Plants, People, Planet Supporting Information

Article title: What lies behind a fruit crop variety name? A case study of the barnī date palm from al-‘Ulā oasis, Saudi Arabia

Authors: Muriel Gros-Balthazard, Vincent Battesti, Jonathan M. Flowers, Sylvie Ferrand, Matthieu Breil, Sarah Ivorra, Jean-Frédéric Terral, Michael Purugganan, Rod A. Wing, Nahed Mohammed, Yann Bourgeois.

The following Supporting Information is available for this article:

Fig. S1 Pictures of barnī date palms.

Fig. S2 Pictures of the date palms groups of clones used to calibrate the relatedness analyses.

Fig. S3 Consensus tree of 35 date palms based on genetic distance calculated using genotype likelihoods across 10,742 sites.

Fig. S4 Relationship between the fraction of sites covered by one or more reads in both sequenced genomes and the King-robust kinship estimator.

Fig. S5 Ancestry coefficients inferred in 88 date palms using a randomly downsampled set of 28,406 Single Nucleotide Polymorphisms (SNPs).

Fig. S6 Cross-validation error of the admixture model.

Fig. S7 Chloroplast DNA tree with uncorrected distances reconstructed using the Neighbor-joining method.

Dataset S1 List of Phoenix spp. accessions, their geographic origin and source.

Dataset S2 Sequencing and read alignment metrics.

Dataset S3 Kinship among 35 date palms assessed using 1,515,793 sites across the genome.

Dataset S4 Coordinates in the Principal Component Analysis of 88 date palms including three from al-‘Ulā.

Dataset S5 Summary of D-tests of admixture in date palm and Phoenix theophrasti.
**Dataset S6** Summary of $f_d$-statistics for testing of admixture in date palm and wild *Phoenix* species.

**Dataset S7** Summary of $f_d$-ratio statistics for inferring the proportion of *Phoenix theophrasti* ancestry in the North African, ancient Judean and ḏakar (al-'Ulā) date palm genomes.

**Methods S1** Detailed protocol for sequencing read processing, genome alignment, variant calling and Single Nucleotide Polymorphisms (SNPs) filtering.

**Methods S2** Admixture tests and ancestry proportions.

**Notes S1** Date palms growing situations in al-‘Ulā region.

**Notes S2** Categorizing the production of the barnī date palm.

**Notes S3** On the relationship between coverage and the King-robust kinship estimator.

**References present in supporting information**
Fig. S1 Pictures of barnī date palms. (a) A young barnī date palm, with two growing offshoots still attached to its base (other offshoots have had their leaves cut off). To their left are two offshoots that are in their first year of planting (wrapped with dry palms to protect these young plants). Photo taken at a Qaraher farm, north of al-‘Ula. Nov. 11th 2019; (b) The date palm sampled as barnī_00024, photographed in low angle, showing the arrangement of its leaves.
around its terminal bud and its leaflets on its palms. In the palm grove of al-Oziyāt, north of al-‘Ulā. Nov. 10th 2019; (c) A barnī date palm in a new 21st c. palm grove, northeast of al-‘Ulā, displaying young inflorescences (future date bunches) some of which are still attached by a link: the farmer has slipped into the inflorescences male spikelets still visible (top left) for pollination. April 6th, 2019. Pictures: Vincent Battesti.
Fig. S2 Pictures of the date palms groups of clones used to calibrate the relatedness analyses. (a) The mother plant barnī 00036, sampled twice (barnī_00036 and barnī_00036A), and its offshoots (00036r1 and 00036r2), (b) The male ḏakar (ḏakar_00254) and its two sampled offshoots (00169r1 and 00169r2) (c) The offshoots of ‘aselā, of which two were sampled
(00169r1 and r2) in front of the mother palm (‘aselā_00169A). Pictures: Vincent Battesti.
**Fig. S3** Consensus tree of 35 date palms based on genetic distance calculated using genotype likelihoods across 10,742 sites. It was obtained by computing a consensus tree where only clades found in 90% of the 100 bootstrap replicates are displayed. The non-consensus tree where genetic distances are visible can be found in Fig. 4c.
**Fig. S4** Relationship between the fraction of sites covered by one or more reads in both sequenced genomes and the King-robust kinship estimator. (a) Kinship between the sample barni_00010 and the other 22 barnī; b) Kinship between barnī_Oman and the other 22 barnī; c) Kinship among the 22 barnī accessions. Correlations (adjusted $R^2$) were calculated using a linear mixed model and asterisks indicate p value < 0.01.
Fig. S5 Ancestry coefficients inferred in 88 date palms using a randomly downsampled set of 28,406 Single Nucleotide Polymorphisms (SNPs). Cross-validation error plot may be found in Fig. S6.
**Fig. S6** Cross-validation error of the admixture model. The corresponding ancestry plot may be found in Fig. 5a.
Fig. S7 Chloroplast DNA tree with uncorrected distances reconstructed using the Neighbor-
joining method. Support values represent the percent of replicates supporting each node from 100 bootstrap replicates. The tree was rooted with *Phoenix reclinata* (PREC1). All three date palms from al-‘Ulā (in black) are found within a clade comprising the West Asian date palms (in blue) and bear the so-called oriental chlorotype. North African date palms are written in yellow, ancient Judean date palms in pink, *Phoenix theophrasti* in light blue, and *Phoenix sylvestris* in grey.
**Methods S1** Detailed protocol for sequencing read processing, genome alignment, variant calling and Single Nucleotide Polymorphisms (SNPs) filtering.

**Read processing and genome alignment.** Reads were demultiplexed and those passing Illumina quality control filters were processed with Trimmomatic (Bolger et al., 2014) v. 0.39 to remove contaminating adapter sequences. For adapter removal, we used the adapter and Nextera transposase sequence database included with Trimmomatic (v. 0.39) downloaded 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.

We aligned reads to the reference genome from Hazzouri et al., 2019. This genome is a PacBio/Illumina genome assembly from a male date palm from a fourth-generation backcross (BC4) with a female of the Barhee cultivar. Its size is ~772 Mb and it contains 18 primary contigs (49.9% of total length) along with 2,371 unplaced contigs. The nuclear and chloroplast genomes were combined to form a single modified reference sequence that was used for variant calling. Site filtering was nonetheless carried out separately.

Processed reads were aligned to this unmasked genome using bwa mem (Li, 2013) 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 SamSort from Picard-tools (https://github.com/broadinstitute/picard) v. 2.8.2 to coordinate-sort the alignments. We used MarkDuplicates (Picard-tools) to flag duplicate read pairs. Processed alignments were summarized with samtools v. 1.9 (Li et al., 2009) using the stats option.

**Variant calling and Single Nucleotide Polymorphisms (SNPs) filtering.** SNP-calling and genotyping was performed with the Genome Analysis Toolkit (GATK) v. 4.2.0 HaplotypeCaller (McKenna et al., 2010) run in Genomic Variant Call Format (GVCF) mode followed by joint-genotyping with GenotypeGVCFs. We used the option -all-sites to include not only variant sites but also loci found to be non-variant after genotyping.

The sites were filtered using GATK v4.2.0, Picard-tools v.2.23.8, bcftools v.1.14 (Li, 2011) and vcf tools v.0.1.16 (Danecek et al., 2011). For the nuclear genomes, for both variant and invariant sites, we restricted analysis to the non-repetitive fraction of the genome assembly by excluding...
SNPs in repetitive regions identified during the annotation of the genome assembly and to the regions that are not associated to sex determination (Gros-Balthazard et al., 2021; Hazzouri et al., 2019). We removed indels, excluded multi-allelic SNPs and sites having an average depth below 10. We set as missing genotypes showing a depth below 10 or above 50. We also excluded SNPs meeting the following conditions: within 6bp of indel polymorphisms, strand bias estimated using Fisher’s exact test (FS) > 60.0, strand bias estimated by the symmetric odds ratio test (SOR) > 3.0, quality by depth (QD) < 8.0, mapping quality (MQ) < 40.0, mapping quality rank sum (MQRankSum) < -3.0, read position rank sum (ReadPosRankSum) < -1.5, base quality rank sum (BaseQRankSum) < -8.0, quality (QUAL) < 30. We filtered out genotypes having genotype quality (GQ) < 20.0 and SNPs with a genotype call rate < 80%. This procedure yielded a filtered site set of 105,236,083 bases including 1,007,281 SNPs that served as the basis for subsequent analyses.

**Reconstruction of a chloroplast DNA tree**

We called Single Nucleotide Polymorphisms (SNPs) and genotypes for the chloroplast genome (cpDNA) (Hazzouri et al. 2019) using the same workflow as the nuclear genome, but applied a unique set of filters to obtain the final call set. The large inverted repeat (IR) regions characteristic of the date palm cpDNA (Yang et al. 2010) were identified by BLAST and excluded from further consideration. A separate BLAST analysis was conducted to identify regions of the cpDNA that share a high degree of similarity with the mitochondrial genome (mtDNA) (Fang et al. 2012). Any region of the cpDNA that BLAST to the mtDNA (Fang et al. 2012) with significance of 1e-6 or less was excluded from the analysis. Hard filters were then applied such that SNPs meeting any of following conditions were removed quality by depth (QD) < 2.0, strand bias estimated using Fisher’s exact test (FS) > 60.0, mapping quality (MQ) < 40.0, mapping quality rank sum (MQRankSum) < 12.5, ReadPosRankSum < -8.0 (see GATK website for tag definitions). SNPs within 10 bp of an indel in the raw (unfiltered) indel call set were also removed. Finally, we excluded any site with missing data or with one or more heterozygote genotype calls as they may represent instances of heteroplasmy or genotyping artifacts. Remaining homozygous genotypes were converted to haploid to produce a sequence alignment with 52 SNPs. This alignment was used to construct a cpDNA tree with uncorrected distances using the Neighbor-
joining method with the phangorn (v. 2.8.1) package in R.
Methods S2 We tested for the presence and extent of admixture between date palms and *Phoenix theophrasti* using the R package *admixr* (Petr et al., 2019) v.0.9.1, which provides an implementation of ADMIXTOOLS (Patterson et al., 2012). We used a script available online to generate the input in EIGENSTRAT format from our VCF (https://github.com/joanam/scripts/blob/master/convertVCFtoEigenstrat.sh). We ran the tool using the full Single Nucleotide Polymorphisms (SNPs) set.

We focused on determining whether al-‘Ulā date palms show evidence of gene flow with *P. theophrasti* as evidenced in North African cultivars, a few West Asian date palms, and most of the ancient Judean date palms (Flowers et al., 2019; Gros-Balthazard et al., 2021). We use the same pattern of test, fitting the phylogenetic relationships among *Phoenix*, as in Gros-Balthazard et al., 2021.

We calculated Patterson’s *D* also known as the ABBA-BABA test (Green et al., 2010; Patterson et al., 2012). The implementation of the test in *admixr* requires a rooted and asymmetric four population tree in the form (((W, X), Y), Z), where Z is an outgroup to a clade formed by W, X and Y, and Y is an outgroup to a clade formed by W and X. The test is used to evaluate if the data is inconsistent with the null hypothesis that the tree is correct and that there is no gene flow between Y and either W or X. It is based on comparing the proportions of BABA and ABBA sites patterns observed in the data. *D* is calculated as follow: $D = (n_{BABA} - n_{ABBA})/(n_{BABA} + n_{ABBA})$. We also calculated the $f_4$-statistic, which is very similar to *D* except that the denominator is the total number of sites (Patterson et al., 2012). Gene flow between populations W and Y leads to an increase of shared alleles between populations resulting in an elevated number of BABA sites and thus a positive *D*-statistic and a positive $f_4$-statistic. Here, we tested for gene flow between North African, ancient Judean and al-‘Ulā date palms individually (test sample) and *P. theophrasti*. We conducted the tests separately including *P. reclinata* as the outgroup, following a tree reconstruction (Gros-Balthazard et al., 2021). We therefore estimated the *D*-statistics using the following tree: (((test sample, West Asian date palms), *P. theophrasti*), *P. reclinata*). We used all samples together for each population/species and after removing admixed samples, as identified in Gros-Balthazard et al. 2021 (PDAC64, PTHE2, PTHE3, PTHE4, PTHE26, PTHE27, PSYL7 and PSYL44). Significance was tested with block
jackknife with each linkage groups and unplaced contigs treated as a separate block. The standard error of the statistics was used to calculate a standard score (Z-score = statistics / standard error), and we used an absolute Z-score >2 to assess statistically significant results. Significance was tested with block jackknife with each linkage groups and unplaced contigs treated as a separate block. The standard error of the statistics was used to calculate a standard score (Z-score = statistics / standard error), and we used an absolute Z-score >2 to assess statistically significant results.

To infer the proportion of ancestry in the date palms displaying evidence of admixture based on the $D$ and $f_4$-statistic, we calculated the $f_4$-ratio statistics, as described in Patterson et al. 2012. More precisely, we defined the following tree $(((A,B), C), O)$. The $f_4$-ratio is $f_4(A,O;X,C)/f_4(A,O;B,C)$, where $X$ is the test sample, $A$ is a sister species, namely $P. sylvestris$, $B$ and $C$ are the mixing populations, namely West Asian date palms and $P. theophrasti$, respectively, and $O$ is the outgroup, that is $P. reclinata$. We used all samples together for each population/species and after removing admixed samples (see above; PDAC64, PTHE2, PTHE3, PTHE4, PTHE26, PTHE27, PSYL7 and PSYL44). The resulting alpha value corresponds to the ancestry proportion of $B$ in $X$, and by extension $1$-alpha provides the proportion of $C$ in $X$, therefore the proportion of $P. theophrasti$ in the test sample $X$. Negative values are uninformative. Significance was tested with block jackknife with each linkage groups and unplaced contigs treated as a separate block. The standard error of the statistics was used to calculate a standard score (Z-score = statistics / standard error), and we used an absolute Z-score >2 to assess statistically significant results.
Notes S1 Date palms growing situations in al-‘Ulā region.

We consider four distinct date palms growing situations in al-‘Ulā region (Battesti & Marty, in prep.), of which a quick summary is given below:

- The old oasis palm grove of al-‘Ulā (the former “core area” according to the terminology of the Royal Commission for AlUla (RCU) administration, and now the palm grove of the so-called “Cultural Oasis District” according to the new terminology of the RCU administration): it is the historical center of oasian production of dates and agrobiodiversity of date palm (dozens of varieties) in the region. Date palms were cultivated by sedentary (ḥaḍāraī) social groups (mainly al-‘Alawī) in palm grove small enclosed gardens (basāṭīn). They do still exist and produce, but a part of them is in bad shape and the other part reorganized. The classic form of farming that is carried out there is layered polyculture with an elaborate soil design work.

- The old Bedouin palm groves: it is the historical way for Bedouin social groups, with a pastoral way of life, to produce a (small) part of their food needs in dates (other historical means are the submission of oasis people, purchase/exchange and predation). These small palm groves are located in the pastoral tribal lands outside of the area of the sedentary oases, scattered in desert wadis. They have a range of labor investment levels, from nearly none in the picking palm groves (few date palms cultivated with neither fruit trees nor annual crops growing beneath) to small gardens around a spring with some fruit trees. The date palm agrobiodiversity is, moreover, restricted to the two local elite varieties, barnī and ḥelwa (ḥamrā’). These Bedouin palm groves do still exist.

- The modern 20th century palm groves: these palm groves were the first to be established outside the historical area of the old al-‘Ulā palm grove, but in its immediate (north and south) vicinity. These palm groves were mostly created by sedentary oasis people, or by Bedouin groups already settled in al-‘Ulā. The gardens or farms (mazāra‘) no longer have organic forms, but are linear properties, with date palms lined up planted at the same time, thus having the same age and height. Besides a few fruit trees, date monoculture is the rule, and these farms depend heavily on the only variety for export, the barnī. The purpose is clearly no longer local food but export income.
The new 21st century palm groves: these palm groves are the most recent and have been established well beyond the first perimeter of old and modern palm groves. The spirit that governs the exploitation of these farms (mazāra) is generally that of agricultural entrepreneurship, particularly Bedouin, which does not negate forms of small varietal collections of date palms (especially exotic, sometimes also local especially when the owner is ‘Alawī) and a crop diversification into citrus fruit or moringa, for instance.
Categorizing the production of the barnī date palm.

In al-‘Ulā, the barnī variety is the only one that really benefits from such precision of quality categories (mabrūm for the best dates, mašrūk for the second grade, and ‘ādī for the remains), which some still refine, with a #1bis quality, the “(mabrūm) asfar” (which look like mabrūm, but whose seeds are small and poorly formed), or more recently the mabrūm jumbo (for their superior dimensions), term coined by wholesalers and commercial largescale farmers. The other date varieties, which we currently estimate at n≈99 in al-‘Ulā (work in progress), are locally valued for their differences, but their respective productions are just sorted into good and bad dates with no specific names. Some local farmers logically argue that this refinement in the categorization of barnī qualities is simply a result of the market.

The data regarding the relative proportions of the different qualities of dates on a barnī palm fluctuates widely, depending on the type of palm grove and farm. This example can be considered correctly in the average: a producer in al-Oziyāt (northern part of al-‘Ulā) gave the year 2018 as bad with 40% mabrūm and 60% mašrūk and ‘ādī, while he deemed the year 2019 as good with an inverse ratio of 60% and 40%.

The commitment to a high production of mabrūm is easily explained from a cash crop perspective: the market values are very different. For example (in 2019), mabrūm was priced at 24 riyals/kg, mašrūk at 10 riyals/kg, and ‘ādī was left for 2.5 riyals/kg. For years 2020 and 2021, prices were greatly impacted downward with the Covid-19 pandemic: less export, because fewer foreign wholesale buyers came to buy (visa and vaccination certificate problem) and absence of the millions of pilgrims (the pilgrimage was suspended by the kingdom) who usually buy in the holy city of Medina, before their return home, large quantities of “the date of al-Madīna al-munawara”, which is for some part the barnī from al-‘Ulā). As a matter of fact, if our focus area is expanded to the entire province of Medina (Saudi Arabia), there seem to be three varieties of barnī (Aleid et al., 2015): from Medina, from al-‘Aīṣ—as we mentioned, putatively barniyat al-‘aīṣ cultivated also in al-‘Ulā—, and from al-‘Ulā. Whether they are homonyms or genetically the same variety cannot be certified at this point and will be assessed in the second phase of our Al-‘Ulā DPA (Date Palm Agrobiodiversity) Project (al-‘Ulā DPA).
Notes S3 On the relationship between coverage and the King-robust kinship estimator.

We investigated why the King-robust kinship estimator is lower within the barnī accessions we believe to be clones compared to that found within each quartet/triplets of known clones. The latter being sequenced at a higher coverage, we hypothesized that it could affect the estimator, although we used genotype likelihoods rather than calls and a relatedness estimator presumably robust to Single Nucleotide Polymorphisms (SNPs) ascertainment biases linked to low coverage (Waples et al., 2019). Based on the estimated kinships (Fig. 4ab; Dataset S3), we found three types of relationships among the barnī samples: 1) barnī_00010 is unrelated to other barnī, 2) barnī_Oman is unrelated to other barnī, and 3) all the other barnī are clones. We therefore visualized the relationship between the pairwise coverage (fraction of sites covered by 1 or more reads in both sequenced genomes used for the calculation) and the resulting King-robust kinship estimator. We calculated the correlations using a linear mixed model (lm function in stats R package) and indeed found that pairwise coverage affects the estimator in the three cases (adjusted $R^2$ ranged from 0.851 to 0.937, p-values < 0.01).
References present in supporting information


