nature biotechnology

De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*)

Eman K Al-Dous[1,10], Binu George[1,10], Maryam E Al-Mahmoud[1], Moneera Y Al-Jaber[1], Hao Wang[2], Yasmeen M Salameh[1], Eman K Al-Azwani[1], Srinivasa Chaluvadi[2], Ana C Pontaroli[2,9], Jeremy DeBarry[2], Vincent Arondel[3], John Ohlrogge[4], Imad J Saie[5], Khaled M Suliman-Elmeer[6], Jeffrey L Bennetzen[2], Robert R Kruegger[7] & Joel A Malek[1,8]

Date palm is one of the most economically important woody crops cultivated in the Middle East and North Africa and is a good candidate for improving agricultural yields in arid environments. Nonetheless, long generation times (5–8 years) and dioecy (separate male and female trees) have complicated its cultivation and genetic analysis. To address these issues, we assembled a draft genome for a Khalas variety female date palm, the first publicly available resource of its type for a member of the order Arecales. The ~380 Mb sequence, spanning mainly gene-rich regions, includes >25,000 gene models and is predicted to cover ~90% of genes and ~60% of the genome. Sequencing of eight other cultivars, including females of the Deglet Noor and Medjool varieties and their backcrossed males, identified >3.5 million polymorphic sites, including >10,000 genic copy number variations. A small subset of these polymorphisms can distinguish multiple varieties. We identified a region of the genome linked to gender and found evidence that date palm employs an XY system of gender inheritance.

The date palm is one of the oldest cultivated trees in the world, with evidence of domestication dating back >5,000 years[1]. The discovery of dates in the tombs of pharaohs and in neolithic sites dating from 7,000 to 8,000 years ago[2] demonstrates the historical significance of the species to human nutrition. Date palm trees are critical to agriculture in many hot and arid regions, and dates are the most important agricultural product of many countries in the Arabian Gulf. Total global production of dates in 2007 was 6.9 million tons (http://faostat.fao.org/).

However, date palm biotechnology faces many challenges, including long generation times, the inability to simply distinguish between the many varieties of date palm and the inability to distinguish female from male trees at an early stage. There are >2,000 date varieties with differences in color, flavor, shape, size and ripening time[3] and the genetic component of gender determination is not well understood[4]. Specifically, date palms take 5–8 years after planting to flower, the earliest point at which male and female trees can be distinguished. Especially because date palm orchards, which primarily comprise fruit-bearing female trees, can be rapidly ravished by disease, the ability to quickly replant orchards from seeds or seedlings known to be female would be of great benefit. There are no easily distinguishable sex chromosomes in date palm, despite some cytological evidence that they exist[5]. As biochemical studies have yielded little insight into how to identify the genders of immature plants[6], the identification of DNA sequences or sequence polymorphisms that are gender specific offers a promising alternative to efficiently determine date palm gender.

In large part owing to its long generation time, there are few genetic resources for date palm. The most extensive is a backcrossing program initiated in California in the 1940s[7], which required >30 years to generate. To our knowledge, there is no publicly available physical or genetic map for the genome of any date palm, and at the outset of this project only ~100 kbp of nuclear date palm DNA sequences were found in GenBank (http://www.ncbi.nlm.nih.gov/, March 1, 2009) (**Fig. 1**). To provide date palm researchers with the additional resources needed for comprehensive efforts to study and improve this important crop, we used massively parallel sequencing to assemble a draft genome sequence of date palm. Our analysis of nine varieties reveals polymorphisms that should provide an invaluable resource for the date palm community to identify ways to predict plant gender, maintain genetic diversity and improve traits such as fruit quality and ripening time.

RESULTS

We sequenced and assembled the genome of a female tree because, as the fruit-producing trees, female plants are of greater agricultural significance than male trees. The Khalas cultivar we selected to work with is traditionally regarded as the quintessential date variety with very high fruit quality.

[1]Genomics Core, Weill Cornell Medical College in Qatar, Doha, Qatar. [2]Department of Genetics, University of Georgia, Athens, Georgia, USA. [3]Laboratoire de Biogenèse Membranaire, CNRS UMR, Université V. Segalen Bordeaux, Bordeaux, France. [4]Department of Plant Biology, Michigan State University, East Lansing, Michigan, USA. [5]Agricultural and Water Research, Ministry of Environment, Doha, Qatar. [6]Biotechnology Centre, Ministry of Environment, Doha, Qatar. [7]USDA-ARS National Clonal Germplasm Repository for Citrus & Dates, University of California, Riverside, California, USA. [8]Department of Genetic Medicine, Weill Cornell Medical College in Qatar, Doha, Qatar. [9]Present address: EEA Balcarce, Instituto Nacional de Tecnología Agropecuaria, Balcarce, Argentina. [10]These authors contributed equally to this work. Correspondence should be addressed to J.A.M. (jom2042@qatar-med.cornell.edu).

Received 17 February; accepted 29 March; published online 29 May 2011; doi:10.1038/nbt.1860

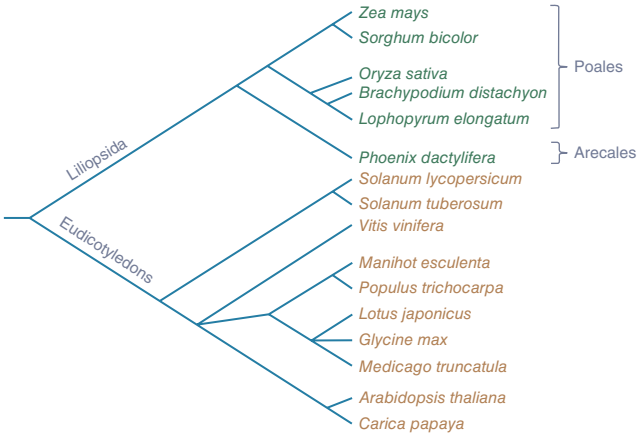

Figure 1 Taxonomic tree of selected crops for which genome sequences are available. Date palm is the first member of the order Arecales and the family Arecaceae for which a draft genome sequence is available. Other monocotyledonous plants (class Liliopsida) for which genome sequences are available are mainly grasses (order Poales). The tree was constructed in the Interactive Tree Of Life (http://itol.embl.de/) from taxonomy numbers in NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/).

Genome sequencing and assembly

The date palm genome contains 18 pairs of chromosomes[5] and our analysis suggests a genome size of ~658 Mb (**Supplementary Notes**). We undertook *de novo* next-generation sequencing of the date palm genome with the expectation that intragenic regions would have few large repeats, as is true in the similarly small genomes of other monocotyledonous crops, such as rice[8] and sorghum[9]. If this is the case in date palm, we reasoned that most genic regions should assemble uninterrupted by repeats, thus allowing a relatively unbiased view of the gene space. To this end, we used the Genome Analyzer IIx to generate sequences 36–84 bp in length from genomic fragments of ~170 bp or ~370 bp. We assembled the genome by using the SOAPdenovo genome assembler[10], which can use paired-end information for resolving repeats and has been used for other large genomes[11]. We used the SOAP Correction Tool to correct sequence reads before assembly and closed gaps where possible with the SOAP GapCloser.

The assembly stage used 526,443,374 sequences as input (**Supplementary Fig. 1**). This yielded an N50 contiguous sequence (contig), the shortest length of contigs contributing more than half of assembled sequence, of 6,441 bp and a scaffold N50 size of 9,339 bp when scaffolds <500 bp were excluded. We further joined SOAPdenovo scaffolds into larger scaffolds with 28.6× physical coverage from type III restriction enzyme libraries (2,000–5,000 bp)[12] using the BAMBUS software[13]; at least three longer mate-pair links were required to join contigs to scaffolds. This resulted in 57,277 scaffolds with an N50 size of 30,480 bp spanning 381 Mb of sequence. Post-assembly matching of sequences revealed a sequence redundancy of 53.4× from reads with an average length of 64 bp. This coverage is greater than the theoretically determined minimum for a high-quality assembly using reads of this length[10]. With a heterozygous genome, it is possible for the assembler to have split alleles and assembled them separately. This would result in contigs with half the sequence coverage of the genome average. However, distribution of coverage on the assembly showed no secondary peak at half the mean coverage (**Supplementary Notes**, **Supplementary Figs. 2** and **3**), indicating that assembly of separate haplotypes is most likely localized to short regions. With this short-read strategy, contigs broken by short repeats are joined to a scaffold by paired-end information. Large repetitive regions are expected to be intractable with this approach and are not included in the assembly.

To investigate the accuracy and completeness of the full genome assembly, we next made comparisons to fully sequenced genomic DNA regions from both the Deglet Noor cultivar and other Khalas cultivars (**Supplementary Table 1** and **Supplementary Fig. 4**).

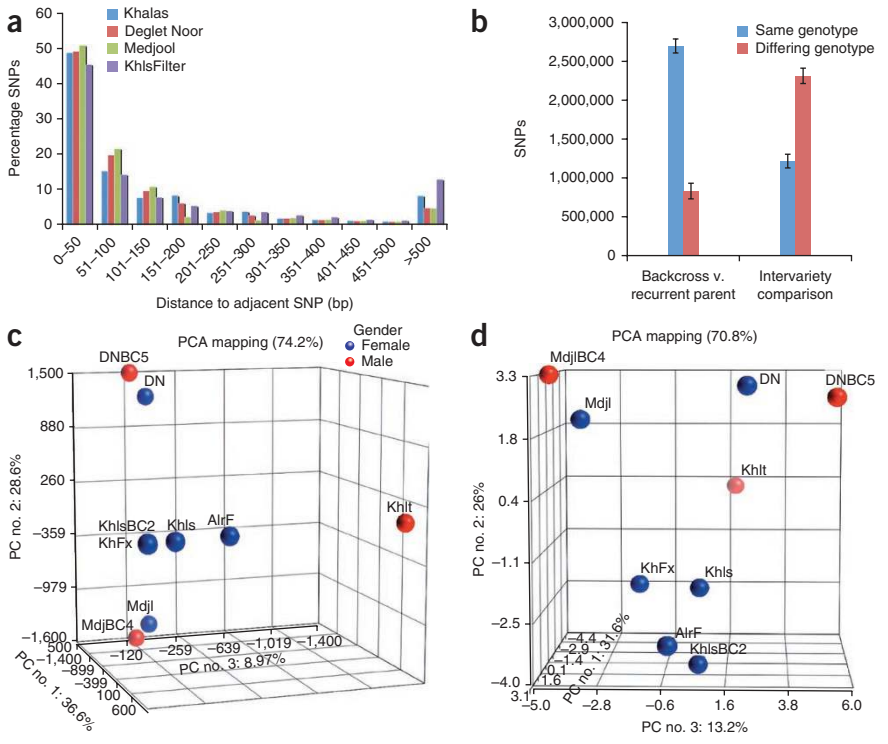

Figure 2 Date palm SNP analysis. SNPs were compared between parental alleles of the Khalas reference genome and different varieties. (**a**) The distance between parental allele SNPs in Khalas is not normally distributed. The skewed distribution of adjacent SNP distances demonstrates the occurrence of high and low polymorphism islands in the genome. About 49% of SNPs occur within 50 bp of another SNP. This trend was maintained even after removing SNPs likely to be in repetitive regions (KhlsFilter). (**b**) Backcrossed varieties of date palm on average show high levels of similarity to their recurrent parent with the number of generations of backcrossing (ranging from backcross 1 to 5 generations) having little effect on similarity levels (error bars are quite small). Intervariety comparisons show significantly more sites with different genotypes. (**c**) Principal component analysis (PCA) of sequenced genomes based on 3.5 million polymorphic sites. Khalas and backcrossed variants are essentially on top of each other. DN, Deglet Noor; MdjI, Medjool, BC, backcross; AlrF, AlrijalF; Khls, Khalas; Khlt, Khalt. (**d**) PCA of sequenced genomes based on 32 decision tree–selected polymorphic sites reveals little loss of discrimination quality with much reduced genotyping required. KhFx, Khalas x Khalas F1.

Table 1 Date palm genomes sequenced in this study

Date palm cultivar	Presumed origin	Site of collection	Gender	Sequence coverage
Khalas	Arabia	Qatar	Female	53.4×
Khalas × Khalas F1	California	California	Female	19.3×
Khalas BC2	California	California	Female	10.1×
Deglet Noor	North Africa	California	Female	19.8×
Deglet Noor BC5	California	California	Male	19.6×
Medjool	North Africa	California	Female	14.9×
Medjool BC4	California	California	Male	13.1×
Khalt	Qatar	Qatar	Male	11.2×
AlrijalF	Qatar	Qatar	Female	10.7×

Three commercially important female genomes together with various backcrosses and uncultivated varieties were sequenced to various levels of sequence redundancy to permit SNP identification.

We used Sanger technology to completely sequence six fosmids containing Deglet Noor inserts. Scaffolds from the assembly aligned to 60% of the mainly gene-rich total fosmid sequence, giving an indication of the completeness of the draft genome sequence. Analysis of the fosmid sequence not captured in the full genome assembly revealed that the majority of these regions are highly repetitive (**Supplementary Notes** and **Supplementary Table 2**). This investigation indicates that gene-rich regions were reconstructed more effectively than regions rich in transposable elements, and genes were recovered at much higher frequencies than repeat sequences.

We further compared the genome assembly to 109,244 contigs of assembled date palm expressed sequence tags (ESTs) (unpublished data). Using BLAT[14], 72% of EST contigs matched at least 90% of their length, whereas 86% of high-quality EST bases could be aligned to the reference sequence with a minimum of 98% sequence identity. Furthermore, using the CEGMA pipeline[15], which checks for full-length models of core genes, 94% of core eukaryotic genes were found in the assembly, and 71% of these were recovered as full-length gene models. Taken together, the data suggest that our assembly describes ~90% of date palm genes and ~60% of the full date palm genome sequence. The uncaptured regions of the genome are likely to be highly repetitive and thus intractable to the assembly approach used.

Genome annotation

Repeat masked scaffolds were passed to the Fgenesh++ pipeline for both *de novo* and homology-based gene prediction[16]. A total of 28,890 gene models were predicted. Of these, 25,059 predicted protein-encoding genes had significant BLAST similarity to proteins from other organisms in the nonredundant (NR) database at the National Center for Biotechnology Information (NCBI). Gene ontology information was assigned using BLAST2GO[17]. GC content within coding DNA sequence was 47.6%, whereas the entire assembled genome has a GC content of 38.5%.

The top BLAST hits for 9,022 of the predicted proteins for date palm matched predicted proteins from *Vitis vinifera*, a eudicotyledonous crop, followed by 5,094 top matches to predicted proteins from the monocotyledonous crop *Oryza sativa*. This higher protein sequence similarity between the two less phylogenetically related plants (the monocotyledonous date palm and the eudicotyledonous grapevine) has been observed by others in gene families from oil palm[18] and for oil palm ESTs[19]. Initial suggestions are that the grasses are a more diverged monocotyledonous group than date palm; however, additional studies will be required to fully explain this observation.

We found a total of 2,949 gene models (10% of those predicted) with high homology to genes encoding transposable elements. Among them, the protein-coding regions of 2,097 models matched transposable element proteins (BLASTP, E-value < 10^{-5}). The other 852 models matched predicted transposable element genes in their intron regions. These transposable elements within genes are likely to be found at low copy numbers in the genome, or they would not have been assembled. Overall, 55,855 sequences identified in the full genome assembly had characteristics of transposable elements. Some of these, including a few long terminal repeat (LTR) retrotransposons (45 families) and the tiny transposable elements called MITEs (35 families) were identified using structural criteria[20–22]. Representative sequences of these MITE and LTR families are presented in the **Supplementary Notes**. However, most were found by homology to known transposable element proteins. The transposable elements found in the full genome assembly were compared with the raw genomic sequence data. As expected, because of the inability of short reads to resolve long repeats, many more transposable element–related sequences were identified in the raw shotgun data than in the assemblies (**Supplementary Table 3**). The most abundant transposable elements identified in date palm, LTR retrotransposons of the Copia (~3.1% of reads) and Gypsy (~1.4% of reads) superfamilies, were found to occur 50-fold (0.062%) and 25-fold (0.056%) less frequently in assembled reads than in shotgun reads, respectively. The most abundant DNA transposable elements are the CACTA elements (0.03% of shotgun reads) (**Supplementary Table 3**). Because only predicted protein homologies were used to identify transposable elements and because all transposable elements contain extensive noncoding DNA, we expect that the vast majority of the transposable element–related DNA in the date palm genome assembly was missed by this approach (**Supplementary Notes**).

Polymorphism and comparative genomics

Using massively parallel sequencing on a cultivar of date palm with no documented inbreeding allowed us to detect a large number of parental allelic differences (**Supplementary Fig. 1**). Using BWA[23] and SAMTOOLS[24] software, we called 1,748,109 single-nucleotide polymorphisms (SNPs) in 381 Mb of sequence, yielding a heterozygosity rate of 0.46% or 1 SNP/217 bp. However, the distribution was significantly skewed, with 49% of SNPs being found within 50 bp of another SNP (**Fig. 2a**). These results were observed even when suspected repetitive regions, including ends of contigs and high sequence-coverage regions, were excluded from the analysis. These results suggest that there are islands of higher polymorphism within the genome, and this observation is important to subsequent large polymorphism analysis. A total of 100,019 of the parental SNPs

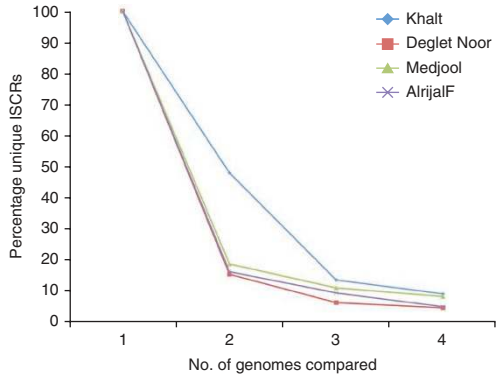

Figure 3 Analysis of imbalanced sequence count regions (ISCRs) among date palm genomes. Numbers of unique ISCRs remaining in each genome after comparison with other genomes are shown. Only nonbackcrossed genomes were considered to avoid bias from inbreeding. Approximately 7% of ISCRs were unique to any single genome, whereas the majority were observed in at least one other genome.

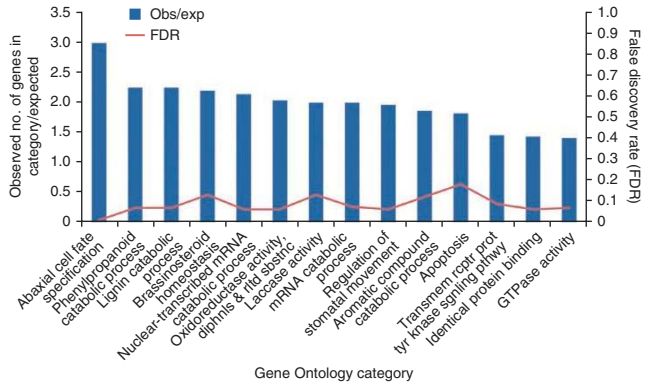

Figure 4 Enrichment of Gene Ontology categories for genes covered by imbalanced sequence count regions (ISCRs). Gene Ontology categories from genes covered by ISCRs in at least two genomes were analyzed for enrichment. Gene counts in each category were normalized to total gene counts in either the genome or ISCRs. A false discovery rate (FDR) of 0.2 was applied and only categories showing at least twofold enrichment in the ISCRs are reported.

occurred within a predicted gene-encoding sequence, and 53,890 of these cause an amino acid change. This yields a nonsynonymous-to-synonymous SNP ratio of 1.17; a ratio similar to that of 1.2 reported in rice[25].

To better characterize polymorphism in date palm from a biotechnology perspective, we sequenced to varying levels of coverage the genomes of representative male and female plants from the most popular commercial varieties Deglet Noor and Medjool, and the genome of a female belonging to the noncommercial AlrijalF variety (**Table 1**). Additionally, to characterize possible gender differences, we sequenced genomic DNA from two backcrossed males, two backcrossed females and one nonbackcrossed male (**Table 1**). We identified 3,518,029 SNPs in 381 Mb that were polymorphic in at least one of the sequenced genomes. The genotypes of all sequenced genomes were documented at these sites. As expected, genotypes were much more conserved between the backcrossed genomes and their recurrent (the parent maintained at each crossing with its progeny) parents than between different varieties (**Fig. 2b**). Indeed, clustering of the genomes by the genotypes at 3.5 million locations revealed the close relationship of the backcrossed trees and their recurrent parents (**Fig. 2c**). Moreover, the genome of an individual of the Khalas variety collected in Qatar clustered very close with trees backcrossed to a California Khalas plant believed to have been imported from Arabia almost 100 years ago[26]. We used a decision-tree algorithm[27] to identify a minimal five SNPs capable of distinguishing the nine varieties for which genomic sequencing data were available (**Supplementary Table 4**). A total of 32 SNPs (**Supplementary Table 4**) were highly informative in discriminating the varieties and may be helpful in discriminating other date palm varieties. Using just these 32 SNPs to separate the varieties provided little loss of discrimination power, with the top three principal components decreasing from 74% to 71% when comparing results with 3.5 million versus the core 32 SNPs (**Fig. 2d**). An additional four genomes were genotyped at these 32 SNPs and clustering analysis revealed their capacity to

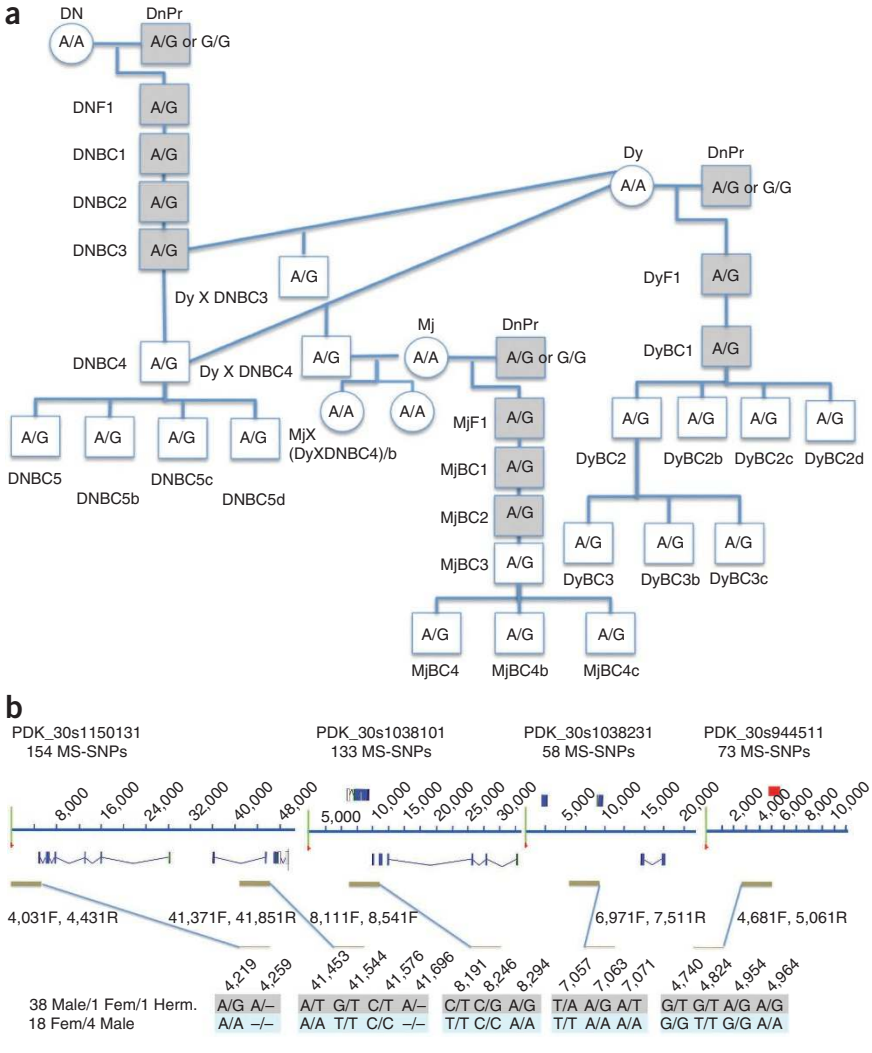

Figure 5 Pedigree and genotype information for gender-discriminating regions. Date palms of known genealogy were genotyped at multiple gender-discriminating regions. (**a**) A section of the full pedigree used for linkage analysis showing the complex relationship of the trees. DN, Deglet Noor; Dy, Dayri; Mj, Medjool; BC, backcross; DnPr, initial donor parents. Gray boxes indicate an unknown but theoretically determined genotype. The genotype in each individual is the genotype found at the first gender-discriminating SNP that was genotyped. Segregation of heterozygosity with the male phenotype is clear. (**b**) Genotypes from four scaffolds (scales with exons annotated as blue ticks and repeats as red rectangles) with the largest number of male-specific SNPs (MS-SNPs). Genotypes from selected regions (tan rectangles) are presented with their scaffold base pair location above each genotype. F and R indicate on which strand (forward or reverse) primers were designed to amplify the selected region. The number observed (both empirically and theoretically) for each gender in each genotype is included. Fem, female; herm., hermaphrodite. Heterozygous SNP calls are shaded gray whereas homozygous calls are shaded blue.

distinguish between the varieties (**Supplementary Fig. 5**). This set of SNPs provides a starting point for developing DNA markers capable of discrimination between the >2,000 varieties of date palm.

Large-scale polymorphisms, including copy number variations (CNVs), can be detected from sequence data by identifying regions where the observed number of matching sequences from a genome significantly deviates (either up or down) from the expected number (**Supplementary Fig. 1**). By matching sequences from each genome to the Khalas reference, gene-sized regions with significantly imbalanced counts of sequences were detected using CNV-SEQ software[28]. We term these 'imbalanced sequence count regions' (ISCRs) to distinguish them from more rigorously proven CNVs. As with the SNP data, extensive conservation of ISCRs was observed between backcrossed genomes and their recurrent parent (data not shown). Subsequent analysis was restricted to nonbackcrossed genomes to avoid duplication of results from inbreeding. A total of 10,388 ISCRs were detected that both overlap a predicted gene-coding region and occur in at least two genomes (**Supplementary Table 5**). At most, 10% of ISCRs were unique to a given genome (**Fig. 3**).

Whereas uneven distribution of polymorphisms interspersed with high sequence conservation in gene regions[29,30] may lead to false ISCR detection, modeling (**Supplementary Notes**) suggests that most of these ISCRs are real. Furthermore, quantitative PCR (qPCR) of five ISCRs on the four test genomes (20 different tests), gave 16 results consistent with expectation (amplified or deleted). Visual inspection of the sequence alignment in the four ISCR regions that failed to be validated revealed that, in some cases, sequence coverage variability is due to very high sequence polymorphism rather than absolute loss of sequence.

Genes exhibiting ISCRs in at least two genomes were analyzed for enrichment of Gene Ontology categories using the GOSSIP package within BLAST2GO[17], and enrichment was found in certain functional categories (**Fig. 4**). Interestingly, the categories for lignin, laccase and phenylpropanoid metabolism were overrepresented in ISCR regions. Genes in these processes are important in fruit flavor and ripening[31], two of the most distinguishable differences between date palm varieties, and may thus be of value in understanding the genetic regulation of the commercially relevant properties of date palm fruit. The large number of ISCRs between the genomes analyzed here are not entirely unexpected. It has been shown that a significant amount of variation among genomes is related to insertions and deletions[32,33] and, in some plants, this amounts to 10–20% of genome variation between cultivars[34,35].

No ISCRS were found to segregate with gender. Recognizing that comparing all genomes to the Khalas female genome could only identify female-specific sequences, we attempted to assemble male-specific sequences. We assembled reads from the male Deglet Noor BC5 genome. Very short contigs were expected because sequence redundancy (20×) was low, but this served as a first check for male-specific sequences. Sequences from the female genomes were matched to the Deglet Noor BC5 male contigs. All contigs were found to have significant sequence coverage from at least one of the six female genomes. Annotation of the short contigs revealed high frequencies of LTR retrotransposons, but no distinguishable male-specific genes.

Identification of gender-linked scaffolds

We scanned the 3.5 million SNP genotypes in the male and female genomes to identify polymorphisms that segregate with gender (**Supplementary Fig. 1**). The observed results best fit an XY sex-determination model with males being the heterogametic sex. Applying a heterogamete male model, we observed 1,605 SNPs that segregated with gender. Of these, 923 (58%) localized to 344 kb within 24 scaffolds spanning 602 kb (**Supplementary Table 6**). Specifically, all male genomes shared mainly the same heterozygous genotypes, whereas female genomes shared mainly the same homozygous genotypes in these scaffolds (**Fig. 5**). The scaffolds are broken by gaps that probably contain substantial amounts of repetitive DNA.

Analyzing two scaffolds with the most gender-segregating SNPs, we observed an approximate threefold difference in divergence from the reference sequence between male and female haplotypes. Furthermore, an almost 30-fold difference was observed between the number of male and female heterozygous SNPs within these regions. The genotypes of 867 polymorphic sites were recorded for all genomes in these regions. Comparison of the Deglet Noor and Medjool females to the Khalas female reference revealed that 253 and 271 sites differed from the Khalas reference and only 24 (9%) and 19 (7%) sites were heterozygous, respectively. At the same positions, their backcrossed males showed 736 and 770 sites differing from the Khalas reference, and 584 (79%) and 578 (75%), respectively, of these were heterozygous. The significantly higher heterozygosity levels (χ^2 = 893.6 and 767.7, respectively, 1 d.f., $P < 0.0001$) in the males represents an approximately threefold greater heterozygosity in these regions when compared to the rest of the genome. The females have significantly lower heterozygosity in these regions with respect to the rest of the genome (χ^2 = 435.9 and 410.2, 1 d.f., $P < 0.0001$), resulting in an ~14-fold lower heterozygosity in these regions relative to the rest of the sequenced genome. This pattern of sequence degeneration between male and female haplotypes may be indicative of reduced recombination between the male and female haplotypes, which is a step that may be critical to the development of gender-specific regions[36,37]. In these two scaffolds, we observed seven exons in three of the four annotated genes (**Fig. 5b**) that contained unusually long introns, ranging between 4 kb and 13.1 kb (compared to an average of <200 bp for most flowering plant introns). Longer introns occur more frequently in regions of low recombination in *Drosophila melanogaster*[38].

To determine whether the observed differences in heterozygosity are indeed linked to gender, we selected short regions from the four scaffolds with the largest number of segregating SNPs for genotyping in a pedigree containing six date palm female varieties and their 28 progeny (**Supplementary Table 7**). Genotyping results indicate that these four scaffolds are linked to each other with no recombination between them (**Fig. 5b**), suggesting they likely localize to the same region of the genome. Using only empirically determined genotypes on males and females (excluding the rare hermaphrodites), the genotyped scaffolds significantly link to gender with a $\log_{10}$ odds (LOD) score of 5.3 (recombination frequency of 0.07), with only two males showing recombination. Furthermore, as backcrossed plants were used in the pedigree, theoretically determined genotypes of donor parents (**Supplementary Methods**) can be included, improving the LOD score to 8.9 (recombination frequency of 0.05) (**Fig. 5a**). Genotyping of date palms outside the pedigree was consistent with the trend of male heterozygosity and female homozygosity. Of 63 empirically and theoretically genotyped males and females, only 5 did not give the expected genotype (**Fig. 5b**). Additionally, one male was observed to be homozygous for the male-specific allele (**Supplementary Table 7**). Predicted genes in this region (**Supplementary Table 6**) include one encoding rcd-1, 'required for cell differentiation homolog', a Myb family member, and a gene predicted to encode a prenyltransferase of the rab geranylgeranyl transferase family. Interestingly, the rcd-1 gene is important in sexual development in yeast[39] and interacts with c-Myb[40]. Moreover, it has been shown that cell differentiation control in date

palm floral development is critical to sex organ development[41] and that there are sets of MADS box genes, which control flower development, and require prenylation for correct function[42]. We observed multiple nonsynonymous polymorphisms between the male and female haplotypes in these genes, although none were certain to be deleterious to protein function.

DISCUSSION

We present, to our knowledge, the first publicly available draft of the nuclear genome for a member of the palm family (Arecaceae) and indeed the entire order Arecales. Date, oil and coconut palms are important crops in several developing countries, and this sequence provides a resource that may be vital for their improvement. For instance, it would have been extremely difficult to identify the gender-specific SNP markers we report without the availability of a draft genome sequence. Despite the limitations of short-read assembly in handling heterozygous and repetitive regions, we obtained gene regions with contiguity similar to other draft genome sequences[43,44] by using paired-end libraries of varying sizes. The approach focused on assembling the gene-containing regions of the date palm by relying on the observation that most plants have fewer repeat sequences within genes than in extragenic regions. In terms of a framework used to classify the quality of plant genome assemblies[45], we regard this as a high-quality draft genome, with its implied benefits and caveats. The next step in the improvement of this sequence should be its anchoring to physical and genetic maps. However, the utility of the current assembly is best revealed by its promise for beginning to answer pressing needs in date palm improvement.

The capacity to use genetics to differentiate between cultivars and predict the gender of immature trees are perhaps the two most immediate challenges in applying biotechnology to date palm cultivation and improvement. Annotation of the current assembly has dramatically improved our knowledge of the gene content and allelic variation of date palm. Sequence data from multiple genomes have provided the largest resource of polymorphic markers to date. A small subset of these new markers can serve as a starting point for a set capable of distinguishing the >2,000 date palm varieties.

We have sequenced three of the top date palm varieties that are important in three regions of date palm production: Khalas, favored in Arabia; Deglet Noor, favored in North Africa; and Medjool, increasingly favored in California[26]. This resource will allow future comparisons of traits, such as fruit quality and ripening time, which vary among these favored varieties. Sequencing of the backcrossed males, a unique resource in any long-generation plant, allowed us to begin to dissect genomic differences between male and female date palms. The scaffolds we have identified to be strongly linked to gender may form the basis of a DNA marker-based gender test for use at the seed and/or seedling stages. These regions should be studied further to identify a possible specific mutation, mutations or other gene content difference that determines plant gender.

For millennia date palm cultivation of favored female varieties has taken the form of offshoot propagation. More recently, somatic embryogenesis has been used to propagate favored varieties. It has been essentially impossible to grow a specific female date palm variety from seed because seedling-grown fruit quality is too different from the mother to be economically useful. By combining the findings presented here with the backcrossed genetic resources that have been generated for decades[7], we may soon have access to seeds of backcrosses that are identified as being female before germination and genotyped at trait loci to show similarity to the original mother. Our results lay the foundation for future date palm research at the genomic level by providing the first genome-wide gene set, the first genome-wide multivariety polymorphism set and the first gender-linked regions for this species.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. Data from this Whole Genome Shotgun project have been deposited at DDBJ/EMBL/GenBank (http://www.ncbi.nlm.nih.gov/) under the accession no. ACYX00000000. The version described in this paper is the second version, ACYX02000000. Date Palm fosmid sequences have been submitted to DDBJ/EMBL/GenBank as follows: 9A12F7 under accession no. JF313259, 9B12 under accession no. JF313260, 9H12 under accession no. JF313261, E2 under accession no. GU183367, R1 under accession no. GU183365 and D6 under accession no. GU183366. Short-read sequence data have been deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession no. SRA029799, study accession no. SRP005625. SNPs have been submitted to dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) under the handle 'WCMCQ-GENOMICS', submitter batch id 'palmqatar1'. Assembly, polymorphism and annotation data are available at http://qatar-weill.cornell.edu/research/datepalmGenome/download.html.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by Qatar Foundation Biomedical Research Program Grant funding. We thank G. Parra for running the CEGMA pipeline on our initial assembly. We thank the Cornell Center for Academic Computing for providing hardware systems support during the assembly stage. We thank K. Machaca and R. Crystal for critical review of the manuscript.

AUTHOR CONTRIBUTIONS

E.K.A.-D. extracted genomic DNA, created libraries, sequenced the genome and assisted with the manuscript writing. B.G. conducted SNP, CNV and annotation analysis. M.E.A.-M. genotyped gender-discriminating regions. E.K.A.-A. and Y.M.S. assisted in genome sequencing, conducted qPCR validation of CNVs and helped write the manuscript. M.Y.A.-J. cloned, sequenced and analyzed sequences from standard sequencing technology for comparison to the next generation data. I.J.S. and K.M.S.-E. maintained the tree tissue culture and cultivar data on the sequenced trees. H.W., S.C., A.C.P., J.D. and J.L.B. constructed the fosmid library, sequenced date palm fosmids, provided transposable element annotation and generated comparative analyses and genome size predictions. J.O. and V.A. constructed EST libraries and provided DNA sequence from ESTs. H.W. and J.L.B. also helped write the manuscript. R.R.K. maintained and provided the date palm genetic resource including the pedigree information and assisted in phenotyping of the date palms. J.A.M. conceived and planned the project, created libraries, analyzed for gender-specific regions, assembled and annotated the genome and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Zohary, D. & Spiegel-Roy, P. Beginnings of fruit growing in the old world. *Science* **187**, 319–327 (1975).
2. Kwaasi, A.A.A. Date palms. in *Encyclopedia of Food Sciences and Nutrition*. 2nd edn. (ed. Caballero, B.) 1730–1740 (Elsevier Science, 2003).
3. Al-Farsi, M.A. & Lee, C.Y. Nutritional and functional properties of dates: a review. *Crit. Rev. Food Sci. Nutr.* **48**, 877–887 (2008).
4. Ainsworth, C., Parker, J. & Buchanan-Wollaston, V. Sex determination in plants. *Curr. Top. Dev. Biol.* **38**, 167–223 (1998).

5. Siljak-Yakovlev, S. *et al.* Chromosomal sex determination and heterochromatin structure in date palm. *Sex. Plant Reprod.* **9**, 127–132 (1996).
6. Qacif, N., Baaziz, M. & Bendiab, K. Biochemical investigations on peroxidase contents of male and female inflorescences of date palm (*Phoenix dactylifera* L.). *Sci. Hortic. (Amsterdam)* **114**, 298–301 (2007).
7. Barrett, H.C. Date breeding and improvement in North America. *Fruit Varieties Journal* **27**, 50–55 (1973).
8. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* **296**, 79–92 (2002).
9. Paterson, A.H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
10. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
11. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
12. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
13. Pop, M., Kosack, D.S. & Salzberg, S.L. Hierarchical scaffolding with Bambus. *Genome Res.* **14**, 149–159 (2004).
14. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
15. Parra, G. *et al.* Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297 (2009).
16. Solovyev, V. *et al.* Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7** (Suppl 1), S10 (2006).
17. Conesa, A. & Götz, S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008**, 619832 (2008).
18. Adam, H. *et al.* MADS box genes in oil palm (*Elaeis guineensis*): patterns in the evolution of the SQUAMOSA, DEFICIENS, GLOBOSA, AGAMOUS, and SEPALLATA subfamilies. *J. Mol. Evol.* **62**, 15–31 (2006).
19. Jouannic, S. *et al.* Analysis of expressed sequence tags from oil palm (*Elaeis guineensis*). *FEBS Lett.* **579**, 2709–2714 (2005).
20. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–268 (2007).
21. McCarthy, E.M. & McDonald, J.F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362–367 (2003).
22. Han, Y. & Wessler, S.R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* (2010).
23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
24. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
25. McNally, K.L. *et al.* Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. USA* **106**, 12273–12278 (2009).
26. Hodel, D.R., Johnson, D.V. & Nixon, R.W. *Dates—Imported and American Varieties of Dates in the United States* (ANR Publications, 2007).
27. Zhang, H., Wang, M. & Chen, X. Willows: a memory efficient tree and forest construction package. *BMC Bioinformatics* **10**, 130 (2009).
28. Xie, C. & Tammi, M.T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
29. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **101**, 12404–12410 (2004).
30. Yu, J. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38 (2005).
31. Singh, R., Rastogi, S. & Dwivedi, U.N. Phenylpropanoid metabolism in ripening fruits. *Comprehensive Reviews in Food Science and Food Safety* **9**, 398–416 (2010).
32. Britten, R.J. *et al.* Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl. Acad. Sci. USA* **100**, 4661–4665 (2003).
33. Springer, N.M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734 (2009).
34. Ding, J. *et al.* Highly asymmetric rice genomes. *BMC Genomics* **8**, 154 (2007).
35. Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997–1002 (2005).
36. Charlesworth, B. & Charlesworth, D. A model for the evolution of dioecy and gynodioecy. *Am. Nat.* **112**, 975–997 (1978).
37. Bergero, R. & Charlesworth, D. The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* (Personal edition) **24**, 94–102 (2009).
38. Carvalho, A.B. & Clark, A.G. Intron size and natural selection. *Nature* **401**, 344 (1999).
39. Okazaki, N. *et al.* Novel factor highly conserved among eukaryotes controls sexual development in fission yeast. *Mol. Cell. Biol.* **18**, 887–895 (1998).
40. Haas, M. *et al.* c-Myb protein interacts with Rcd-1, a component of the CCR4 transcription mediator complex. *Biochemistry* **43**, 8152–8159 (2004).
41. Daher, A. *et al.* Cell cycle arrest characterizes the transition from a bisexual floral bud to a unisexual flower in *Phoenix dactylifera*. *Ann. Bot. (Lond.)* **106**, 255–266 (2010).
42. Yalovsky, S. *et al.* Prenylation of the floral transcription factor APETALA1 modulates its function. *Plant Cell* **12**, 1257–1266 (2000).
43. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* **296**, 79–92 (2002).
44. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
45. Chain, P.S.G. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236–237 (2009).

ONLINE METHODS

Genomic libraries and sequencing. Date palm genomic DNA was extracted from leaves obtained from farmed trees in the Doha, Qatar, area or at the US Department of Agriculture collection in Riverside, California. The Khalas female had been grown from well-documented plant tissue culture. The Alrijal female and Khalt male were seed grown but otherwise of unknown descent. DNA was extracted from the fresh leaves of date palm trees using the Wizard Genomic DNA preparation kit (Promega). Leaves used for preparation of DNA employed in generating the Deglet Noor fosmid library were derived from the seedling of a single germinated seed.

Library construction for the short-paired libraries was conducted according to the manufacturer's protocol (Illumina). Two paired libraries of average insert size 172 bp and 370 bp were used. Longer mate-pair libraries were constructed using a linker sequence-modified version of the Type III restriction enzyme *Eco*P15I library method as described[12], producing 25–27 bp from either end of a DNA molecule. Fosmid library construction in vector pCC1FOS (Epicentre) was done as previously described[46].

The genome was assembled and scaffolded using SOAPdenovo v1.4 (ref. 10) with a k-mer of 31. Scaffolding using type III restriction libraries was conducted in BAMBUS[13] using 60 Ns to designate a scaffold gap.

Annotation. A repeat masked version of the genome was used for gene prediction. Ten million random short reads were assembled to create an initial repetitive region database to screen against the sequence data using REPEATMASKER (http://www.repeatmasker.org/). Previously trained monocot gene prediction parameters were used with the Fgenesh++ pipeline, and the entire plant section of REFSEQ was used as input for homology searches. Genes in the gender-specific region were manually curated. Functional annotation was carried out using a local implementation of the BLAST2GO[17] software. All predicted genes were searched using BLASTP (e-value cutoff of 10^{-5}) against the NR database at NCBI and also searched using the INTERPRO database at the European Bioinformatics Institute. Functional assignments, Gene Ontology and Enzyme Commission numbers were assigned whenever possible.

For the fosmid sequences, predicted ORFs were searched against the GenBank NR nt and EST databases using BLASTN and against the NR database using BLASTX. A cutoff value of e^{-10} was used as the significance similarity threshold for the comparison.

Transposable element identification. Transposable element identification and quantification were by a series of complementary approaches. Small non-coding transposable elements such as MITEs were found by MITE-Hunter[22] and RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html). Protein-coding transposable elements were mainly identified by homology to transposable element-encoded proteins using BLASTX and required e-value of 10^{-5} between predicted peptides. Intact LTR retrotransposons were found using LTR_FINDER[20] and LTR_STRUC[21]. Once transposable elements were identified, their multiple copies were found by homology in the full genome assembly and in the shotgun reads.

Polymorphism detection. SNPs were called by matching the original shotgun sequences to the *de novo* assembly reference sequence using BWA[23] and documenting regions, using SAMTOOLS[24],where it was apparent that the reads represented two alleles (**Fig. 1**). For SNP calling in Khalas only the longest paired-end sequences were used, resulting in 29.3× (of a total 53.4× used in assembly) coverage of 84 bp sequences. To avoid calling SNPs due to low-quality sequence or collapsed repetitive sequence, a SNP was required to have at least fivefold and not more than 70-fold coverage. For filtered SNP analysis in the measure of inter-SNP distance (**Fig. 2a**), 500 bp from either end of a contig and regions with >38× or <20× (1.3× and 0.7× the mean sequence coverage) were removed.

ISCRs were detected using CNV-SEQ[28]. The window size for a detectable ISCR with an absolute $\log_2$ value of ≥0.6 ranged in size from 800 bp to 1,000 bp depending on depth of sequence coverage for the test genome. To be conservative, a universal window size of 1,600 bp was set to call an ISCR. This was >1.5× larger than the window size required for statistically significant ISCR calling. At least three adjacent windows were required before annotating the region as an ISCR. Global normalization was used to take into account the lack of chromosome-sized contigs.

ISCRs were annotated by documenting all locations of an ISCR in each sequenced genome. If the regions between any two genomes overlapped this was collapsed and considered one ISCR region. All genomes were then documented for their level of sequence variation in these ISCR regions. Only those ISCRs that overlapped a coding region were documented.

Polymorphisms linked to gender were detected by scanning the genotypes of all genomes at the 3.5 million documented polymorphic sites. Scaffolds were identified that had more than ten gender-segregating SNPs.

Statistical analysis. LOD scores were calculated as described[47]. Gene Ontology enrichment was calculated using the GOSSIP algorithm within the BLAST2GO package[17] which provides false-discovery rates. χ^2 analysis was conducted using expected numbers of heterozygote SNPs based on the entire genome in a contingency table with heterozygous or homozygous as the two categories. Of all recorded genotyped positions in the genome, males were on average heterozygous in 25% whereas females were heterozygous in 36% of positions. In the suspected gender-linked scaffolds, all genotyped polymorphic sites in Deglet Noor and Medjool females and their backcrossed males were documented for homozygous or heterozygous changes and these used for the observed numbers.

Principal component analysis of the cultivar genotypes was carried out using the Partek Genomics Suite (Partek). Genotypes were transformed to numeric genotypes with 1 representing homozygous matching the Khalas reference, 2 representing heterozygous and 3 representing homozygous difference to the Khalas reference. The decision tree algorithm within the Willows package[27] was used to find the best cultivar-discriminating SNPs. The top 1,000 most informative SNPs were selected based on a showing of all three possible alleles (AA, AB, BB) in the nine sequenced genomes. From this set, the decision tree algorithm was used to select the fewest number of SNPs that could distinguish the nine sequenced varieties. Though only five SNPs were enough to separate all nine genomes, the backcrossed genomes did not always cluster with their recurrent parents accurately. SNPs with the most distinguishing power in the decision tree (32 SNPs) were chosen to provide a set from which a future subset can be selected once testing in a much larger and more diverse population is completed.

46. Pontaroli, A.C. *et al.* Gene content and distribution in the nuclear genome of Fragaria vesca. *The Plant Genome* **2**, 93–101 (2009).
47. Lathrop, G.M. & Lalouel, J.M. Easy calculations of lod scores and genetic risks on small computers. *Am. J. Hum. Genet.* **36**, 460–465 (1984).

doi:10.1038/nbt.1860