PNAS
www.pnas.org

Supplementary Information for

Cross-species hybridization and the origin of North African date palms

Jonathan M. Flowers, Khaled M. Hazzouri, Muriel Gros-Balthazard, Ziyi Mo, Konstantina Koutroumpa, Andreas Perrakis, Sylvie Ferrand, Hussam S. M. Khierallah, Dorian Q. Fuller, Frederique Aberlenc, Christini Fournaraki and Michael D. Purugganan

Michael D. Purugganan
Email: mp132@nyu.edu

This PDF file includes:

Supplementary Text
Supplementary Materials and Methods
Figs. S1 to S8
Tables S1 to S11
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Dataset S1

Supplementary Information Text

Archaeological evidence for date palms (*Phoenix* spp.). The archaeobotanical evidence for the distribution of date palms or date fruits over time has been compiled in the companion document Dataset S1 as part of the Old World Crops Archaeobotanical Database (OWCAD) generated at UCL as part of the European Research Council funder research project on "Comparative Pathways to Agriculture" (ERC # 323842). This database consists of presence/absence data for a range of crops and economic taxa in archaeological sites, some broken into multiple phases, together with georeferences, dating evidence and a grade of evidential quality. The database covers all of Africa and Asia, with more selective coverage of Europe, and it is especially suited to tracking crops in time and space at broad scale, as illustrated, for example, at a Pan Asian scale for key cereals in (1).

The distribution of data in the database provides a visual assessment of geographical coverage of archaeobotanical data. For example supplementary Figure S6 plots all sites from the countries that include *Phoenix* finds (Morocco, Tunisia, Libya, Mali, Egypt, Sudan, Yemen, Oman, United Arab Emirates, Bahrain, Saudi Arabia, Iraq, Iran, Israel, Palestine, Syria, Pakistan, India, Tajikistan). Sites that contain some crop evidence are indicated with open circles, while those with reported *Phoenix* evidence are shown with blue triangles. This highlights the meager coverage of archaeobotanical data in the Sahara and Northern Africa, parts of Iran, Pakistan and Central Asia. Nevertheless, we still regard current evidence of *Phoenix* archaeology as informative.

Figure S7 plots the occurrences of archaeological date finds, referred to *P. dactylifera,* from this database in millennium time bands. Finds of wild sister species, including probable *P. theophrasti* in early western Asia and *P. sylvestris* in some Indian sites are mapped in Figure S8, along with the presumably intrusive date stones from Takarkori. Full details of the distribution of archaeological *Phoenix* in time and space, and which have been or ought to be referred to *P. theophrasti* and *P. sylvestris* are detailed in Dataset S1, followed by a full list of references.

There is additional evidence not plotted in Figure S7 or included Dataset S1 relating to ancient Egypt and southern Mesopotamia. Egypt has seen ~two centuries of archaeological exploration and large quantities of chance finds of plant remains have been recovered, reported, and deposited in museums. In many cases these are from tombs but also these are often poorly recorded as to provenance, and as such it is often hard to be certain that these finds are securely dated. A comprehensive catalogue of such evidence is provided by (2). The compilation provided here included material from the most secure contexts in Egypt (as judged by DQF) and material that comes from more recent systematic sampling. It nevertheless provides a representative overview of the Egyptian evidence for date palm. Out of 142 reports listed in (2) only 18 are older than Middle Kingdom (i.e. before 4000 ybp), while 110 are form the New Kingdom or later (i.e. after 3600 ybp), indicating the widespread establishment of date palm cultivation in Egypt between the Middle Kingdom and New Kingdom. This is reflected in another line of evidence, the art historical record, represented by scene on tomb walls, in which date palms are a regular part of garden scene from the New Kingdom onwards (3,4). Similarly, early cultivation and dates in southern Iraq are indicated by inclusion in the early pictographic script, by depictions on seals and other art from the Late Uruk (Warka) and Early Dynastic period that show date palms, indicating aspects of management (5). The artistic record agrees with the archaeobotanical evidence for the early establishment of date palm cultivation in southern Mesopotamia and its later establishment in Egypt.

The table of archaeobotanical data in Dataset S1 includes a few conventions on data quality. Confidence in the georeferences is graded on a scale from 1 to 3, with three being the most precise. The grade of 3 can be regarded as ±1km, the grade of 2 as ±10km and the grade 1 as ±100km. Sample quality is also graded based on the conventions of (6), with 1= haphazard unsystematic sampling; 2= some systematic

sampling but insufficient reporting of detailed to allow full reanalysis of the data; an 3= full data with sample by sample quantitative data available. Finally quality of data evidence is graded. It is worth noting that very few *Phoenix* remains are directly dated by AMS radiocarbon, and they are therefore dated by association with other directly dated seeds (indicated by *AMS*) in Dataset S1 or other radiocarbon dates, on charcoal or bone (indicated by C14), or simply by associated artefactual material and regional chronologies (ass.). Dates are indicated in terms of likely earliest and latest dated by phase as well as the median between these, which can usually be regarded as the statistically most probable. Dates BCE are given as negative numbers and dates CE as positive.

Supplementary Materials and Methods

Sampling. Date palm samples were obtained from various sources worldwide including 59 reported previously (7), and new samples from Pakistan (Gajar, Hawawiri, Otaquin), Iraq (Manjouma), Libya (Hamria, Barmel), and Morocco (Kamla, Bousl Khine, Raslatmar, Jihl, Boufkouss Rarass, and a Khalte sample). Wild *Phoenix* samples included seven *P. sylvestris*, six *P. canariensis*, and one *P. reclinata* collected from ornamental gardens in southern Europe or from specimens propagated from wild-collected seed (Table S1). Two *P. atlantica* samples were collected from the Cape Verde Islands (8). *Phoenix theophrasti* samples included 15 collected from natural populations in Crete, Greece, two samples from a putatively wild population in Epidaurus, Greece, and one sample from a possible hybrid population in Gölköy, Turkey.

Library preparation and genome sequencing. Genomic DNA was extracted from either leaf or fruit mesocarp/epicarp tissue (Table S1) and 2 X 100 paired-end libraries constructed with Nextera or TruSeq library preparation protocols and sequenced on an Illumina HiSeq 2000 or 2500 system according to the manufacturer's protocols. The date palm draft genome assembly (9) and annotation was downloaded from the National Center for Biotechnology Information (NCBI) on February 28, 2016. This genome is a female assembly that contains the scaffold sequences of RefSeq version DPV01 from (9), the mitochondrial genome (10; RefSeq ID: NC_016740.1) and the chloroplast genome (11; RefSeq ID: NC_013991.2). The nuclear, mitochondrial and chloroplast genomes were combined to form a single modified reference sequence that was used in all subsequent steps.

Raw read and alignment processing. Reads were demultiplexed and those passing Illumina quality control filters were processed with Trimmomatic (12; v. 0.36) to remove contaminating adapter sequences. For adapter removal, we used the adapter and Nextera transposase sequence database included with the Trimmomatic (v. 0.32) download with the following setting ILLUMINACLIP:<adapter library>:2:30:10 and only reads pairs where both reads in a pair were 76 bp or longer following trimming were retained for subsequent steps.

Processed reads were aligned to the unmasked date palm reference genome using bwa mem (13; v. 0.7.15-r1140). The bwa mem aligner was run with the -M option to mark supplementary reads (0x800 bitwise flag) as secondary (0x100). Sample alignments were processed with FixMateInformation (Picard-tools v. 2.8.2; http://broadinstitute.github.io/picard) to ensure consistency in paired-read information, SamSort (Picard-tools v. 2.8.2) to coordinate-sort the alignments, MarkDuplicates (Picard-tools v. 2.8.2) to flag duplicate read pairs, and with GATK IndelRealignerTargetCreator/IndelRealigner tool (14; GATK v. 3.7-0) to realign reads in indel regions. Sample alignments were validated at each step using ValidateSam (Picard-tools v. 2.8.2) to ensure no errors in production. Processed alignments were summarized with CollectAlignmentSummaryMetrics (Picard-tools v. 2.8.2) and Samtools (15; Table S2).

SNP-calling and genotyping. SNP-calling and genotyping was performed with the GATK (GATKv. 3.7-0) HaplotypeCaller run in GVCF mode followed by joint-genotyping with GenotypeGVCFs (16). Reads were filtered from the HaplotypeCaller step to exclude those with a mapping quality less than 20 and to exclude those marked as PCR duplicates or secondary alignments (see above). This approach yielded 39,476,646 SNPs and 5,290,078 indels across all samples.

We restricted analysis to the non-repetitive fraction of the genome assembly by excluding SNPs in regions masked by RepeatMasker (http://www.repeatmasker.org). Additional SNP filtering was conducted by applying hard filters to the raw variants. Thresholds were determined by considering GATK guidelines, considering the impact of thresholds on the transition:transversion ratio (14), and drawing on the approaches of comparable re-sequencing studies of non-model organisms and their relatives. For example, we observed a dependence of the proportion of called heterozygotes on depth in the raw variant calls as expected if spurious SNPs called in regions of the draft assembly with collapsed repeats (17). We therefore tailored our filtering thresholds to minimize this dependency by filtering the raw call set to exclude SNPs with low (< 800) and high depth (> 2200) summed across samples. We also excluded multi-allelic SNPs, SNPs within 6 bp of indel polymorphisms, SNPs with a genotype call rate < 85%, and SNPs meeting the following conditions: FS > 60.0, SOR > 3.0, QD < 8.0, MQ < 40.0, MQRankSum < -3.0, ReadPosRankSum < -1.5, BaseQRankSum < -8.0 (see https://software.broadinstitute.org/gatk/ for tag definitions). This procedure yielded a filtered call set of 14,402,469 SNPs that served as the basis for all analysis.

SNPs from the chloroplast and mitochondrial genomes were called using the same GATK HaplotypeCaller/joint-genotyping work flow. SNPs were called with ploidy set to diploid to identify heterozygous genotypes attributable to heteroplasmy (18) or insertions of either plastid genome into nuclear DNA. We applied both SNP and genotype filters to the mitochondrial and chloroplast call sets. Genotypes were set to missing if the Phred-scaled genotype quality (GQ) was less than 20. SNPs were filtered by excluding SNPs in which any sample had a heterozygous or missing genotype, excluding SNPs in the region of the chloroplast genome that is duplicated in the mitochondrial genome based on coordinates reported (19), and excluding sites found in repeat regions reported in (19,20). We then applied SNP filters to the cpDNA and mtDNA SNPs with thresholds modified after (21). SNPs meeting the following criteria were excluded: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, sites within 10 bp of called indels, and sites reported as repeat regions. This filtering strategy reduced the total number of raw chloroplast SNPs from 436 to 121 and mitochondrial from 6,019 to 760.

Phylogeny Reconstruction. Neighbor-joining trees were generated for mtDNA, cpDNA and selected introgressed regions using the JC69 model of nucleotide substitution with the ape and phangorn packages in R. Bootstrap support values for branches were calculated with 1000 resampling iterations and output trees produced with Dendroscope with branches with less than 50% support collapsed. Maximum likelihood phylogenies were constructed with Randomized Axelerated Maximum Likelihood (RAxML; 22). The phylogeny based on SNPs in the nuclear genome was produced with the Generalized Time Reversible (GTR) substitution model with Gamma rate heterogeneity (-m GTRGAMMA) using 33,505 variable sites, but excluding sites with heterozygotes. An ascertainment bias correction (-m ASC_GTRGAMMA --asc-corr=lewis) was applied to likelihood calculations to prevent overestimation of branch lengths and biases in tree topology. The number of bootstrap iterations was determined by the automatic majority-rule consensus tree criterion for bootstrap convergence (-# autoMR). Output trees were produced with Dendroscope. Maximum Likelihood cpDNA and mtDNA trees used the same settings but are based on all SNPs in the filtered datasets for these genomes.

Population statistics. Statistics nucleotide diversity (π), Watterson's θ (θ_w) and Tajima's D were calculated for each population or species in 5 kb non-overlapping intervals using ANGSD (v. 0.917)

using sample BAM alignments as input. Estimates were obtained for each population by excluding probable hybrid individuals and by filtering out reads with mapping quality < 20 and base quality < 20. F_{st} was calculated in the same 5 kb non-overlapping intervals from the filtered SNP call set using vcftools (23; v. 0.1.14) analysis by excluding probable hybrid *P. theophrasti* samples in comparisons with *P dactylifera* and varieties of date palm from Egypt and Sudan in F_{st} estimates between date palm populations.

Population clustering. Model-based clustering of genotypic data was performed with STRUCTURE (24). The filtered SNP dataset was randomly sub-sampled to include ~30,000 SNPs to limit the effects of linkage on the analysis. STRUCTURE was then run with the Admixture model with correlated allele frequencies without including geographic or species-membership information for K = 1 to K = 8 with chain lengths between 750,000 and 1,000,000 steps and burn in of 200,000 steps. Analyses were then repeated by running the Admixture model with independent allele frequencies. A second set of "hierarchical" analyses was run separately on species pairs (i.e., date palms and a wild relative *P. sylvestris*, *P. theophrasti*, or *P. canariensis*) again for both correlated and independent allele frequency models. Admixture proportions were monitored for consistency across replicates and the run with the highest maximum likelihood run at each K are presented (Fig. 2). Additional summary metrics were calculated with STRUCTURE Harvester (25). Analysis of the full set of samples was repeated with a second set of ~30,000 random SNPs to confirm the results were not sensitive to a particular set of SNPs. All outputs were qualitatively similar between the two SNP sets.

Admixture tests. Tests for admixture were conducted by selecting a subset of samples from each population. We selected six samples from Libya, Tunisia, Algeria, and Morocco for our North African sample, six Middle Eastern, six *P. sylvestris* (excluding a probable date palm hybrid from Faisalabad, Pakistan), six *P. theophrasti* (excluding any putative *P. dactylifera* X *P. theophrasti* hybrids), six *P. canariensis*, and one *P. reclinata*. In a separate set of analysis we defined a population consisting of four Egyptian samples. All tests of admixture were performed with the Popstats software (https://github.com/pontussk/popstats).

We used the *D*-statistic to test for admixture between *Phoenix* wild relatives and *P. dactylifera* populations (26,27). Since we were interested in gene flow between date palm and its wild relatives, we focused on tests that included North African and Middle Eastern populations and a wild relative as the test population. In these tests, P1 and P2 are sister taxa, P3 is a test population, and P4 is an outgroup, which corresponds to the notation *D*(P1,P2,P3,O). We note that this is equivalent to the notation *D*(O,P3;P1,P2) used by the Popstats software.

We performed tests where P1 and P2 are Middle Eastern and North African populations, respectively, and wild relatives *P. sylvestris, P. theophrasti*, or *P. canariensis* were the test population. *Phoenix canariensis* or *P. reclinata* were included as outgroups. *D*-tests were performed with the approach of (27) for SNP data. Significance was assessed by block jackknife by treating each scaffold as a block and weighting each block by the number of SNPs. The standard error (SE) of the test statistic was used to define a Z-score (D/SE; 26). For the tests presented, we excluded scaffolds shorter than 800 kb in our analysis as inclusion of smaller scaffolds led to smaller standard errors and inflated |Z|. |Z| used to assess significance may be over-estimated in some cases owing to the constraints on increasing the jackknife block sizes given the current state of the draft assembly.

We employed the f_3 statistic to test for admixture within the shared genetic drift framework of (28-30). The f_3-statistic tests for admixture among three populations. In the no-admixture case, the f_3 test of the form f_3(Px;P1,P2) measures the branch length in a population phylogeny between Px and the internal node of the unrooted tree. The statistic in this case is expected to be greater than zero. Negative f_3 is

indicative of Px having a mixed ancestry from P1 and P2 or populations closely related to them. Significance was assessed using the same approach as the *D*-statistic described above.

Population modeling. *TreeMix* is a software for modeling population history as a directed acyclic graph with both split and mixture events (31). *TreeMix* constructs a population tree that maximizes the composite likelihood of the observed covariance in allele frequencies among populations. It then sequentially adds migration edges to connect pairs of populations that show a relative excess of allele frequency covariance and are therefore poor fits to the strict tree model. For this analysis, we chose six samples from each *P. canariensis*, *P. theophrasti*, *P. sylvestris*, North African and Middle Eastern date palm and the single *P. reclinata* sample. For each set of analyses, we ran *TreeMix* with either zero, one, or two migration events and specified either *P. reclinata* as the root taxon. For models from which we dropped *P. reclinata*, we included *P. canariensis* as the root. We confirmed that the inferred population tree and migration edges are robust to different input taxa and to different block sizes incorporated to account for linkage disequilibrium among linked SNPs.

Ancestry proportions. The f_4 statistic is closely related to *D*-tests differing only in the denominator of the two statistics (29). Ancestry proportions in an admixed population can be estimated by calculating a ratio of appropriate f_4 statistics assuming a specific phylogeny (28-30). When estimating ancestry proportions in North African date palms with the f_4-ratio approach, we assume that *P. dactylifera* and *P. sylvestris* are sister species and that *P. canariensis* (or *P. reclinata*) is an outgroup to *P. dactylifera*, *P. sylvestris*, and *P. theophrasti*. We then estimate the proportion of North African date palm ancestry that traces to the Middle Eastern date palm population as:

$$\alpha = f_4(\textit{sylvestris}, \textit{canariensis}; \textit{dactylifera}_{\mathrm{NAF}}, \textit{theophrasti}) / f_4(\textit{sylvestris}, \textit{canariensis}; \textit{dactylifera}_{\mathrm{ME}}, \textit{theophrasti})$$

where $\textit{dactylifera}_{\mathrm{NAF}}$ represents North African date palm and $\textit{dactylifera}_{\mathrm{ME}}$ represents the Middle East. The proportion of North African ancestry that traces to *P. theophrasti* in this context is defined as 1 - α.

The ancestry of Egyptian samples was calculated separately by replacing North African date palm samples with those from Egypt in the above f_4-ratio calculations. In addition, we repeated the North African and Egyptian f_4-ratio calculations by replacing *P. canariensis* with *P. reclinata* to assess the robustness of the ancestry estimates to outgroup species. Standard errors of f_4-ratios were estimated with the weighted block jackknife approach described for the *D*-statistic. f_4-ratio estimation was performed with Popstats on the set of scaffolds 800 kb or larger. Estimates of ancestry proportions from *TreeMix* are based on the mixture weights on the migration edges (31).

Identification of *P. theophrasti*-like and Middle-Eastern-like alleles
We identified SNPs that are fixed between *P. theophrasti* and the Middle Eastern date palm population. Samples used to infer fixations were all Middle Eastern date palm and *P. theophrasti* (excluding probable hybrid individuals). We then evaluated whether the North African population was fixed for the Middle Eastern-*like* allele, fixed for the *P. theophrasti*-*like* allele, or polymorphic for both.

Population statistics in introgressed regions. The introgression fraction, f_d, was obtained from the filtered SNP call set in non-overlapping intervals of 5 kb with a script reported in (32). f_d was calculated for the ABBA-BABA configuration *D*(P1=Middle East,P2=North Africa,P3=*theophrasti*,O=*reclinata*). Introgressed tracts were defined as two or more consecutive intervals of 5 kb with f_d in the upper 10^{th} percentile of the genomewide distribution. Comparisons of population statistics across f_d bins were obtained by subsetting intervals with $D < 0$ (in the above configuration) as these regions show no evidence of excesses of shared derived alleles between P2 and P3 and f_d is therefore not a quantitatively meaningful measure of introgression (32). The remaining intervals were then ranked by f_d and binned into

20 bins with the same number of 5 kb intervals in each bin (see Fig. S5). Population summary statistics were then summarized for each bin.

Local haplotype phasing. Regions of the genome that showed introgression signatures were phased locally to build phylogenetic trees of local haplotypes, genotypes on scaffolds that showed a signatures of introgression were phased by Beagle (33; v. 4.1). Beagle performs haplotype inference using a Hidden Markov Model (HMM) of localized haplotype-clusters and applies stochastic expectation-maximization (EM) to iteratively improve the likelihood of the inferred haplotype pairs. Neighbor-joining trees of selected phased regions were then generated following the same approach outlined above.

LD decay. LD was calculated using vcftools (23; v0.1.14) with command line options --geno-r2 --maf 0.1 --ld-window-bp 100000 to calculate LD for unphased SNPs, exclude SNPs with minor allele frequency < 10%, and exclude pairs of sites further than 100 kb apart. Twenty North African samples and 20 randomly selected Middle Eastern samples were included for the calculation of North African and Middle Eastern population LD, respectively. All LD analyses were based on 516 scaffolds in the genome assembly that exceed 200 kb in length. To reduce the total number of pairwise comparisons, the filtered SNP call set was down-sampled to keep 33% of all sites.

LD decay curves were plotted by nonlinear least squares (nls) regression using an approach adapted from (34), which fits LD data to the following model (35):

$$\mathbb{E}\left(r^2\right) = \left[\frac{10+C}{(2+C)\,(11+C)}\right]\left[1+\frac{(3+C)\left(12+12C+C^2\right)}{n\,(2+C)\,(11+C)}\right]$$

where n is the sample size and C, the parameter to be estimated, represents the product of the population recombination parameter and the distance in base pairs. Half decay distance was estimated by taking the distance at which the value of the curve is half of its maximum value (i.e. at 1 bp).

Private alleles. We identified private polymorphisms in each population or species using the filtered SNP call set. We define a private polymorphism as a SNP segregating for an allele ("private allele") that is restricted in its distribution to a particular focal population or species in our analysis. Private polymorphisms were discovered by defining a focal set of samples and then identifying SNPs in which one of the two alleles is restricted to that set at the exclusion of the other samples in the analysis (the "non-focal set"). We define private fixations as an allele observed at 100% frequency in the focal population or species, but not observed in the non-focal set.

In all private allele analysis, we excluded wild relative samples that appear to be inter-specific hybrids (1 *P. sylvestris*, 5 *P. theophrasti*), all samples from Egypt/Sudan (which are highly admixed between date palm populations), and date palm cultivars from more recently established production areas in Pakistan, which in some cases show admixture between Middle East and North African populations (e.g., the Aseel variety). For analysis of private alleles in North African date palm and *P. dactylifera* (Middle East + North Africa) we also excluded the two *P. atlantica* samples which are similar to North African date palm as the inclusion of these samples in the non-focal set reduces the number of private polymorphisms in North Africa and *P. dactylifera*.

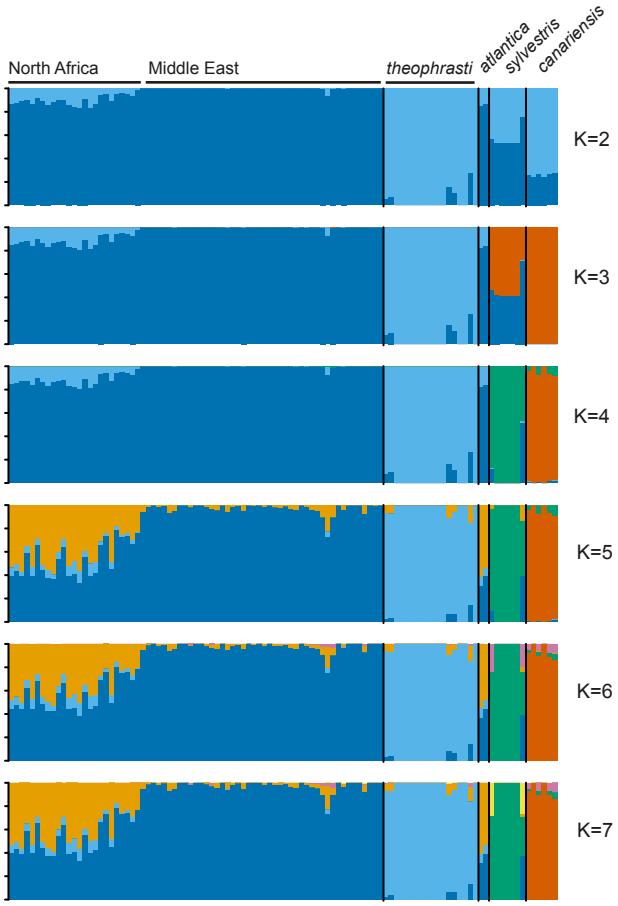

Fig. S1. STRUCTURE analysis with correlated frequency model.

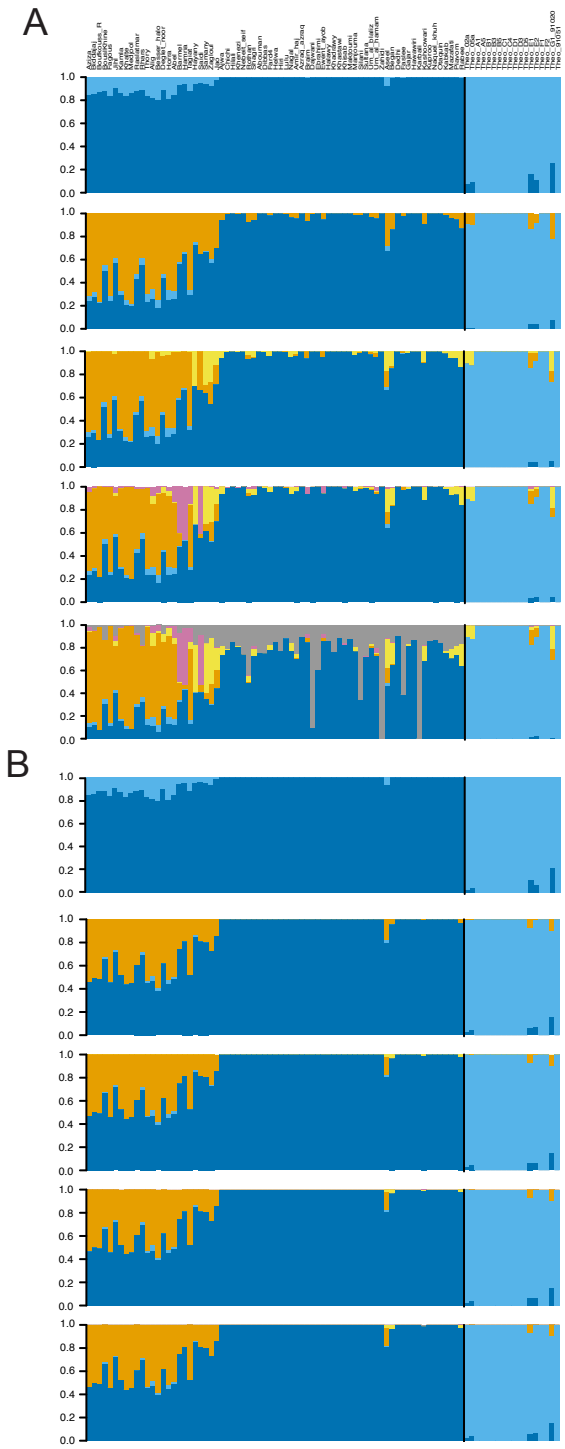

Fig. S2. Pairwise STRUCTURE analysis of *P. theophrasti* and date palm. (A) The correlated allele frequency model (K = 2 - 6), and (B) independent allele frequency model with date palm and *P. theophrasti* samples only (K = 2 - 6).

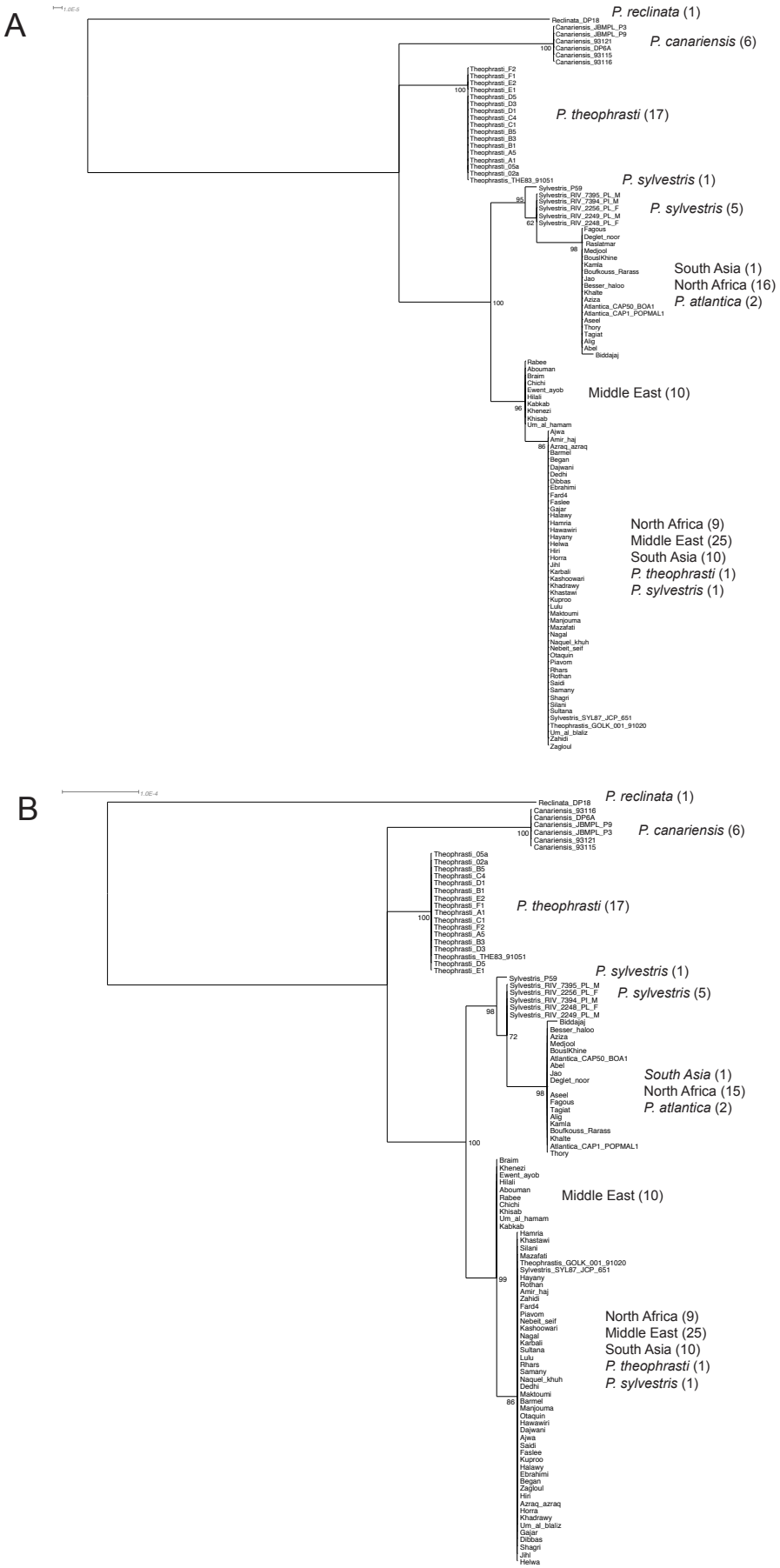

Fig. S3. Phylogeny of *Phoenix* species based on whole genome re-sequencing of chloroplast DNA. The number of samples from each date palm region or *Phoenix* wild relative is shown. (A) Neighbor-joining tree based on JC69-corrected distances. Node support values are the percent of bootstrap replicates supporting the node. (B) Maximum Likelihood tree. Node support values are the percent of bootstrap replicates supporting the node. Nodes with less than 50% support have been collapsed in both (A) and (B).

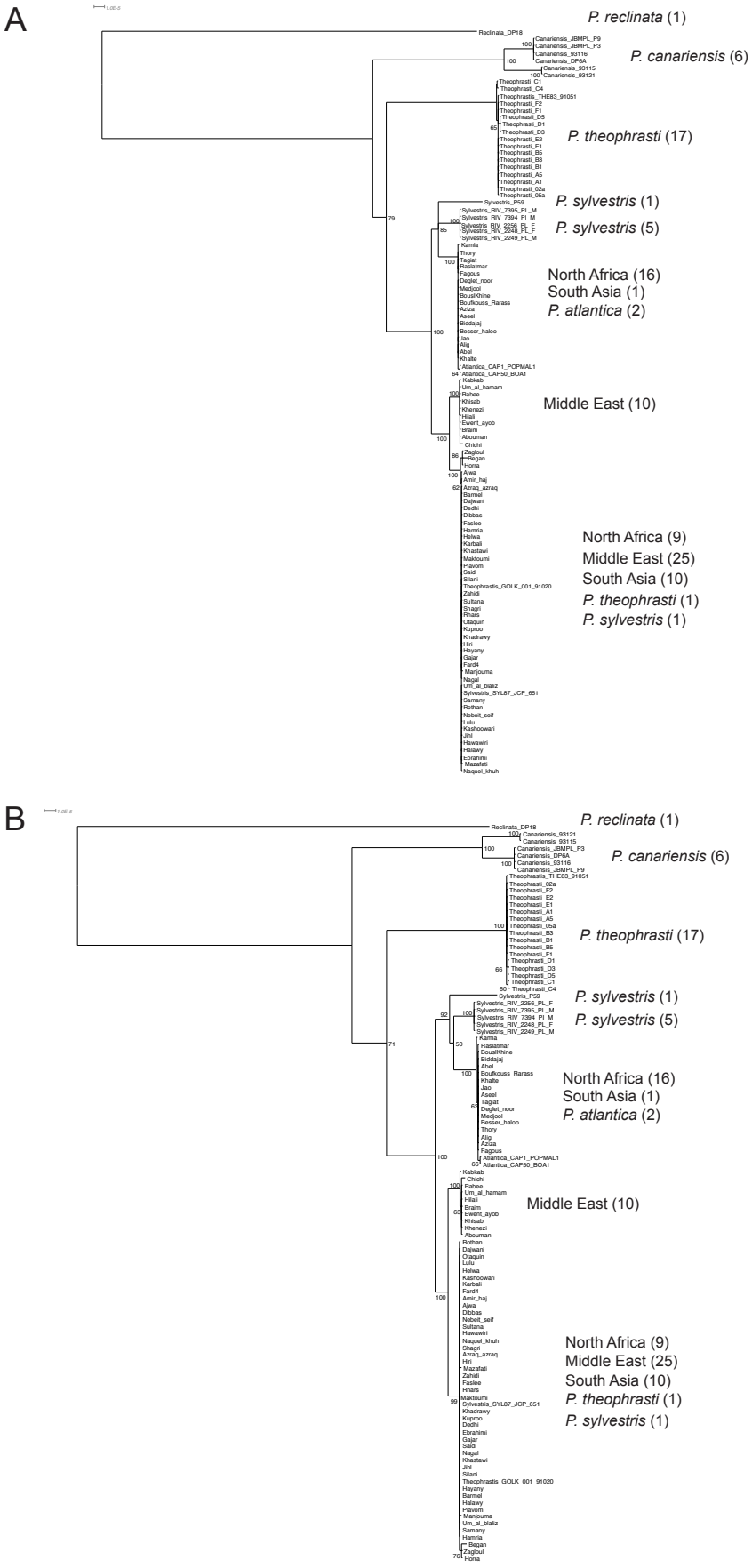

Fig. S4. Phylogeny of *Phoenix* species based on whole genome re-sequencing of mitochondrial DNA. The number of samples from each date palm region or *Phoenix* wild relative is shown. (A) Neighbor-joining tree based on JC69-corrected distances. Node support values are the percent of bootstrap replicates supporting the node. (B) Maximum Likelihood tree. Node support values are the percent of bootstrap replicates supporting the node. Nodes with less than 50% support have been collapsed in both (A) and (B).

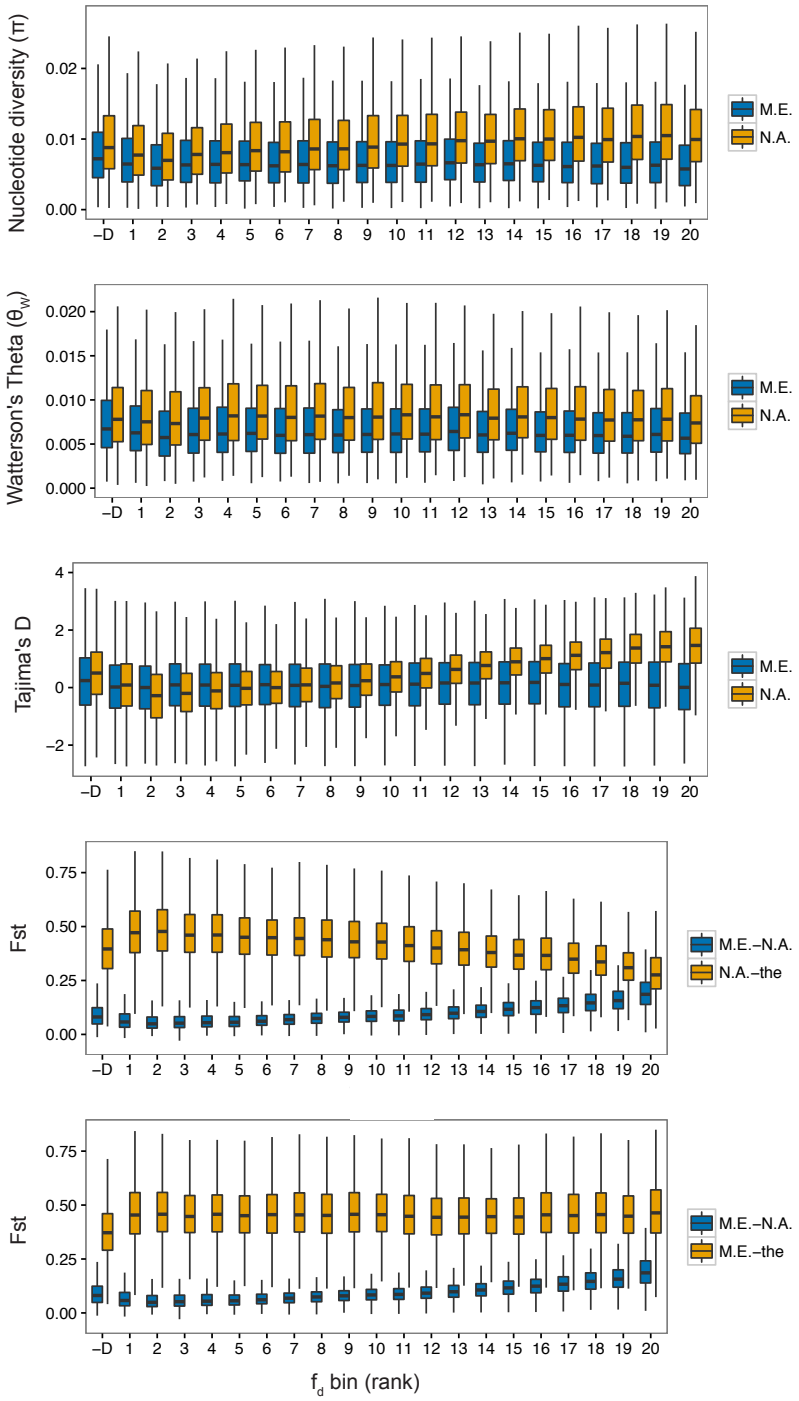

Fig. S5. Population genetic summary statistics in rank-ordered introgression fraction bins (f_d) for the population configuration D(Middle East, North Africa, *P. theophrasti*,*P.reclinata*). f_d was calculated in 5 kb intervals and bins with positive *D* binned according to percentile such that each bin has approximately the same number of genomic intervals. The bin labelled -D are those intervals where D is negative and which f_d has no meaningful quantification of introgression (32) and may not contain same number of intervals as bins with positive D. Boxplots were then generated for various population genetic statistics in each f_d bin. M.E. = Middle Eastern date palm, N.A = North African date palm, the = *P. theophrasti*.

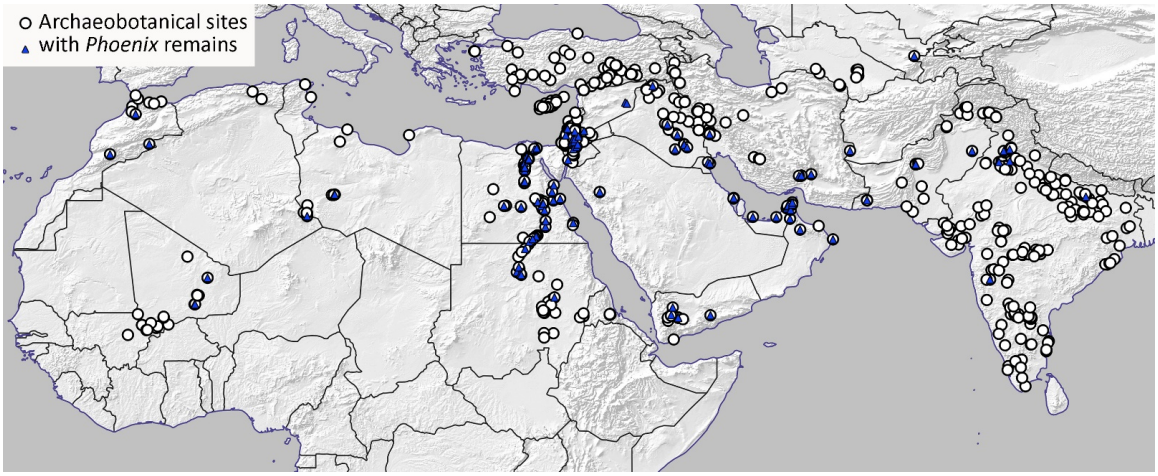

Fig. S6. Distribution of *Phoenix* archaeological reports from the relevant range of dates among all sites with archaeobotanical data for a selection of countries.

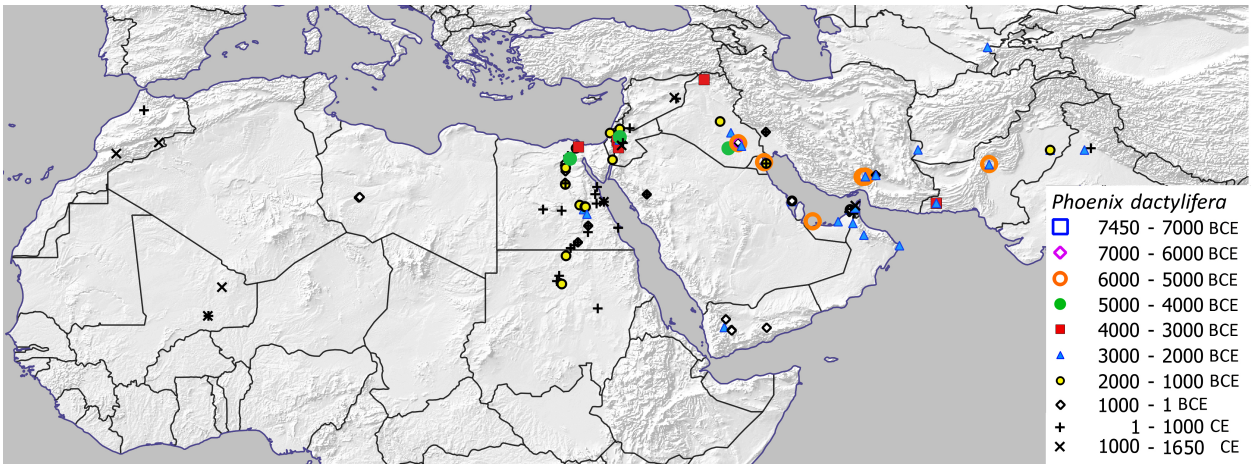

Fig. S7. A map of archaeological finds of *P. dactylifera* differentiated by age.

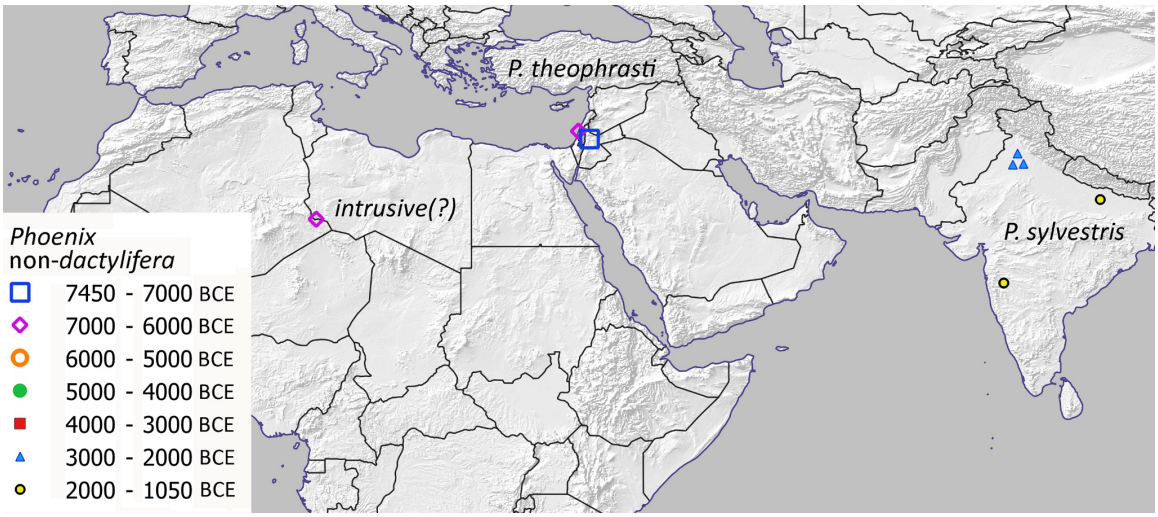

Fig. S8. Distribution of archaeological finds excluded from Fig. S7, including *P. sylvestris, P. theophrasti,* and probably intrusive Takarkori find.

Table S1. Sample information.

Sample[*]	Species	Origin[a]	Collecting Locale	Sex	Source[b]	Tissue
Kamla	dactylifera	Morocco	Morocco	F	market	fruit
Khalte	dactylifera	Morocco	Morocco	F	market	fruit
Bousl Khine	dactylifera	Morocco	Morocco	F	market	fruit
Raslatmar	dactylifera	Morocco	Morocco	F	market	fruit
Jihl	dactylifera	Morocco	Morocco	F	market	fruit
Boufkouss Rarass	dactylifera	Morocco	Morocco	F	market	fruit
Aziza	dactylifera	Morocco	Morocco	F	market	fruit
Fagous	dactylifera	Morocco	Morocco	F	market	fruit
Biddajaj	dactylifera	Morocco	Morocco	F	market	fruit
Medjool	dactylifera	Morocco	UAE	F	DPTCL[c]	leaf
Thory	dactylifera	Algeria	California, USA	F	USDA[d]	leaf
Rhars	dactylifera	Algeria	Arizona, USA	F	ASU[e]	leaf
Deglet Noor	dactylifera	Algeria	Tunisia	F	TCD[f]	leaf
Alig	dactylifera	Tunisia	Tunisia	F	TCD	leaf
Besser Haloo	dactylifera	Tunisia	Tunisia	F	TCD	leaf
Horra	dactylifera	Tunisia	California, USA	F	USDA	leaf
Abel	dactylifera	Libya	Libya	F	market	fruit
Tagiat	dactylifera	Libya	Libya	F	market	fruit
Hamria	dactylifera	Libya	Libya	F	market	fruit
Barmel	dactylifera	Libya	Libya	F	market	fruit
Hayany	dactylifera	Egypt	California, USA	F	USDA	leaf
Samany	dactylifera	Egypt	California, USA	F	USDA	leaf
Saidi	dactylifera	Egypt	California, USA	F	USDA	leaf
Zagloul	dactylifera	Egypt	Syria	F	AECS[g]	leaf
Jao	dactylifera	Sudan	UAE	F	market	fruit
Chichi	dactylifera	Saudi Arabia	UAE	F	DPTCL	leaf
Hilali	dactylifera	Saudi Arabia	California, USA	F	USDA	leaf
Rothan	dactylifera	Saudi Arabia	UAE	F	ICBA[h]	leaf
Shagri	dactylifera	Saudi Arabia	UAE	F	ICBA	leaf
Khenezi	dactylifera	Saudi Arabia	UAE	F	DPTCL	leaf

Nebeit Seif	dactylifera	Saudi Arabia	UAE	F	DPTCL	leaf
Ajwa	dactylifera	Saudi Arabia	UAE	F	ICBA	leaf
Dibbas	dactylifera	UAE	UAE	F	DPTCL	leaf
Helwa	dactylifera	UAE	UAE	F	DPTCL	leaf
Hiri	dactylifera	UAE	UAE	F	DPTCL	leaf
Fard #4	dactylifera	UAE	California, USA	M	USDA	leaf
Lulu	dactylifera	UAE	UAE	F	DPTCL	leaf
Abouman	dactylifera	UAE	UAE	F	DPTCL	leaf
Nagal	dactylifera	UAE	UAE	F	market	fruit
Maktoumi	dactylifera	Iraq	UAE	F	DPTCL	leaf
Khadrawy[l]	dactylifera	Iraq	California, USA	F	USDA	leaf
Khastawi	dactylifera	Iraq	Arizona, USA	F	ASU	leaf
Sultana	dactylifera	Iraq	UAE	F	DPTCL	leaf
Um al hamam	dactylifera	Iraq	Iraq	F	DPRU[i]	leaf
Um al blaliz	dactylifera	Iraq	Iraq	F	DPRU	leaf
Ewent ayob	dactylifera	Iraq	Iraq	F	DPRU	leaf
Azraq azraq	dactylifera	Iraq	Iraq	F	DPRU	leaf
Ebrahimi	dactylifera	Iraq	Iraq	F	DPRU	leaf
Dajwani	dactylifera	Iraq	Iraq	F	DPRU	leaf
Silani	dactylifera	Iraq	Iraq	F	DPRU	leaf
Khisab	dactylifera	Iraq	California, USA	F	USDA	leaf
Halawy	dactylifera	Iraq	California, USA	F	USDA	leaf
Zahidi	dactylifera	Iraq	California, USA	F	USDA	leaf
Amir Haj	dactylifera	Iraq	California, USA	F	USDA	leaf
Manjouma	dactylifera	Iraq	Iraq	F	DPRU	leaf
Braim[l]	dactylifera	Iraq	Arizona, USA	F	ASU	leaf
Kabkab (red)	dactylifera	Iran	Syria	F	AECS	leaf
Mazafati	dactylifera	Iran	Qatar	F	WCMC[j]	fruit
Piavom	dactylifera	Iran	Qatar	F	WCMC	fruit
Rabee	dactylifera	Iran	Qatar	F	WCMC	fruit
Kashoowari	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI[k]	leaf
Dedhi	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Naquel Khuh	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Aseel	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf

Kuproo	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Began	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Faslee	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Karbali	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Gajar	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Hawawiri	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Otaquin	dactylifera	Pakistan	Sindh, Pakistan	F	DPRI	leaf
Sylvestris [RIV 2256 PL]	sylvestris	-	California, USA	F	USDA	leaf
Sylvestris [RIV 7394 PI]	sylvestris	-	California, USA	M	USDA	leaf
Sylvestris [RIV 2248 PL]	sylvestris	-	California, USA	F	USDA	leaf
Sylvestris [RIV 7395 PL]	sylvestris	-	California, USA	M	USDA	leaf
Sylvestris [RIV 2249 PL]	sylvestris	-	California, USA	M	USDA	leaf
Sylvestris [P59]	sylvestris	-	Valencia, Spain	?	garden	leaf
Sylvestris [SYL87 JCP 651]	sylvestris	-	Faisalabad, Pakistan	M	garden	leaf
Canariensis [JBMPL P3]	canariensis	-	Montpelier, France	F	garden	leaf
Canariensis [JBMPL P9]	canariensis	-	Montpelier, France	M	garden	leaf
Canariensis [93115]	canariensis	-	Sanremo, Italy	F	garden	leaf
Canariensis [93116]	canariensis	-	Sanremo, Italy	M	garden	leaf
Canariensis [93121)	canariensis	-	Sanremo, Italy	M	garden	leaf
Canariensis [DP6A]	canariensis	Gran Canaria	California, USA	?	wild	leaf
Atlantica [CAP1 POPMAL1]	atlantica	Maio I.	Maio I.	F	wild	leaf
Atlantica [CAP50 BOA1]	atlantica	Boa Vista I.	Boa Vista I.	F	wild	leaf
Reclinata [DP18]	reclinata	Rwanda	California, USA		USDA	leaf
Theophrasti [THE83 91051]	theophrasti	Crete, Greece	Sanremo, Italy	?	garden	leaf
Theophrasti [GOLK001 91020]	theophrasti	Golkoy, Turkey	Sanremo, Italy	?	garden	leaf
Theophrasti [02a]	theophrasti	Epidaurus	Epidaurus	?	putative wild	leaf
Theophrasti [05a]	theophrasti	Epidaurus	Epidaurus	F	putative wild	leaf
Theophrasti [A1]	theophrasti	White Lake	White Lake	?	wild	leaf
Theophrasti [A5]	theophrasti	Chrisoskalitissa	Chrisoskalitissa	F	wild	leaf
Theophrasti [B1]	theophrasti	Preveli	Preveli	F	wild	leaf
Theophrasti [B3]	theophrasti	Preveli	Preveli	?	wild	leaf
Theophrasti [B5]	theophrasti	Preveli	Preveli	F	wild	leaf
Theophrasti [C1]	theophrasti	Maridaki	Maridaki	F	wild	leaf
Theophrasti [C4]	theophrasti	Maridaki	Maridaki	F	wild	leaf

Theophrasti [D1]	theophrasti	Vai	Vai	F	wild	leaf
Theophrasti [D3]	theophrasti	Vai	Vai	M	wild	leaf
Theophrasti [D5]	theophrasti	Vai	Vai	F	wild	leaf
Theophrasti [E1]	theophrasti	Almyros	Almyros	?	wild	leaf
Theophrasti [E2]	theophrasti	Almyros	Almyros	?	wild	leaf
Theophrasti [F1]	theophrasti	Drapano	Drapano	M	wild	leaf
Theophrasti [F2]	theophrasti	Drapano	Drapano	F	wild	leaf

[*]Internal identifiers are provided in brackets where applicable
[a]traditionally-recognized country of origin for the variety, cultivar, or uncultivated sample
[b]samples indicated as "garden" were sampled from ornamental gardens
[c]Date Palm Tissue Culture Laboratory, United Arab Emirates University, Al Ain, UAE
[d]United States Department of Agriculture, Thermal, California, USA
[e]Arizona State University Date Palm Collection, Tempe, AZ, USA
[f]Technical Center of Dates, Ministry of Agriculture, Kebili, Tunisia
[g]Department of Molecular Biology and Biotechnology, Atomic Energy Commission of Syria, Damascus, Syria
[h]International Center for Biosaline Agriculture, Dubai, UAE
[i]Date Palm Research Unit, College of Agriculture, University of Baghdad, Baghdad, Iraq
[j]Genomics Core Laboratory, Weill Cornell Medical College in Qatar, Doha, Qatar
[k]Date Palm Reasearch Institute, Sindh, Pakistan
[l]sample information in Hazzouri et al. (2015) was incorrect

Table S2. Sequencing and read alignment metrics.

sample[a]	Reads mapped	Reads unmapped	proportion mapped	depth	coverage breadth[b]	mismatch rate[c]
Abel	150525640	1717101	0.99	22.3	0.88	0.01
Abouman	263463003	3393585	0.99	37.88	0.89	0.0098
Ajwa	132578419	1620879	0.99	19.8	0.88	0.0086
Alig	141920612	1820595	0.99	20.95	0.88	0.0103
Amir_haj	352510217	8852814	0.98	49.57	0.89	0.0094
Aseel	160580565	3485016	0.98	23.69	0.89	0.0108
Atlantica [CAP1 POPMAL1]	195667937	2750604	0.99	28.95	0.88	0.0153
Atlantica [CAP50 BOA1]	176266558	2649869	0.99	26.21	0.88	0.0151
Aziza	141765589	1841552	0.99	21.31	0.88	0.0125
Azraq azraq	75701329	974123	0.99	11.76	0.88	0.009
Barmel	62519936	843094	0.99	9.77	0.87	0.01
Began	209037901	3351420	0.98	30.57	0.89	0.0096
Besser haloo	160530252	2030261	0.99	22.85	0.88	0.0101
Biddajaj	187951211	3470057	0.98	27.47	0.88	0.013
Boufkouss Rarass	66747320	919014	0.99	10.47	0.87	0.0109
BouslKhine	76771594	1614322	0.98	11.99	0.87	0.0105
Braim	154146979	2285136	0.99	22.54	0.88	0.0098
Canariensis [93115]	181999222	3793091	0.98	25.84	0.86	0.0223
Canariensis [93116]	198573506	4089848	0.98	27	0.85	0.022
Canariensis [93121]	180054559	3715100	0.98	26.13	0.86	0.0219
Canariensis [DP6A]	100650053	1971197	0.98	14.71	0.83	0.0183
Canariensis [JBMPL P3]	190496371	3831199	0.98	27.55	0.86	0.022
Canariensis [JBMPL P9]	172643248	3211051	0.98	24.71	0.86	0.0213
Chichi	102551251	1076301	0.99	15.17	0.88	0.0082
Dajwani	69557483	913802	0.99	10.84	0.88	0.0095
Dedhi	80903437	1379916	0.98	12.51	0.88	0.0088
Deglet_noor	137190335	1926557	0.99	20.37	0.88	0.0103
Dibbas	68088290	743069	0.99	10.04	0.87	0.0085
Ebrahimi	79483973	1281262	0.98	12.31	0.88	0.0099

Ewent_ayob	72542802	1048573	0.99	11.26	0.88	0.0097
Fagous	122690639	2314674	0.98	18.45	0.88	0.0129
Fard4	172547304	2388039	0.99	25.48	0.88	0.0093
Faslee	170387024	1882082	0.99	25.44	0.89	0.0089
Gajar	75888064	1128411	0.99	11.74	0.88	0.0089
Halawy	169137087	1931453	0.99	29.16	0.89	0.0086
Hamria	50541295	11005690	0.82	7.86	0.86	0.0095
Hawawiri	69844164	1035048	0.99	10.83	0.87	0.0088
Hayany	406488835	8789389	0.98	56.18	0.89	0.0099
Helwa	94299676	1007866	0.99	13.93	0.88	0.0081
Hilali	40604623	411689	0.99	6.26	0.85	0.0077
Hiri	34289222	381182	0.99	5.26	0.83	0.0085
Horra	156251978	2241102	0.99	22.67	0.88	0.0106
Jao	148401276	1943698	0.99	22.51	0.88	0.0094
Jihl	64114457	958979	0.99	10.08	0.87	0.0101
Kabkab	58092939	503205	0.99	8.37	0.84	0.0091
Kamla	63738218	938921	0.99	10.01	0.87	0.0109
Karbali	179479963	3777412	0.98	26.62	0.89	0.0089
Kashoowari	139019233	3777854	0.97	20.74	0.88	0.0091
Khadrawy	184213949	2256609	0.99	27.12	0.89	0.0087
Khalte	104310967	1594730	0.98	16.03	0.88	0.0106
Khastawi	163887349	2225925	0.99	23.89	0.88	0.009
Khenezi	179498132	2234006	0.99	25.55	0.88	0.0093
Khisab	172731619	1899587	0.99	29.68	0.89	0.0092
Kuproo	156202804	2893033	0.98	22.97	0.88	0.0097
Lulu	155227536	2007456	0.99	22.51	0.88	0.0091
Maktoumi	98832410	1513749	0.98	14.33	0.88	0.0095
Manjouma	78983294	937515	0.99	12.19	0.88	0.0079
Mazafati	250396814	3000638	0.99	36.83	0.89	0.009
Medjool	119139190	1508631	0.99	18.06	0.88	0.0103
Nagal	92313798	1282876	0.99	13.83	0.84	0.0111
Naquel_khuh	131153588	5536754	0.96	19.64	0.89	0.0085
Nebeit_seif	188301301	2086189	0.99	27.62	0.89	0.0113
Otaquin	63805023	1014321	0.98	9.88	0.87	0.0086

Piavom	105409353	1336422	0.99	15.91	0.88	0.0091
Rabee	83212377	1085128	0.99	12.74	0.88	0.0089
Raslatmar	72563689	2521136	0.97	11.36	0.88	0.0106
Reclinata [DP18]	81291572	1659108	0.98	12.34	0.84	0.0196
Rhars	162357815	2288204	0.99	23.96	0.89	0.0102
Rothan	150201486	2194572	0.99	22.44	0.88	0.0092
Saidi	142965921	1849339	0.99	20.92	0.88	0.0101
Samany	297430415	9150084	0.97	42.74	0.89	0.0097
Shagri	123694196	1550731	0.99	18.7	0.88	0.0091
Silani	110917413	1849289	0.98	16.76	0.89	0.0107
Sultana	163448822	2197739	0.99	23.82	0.88	0.0094
Sylvestris [P59]	178874185	2967238	0.98	26.05	0.86	0.0201
Sylvestris [RIV 2248 PL F]	168214899	2904230	0.98	23.68	0.86	0.0191
Sylvestris [RIV 2249 PL M]	190588868	3373815	0.98	26.72	0.86	0.0203
Sylvestris [RIV 2256 PL F]	183684615	3078551	0.98	26.88	0.86	0.0189
Sylvestris [RIV 7394 PI M]	177363543	3040518	0.98	25.65	0.86	0.0193
Sylvestris [RIV 7395 PL M]	187297438	3212398	0.98	27.03	0.86	0.0194
Sylvestris [SYL87 JCP 651]	178271416	2558327	0.99	26.16	0.88	0.0158
Tagiat	227371304	2853356	0.99	33.52	0.88	0.0108
Theophrasti [02a]	218870549	3589098	0.98	33.68	0.86	0.017
Theophrasti [05a]	196875873	3295053	0.98	29.43	0.85	0.0167
Theophrasti [A1]	133946661	2318399	0.98	21.12	0.85	0.0165
Theophrasti [A5]	162169468	2553814	0.98	24.81	0.84	0.0162
Theophrasti [B1]	186113518	3360918	0.98	28.87	0.85	0.0167
Theophrasti [B3]	187302115	3485241	0.98	29.18	0.85	0.0169
Theophrasti [B5]	186817680	3926283	0.98	28.76	0.85	0.017
Theophrasti [C1]	199183167	3350046	0.98	30.49	0.85	0.0165
Theophrasti [C4]	179182863	3167466	0.98	27.97	0.85	0.0164
Theophrasti [D1]	188890651	3174149	0.98	29.1	0.85	0.0165
Theophrasti [D3]	173475133	3077536	0.98	27.21	0.85	0.0167
Theophrasti [D5]	196887547	3343095	0.98	30.5	0.85	0.0166
Theophrasti [E1]	167492886	2891138	0.98	25.98	0.87	0.0158
Theophrasti [E2]	210170838	3560309	0.98	32.74	0.86	0.0158
Theophrasti [F1]	59381347	912539	0.98	9.53	0.82	0.0154

Theophrasti [F2]	56127625	894886	0.98	9.11	0.82	0.0154
Theophrasti[GOLK001 91020]	184608035	3379240	0.98	26.88	0.87	0.0183
Theophrasti [THE83 91051]	167090719	2964512	0.98	24.46	0.85	0.0193
Thory	148662341	2030735	0.99	21.68	0.88	0.0107
Um al blaliz	78240760	1189431	0.99	12.14	0.88	0.01
Um al hamam	110025473	1471829	0.99	16.63	0.88	0.0104
Zagloul	171154057	2003873	0.99	24.75	0.88	0.0107
Zahidi	293148570	8159624	0.97	41.82	0.89	0.0086

[a]internal identifiers are provided in brackets where applicable
[b]proportion of bases in reference genome covered by at least one read
[c]This colum contains the “PF_MISMATCH_RATE” output from Picard CollectAlignmentSummaryMetrics

Table S3. Population statistics in cultivated date palm and its wild relatives. Statistics were estimated in non-overlapping 5 kb intervals using sample short read alignments as input to ANGSD. See SI Materials and Methods for additional details.

Population	θ_W(mean $\pm$ sd)	θ_W (median)	π (mean $\pm$ sd)	π (median)	Tajima'sD (mean $\pm$ sd)	Tajima' D (median)
Middle East	0.0083 (0.0058)	0.0067	0.0084 (0.0062)	0.0069	0.0645 (1.0554)	0.0221
Egypt/Sudan	0.0095 (0.0067)	0.008	0.0098 (0.0072)	0.0081	0.1976 (1.3078)	0.2512
North Africa	0.0106 (0.0069)	0.0087	0.0115 (0.0074)	0.0098	0.4513 (0.9883)	0.4273
P. sylvestris	0.0094 (0.0088)	0.0064	0.0105 (0.0104)	0.0069	0.5221 (1.3851)	0.6663
P. theophrasti	0.0053 (0.0074)	0.002	0.0072 (0.0103)	0.0025	1.111 (1.2604)	1.1877
P. canariensis	0.0116 (0.0114)	0.0077	0.0117 (0.0134)	0.0064	-0.5062 (1.4786)	-0.3652

θ_W and π are per site estimates

Table S4. STRUCTURE analysis of *Phoenix* species with independent allele frequencies. Analyses were conducted with MCMC with burn in of 200,000 steps and chain lengths of 1,000,000 steps for K = 1-5 and 750,000 steps for K = 6-8.

Species[a]	K	Reps	MeanLnP(K)	Stdev	LnP(K)	Ln'(K)	ΔK
d,t,a,s,c	1	5	-1936012.66	11.7343	NA	NA	NA
d,t,a,s,c	2	5	-1432066.12	10.1593	503946.54	309836.02	30497.68037
d,t,a,s,c	3	5	-1237955.6	80.9156	194110.52	177584.12	2194.682398
d,t,a,s,c	4	5	-1221429.2	37475.5014	16526.4	50676.32	1.352252
d,t,a,s,c	5	5	-1154226.48	115.9024	67202.72	67206.14	579.85092
d,t,a,s,c	6	5	-1154229.9	75.6712	-3.42	133.04	1.758133
d,t,a,s,c	7	5	-1154366.36	117.9535	-136.46	236.5	2.005028
d,t,a,s,c	8	5	-1154266.32	233.6165	100.04	NA	NA

[a]*Phoenix* species included in analysis (c=*canariensis*, t=*theophrasti*, s=*sylvestris*, a=*atlantica*, d=*dactylifera*)

Table S5. STRUCTURE analysis of *Phoenix* species results with correlated allele frequencies. Analyses were conducted with MCMC with burn in of 200,000 steps and chain lengths of 1,000,000 steps for K = 1-5 and 750,000 steps for K = 6-8.

Species[a]	K	Reps	MeanLnP(K)	Stdev	LnP(K)	Ln'(K)	ΔK
d,t,a,s,c	1	5	-1935321.24	7.735179	NA	NA	NA
d,t,a,s,c	2	5	-1402077.66	11.140153	533243.58	300385.88	26964.25181
d,t,a,s,c	3	5	-1169219.96	38.385974	232857.7	165484.52	4311.06737
d,t,a,s,c	4	5	-1101846.78	48572.62954	67373.18	11398.6	0.234671
d,t,a,s,c	5	5	-1023075	17227.20709	78771.78	41443506.4	2405.700831
d,t,a,s,c	6	5	-42387809.62	57781791.42	-41364734.62	78717412.96	1.362322
d,t,a,s,c	7	5	-5035131.28	7645389.286	37352678.34	68929681.62	9.01585
d,t,a,s,c	8	5	-36612134.56	73687383.83	-31577003.28	NA	NA

[a]*Phoenix* species included in analysis (c=*canariensis*, t=*theophrasti*, s=*sylvestris*, a=*atlantica*, d=*dactylifera*)

Table S6. Hierarchical STRUCTURE analysis with *Phoenix* species pairs. All analyses were conducted with MCMC chain lengths of 1,000,000 steps with burn in of 200,000.

Species[a]	Model[b]	K	Reps	MeanLnP(K)	Stdev	LnP(K)	Ln'(K)	ΔK
d,t	correlated	1	5	-1382496.34	6.0665	NA	NA	NA
d,t	correlated	2	5	-898512.58	5.7456	483983.76	447617.92	77906.11482
d,t	correlated	3	5	-862146.74	81.8886	36365.84	33791.36	412.650351
d,t	correlated	4	5	-859572.26	181.062	2574.48	4707.16	25.997502
d,t	correlated	5	5	-852290.62	192.5436	7281.64	79870.02	414.815303
d,t	correlated	6	5	-924879	179598.4296	-72588.38	NA	NA
d,t	independent	1	5	-1382719.96	9.5981	NA	NA	NA
d,t	independent	2	5	-924093.16	10.8992	458626.8	439922.06	40362.70978
d,t	independent	3	5	-905388.42	138.192	18704.74	18520.52	134.020158
d,t	independent	4	5	-905204.2	170.7	184.22	65.88	0.38594
d,t	independent	5	5	-905085.86	173.3255	118.34	72.86	0.420365
d,t	independent	6	5	-905040.38	270.0099	45.48	NA	NA
d, s	correlated	1	5	-1072630.06	14.3055	NA	NA	NA
d, s	correlated	2	5	-910251.5	19.7673	162378.56	100871.78	5102.969467
d, s	correlated	3	5	-848744.72	8.697	61506.78	58995.94	6783.517938
d, s	correlated	4	5	-846233.88	319.7293	2510.84	NA	NA
d, s	independent	1	5	-1073323.54	9.0768	NA	NA	NA
d, s	independent	2	5	-933198.8	19.2005	140124.74	110636.1	5762.140569
d, s	independent	3	5	-903710.16	96.3141	29488.64	29482.62	306.108957

d, s	independent	4	5	-903704.14	130.8617	6.02	NA	NA
d, c	correlated	1	5	-1145690.36	4.8911	NA	NA	NA
d, c	correlated	2	5	-887905.46	11.5641	257784.9	199014.66	17209.72356
d, c	correlated	3	5	-829135.22	97.8919	58770.24	274468.48	2803.790371
d, c	correlated	4	5	-1044833.46	308448.6054	-215698.24	NA	NA
d, c	independent	1	5	-1146389.82	11.6106	NA	NA	NA
d, c	independent	2	5	-914142.86	24.4113	232246.96	208875.42	8556.495106
d, c	independent	3	5	-890771.32	86.9131	23371.54	23480.9	270.165256
d, c	independent	4	5	-890880.68	164.8779	-109.36	NA	NA

[a]*Phoenix* species included in analysis (c=*canariensis*, t=*theophrasti*, s=*sylvestris*, d=*dactylifera*)
[b]allele frequency model

Table S7. Summary of D-tests of admixture in date palm and wild *Phoenix* species. Analysis is based on 800 kb or larger scaffolds. Population abbreviations are Rec=*P. reclinata*, Can=*P. canariensis*, The=*P. theophrasti*, Syl=*P. sylvestris*, Me= *P. dactylifera* (Middle East), Af= *P. dactylifera* (North Africa)

D(P1,P2,P3,O)[a]	D(A,B;X,Y)[b]	*D*	SE	Z-score	Sites	Blocks
D(Me,Af,Can,Rec)	D(Rec,Can;Me,Af)	0.0496	0.0076	6.5106	5201076	114
D(Me,Af,The,Rec)	D(Rec,The;Me,Af)	0.5795	0.0155	37.2416	5212567	114
D(Me,Af,Syl,Rec)	D(Rec,Syl;Me,Af)	-0.1968	0.0107	-18.2776	5200071	114
D(Me,Af,The,Can)	D(Can,The;Me,Af)	0.5750	0.0157	36.6246	5417214	114
D(Me,Af,Syl,Can)	D(Can,Syl;Me,Af)	-0.2245	0.0109	-20.5572	5401578	114
D(Me,Eg,Can,Rec)	D(Rec,Can;Me,Eg)	0.0229	0.0073	3.1205	5201001	114
D(Me,Eg,The,Rec)	D(Rec,The;Me,Eg)	0.3162	0.0274	11.5174	5212492	114
D(Me,Eg,Syl,Rec)	D(Rec,Syl;Me,Eg)	-0.0713	0.0105	-6.7544	5199996	114
D(Me,Eg,The,Can)	D(Can,The;Me,Eg)	0.3142	0.0275	11.4011	5417133	114
D(Me,Eg,Syl,Can)	D(Can,Syl;Me,Eg)	-0.0853	0.0115	-7.3725	5401497	114

[a]notation used in this manuscript (see main text and Supplementary Materials and Methods)
[b]notation adopted in Popstats

Table S8. Summary of f_3 tests of admixture in *Phoenix*.

Reference[a]	Reference	Test	f_3	SE	Z-score	Sites	Blocks
Can	The	Af	0.3940	0.0198	19.8481341554	5417214	114
Can	Syl	Af	0.3278	0.0106	30.8714085537	5401578	114
Can	Me	Af	-0.0068	0.0044	-1.53453598213	5426623	114
The	Syl	Af	0.1461	0.0144	10.1319179079	5420077	114
The	Me	Af	-0.1547	0.0078	-19.8010654255	5446954	114
Syl	Me	Af	0.0470	0.0047	9.98759435075	5429487	114
Can	The	Me	1.0501	0.0203	51.6515596788	5417215	114
Can	Syl	Me	0.6612	0.0144	45.8828550979	5401579	114
Can	Af	Me	0.2551	0.0121	20.9516945251	5426623	114
The	Syl	Me	0.6124	0.0142	43.0248181474	5420078	114
The	Af	Me	0.4693	0.0231	20.2505322114	5446954	114
Syl	Af	Me	0.1769	0.0084	20.9796711917	5429487	114
Can	The	Syl	1.4643	0.0510	28.7000406841	5392170	114
Can	Af	Syl	1.0216	0.0394	25.8832571533	5401578	114
Can	Me	Syl	0.9254	0.0367	25.1703221768	5401579	114
The	Af	Syl	1.3476	0.0503	26.7814032312	5420077	114
The	Me	Syl	0.9867	0.0371	26.5743008083	5420078	114
Af	Me	Syl	1.5257	0.0499	30.5289882338	5429487	114
Can	Syl	The	7.0383	0.3977	17.6945835456	5392170	114
Can	Af	The	5.6826	0.3282	17.3099792149	5417214	114
Can	Me	The	6.7754	0.3838	17.6509054061	5417215	114
Syl	Af	The	7.5278	0.4243	17.7397313993	5420077	114
Syl	Me	The	9.0215	0.4977	18.124694517	5420078	114
Af	Me	The	9.7511	0.5400	18.0561612535	5446954	114
The	Syl	Can	2.1101	0.1213	17.3814614312	5392170	114
The	Af	Can	2.5294	0.1383	18.2816038789	5417214	114
The	Me	Can	2.1915	0.1236	17.719634529	5417215	114
Syl	Af	Can	2.6749	0.1489	17.9615447772	5401578	114
Syl	Me	Can	2.7977	0.1537	18.1998852504	5401579	114

Af	Me	Can	3.4463	0.1817	18.9582892881	5426623	114
Af	Me	Eg	0.0111	0.0046	2.40519819898	5456282	114

[a]Species/population abbreviations (Can = *P. canariensis*, The = *P. theophrasti*, Syl = *P. sylvestris*, Af = North Africa, Me = Middle East)

Table S9. Summary of admixture modeling with *TreeMix*.

Taxa[a]	root	m	block size (SNPs)	edge	mixture weight[b]	% variance explained
r,t,s,af,me	r	0	1200	-	-	0.9865042
r,t,s,af,me	r	1	1200	t->af	0.154444	0.9996024
r,t,s,af,me	r	2	1200	t->af	0.159249	
				r->s	0.0409515	1
r,t,s,af,me	r	0	2500	-	-	0.9865042
r,t,s,af,me	r	1	2500	t->af	0.15467	0.9996024
r,t,s,af,me	r	2	2500	t->af	0.159208	
				s->r	0.0662112	1
r,c,t,s,af,me	r	0	1500	-	-	0.9858358
r,c,t,s,af,me	r	1	1500	t->af	0.157475	0.9988405
r,c,t,s,af,me	r	2	1500	t->af	0.154713	
				s->r	0.139744	0.9995955
r,c,t,s,af,me	r	0	3000	-	-	0.9858358
r,c,t,s,af,me	r	1	3000	t->af	0.157475	
r,c,t,s,af,me	r	2	3000	t->af	0.154713	
				s->r	0.138765	0.9995959

[a]Taxa included in model. Labels are r = *P. reclinata*, t = *P. theophrasti*, s = *P. sylvestris*, af = North Africa, me = Middle East
[b]Mixture weights for m = 2 are the weight of the two migration edges.

Table S10. Ancestry estimates by the f_4-ratio approach. Population abbreviations are Rec=*P. reclinata*, Can=*P. canariensis*, The=*P. theophrasti*, Syl=*P. sylvestris*, Me= *P. dactylifera* (Middle East), Af= *P. dactylifera* (North Africa)

Ratio	α*	SE	Sites	Blocks
f_4(Syl,Can;Af,The)/f_4(Syl,Can;Me,The)	0.8212	0.0101	5392169	114
f_4(Syl,Rec;Af,The)/f_4(Syl,Rec;Me,The)	0.8198	0.0105	5191207	114
f_4(Syl,Can;Egypt,The)/f_4(Syl,Can;Me,The)	0.9474	0.0074	5392088	114
f_4(Syl,Rec;Egypt,The)/f_4(Syl,Rec;Me,The)	0.9489	0.0077	5191132	114

*estimates are based on scaffolds 800 kb or longer

Table S11. Summary of private alleles in date palm and its wild relatives. See SI Materials and Methods for definitions of private SNPs and private Fixations.

Population	Private SNPs	SNPs	% Private	Private Fixations
Middle East (n=35)	738739	5493748	13.446903643924	0
North Africa (n=20)	390690	6844071	5.708444579257	0
P. canariensis (n=6)	1649668	3502073	47.1054715307191	327352
P. dactylifera (n=55)	3362180	7891182	42.606798322482	3842
P. reclinata (n=1)	371400	622028	59.7079231160012	877042
P. sylvestris (n=6)	1330458	3083543	43.1470551894363	156064
P. theophrasti (n=13)	162935	1024500	15.9038555392875	72760

References

1. Stevens, CJ, Murphy, C, Roberts, R, Lucas, L, Silva, F., and Fuller, DQ (2016) Between China and South Asia: A Middle Asian corridor of crop dispersal and agricultural innovation in the Bronze Age. *The Holocene* 26(10):1541-1555.
2. De Vartavan, C. and Asensi Amoros, V. (1997) *Codex of Ancient Egyptian Plant Remains.* Triade exploration, London
3. Brewer, D. J., D. B. Redford, S. Redford (1994) Domestic plants and animals. The Egyptian Origins. Warminster: Aris and Phillips
4. M.A. Murray, M. A. (2000) Fruits, vegetables, pulses and condiments. In: P.T. Nicholson, I. Shaw (Eds.), *Ancient Egyptian Materials and Technology*, Cambridge University Press, Cambridge, pp. 609-655
5. Miller, Naomi F., Philip Jones, and Holly Pittman (2016) Sign and image: Representations of plants on the Warka Vase of early Mesopotamia. *Origini* 39: 53-73
6. Fuller, Dorian Q and Weber, Steven A. (2005) Formation Processes and Paleoethnobotanical Interpretation in South Asia. *Journal of Interdisciplinary Studies in History and Archaeology*2(1): 91-114
7. Hazzouri KM, et al. (2015) Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. 2015. *Nat Commun* 6:8824.
8. Henderson et al. (2006) Genetic isolation of Cape Verde Island *Phoenix atlantica* (Arecaceae) revealed by microsatellite markers. *Conserv Genet* 7(2):213-223.
9. Al-Mssallem IS, et al. 2013. Genome sequence of the date palm *Phoenix dactylifera* L. *Nat Commun* 4:2274.
10. Fang Y, et al. (2012) A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera L.*) mitochondrial genome. *PLoS ONE* 7(5): e37164.
11. Yang M, et al. (2010) The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS ONE* 5(9): e12762
12. Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120.
13. Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997v1 [q-bio.GN].
14. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.
15. Li H, et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078-9.
16. Poplin R, et al. (2017) Scaling accurate genetic variant discovery to tens of thousands of sample. *bioRxiv* doi:10.1101/201178.
17. Ramu P, et al. (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat Genet* 49:959-963.
18. Sabir JSM, 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. Fang Y, Wu H, Zhang T, Yang M, Yin Y, Pan L, et al. (2012) A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera L.*) mitochondrial genome. *PLoS ONE* 7(5): e37164. https://doi.org/10.1371/journal.pone.0037164
20. Yang M, et al. (2010) The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS ONE*. 9:e12762.
21. Van der Auwera et al. (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*: John Wiley & Sons, Inc.
22. Stamatakis, A. (2014) Raxml version 8: a tool for phylogenetic analysis and post-analysis

of large phylogenies. *Bioinformatics, 30*(9), 1312–1313.
23. Danecek P, et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27(15):2156-2158.
24. Pritchard JK, Stephens M, Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 55:945-959.
25. Earl, Dent A. and vonHoldt, BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet Resources* 4(2):359-361.
26. Green RE, Krause J, Briggs AW, et al. (2010) A draft sequence of the Neandertal genome. *Science* (56 co-authors). 328(5979):710-722.
27. Durand EY, Patterson N, Reich D, Slatkin M. (2011) Testing for ancient admixture between closely related populations. *Mol Biol Evol* 28(8):2239–2252.
28. Reich D, Thangaraj K, Patterson N, Price Al, Singh L (2009) Reconstructing Indian population history. *Nature* 461:489-494.
29. Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192:1065-1093.
30. Peter BM (2016) Admixture, population structure, and F-statistics. *Genetics* 202:1485-1501.
31. Pickrell JK, Pritchard JK. (2012) Inference of populations splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11): e1002967.
32. Martin SH, Davey JW, Jiggins CD (2015) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol* 32(1):244-257.
33. Browning SR, Browning BL. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Human Genet* 81(5):1084–1097
34. Marroni F, et al. (2011) Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genet Genomes* 7:1011-1023.
35. Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol* 33:54-78.