% respectively; Table S2). Although a strong genetic differentiation between African and Middle Eastern date palms has been reported previously [7–10], the source of this differentiation remained unclear. Since both the inferred population tree (Figure 1D) and the fitted admixture model (Figure 1B, preferred K = 3; Figure S2) identified African cultivated accessions as a sister clade to both wild individuals and Middle Eastern cultivated accessions, our data strongly suggest that at least one ancestral gene pool contributed uniquely to African date palms. However, whether this is reflective of (1) an independent (primary) domestication event in Africa or (2) the crossing of previously domesticated individuals imported from the Middle East with local wild individuals (referred to as secondary domestication or diversification) cannot be resolved from these data. We note that a primary domestication event appears unlikely given the highly similar shape of seeds in African and Middle Eastern cultivars (Figure S1).

To shed additional light on the domestication history of date palms, we sequenced three candidate wild date palms each from a different population to an average depth of 13.3× genome-wide (Table 1; Table S1; see Data S1B). For comparison, we also sequenced one African and one Middle Eastern cultivar with the same protocol and complemented our data with 14 accessions available in GenBank [17, 18] (Table 1; Table S1; see Data S1B). Unexpectedly, we detected two pairs of (male) clones among the Middle Eastern date palms from Saudi Arabia obtained from Sabir et al. [18] and therefore removed the sample with less data for each pair (Sukkariat Qassim and Shalaby, respectively) from any downstream analyses.

The whole-genome sequencing data confirmed that wild individuals and the African and Middle Eastern cultivated date palms represent distinct populations. First, the three populations differed strongly in their genetic diversity, with the highest diversity again found among African (nucleotide diversity 0.55%, Table 2) and the lowest among Middle Eastern cultivated accessions (0.34%, Table 2), in line with previous evidence [8]. The diversity of wild individuals was intermediate (0.43%, Table 2) but higher than that of Middle Eastern cultivated date palms. Second, all three populations formed their own non-overlapping cluster in a genome-wide PCA (Figure 2A). Third, pairwise genetic differentiation was significant and much higher between African and Middle Eastern cultivated date palms (22.92%) than between the latter and wild samples (16.45%), as was observed in the microsatellite data (Table S2). Finally, multiple sequentially Markovian coalescent (MSMC) analysis [19] inferred distinct population size trajectories for the wild samples, the African cultivated, and three representative Middle Eastern cultivated date palms (Figure 2B). The marked and robust differences in these trajectories were also observed when running the analysis with only two individuals per population or for each sample individually, albeit some variation between samples (Figure S3).

We next inferred an individual-based phylogeny using ExaML [20] to quantify the evolutionary relationship among the samples. In line with the population structure revealed by the few microsatellites that we genotyped in many more individuals (Figure 1B), wild individuals clustered at the base of the Middle Eastern cultivated samples, and the African cultivars appeared as a sister clade to all other *P. dactylifera* samples (Figure 2C). Interestingly, however, clustering was not perfect, suggesting potential gene flow between African and Middle Eastern cultivars after domestication. The Middle Eastern cultivar Moshwaq Al-Riyad, for instance, clustered with African cultivars. Similarly, the Egyptian cultivar Siwi clustered basal to wild and Middle Eastern cultivated accessions, and not within the African cluster. The intermediate placement of the cultivar Siwi was also visible in our

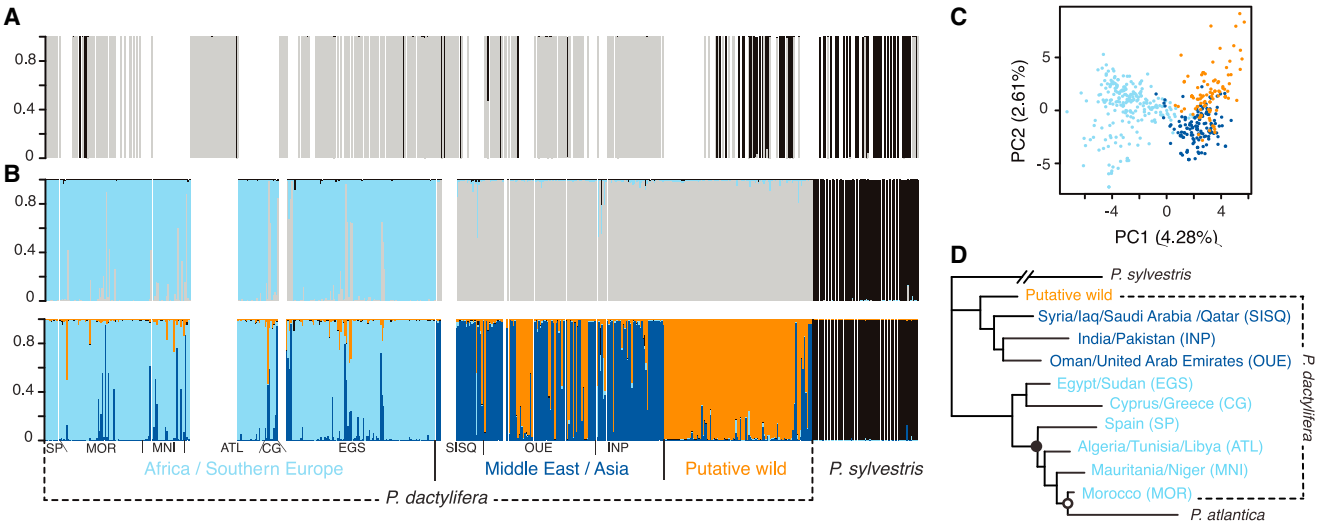

Figure 1. Date Palm Population Structure as Inferred from Seed Morphology and Microsatellites

(A) Mixture proportions based on seed shapes of 348 *Phoenix* samples modeled as a mixture of two normal distributions.

(B and C) Admixture proportions with K = 3 (B, top) and K = 4 (B, bottom) and principal component analysis (C, variance explained in parentheses) of 532 *Phoenix* samples inferred at 17 microsatellite markers. Results in (B) are "stacked" underneath the corresponding sample from (A).

(D) Neighbor-joining tree of the same samples grouped by geographic location and setting *P. sylvestris* accessions as outgroup. White and black circles indicate nodes with >50% and >95% bootstrap support, respectively.

Color coding: black, *P. sylvestris* and *P. atlantica*; dark blue, Middle Eastern/Indian/Pakistan cultivated date palms; light blue, African/South European date palms; orange, putative wild date palms. See also Figures S1, S2, and S4 and Table S2.

MSMC analysis, in which its population size trajectory was found to be intermediate between those of African and Middle Eastern cultivated samples (Figure S3), as well as in the PCA analysis that placed it much closer to Middle Eastern cultivated accessions than the other African samples (Figure 2A).

Some evidence for gene flow between Middle Eastern and African cultivars was previously reported [8, 10]. For instance, we recently reported an east-west cline in the frequency of the chloroplastic haplotypes 3 and 4, indicative of the exchange of maternal lineages between the two populations [21]. Here we genotyped the same chloroplastic marker in 264 cultivated date palms across the entire distribution range and in 102 wild and 58 *P. sylvestris* individuals. We found haplotype 3 in cultivated individuals from all regions, but not in any wild or *P. sylvestris* individuals (Figure S4), a finding that is difficult to explain without female-mediated gene flow. Recently, Hazzouri et al. [8] reported an admixture analysis from genome-wide data of 62 cultivars in which several African cultivars, and in particular those samples geographically close to the Middle East (e.g., in Egypt or Sudan), exhibited variable proportions of Middle Eastern ancestry.

Here we benefited from the wild individuals to specifically test for past events of gene flow by inferring a population graph allowing for both population splits and mixtures with TreeMix [22]. The inferred graph (Figure 2D; Figure S3) revealed strong evidence for a primary domestication event in the Middle East, followed by introgression from wild individuals in Africa (secondary domestication event in Africa): the wild population included in this analysis was inferred to be basal to both Middle Eastern and African cultivated accessions, but African cultivars were inferred to have received about a third of their genome from an unsampled source population sister to all sampled *P. dactylifera* populations, most likely representing the African population prior to domestication. This result was robust to the clustering of neighboring SNPs into blocks, the exclusion of the cultivar Siwi from the African population, and restricting the analysis to only coding SNPs or SNPs at least 10 kbp away from genes (Figure S3). We further confirmed the finding of a secondary domestication event in Africa by inferring demographic parameters of a flexible admixture model in which models of an independent

Table 2. Nucleotide Diversity π, Watterson's Theta Estimator θ_W, and an Estimate of the Selective Constraint C

Site Class	%π	%θ_W	C
African Date Palms (6 Haplotypes)			
Nonsynonymous	0.220 (0.001)	0.217 (0.001)	0.646 (0.002)
Synonymous	0.621 (0.002)	0.611 (0.002)	–
Intronic	0.476 (0.0004)	0.470 (0.0004)	0.234 (0.003)
Intergenic	0.599 (0.001)	0.587 (0.001)	0.037 (0.004)
Middle Eastern Date Palms (22 Haplotypes)			
Nonsynonymous	0.142 (0.001)	0.129 (0.0004)	0.641 (0.002)
Synonymous	0.395 (0.002)	0.351 (0.001)	–
Intronic	0.298 (0.0003)	0.283 (0.0002)	0.246 (0.003)
Intergenic	0.348 (0.0005)	0.357 (0.0005)	0.119 (0.004)
Wild Date Palms (6 Haplotypes)			
Nonsynonymous	0.169 (0.001)	0.166 (0.001)	0.640 (0.002)
Synonymous	0.470 (0.002)	0.453 (0.002)	–
Intronic	0.354 (0.0003)	0.345 (0.0003)	0.246 (0.004)
Intergenic	0.457 (0.001)	0.444 (0.001)	0.027 (0.005)

Means and standard deviations (in parentheses) across sites are shown. Intergenic regions include all sites >10 kb from known genes.

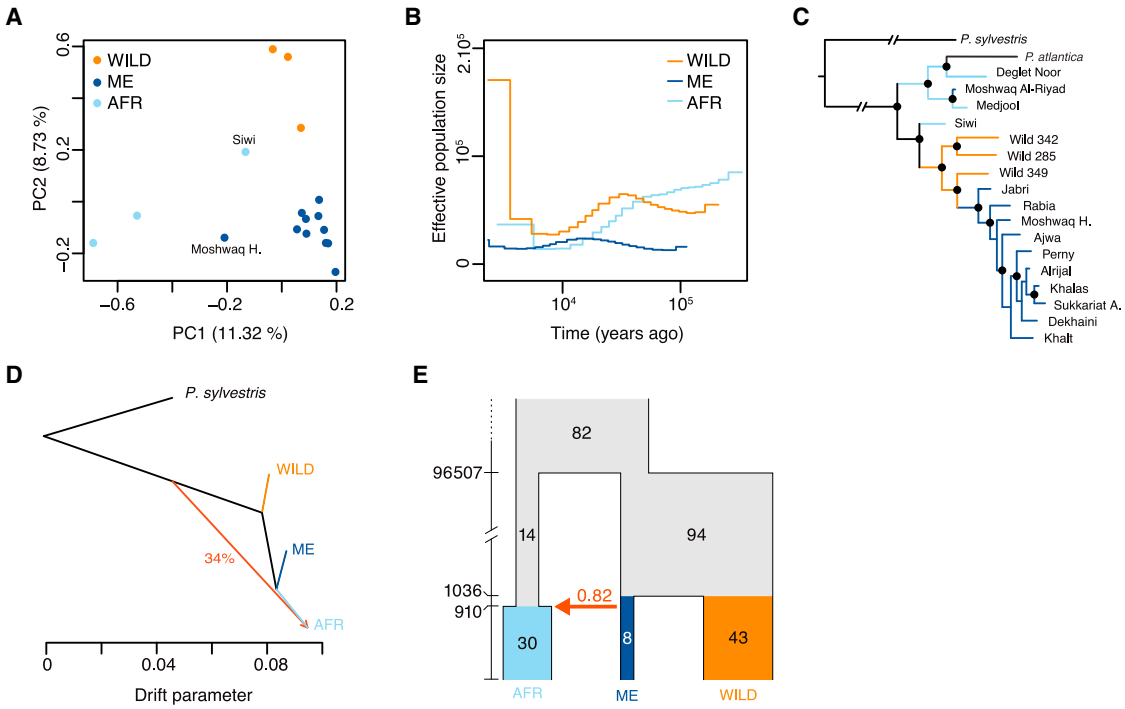

Figure 2. Population Genetic Analyses of Whole-Genome Data

(A) Principal component analysis of genotype likelihoods at ~7 million SNPs. The variance explained by each principal component (PC) is given in parentheses.

(B) Historical effective population sizes inferred using the multiple sequentially Markovian coalescent (MSMC) method from three representative samples of each population.

(C) Phylogenetic tree of genotypes at ~7 million SNPs setting *P. sylvestris* as outgroup. Black circles indicate nodes with >95% bootstrap support.

(D) Population graph with one migration edge (strength 34%) inferred with TreeMix when setting *P. sylvestris* as outgroup.

(E) Maximum composite likelihood estimates of demographic parameters inferred with fastsimcoal2 and drawn to scale. Population sizes are indicated in thousands and times in generations, as inferred assuming a mutation rate of 2.5×10^{-8}. WILD, wild date palms; AFR, African cultivated date palms; ME, Middle Eastern cultivated date palms. See also Figure S3, Table S3, and Methods S1.

and a secondary domestication event in Africa were nested, depending on the admixture strength. Despite this flexibility, the maximum composite likelihood estimates obtained with fastsimcoal2 [23] indicated a massive contribution of 82% of the Middle Eastern cultivars to the modern genomic makeup of African cultivars (Figure 2E), corroborating the inference of a secondary domestication event in Africa.

The Middle East has long been proposed as a primary center of date palm domestication [24]. The oldest archaeological remains attesting to the use of date palms were found in the Arabian Peninsula and date back to the second half of the 6th millennium BCE [25, 26]. The earliest evidence for date palm cultivation in the 3rd millennium BCE was all found around the Persian Gulf, including Oman [5]. Finally, the only fossil evidence of date palms predating cultivation was also found in the Middle East, suggesting that date palms have been present in the region for at least 30,000 years [27, 28]. In contrast, ancient date palm remains from Africa are concentrated in the northeast (Libya and Egypt) and are younger (mostly from the 2nd millennium BCE); iconographic and lexicographic evidence, however, may point to the presence of the date palm in Egypt prior to its cultivation [29, 30]. The lack of fossils and the sparse archeological evidence from Africa may reflect only the current state of research in that region. Indeed, our results suggest an important role of the ancestral African gene pool, and hence that wild date palms were present in Africa prior to domestication. Although no wild date palms from Africa are currently known, many uncultivated populations scattered in North Africa and Spain have been reported [12] that may now warrant verification.

Weak Selection during Domestication

We next explored whether domestication affected the strength of selection acting on the different populations, assuming synonymous sites to be evolving neutrally, using three different methods. First, we quantified the selective constraint of different functional classes of sites by a direct comparison of nucleotide diversity (Table 2). Second, we used *DFE-alpha* [31] to infer the distribution of fitness effects (DFE) of nonsynonymous sites from their site-frequency spectrum (SFS) and the expected SFS under a two-epoch population size change model fitted to the SFS of synonymous sites (Methods S1). Finally, we estimated the DFE of nonsynonymous sites using *DoFE* [32], which makes no assumption regarding the demographic history and instead compares the nonsynonymous SFS to a parametric function directly fitted on the synonymous SFS (Table S3). Interestingly, none of these methods revealed any difference in selection strength between wild date palms and the African and Middle Eastern cultivated date palms, suggesting a low cost of

domestication, in line with the weak domestication syndrome of date palms and perennials in general [4]. However, we note that individual genes might very well show a strong signature of selection [8], but their detection will require much larger sample sizes than analyzed here.

Phylogenetic Considerations

Finally, our data also resolve two controversial aspects about the taxonomy of the genus *Phoenix*. First, ongoing and male-mediated hybridization between *P. dactylifera* and the sister species *P. sylvestris* has been suggested [33]. However, neither our TreeMix analysis (Figure 2D; Figure S3) on whole-genome data nor the extensive survey based on microsatellites (Figures 1B and 1C) revealed any sign of admixture, even among our Indian samples that were collected in areas where both species occur sympatrically and are known to be interfertile. Instead, the two species appear as highly differentiated clusters (F_{ST} of 29.0%) in both an admixture analysis (Figure 1B; Figure S2) and a PCA (Figure S2). Second, our data show that *Phoenix atlantica* is not a separate species and is genetically indistinguishable from the African *P. dactylifera* samples. *P. atlantica* was originally described as a species endemic to the Cape Verde islands [34], but that designation has been disputed [35, 36]. Here, we genotyped 37 samples from different Cape Verde islands (Table 1; Table S1) for 17 autosomal microsatellites and a chloroplastic marker (See Data S1A) and sequenced the whole genome of one sample (Table 1; Table S1; Data S1B). In all cases, *P. atlantica* samples cluster consistently within *P. dactylifera* samples from Africa regardless of the analytical method (Figure 1D; Figure 2; Figure S2) and display no private allele (Table S2), suggesting that modern *P. atlantica* individuals derive from African *P. dactylifera* individuals brought to the Cape Verde islands rather recently.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Uncultivated, putatively wild date palms
  - Cultivated date palms
  - Sister species
  - Samples used for whole-genome sequencing
  - Archaeological samples
- METHOD DETAILS
  - Morphometric analysis
  - DNA extraction
  - Microsatellite genotyping
  - Whole-genome sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Seed morphometric analysis
  - Analysis of microsatellite data
  - Bioinformatic analysis of whole-genome sequencing data
  - Genome annotation
  - Analyses of whole-genome sequencing data
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, three tables, one dataset, and one methods file and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.06.045.

AUTHOR CONTRIBUTIONS

Conceptualization: M.G.-B., J.-C.P., J.-F.T., D.W. Morphometric analyses of current and ancient material: M.G.-B., C.N., S.I., R.C, M.T., V.B., D.W. Map designs: L.P. Microsatellite analyses: M.G.-B., Y.V., S.S. J.-C.P. D.W. Whole-genome analyses: M.G.-B., M.G., A.K., D.W. Manuscript writing: M.G.-B., D.W. Supervision: J.-C.P., J.-F.T., D.W.

ACKNOWLEDGMENTS

Dr. Jean-Christophe Pintaud passed away before the submission of the final version of this manuscript. This article is dedicated to our colleague and friend who died prematurely. We are immensely grateful to Jean-Christophe Pintaud.

We thank the Centre de Recherches sur l'Elevage et le Pâturage, Kébili, Tunisia; the Centre Régional de Recherches sur l'Agriculture d'Oasis (Ministry of Agriculture), Degache, Tunisia; and the Research Department of the Ministry of Agriculture of Oman, for issuing permissions for this study. We acknowledge Susi Gomez and Michel Ferry (Phoenix Station, Elche, Spain) for granting access to the Elche gardens, Sally Henderson and William Baker for providing *P. atlantica* material, Mohammed Aziz El-Houmaizi (Oujda University, Morocco), and Robert Castellana. We are grateful to the many collaborators who gave us permission to collect seeds and leaves on their private lands. We acknowledge Tosso Leeb for assistance in library preparation and sequencing; Megan Bowman, Ning Jiang, Katharina Hoff, and Carson Holt for their support in establishing the gene annotation pipeline; and Vincent Bonhomme for support in conducting the elliptic Fourier transform analysis. This work was supported by the Agence Nationale de la Recherche "PHOENIX" (ANR-06-BLAN-0212) and "FRUCTIMEDHIS" (ANR-07-BLAN-0033) programs, PhD fellowships from the French Ministry of Higher Education and Research granted to M.G.-B., and Swiss National Science Foundation grants PZ00P3_142643 and 31003A_149920 to D.W. This article is ISEM contribution ISEM 2017-111.

Received: February 23, 2017
Revised: May 10, 2017
Accepted: June 19, 2017
Published: July 13, 2017

REFERENCES

1. Harlan, J.R. (1976). Genetic resources in wild relatives of crops. Crop Sci. *16*, 329.
2. Vavilov, N. (1992). Origin and Geography of Cultivated Plants (Cambridge University Press).
3. Meyer, R.S., DuVal, A.E., and Jensen, H.R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol. *196*, 29–48.
4. Miller, A.J., and Gross, B.L. (2011). From forest to field: perennial fruit crop domestication. Am. J. Bot. *98*, 1389–1414.
5. Tengberg, M. (2012). Beginnings and early history of date palm garden cultivation in the Middle East. J. Arid Environ. *86*, 139–147.
6. Chao, C.T., and Krueger, R.R. (2007). The date palm (*Phoenix dactylifera* L.): Overview of biology, uses and cultivation. HortScience *42*, 1077–1082.
7. Zehdi-Azouzi, S., Cherif, E., Moussouni, S., Gros-Balthazard, M., Abbas Naqvi, S., Ludeña, B., Castillo, K., Chabrillange, N., Bouguedoura, N., Bennaceur, M., et al. (2015). Genetic structure of the date palm (*Phoenix dactylifera*) in the Old World reveals a strong differentiation between eastern and western populations. Ann. Bot. (Lond.) *116*, 101–112.
8. Hazzouri, K.M., Flowers, J.M., Visser, H.J., Khierallah, H.S.M., Rosas, U., Pham, G.M., Meyer, R.S., Johansen, C.K., Fresquez, Z.A., Masmoudi, K.,

et al. (2015). Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. Nat. Commun. *6*, 8824.

9. Al-Mssallem, I.S., Hu, S., Zhang, X., Lin, Q., Liu, W., Tan, J., Yu, X., Liu, J., Pan, L., Zhang, T., et al. (2013). Genome sequence of the date palm *Phoenix dactylifera* L. Nat. Commun. *4*, 2274.
10. Mathew, L.S., Seidel, M.A., George, B., Mathew, S., Spannagl, M., Haberer, G., Torres, M.F., Al-Dous, E.K., Al-Azwani, E.K., Diboun, I., et al. (2015). A genome-wide survey of date palm cultivars supports two major subpopulations in *Phoenix dactylifera*. G3 (Bethesda) *5*, 1429–1438.
11. Pintaud, J.C., Zehdi, S., Couvreur, T., Barrow, S., Henderson, S., Aberlenc-Bertossi, F., Tregear, J., and Billotte, N. (2010). Species delimitation in the genus *Phoenix* (Arecaceae) based on SSR markers, with emphasis on the identity of the date palm (*Phoenix dactylifera* L.). In Diversity, Phylogeny, and Evolution in the Monocotyledons, O. Seberg, G. Petersen, A. Barford, and J. Davis, eds. (Aarhus University Press), pp. 267–286.
12. Zohary, D., Hopf, M., and Weiss, E. (2012). Domestication of Plants in the Old World, Third Edition (Oxford University Press).
13. Zohary, D., and Spiegel-Roy, P. (1975). Beginnings of fruit growing in the old world. Science *187*, 319–327.
14. Terral, J.-F., Newton, C., Ivorra, S., Gros-Balthazard, M., de Morais, C.T., Picq, S., Tengberg, M., and Pintaud, J.-C.C. (2012). Insights into the historical biogeography of the date palm (*Phoenix dactylifera* L.) using geometric morphometry of modern and ancient seeds. J. Biogeogr. *39*, 929–941.
15. Gros-Balthazard, M., Newton, C., Ivorra, S., Pierre, M.H., Pintaud, J.C., and Terral, J.F. (2016). The domestication syndrome in *Phoenix dactylifera* seeds: Toward the identification of wild date palm populations. PLoS ONE *11*, e0152394.
16. Rohlf, F.J. (1990). Morphometrics. Annu. Rev. Ecol. Evol. Syst. *21*, 299–316.
17. Al-Dous, E.K., George, B., Al-Mahmoud, M.E., Al-Jaber, M.Y., Wang, H., Salameh, Y.M., Al-Azwani, E.K., Chaluvadi, S., Pontaroli, A.C., DeBarry, J., et al. (2011). De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat. Biotechnol. *29*, 521–527.
18. Sabir, J.S.M., Arasappan, D., Bahieldin, A., Abo-Aba, S., Bafeel, S., Zari, T.A., Edris, S., Shokry, A.M., Gadalla, N.O., Ramadan, A.M., et al. (2014). Whole mitochondrial and plastid genome SNP analysis of nine date palm cultivars reveals plastid heteroplasmy and close phylogenetic relationships among cultivars. PLoS ONE *9*, e94158.
19. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.
20. Kozlov, A.M., Aberer, A.J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics *31*, 2577–2579.
21. Pintaud, J.-C., Ludeña, B., Zehdi, S., Gros-Balthazard, M., Ivorra, S., Terral, J.-F., Newton, C., Tengberg, M., Santoni, S., and Boughedoura, N. (2013). Biogeography of the date palm (*Phoenix dactylifera* L., Arecaceae): insights on the origin and on the structure of modern diversity. ISHS Acta Hortic. *994*, 19–36.
22. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. *8*, e1002967.
23. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. PLoS Genet. *9*, e1003905.
24. Gros-Balthazard, M., Newton, C., Ivorra, S., Tengberg, M., Pintaud, J.-C., and Terral, J.-F. (2013). Origines et domestication du palmier-dattier (*Phoenix dactylifera* L.): Etat de l'art et perpectives d'étude. Rev. Ethnoecol. *4*.
25. Beech, M. (2004). Archaeobotanical evidence for early date consumption in the Arabian Gulf. In The Date Palm: From Traditional Resource to Green Wealth (The Emirates Center for Strategic Studies and Research), pp. 11–31.
26. Parker, A.G. (2010). Palaeoenvironmental evidence from H3, Kuwait. In Maritime Interactions in the Arabian Neolithic: Evidence from H3, As-Sabiyah, an Ubaid-Related Site in Kuwait, R.A. Carter, and H.E.W. Crawford, eds. (Brill), pp. 189–201.
27. Solecki, R.S., and Leroi-gourhan, A. (1961). Palaeoclimatology and archaeology in the Near East. Ann. N Y Acad. Sci. *95*, 729–739.
28. Liphschitz, N., and Nadel, D. (1997). Epipalaeolithic (19,000 B.P.) charred wood remains from Ohalo II, Sea of Galilee, Israel. Mitekufat Haeven. J. Isr. Prehist. Soc. *27*, 5–18.
29. Tengberg, M., and Newton, C. (2016). Origine et évolution de la phéniciculture au Moyen-Orient et en Egypte. In Des fruits d'ici et d'ailleurs: Regards sur l'histoire de quelques fruits consommés en Europe, M.-P. Ruas, ed. (Editions Omnisciences), pp. 83–105.
30. Pelling, R. (2005). Garamantian agriculture and its significance in a wider North African context: The evidence of the plant remains from the Fazzan project. J. N. Afr. Stud. *10*, 397–412.
31. Keightley, P.D., and Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics *177*, 2251–2261.
32. Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics *173*, 891–900.
33. Gros-Balthazard, M. (2013). Hybridization in the genus *Phoenix*: A review. Emirates J. Food Agric. *25*, 831–842.
34. Chevalier, A. (1935). Plantes nouvelles de l'archipel des Iles du Cap Vert. Bulletin du Muséum National d'Histoire Naturelle *7*, 137–144.
35. Barrow, S. (1998). A revision of *Phoenix* L. (Palmae: Coryphoideae). Kew Bull. *53*, 513–575.
36. Henderson, S.A., Billotte, N., and Pintaud, J.C. (2006). Genetic isolation of Cape Verde Island *Phoenix atlantica* (Arecaceae) revealed by microsatellite markers. Conserv. Genet. *7*, 213–223.
37. Billotte, N., Marseillac, N., Brottier, P., Noyer, J.-L.L., Jacquemoud-Collet, J.-P.P., Moreau, C., Couvreur, T., Chevallier, M.-H.H., Pintaud, J.-C.C., and Risterucci, M.M. (2004). Nuclear microsatellite markers for the date palm (*Phoenix dactylifera* L.): characterization and utility across the genus *Phoenix* and in other palm genera. Mol. Ecol. Notes *4*, 256–258.
38. Ludeña, B., Chabrillange, N., Aberlenc-Bertossi, F., Adam, H., Tregear, J.W., and Pintaud, J.-C. (2011). Phylogenetic utility of the nuclear genes AGAMOUS 1 and PHYTOCHROME B in palms (Arecaceae): an example within Bactridinae. Ann. Bot. (Lond.) *108*, 1433–1444.
39. Daher, A., Adam, H., Chabrillange, N., Collin, M., Mohamed, N., Tregear, J.W., and Aberlenc-Bertossi, F. (2010). Cell cycle arrest characterizes the transition from a bisexual floral bud to a unisexual flower in *Phoenix dactylifera*. Ann. Bot. (Lond.) *106*, 255–266.
40. Aberlenc-Bertossi, F., Castillo, K., Tranchant-Dubreuil, C., Chérif, E., Ballardini, M., Abdoulkader, S., Gros-Balthazard, M., Chabrillange, N., Santoni, S., Mercuri, A., and Pintaud, J.C. (2014). In silico mining of microsatellites in coding sequences of the date palm (Arecaceae) genome, characterization, and transferability. Appl. Plant Sci. *2*, apps.1300058.
41. Scarcelli, N., Barnaud, A., Eiserhardt, W., Treier, U.A., Seveno, M., d'Anfray, A., Vigouroux, Y., and Pintaud, J.-C.C. (2011). A set of 100 chloroplast DNA primer pairs to study population genetics and phylogeny in monocotyledons. PLoS ONE *6*, e19954.
42. R Core Team (2015). R: A Language and Environment for Statistical Computing. http://www.r-project.org/.
43. Bonhomme, V., Picq, S., Gaucherel, C., and Claude, J. (2014). Momocs: outline analysis using R. J. Stat. Softw. *56*, 1–24.
44. Fraley, C., Raftery, A.E., Murphy, T.B., and Scrucca, L. (2012). mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Tech. Rep. *597*.

45. Dray, S., and Dufour, A.B. (2007). The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. *22*, 1–20.

46. Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics *26*, 419–420.

47. Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics *27*, 3070–3071.

48. Kamvar, Z.N., Tabima, J.F., and Grünwald, N.J. (2014). *Poppr*: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ *2*, e281.

49. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

50. de Mendiburu, F. (2015). Agricolae: Statistical Procedures for Agricultural Research. https://cran.r-project.org/package=agricolae.

51. Rambaut, A. (2009). FigTree v1.4.2. http://tree.bio.ed.ac.uk/software/figtree/.

52. Excoffier, L., and Lischer, H.E.L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. *10*, 564–567.

53. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

54. Earl, D.A., and Vonholdt, B.M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. *4*, 359–361.

55. Joshi, N., and Fass, J. (2011). Sickle: a windowed adaptive trimming tool for FASTQ files using quality (v1.33). https://github.com/najoshi/sickle.

56. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

57. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

58. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

59. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

60. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

61. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356.

62. Korneliussen, T.S., and Moltke, I. (2015). NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics *31*, 4009–4011.

63. Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. *164*, 513–524.

64. Yin, Y., Zhang, X., Fang, Y., Pan, L., Sun, G., Xin, C., Ba Abdullah, M.M., Yu, X., Hu, S., Al-Mssallem, I.S., and Yu, J. (2012). High-throughput sequencing-based gene profiling on multi-staged fruit development of date palm (*Phoenix dactylifera*, L.). Plant Mol. Biol. *78*, 617–626.

65. Bourgis, F., Kilaru, A., Cao, X., Ngando-Ebongue, G.-F., Drira, N., Ohlrogge, J.B., and Arondel, V. (2011). Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning. Proc. Natl. Acad. Sci. USA *108*, 12527–12532.

66. Zhang, G., Pan, L., Yin, Y., Liu, W., Huang, D., Zhang, T., Wang, L., Xin, C., Lin, Q., Sun, G., et al. (2012). Large-scale collection and annotation of gene models for date palm (*Phoenix dactylifera*, L.). Plant Mol. Biol. *79*, 521–536.

67. Newton, C., Gros-Balthazard, M., Ivorra, S., Paradis, L., Pintaud, J.-C., and Terral, J.-F. (2013). *Phoenix dactylifera* and *P. sylvestris* in Northwestern India: A glimpse into their complex relationships. Palms *57*, 37–50.

68. Carter, R.A., and Crawford, H.E.W. (2010). Conclusions. In Maritime Interactions in the Arabian Neolithic: Evidence from H3, As-Sabiyah, an Ubaid-Related Site in Kuwait, R.A. Carter, and H.E.W. Crawford, eds. (Brill), pp. 203–212.

69. Carter, R.A. (2006). Boat remains and maritime trade in the Persian Gulf during the sixth and fifth millennia BC. Antiquity *80*, 52–63.

70. Kuhl, F.P., and Giardina, C.R. (1982). Elliptic Fourier features of a closed contour. Comput. Graph. Image Process. *18*, 236–258.

71. Ballardini, M., Mercuri, A., Littardi, C., Abbas, S., Couderc, M., Ludeña, B., and Pintaud, J.-C. (2013). The chloroplast DNA locus psbZ-trnfM as a potential barcode marker in *Phoenix* L. (Arecaceae). ZooKeys *365*, 71–82.

72. Nei, M., Tajima, F., and Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. *19*, 153–170.

73. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.

74. Petit, R.J., El Mousadik, A., and Pons, O. (1998). Identifying populations for consevation on the basis of genetic markers. Conserv. Biol. *12*, 844–855.

75. Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. *14*, 2611–2620.

76. Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. *13*, 329–342.

77. Smit, A.F.A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0. http://www.repeatmasker.org.

78. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

79. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

80. Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. *38*, e199.

81. Smit, A.F.A., and Hubley, R. (2008). RepeatModeler Open-1.0. http://www.repeatmasker.org.

82. Bao, Z., and Eddy, S.R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. *12*, 1269–1276.

83. Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21* (*Suppl 1*), i351–i358.

84. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

85. UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

86. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. *34*, W435-9.

87. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

88. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. *12*, R22.

89. Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

90. Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. *7* (*Suppl 1*), 1–12.

91. Singh, R., Ong-Abdullah, M., Low, E.-T.L., Manaf, M.A.A., Rosli, R., Nookiah, R., Ooi, L.C.-L., Ooi, S.-E., Chan, K.-L., Halim, M.A., et al. (2013). Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature *500*, 335–339.

92. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

93. Junier, T., and Zdobnov, E.M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics *26*, 1669–1670.

94. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? Heredity (Edinb.) *116*, 362–371.

95. Keightley, P.D., and Eyre-Walker, A. (2010). What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 1187–1193.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
37 *Phoenix atlantica*, leaf samples	[36]	N/A
493 *Phoenix* spp., leaf and/or seed samples	ISEM collection: collection from the team Dynamique de la biodiversité, anthropo-écologie (DBA) at ISEM	N/A
58 *Phoenix* spp., leaf and/or seed samples	IRD Collection	N/A
6 *Phoenix dactylifera*, leaf samples	USDA-ARS California	N/A
4 archaeological seeds of *Phoenix dactylifera*	Robert Carter Collection	N/A
35 *Phoenix dactylifera*, leaf and/or seed samples	Battesti Collection	N/A
Deposited Data		
7 *Phoenix* spp., whole genome sequencing data	GenBank	GenBank: SRP094744
Seed morphometric data	http://www.unifr.ch/biology/research/wegmann/	http://dx.doi.org/10.17632/sc9cpvfnbj.1
Microsatellite data	http://www.unifr.ch/biology/research/wegmann/	http://dx.doi.org/10.17632/sc9cpvfnbj.1
Date palm genome annotation	http://www.unifr.ch/biology/research/wegmann/	http://dx.doi.org/10.17632/c6bjh7mgby.1
Oligonucleotides		
11 microsatellite primer pairs	[37]	mPdCIR085, mPdCIR078, mPdCIR015, mPdCIR016, mPdCIR032, mPdCIR035, mPdCIR057, mPdCIR025, mPdCIR010, mPdCIR063, mPdCIR050
1 microsatellite primer pair	[38]	AG1
1 microsatellite primer pair	[39]	PdCUC3-ssr1
4 microsatellite primer pairs	[40]	mPdIRD13, mPdIRD33, mPdIRD31, mPdIRD40
1 minisatellite primer pair	[41]	psbZ-trnM(CAU)
Software and Algorithms		
R	www.r-project.org [42]	RRID: SCR_001905
Momocs R package	https://cran.r-project.org/web/packages/Momocs/index.html [43]	N/A
mclust R package	https://cran.r-project.org/web/packages/mclust/index.html [44]	N/A
ade4 R package	https://cran.r-project.org/web/packages/ade4/index.html [45]	N/A
pegas R package	https://cran.r-project.org/web/packages/pegas/index.html [46]	N/A
adegenet R package	http://adegenet.r-forge.r-project.org [47]	RRID: SCR_000825
poppr R package	https://cran.r-project.org/web/packages/poppr/index.html [48]	N/A
ape R package	https://cran.r-project.org/web/packages/ape/index.html [49]	N/A
agricolae R package	https://cran.r-project.org/web/packages/agricolae/index.html [50]	N/A
FigTree	http://tree.bio.ed.ac.uk/software/figtree/ [51]	RRID: SCR_008515
Arlequin	http://cmpg.unibe.ch/software/arlequin35/ [52]	RRID: SCR_009051
Structure	http://web.stanford.edu/group/pritchardlab/structure.html [53]	N/A

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Structure Harvester	http://taylor0.biology.ucla.edu/structureHarvester/ [54]	N/A
Sickle	https://github.com/najoshi/sickle [55]	RRID: SCR_006800
Burrows-Wheeler Aligner	http://bio-bwa.sourceforge.net [56]	RRID: SCR_010910
Samtools	http://samtools.sourceforge.net [57, 58]	RRID: SCR_002105
Genome Analysis Tool Kit	https://software.broadinstitute.org/gatk/ [59]	RRID: SCR_001876
Picard tools	https://broadinstitute.github.io/picard/	N/A
VCFtools	http://vcftools.sourceforge.net [60]	RRID: SCR_001235
ANGSD	http://popgen.dk/angsd/index.php/ANGSD [61]	N/A
NGSrelate	http://www.popgen.dk/software/index.php/Main_Page#NgsRelate [62]	N/A
TreeMix	https://bitbucket.org/nygcresearch/treemix/wiki/Home [22]	N/A
Multiple sequentially Markovian coalescent (MSMC)	https://github.com/stschiff/msmc [19]	N/A
fastsimcoal2	http://cmpg.unibe.ch/software/fastsimcoal2/ [23]	N/A
Maker_P	http://www.yandell-lab.org/software/maker-p.html [63]	RRID: SCR_005309
ExaML	https://github.com/stamatak/ExaML [20]	N/A
DFE-alpha	https://sourceforge.net/projects/dfe-alpha/ [31]	N/A
DoFE	http://www.sussex.ac.uk/lifesci/eyre-walkerlab/resources [32]	N/A
Other		
18 DNA samples of *Phoenix* spp.	IRD collection	N/A
Phoenix dactylifera cv. Khalas, genome assembly	[9]	GCA_000413155.1
9 *Phoenix dactylifera*, whole genome sequence data (raw data)	[18]	SAMN02350655, SAMN02351367, SAMN02351372, SAMN02351373, SAMN02351368, SAMN02351370, SAMN02351371, SAMN02351366, SAMN02351369
5 *Phoenix dactylifera*, whole genome sequence data (raw data)	[17]	SAMN00205585, SAMN00205578, SAMN00205584, SAMN00205580, SAMN00205577
RNA sequences	GenBank	GenBank: PRJNA238431
RNA sequences	GenBank [64]	GenBank: PRJNA72467
RNA sequences	GenBank [65]	GenBank: PRJNA66319
RNA sequences	GenBank [66]	GenBank: SRP010344

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Daniel Wegmann (daniel.wegmann@unifr.ch).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study involved a total 665 individuals *Phoenix* spp. accessions: (1) 550 *Phoenix dactylifera* L. individuals, of which 448 were cultivated and 102 were putative wild date palms (Table 1; Table S1), (2) 74 *Phoenix sylvestris* L. and 37 *Phoenix atlantica* A. Chev. samples to represent the two closest sister species [11, 21] (Table 1; Table S1) and (3) four archaeological seeds retrieved from As-Sabiyah, Kuwait (Table 1; Table S1). We collected seeds, leaf fragments, or both where possible, but also included accessions for which we got extracted DNA or sequencing data, as detailed below and in Table S1.

Uncultivated, putatively wild date palms

We sampled 102 uncultivated date palms from nine different populations in Oman along intermittent streams (wadis) at altitudes ranging from 400 to 2000 m. We considered these populations as uncultivated since they are not presently cultivated as suggested by 1) a balanced sex ratio, while in date palm orchards only few males are kept for pollination, 2) no sign of human-induced pollination since many dates were not fertilized, 3) fully formed dates were generally not harvested, 4) offshoots at the base or along the stems were not cut off and 5) the populations located at highest altitude produced no or only very few fruits.

Some of these uncultivated populations could be relicts of the truly wild date palm, but they might also be feral populations. Indeed, some of the sampled populations were close to abandoned (undated) cultivation terraces located on the side slopes of the valleys and may descend from these previously cultivated populations. Moreover, as the region overall comprises many oasis in which date palms are cultivated and archaeological evidence suggests a long tradition of date palm cultivation [5], none of the sampled populations was very far from agricultural activities or human routes and therefore cannot be considered completely isolated from cultivated populations. Finally, some of the palms seemed at least partially tended: some leaves have been cut off, or fire has been set to the base of the stem.

However, their location in a presently accepted domestication center of the date palm [5] makes them the most promising candidates for wild date palms and some of the samples were previously hypothesized to be wild individuals based on their seed features [14, 15]. Here we aim at verifying their status and identifying if some of these palms represent truly wild date palms.

Cultivated date palms

As reference, we also analyzed 448 date palm cultivars and accessions grown from seeds (also called *khalts* in North Africa). The samples were chosen to cover the whole historical date palm distribution that includes North Africa, Southern Europe and the Middle East up to Pakistan and Northwestern India [35]. Cultivars, mostly females, bear a name and are obtained by vegetative multiplication (by offshoots) from an initial individual selected for its agricultural traits. These samples were either obtained in the field from individual farmers or from living collections of institutions (Table S1). Some additional genomic data were directly retrieved from GenBank. The assignment of cultivars to a country of origin was based on information obtained directly from farmers and individuals growing from seeds were considered to originate from the country of sampling.

Sister species

Samples from two species from the genus *Phoenix,* namely *P. sylvestris* and *P. atlantica,* were included in our analysis, as they are considered the closest relatives of the date palm according to chloroplastic data [21]. Leaves and/or seeds from a total of 74 *P. sylvestris* accessions from five populations were sampled in the field in Gujarat and Rajasthan, India (for more details on these samples see Newton et al., 2013 [67]). An additional 37 leaf samples of *P. atlantica* collected in Cape Verde were donated by Sally Henderson and William Baker (Royal Botanic Gardens, Kew, UK). These samples were previously included in a microsatellite analysis [36].

Samples used for whole-genome sequencing

We retrieved 86 Gb of raw reads of 14 date palm accessions from GenBank Sequence Read Archive (SRA) ([17, 18]; Data S1B). We then complemented this data with seven samples sequenced in this study (Table 1; Table S1; Data S1B). The samples to sequence were chosen to complement the available date palm samples retrieved from GenBank and based on results obtained from the nuclear microsatellite analysis described in this paper. Among the GenBank data, two originated from Africa while 12 originated from the Middle East. We complemented this sampling with one cultivar from Oman (Jabri, 0115_JAB1) and one from Egypt (Siwi, 2025_SIW8). Three putatively wild date palms from three different populations were included: 0285_WILD12, 0342_WILD61 and 0349_WILD68. We also sequenced two outgroups, the date palm sister species *Phoenix sylvestris* (1684_SYL51) and *P. atlantica* (1370_ATL46). In total we thus analyzed the full genomes of 21 individuals including 19 date palms (three putatively wild and 16 cultivated samples) and two outgroups.

Archaeological samples

Four mineralized date seeds were retrieved from the coastal site of H3, As-Sabiyah, Kuwait. The site is dated to 5300-4900 BC by radiocarbon and comparative ceramic analysis. Based on the levels where the stones were found (Period 2-4) these dates can further be narrowed to ca. 5200-4900 BC [68]. The levels contained elements of material culture characteristic of both the Arabian Neolithic (chiefly its lithic technology and shell jewelry) and the Mesopotamian 'Ubaid, mainly comprising southern Mesopotamian ceramics of the early 'Ubaid 3 period, and a range of other small ceramic and stone objects usually associated with the Mesopotamian 'Ubaid period. The site is interpreted as a small coastal settlement, possibly seasonally occupied rather than permanently, whose inhabitants engaged in an early maritime trading network that connected the villages of southern Iraq with the herding and fishing communities of the Neolithic Eastern Arabia. This interpretation is supported by the presence of boat remains, a boat model, and a ceramic disc that appears to depict a boat with a two-footed mast [69]. The archaeological contexts of the date stones consisted of mixed occupation debris trampled into the floors and an entrance of the stone-built chambers found at the center of the site [26].

The four archaeological date seeds were discovered in separate contexts and in different chambers: The sample 3292_DAC (Find no. 1705:01) was found in an occupation deposit or trample in the porch area of Chamber 7 dating to the second period. The samples 3289_DAC (Find no. 1029:05) and 3290_DAC (Find no. 1208:02) were found in occupation deposits in Chambers 1

and 23, respectively, both dating to the third period. And finally the sample 3291_DAC (Find no. 1515:01) was found in the lowest deposit in Chamber 11 dating to period 4, which contained either windblown sand or occupation.

The mixed occupation material in those chambers typically contained broken pottery, edible mollusk shells, fish and faunal remains, discarded lithic tools and debitage, and the debris of shell bead manufacture. The overall botanical assemblage of the site was derived from an extensive flotation and sieving program (overseen by Mark Beech) and included 1,116 plant macrofossil remains from 1,164 l of floated spoil, along with the date seeds, which were recovered separately from the dry sieves. Other elements included a very small number of charred cereal grains (four grains, two identifiable as barley), a single charred jujube stone (*Ziziphus spina-christi)* and a variety of seeds interpreted as belonging to wild species, including suspected fodder and potentially medicinal plants. It is possible that the wild species represent a weed assemblage from agricultural fields nearby, but this requires further investigation. The cereals remain the only cereal grains yet identified in any Arabian Neolithic assemblage, though as noted above, trading links with the agricultural communities of southern Iraq were strong, and it cannot be proven whether the cereals were imported or grown locally. Likewise, it is not possible to know whether the date fruits were imported or locally grown. Presently date palms are not grown in the area, which is barren and not considered suitable for agriculture now or in recent historical times, though a small grove was formerly located in the garden of a sheikhly residence around 15 km away. Palm phytoliths were identified at the site but palm fronds, baskets, and other products could also have been brought in as items of trade for local use.

METHOD DETAILS

Morphometric analysis

Seeds of 348 *Phoenix* samples (20 per sample where available; 6,876 seeds in total; Table 1; Table S1) and four archaeological samples (Table 1; Table S1) were subjected to geometric morphometric analysis. For this, seeds were photographed in both dorsal and lateral view and their outline was described using Elliptic Fourier Transforms using Momocs R package [16, 43, 70] as previously described [14].

While the seeds from extant samples were desiccated before the study as explained in Terral et al., 2012 [14], the mineralized archeological seeds were cleaned but not treated otherwise. We note that the mineralization process may have altered the shape of the archeological seeds. While the chosen analysis method is invariant to changes in size, changes in the outline would affect conclusions. While it is impossible to study the effect of mineralization experimentally, multiple lines of evidence suggest that our conclusion is robust. First, the recovered seeds were excellently preserved. Second, charring does affected seed size, but not its shape, as we recently tested experimentally. While the seeds used here were not charred but mineralized, it does nonetheless suggest that even massive environmental impacts that alter seed size do not easily change seed shape. Finally, it appears unlikely that mineralization would affect seed lengths disproportionately to seed width such that the syndrome of elongated seeds observed among modern cultivars got lost through that process to the degree that the shape among wild individuals is perfectly recovered.

DNA extraction

For each sample subjected to genotyping or sequencing, 40 mg of silica-dried leaves were crushed into a fine powder using the bead-mill homogenizer TissueLyser (QIAGEN). Total genomic DNA was then extracted from the obtained leaf powder using DNeasy plant MINI Kit (QIAGEN) with the modification of adding 1% polyvinylpyrrolidone (PVP 40.000) to buffer AP1.

Microsatellite genotyping

We generated microsatellite data for a total of 569 *Phoenix* samples genotyped at 17 autosomal microsatellites (See Data S1A). Amplification reactions were performed with the QIAGEN Multiplex PCR kit following manufacturer's instructions. Amplified products were detected on an ABI prism 3130xl Genetic Analyzer. Samples were prepared by adding 2 μL of diluted PCR products to 17.85 μL of water and 0.15 μL GenSize HD 400 Rox. Fragment size was determined using the GeneMapper 3.7 software (Applied Biosystems). Most samples (n = 425) were additionally genotyped at a chloroplastic minisatellite present in the intergenic spacer psbZ-trnfM-CAU [41] (See Data S1A). This locus tags the two multi-locus chlorotypes previously reported in date palms [21, 71].

Whole-genome sequencing

A paired-end library for each sample was prepared using standard kits (Illumina) and sequenced on a Genome Analyzer II (Illumina) following standard Illumina protocols. A first run of sequencing with all seven samples pooled on a single lane yielded 64.8 Gb of raw sequencing data. However, the reverse reads were of particularly bad quality, most likely due to the lane being particularly over clustered, which prevented the software from confidently differentiating adjacent clusters and let to clusters becoming more diffuse. We therefore performed a second run of sequencing of the same libraries distributed over three lanes shared with libraries from other projects. This yielded additional 89.4 Gb of raw sequencing data. In total we thus generated 154.2 Gb of data corresponding to 620 million reads of 100 bp each, which corresponds to 89 millions reads per sample on average (See Data S1B). All reads were submitted to the NCBI short-read archive under the accession number SRP094744.

QUANTIFICATION AND STATISTICAL ANALYSIS

Seed morphometric analysis

We used the function *Mclust* from the *mclust* package [44] in R [42] to infer mixture components for each sample and archeological seed. We also run a Principal Component Analysis on this data using the *dudi.pca* function from the ade4 R package [45].

Analysis of microsatellite data

Population tree

We inferred a neighbor-joining tree based on the 12 populations defined from taxonomic and geographic information (Table S1). For this we first prepared the data using the functions *read.loci* and *loci2genind* (both from *pegas* package [46]) followed by *genind2genpop* (adegenet package [47]) in R. Nei's genetic distance [72] was then calculated among populations using the function *nei.dist* (poppr package [48]) and a neighbor-joining tree was then generated with *nj* function (ape package [49]). Bootstrap support was calculated using the *aboot* function (poppr package) by resampling 100 times. The tree was rooted with *P. sylvestris* and displayed using Figtree v1.4.2 [51].

Principal Component analysis

Two Principal Component Analyses (PCA), with and without *P. sylvestris* and *P. atlantica* accessions, were performed in R with the function *dudi.pca* (ade4 package [45]). The data were read and converted using *read.loci* and *loci2genind* from pegas package. Missing values were replaced by the mean allele frequencies using *scaleGen* (adegenet package).

Pairwise F_{ST}

Pairwise F_{ST} [73] among populations were calculated and their significance assessed in Arlequin v. 3.5 [52].

Diversity analyses

Allelic richness and private allelic richness were calculated using the rarefaction method [74] in R to reflect variable sample size. Specifically, we inferred the expected allelic and private allelic richness for each population for a sample size corresponding to the smallest number of chromosomes (74) genotyped in any of the populations involved in the comparison by resampling from the data 1,000 times. Differences in allelic richness between populations or groups were assessed on the bootstrapped samples using Tukey's group test as implemented in the function *HSD.test* (*agricolae* package [50]). Observed and expected heterozygosity were calculated in Arlequin v. 3.5 [52].

Population structure and differentiation

Population structure was inferred from nuclear SSRs data using the software STRUCTURE [53] with 10 independent runs of 100,000 iterations burn-in period followed by 1,000,000 MCMC steps. The optimal value of K was determined both by maximizing the log likelihood [53] and by quantifying the rate of change of the log likelihood between successive K values [75] using Structure Harvester [54]. Both approaches yielded that K = 3 clusters best explained the SSRs (Figure S2).

Bioinformatic analysis of whole-genome sequencing data

Read mapping and alignment cleaning

We used the same genotyping pipeline for all of the 21 samples whether their raw reads were retrieved from GenBank SRA or resulted from our Illumina sequencing. Raw reads were first trimmed for quality and length using Sickle v. 1.33 [55] so that each base pair had a minimum quality of 30. Reads that were smaller than 35 bp after trimming were removed. On average, we kept 91.8% of the raw reads for each individual, resulting in 65 to 402 million reads per individual (average: 101 million, see Data S1B).

All remaining reads were then mapped with the Burrows-Wheeler Aligner v. 0.7.2 [56] against the genome assembly of a *Phoenix dactylifera* var. Khalas individual [9] (GenBank: GCA_000413155.1). This genome assembly is highly fragmented with the largest scaffold being 4.5 Mbp. Prior to mapping we thus removed all scaffolds < 1,000 bp, which resulted in a total length of 476,260,015 bp, representing about 71% of the full genome estimated at 671 Mbp [9]. When both single-end and paired-end files were available, they were aligned independently and the resulting BAM files were then merged using Samtools v. 0.1.18 [57, 58]. On average, 80.9 % of the filtered reads were mapped (See Data S1B).

Three cleaning phases were then performed with Picard tools v. 1.80, Samtools 0.1.18 and Genome Analysis Tool Kit (GATK) v. 2.4.9 [59]. First, the mapping quality of unmapped reads in the BAM files were set to 0 (Picard CleanSam). The MD tag was generated (samtools calmd) and the mate information fixed (Picard FixMateInformation). The resulting BAM files were then checked using Picard ValidateSamFile. Finally, duplicates were marked and removed with Picard MarkDuplicates and indels were realigned with GATK. The resulting average coverage was between 7.3 and 37.6x per sample with an average of 12.5x (See Data S1B).

Variant calling and variant filtering

We called genotypes for all 21 *Phoenix* spp. accessions using the GATK HaplotypeCaller and obtained a total of 9,510,154 variants of which 9,114,275 were SNPs. Using GATK, we filtered out all other variant types (indels, mixed) and multi-allelic SNPs to only keep the 9,059,610 bi-allelic SNPs (95.3% of the total variants). Across all samples and variants, only 4.05% of the genotypes were missing.

The variants were then filtered to 1) minimize the effects of sequencing and alignment errors that might bias downstream analyses and 2) exclude regions of the genome that, irrespective of such errors, might show accelerated rates of evolution. The following criteria were employed to filter sites using VCFtools v. 0.1.11 [60] and custom scripts: sites with a quality lower than 30, a mapping quality lower than 30, a depth below 84 or above 1260, a high Fisher strand bias FS higher than 60. Sites close to indels (5 bp) and in cluster of more than three in window size of 8 bp were also filtered out. Overall, 1,804,421 sites or 19.9% were filtered out leading to

a final number of 7,255,189 SNPs. In addition, we set all genotypes with a genotype quality below 20 as missing, leading to 37.5% of missing genotypes.

A Summary of variant number and proportion of missing genotypes before and after filtering of genotypes called in 21 *Phoenix* spp. accessions is given in Data S1C.

Genome annotation

We used the pipeline Maker-P v. 2.31.8 [63] to structurally annotate the date palm reference genome [9], as detailed below.

Identification of repeats

Repeats can produce sequence alignments with high statistical significance to protein regions creating a false homology throughout the genome [76]. Moreover, transposable elements may occur within introns, which might cause a gene predictor to include extra exons as part of this gene. For these reasons it is critical to identify and mask repetitive regions in the genome prior to any annotation.

Although most repetitive sequences may not be present in the reference genome (only 71% of the genome is available in scaffolds larger than 1,000 bp), we still ran a pipeline for identifying repeated sequences in the date palm genome using a combination of three different approaches in order to maximize the opportunity for repeat collection.

First, we used the two library-based approaches using RepeatMasker v. 4.0.5 [77] to mask repeats from the RepBase libraries [78] and RepeatRunner (available from http://www.yandell-lab.org/software/repeatrunner.html) for masking Transposable Elements (TEs) and viral proteins using the TE proteins database. These two approaches are complementary: RepeatMasker may fail to identify highly divergent repeats since it identifies repeats by means of similarity to a nucleotide library of known repeats. RepeatRunner integrates RepeatMasker with BlastX [79], to search a database of repeat encoded proteins (reverse transcriptases, gag, env, etc.). Because protein homologies can be detected across larger phylogenetic distances than nucleotide similarities, this BlastX search allows RepeatRunner to identify divergent protein coding portions of retro-elements and retro-viruses not detected by RepeatMasker. Second, we used the program MITE-Hunter [80] to identify MITEs and < 2kb class 2 TEs. Finally, we used RepeatModeler [81] v. 1.0.8 for *de novo* identification of repeats based on the repetitive nature of TEs and other repeats that result in high copy numbers of a same sequence. To collect sequences reaching excess copy numbers, RepeatModeler calls three *de novo* repeat finder softwares, namely RECON [82] that uses a self-comparison approach, the k-mer based approach RepeatScout [83] and Tandem Repeat Finder (TRF) [84]. Putative repeats are then classified based on their similarity to known TEs.

Among all the putative repeats collected by these pipelines, sequences of true genes were removed using ProtExcluder (a package available in Maker-P) by mapping each putative repeat against the Uniprot database for plant proteins [85] using BlastX and trimming all sequences that matched plants proteins plus 50 bp flanking sequences.

The remaining repeats were then masked in the reference genome using RepeatMasker, directly called by Maker-P during the first step of the annotation. This lead to an increase of the genome excluded from the annotation from 10.3% (initial proportion of Ns in the downloaded genome) to 34.4% and an average gap length of 3,406 bp (individual gaps between 1 and 31,588 bp). The proportion of the genome excluded from the annotation was higher than previously reported [9] (21.3%), likely due to their pipeline only including RepeatScout, LTR_finder, and MITE-Hunter.

Ab initio *gene prediction*

Ab initio methods seek to recognize sequence patterns within expressed genes and the regions flanking them. Protein-coding regions have distinctive patterns of codon statistics and *ab initio* gene predictors produce gene models based on underlying mathematical models describing patterns of intron/exon structure and consensus start signals. Maker-P supports several gene predictor software and we chose Augustus as it is freely available and highly efficient [86].

Because the patterns of gene structure differ from organism to organism, Augustus must be trained for the date palm genome before being used [76]. For this purpose, we generated in a first run of Maker-P a gene model using transcriptomic data retrieved from GenBank (SRA accession: SRR191838). We then used this gene set to train Augustus. To do so, we first aligned the reads from SRR191838 (43,029,576 reads) to the reference genome using TopHat v. 2.0.11 [87]. 74.4% of the reads were mapped. We then used Cufflinks v. 2.2.0 [88] to assemble the transcripts and obtain a *.gtf* file that we converted into *fasta* using gffread from Cufflinks leading to 34,269 sequences. We filtered out transcripts that are smaller to the expected minimum gene size, that is 150 bp according to Al-Mssallem et al., 2013 [9], leading to 34,099 sequences. We ran Maker-P with these transcripts in order to obtain a training gene set. This file (*.gff*) is composed of 15,132 genes. We eventually used this gene set with autoAug.pl, a perl script provided with Augustus to train it. The resulting parameters will therefore be used in the Maker-P pipeline for doing *ab initio* annotation with parameters that are specific to the date palm genome thus greatly enhancing the search.

Alignment and assembly of transcriptomic data for evidence-based gene prediction

We made use of transcriptomic data (20 runs from 4 different projects) retrieved in GenBank to improve the annotation (See Data S1D).

The obtained RNA-seq reads were mapped to the reference genome [9] using different software depending on the data type. We used TopHat [87] v2.0.11 for Illumina reads and BBMap v. 32.15 (http://sourceforge.net/projects/bbmap/) for 454 reads. Cufflinks v. 2.2.0 [88] was then used on the resulting alignment files to create transcript models (*.gtf* file). These transcripts were directly used by Maker-P for running the evidence-based gene prediction.

Running Maker-P

We ran the pipeline Maker-P [63] with the following options to generate the structural annotation of the date palm reference genome [9]: repeats were masked using the libraries generated as previously explained. The *ab initio* gene prediction was performed

with Augustus using the date palm specific parameters generated as described in the previous section. The tRNA option was set to 1 in order to look for tRNA using tRNAscan [89]. EST evidence was provided as the 20 *.gtf* files obtained from the mapping and assembly of transcriptomic data retrieved in GenBank (see above). For protein homology evidence, we provided Maker-P with a *fasta* file containing all available proteins from plants retrieved in UniProt database in December 2014 [85].

Result of the annotation

The resulting annotation consisted of 25,904 genes (150,491 exons), of which most genes are bounded by 5′ and 3′ UTRs (70.1% and 73.5% respectively). The number of genes described here is very close to the 25,059 genes found in Al-Dous et al., 2011 [17], the first reference genome published. For the more recently updated reference genome that we use in this study [9], a total of 41,660 genes were reported. The lower number of genes we detected here is a direct result of 1) limiting our analysis to the 7,752 contigs of at least 1Kb in size, as opposed to the full set of 10,363 contigs used by Al-Mssallem et al., 2013 [9] and 2) the use of the more stringent predication implemented in Maker-P, which requires all *ab initio* predicted genes to be validated by RNA-seq or protein evidence. In contrast, Al-Mssallem et al., 2013 [9] combined EST assemblies (from pyrosequencing data), RNA-seq reads (SOLiD data), plant protein coding genes and protein domain information with *ab initio* predictions obtained with Fgenesh++ [90] without the requirement of cross-validation. While we are likely missing some genes due to our limit to larger contigs, we expect the genes predicted by Al-Mssallem et al., 2013 [9] to contain a large number of false positives due to the lack of cross-validation. This is also indicated by the only 34,802 genes reported for the oil palm (a not too distant relative), for which a high quality genome and extensive RNA-seq data are available [91].

Analyses of whole-genome sequencing data

Relatedness analysis

To gain knowledge into the relationships of date palm accessions and identify clones, we inferred their degree of genetic relatedness. Using NGSrelate [62] we obtained maximum-likelihood estimates of k_0, k_1 and k_2, the probabilities that two individuals share respectively 0, 1 or 2 alleles identical by descent, respectiely, directly from the BAM files. From these, the relatedness *r* was then calculated, based on which we identified two pairs of accessions displaying high relatedness: Rabia/Sukkariat Qassim and Moshwaq Hada Al-Sham/Shalaby ($r = 1$ and 0.83 respectively). We therefore removed the two accessions with the fewest data (Sukkariat Qassim and Shalaby) in the following analyses. All other pairs of accessions displayed $r \leq 0.5$.

SFS calculation, diversity statistics and F_{ST}

We inferred statistics describing genetic diversity by first inferring population specific site frequency spectra (SFS) for different functional classes of sites directly from the BAM files using ANGSD v. 0.613 [61]. To obtain these spectra, we first pooled all samples and used ANGSD to infer the major and minor alleles for each site while using the reference sequence [9] as ancestral site. We next inferred the SFSs while providing the inferred major and minor alleles and limiting the analysis to reads with a minimum mapping quality of 30 bases with a minimum quality score of 20. The inferred SFSs were further folded to the global minor allele by providing the global major allele as the reference allele to ANGSD. These SFSs were then used to estimate expected heterozygosity (He), nucleotide diversity ($\theta\pi$) and Watterson's estimator (θ_W) using custom R scripts. We tested for a significant difference in the genome-wide nucleotide diversity between the Middle Eastern (ME) and wild populations using a bootstrap approach in which we downsampled the ME individuals to sets of 3 and calculated nucleotide diversity for each set as described above. We found only 4 out of 165 possible sets having higher diversity than the observed diversity in the WILD sampled population.

We obtained F_{ST} estimates in a similar way by first inferring 2D SFS with ANGSD following the same protocol and then estimating $F_{ST} = (H_T\text{-}H_S)/H_T$ using custom R scripts, where H_T and H_S are the expected total heterozygosity and the average expected within population heterozygosity, respectively, estimated directly from the 2D SFS.

Inference of population splits and mixture

We used TreeMix v. 1.12 [22] to infer a maximum-likelihood population graph of both population splits and mixtures from the called SNPs (see section 3C for SNP calling) in blocks (windows) of 500 SNPs, excluding sites with more than two alleles as well as those with missing counts in at least one of the populations. We grouped individuals in the three populations African cultivated date palms, Middle-Eastern cultivated date palms and wild date palms and used the *P. sylvestris* sample as outgroup. We found that a graph with a single mixture event almost perfectly explained the data (Figure S3).

To test the robustness of our inference, we repeated this analysis without grouping the SNPs into blocks, by only including coding SNPs or SNPs at least 10Kb from genes, and by excluding the Egyptian Siwi cultivar (2025_SIW8) from the African population, as we found this individual to be admixed between African and Middle-Eastern populations. All these analyses result in virtually identical population graphs (Figure S3).

We next used the maximum composite likelihood approach implemented in fastsimcoal2 [23] to confirm the finding of an admixture origin of the African cultivars. Since model choice is difficult for composite likelihood approaches, we implemented an admixture model (Methods S1) that allowed, depending on the admixture proportion γ, for a primary ($\gamma \approx 0.0$) or a secondary ($0.5 < \gamma < 1.0$) domestication event in Africa, as well as for a single domestication event for both populations (≈ 1.0).

Phylogeny

A phylogenetic tree based on the called genotypes was reconstructed using Exascale Maximum Likelihood (ExaML) v. 1.0.12 [20]. This software implements the RaxML search algorithm for maximum likelihood (ML) based inference of phylogenetic trees [92] on supercomputers using MPI. It uses, as input, a *.phylip.binary* file. We converted the vcf obtained after the filtering step into a phylip file using our own code and subsequently converted it into a *.phylip.binary* using the executable parse available with the RaxML

software v. 7.2.8. The ExaML software requires an initial tree to start the phylogenetic reconstruction. In order to test the impact of the topology of the starting tree, we generated ten such trees using raxmlHPC-PTHREADS-SSE3 and ran ExaML on all of them. We observed that all runs converged on the same topology and a very similar likelihood value, suggesting that the starting tree had no impact on the tree inference.

In order to obtain bootstrap values, we launched 1,000 ExaML inferences on 1,000 starting trees with a substitution model Gamma. We then ran raxmlHPC with the parameter –z and –n in order to compile the presence or absence of each node in the 1,000 maximum-likelihood trees obtained and thus computed bootstrap values. The consensus tree obtained from this last step was rooted with Newick Utilities v1.6 [93] using *P. sylvestris* (1684_SYL51) as an outgroup. We plotted the tree using *nw_display* executable from the same software.

Multiple sequentially Markovian coalescent

We inferred the population size history using multiple sequentially Markovian coalescent (MSMC) [19] for each individual as well as for two and three individuals (two, four, and six haplotypes) from each population using default settings. This software relies directly on BAM files. The analysis was performed using data from the full genome and assuming a generation time of 10 years and a mutation rate of 2.5×10^{-8} per base per generation.

We note that the MSMC analysis infers changes in effective population sizes and is strongly affected by population structure. As was recently shown, MSMC will infer a bottleneck (i.e., large ancestral population sizes) if samples are drawn from the same deme in a structured population, even if the population sizes were constant [94]. This is because the rate of coalescent is decreasing as lineages migrate away from the sampling deme when going backward in time. It seems likely that an admixture event, which can be seen as a unique pulse of migration, leaves a similar signature in that lineages that are separated into different populations (going backward in time) will not coalesce until very late, leading to an increase population size in the past. This is likely explaining the elevated ancestral population size estimated for the African cultivars.

Inference of the distribution of fitness effects

We used the gene annotation coordinates as predicted by Maker-P to annotate (1) each coding genic site as nonsynonymous and synonymous, (2) intronic sites and (3) intergenic regions that are 10Kb away from genes. This was done with custom C++ code that parsed the reference sequence by codon and according to the eukaryote nuclear genetic code. Codon positions where any mutation would change the encoded amino acid (0-fold degenerate) were annotated as nonsynonymous whereas codon positions where all possible mutations would lead to the same encoded amino acid (4-fold degenerate) were annotated as synonymous.

The demography and the distribution of fitness effects of new mutations were inferred with a maximum likelihood method implemented in computer program *DFE-alpha* [31]. In this framework, new mutations are assumed to be unconditionally deleterious and their effects are quantified as the difference in fitness between homozygotes. With *DFE-alpha* a 2-epoch demographic model is fitted to the SFS of a class of sites that is presumably evolving neutrally *aka* the synonymous SFS as calculated above. This model assumes a population of size N_1 at mutation-drift equilibrium which underwent a change in population size to N_2, t generations ago. For reasons of computational efficiency *DFE-alpha* assumes that $N_1 = 100$ and therefore only the relative change in size (N_2/N_1) and scaled time by effective population size (t/N_2) are meaningful in this estimation framework. The selection model assumes a gamma distribution of selection coefficients parameterized by the shape (b) and the mean of the distribution ($E(s)$) and is fitted to the SFS of the focal class of sites that is the nonsynonymous SFS. Previous simulation work has shown that $E(s)$ is very difficult to estimate accurately [95], especially when sample sizes are small, because strongly deleterious variants are extremely unlikely to be found in a small sample and precisely these variants are those that are most informative on $E(s)$. We thus calculated also the proportion of mutations with effects in different $N_eE(s)$ ranges which is a summary of the DFE that has been shown to be accurately estimated in most settings [31, 95].

We also inferred the DFE with an alternative maximum likelihood method implemented in program *DoFE* [32]. The *DoFE* approach makes no assumption regarding the demographic history of the population and instead fits nuisance parameters to the neutral class of sites. The nuisance parameters quantify the distortion of the neutral SFS compared to its equilibrium expectation and are used to appropriately fit the DFE parameters to the focal (nonsynonymous) SFS.

DATA AND SOFTWARE AVAILABILITY

The accession number for our whole-genome data reported in this paper is GenBank: SRP094744. All morphometric data, microsatellite genotypes (http://dx.doi.org/10.17632/sc9cpvfnbj.1), and the gene annotation (http://dx.doi.org/10.17632/c6bjh7mgby.1) generated in this study are available at Mendeley Data.