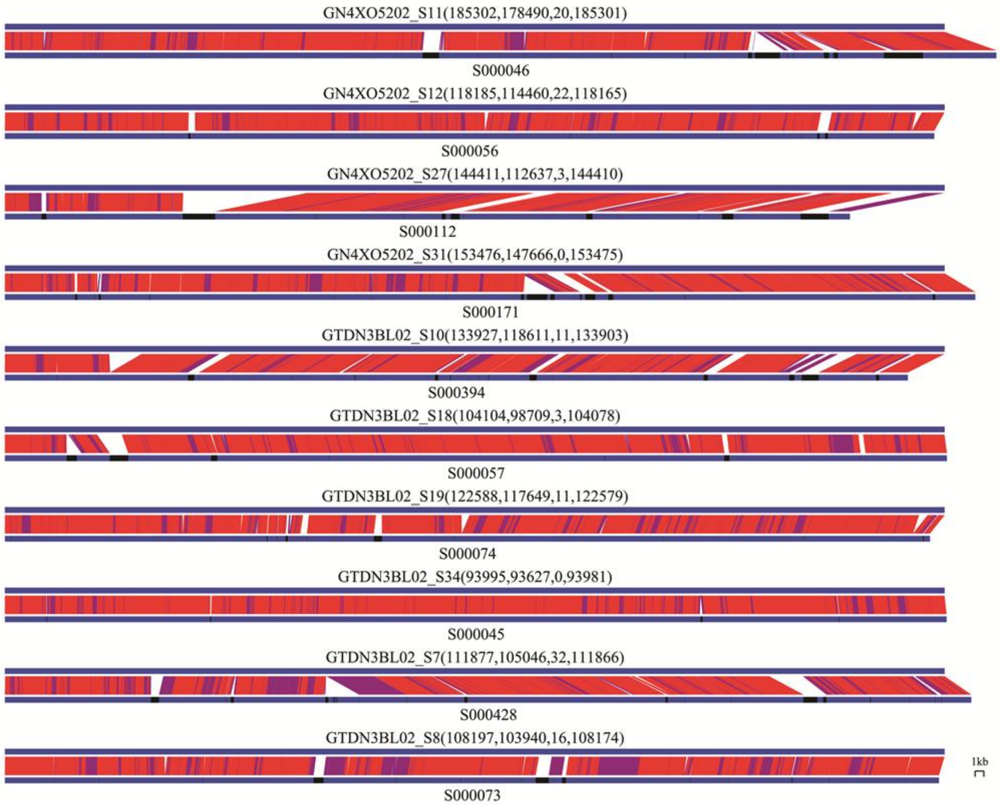
Supplementary Information for
Genome Sequences of the Date Palm *Phoenix dactylifera* L.

Ibrahim S. Al-Mssallem[1,3,4], Songnian Hu[1,2,4], Xiaowei Zhang[1,2,4], Qiang Lin[1,2,4], Wanfei Liu[1,2,4], Jun Tan[1,4], Xiaoguang Yu[1,2], Jiucheng Liu[1,2], Linlin Pan[1,2], Tongwu Zhang[1,2], Yuxin Yin[1,2], Chengqi Xin[1,2,4], Hao Wu[1,2], Guangyu Zhang[1,2], Mohammed M. Ba Abdullah[1], Dawei Huang[1,2], Yongjun Fang[1,2], Yasser O. Alnakhli[1], Shangang Jia[1,2], An Yin[1,2] , Eman M. Alhuzimi[1], Burair A. Alsaihati[1], Saad A. Al-Owayyed[1], Duojun Zhao[1,2], Sun Zhang[1,2], Noha A. Al-Otaibi[1], Gaoyuan Sun[1,2], Majed A. Majrashi[1], Fusen Li[1,2], TALA[1,2], Jixiang Wang[1,2], Quanzheng Yun[1,2], Nafla A. Alnassar[1], Lei Wang[1,2], Meng Yang[1,2], Rasha F. Al-Jelaify[1], Kan Liu[1,2], Shenghan Gao[1,2], Kaifu Chen[1,2], Samiyah R. Alkhaldi[1], Guiming Liu[1,2], Meng Zhang[1,2], Haiyan Guo[1,2], and Jun Yu[1,2]

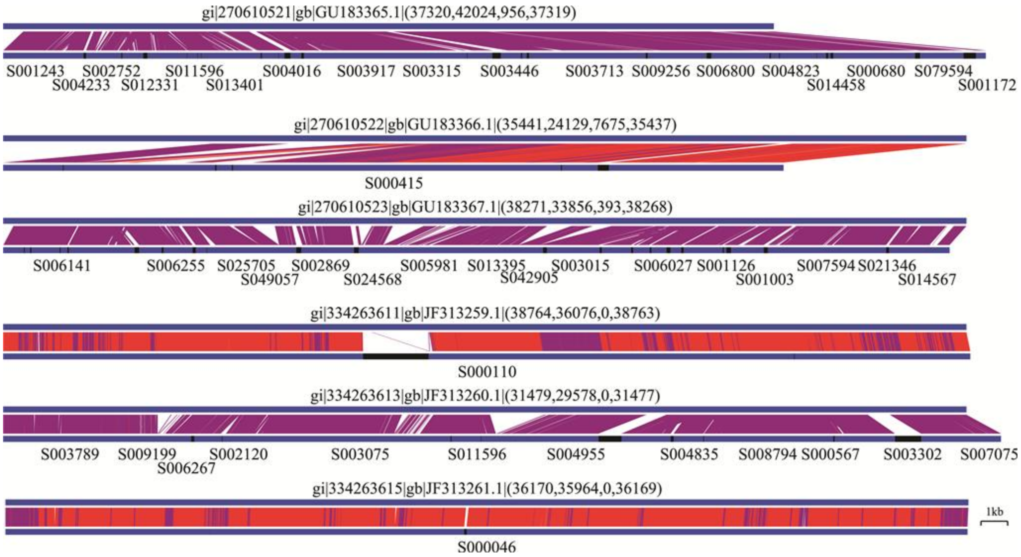
[1] Joint Center for Genomics Research (JCGR), King Abdulaziz City for Science and Technology (KACST) and Chinese Academy of Sciences (CAS), Prince Turki Road, Riyadh 11442, Kingdom of Saudi Arabia.
[2] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, 1-7 Beichen West Road, Beijing 100101, China.
[3] Department of Biotechnology, College of Agriculture and Food Sciences, King Faisal University, Ahssa 31982, Kingdom of Saudi Arabia.
[4] These authors contributed equally to this work.

Correspondence and requests for materials should be addressed to I.S.M. (imssallem@kacst.edu.sa), J.Y. (junyu@big.ac.cn) or S.N.H. (husn@big.ac.cn).

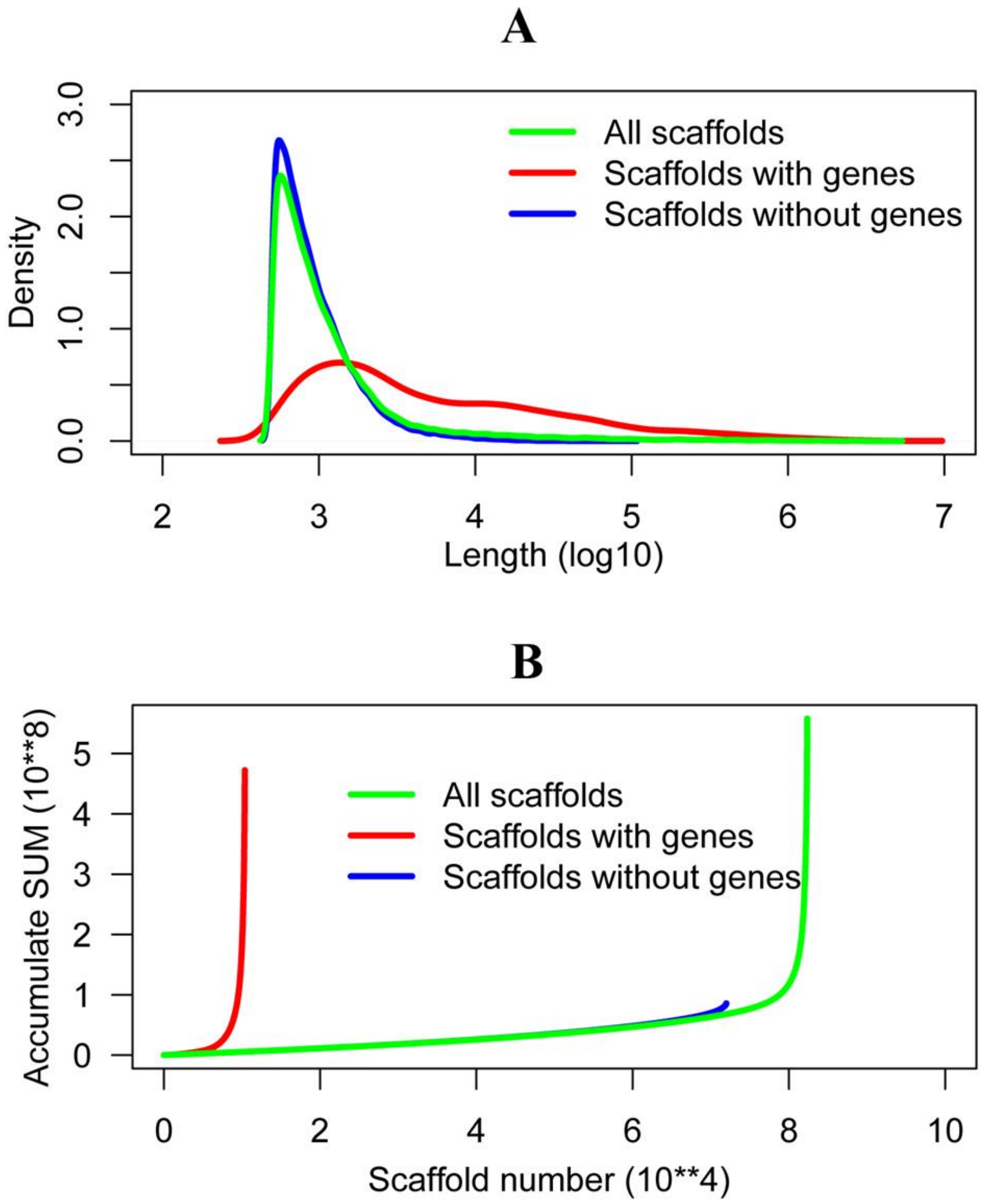
Supplementary Figures


GN4XO5202_S11(185302,178490,20,185301)
S000046
GN4XO5202_S12(118185,114460,22,118165)
S000056
GN4XO5202_S27(144411,112637,3,144410)
S000112
GN4XO5202_S31(153476,147666,0,153475)
S000171
GTDN3BL02_S10(133927,118611,11,133903)
S000394
GTDN3BL02_S18(104104,98709,3,104078)
S000057
GTDN3BL02_S19(122588,117649,11,122579)
S000074
GTDN3BL02_S34(93995,93627,0,93981)
S000045
GTDN3BL02_S7(111877,105046,32,111866)
S000428
GTDN3BL02_S8(108197,103940,16,108174)
S000073
1kb

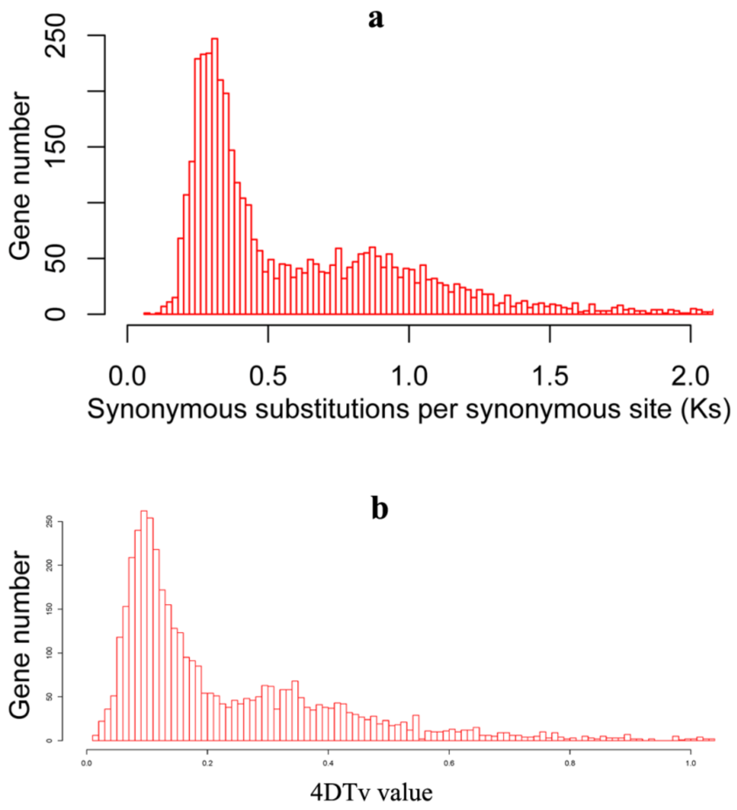
Supplementary Figure S1. Sequence alignments of 10 BAC sequences (top blue lines) with *P. dactylifera* scaffolds (bottom blue lines). Unmapped repetitive sequences are often filled with unassembled Ns (black segments) in scaffolds. Repeats are highlighted in purple to differentiate them from the unique matches (red connecting lines).



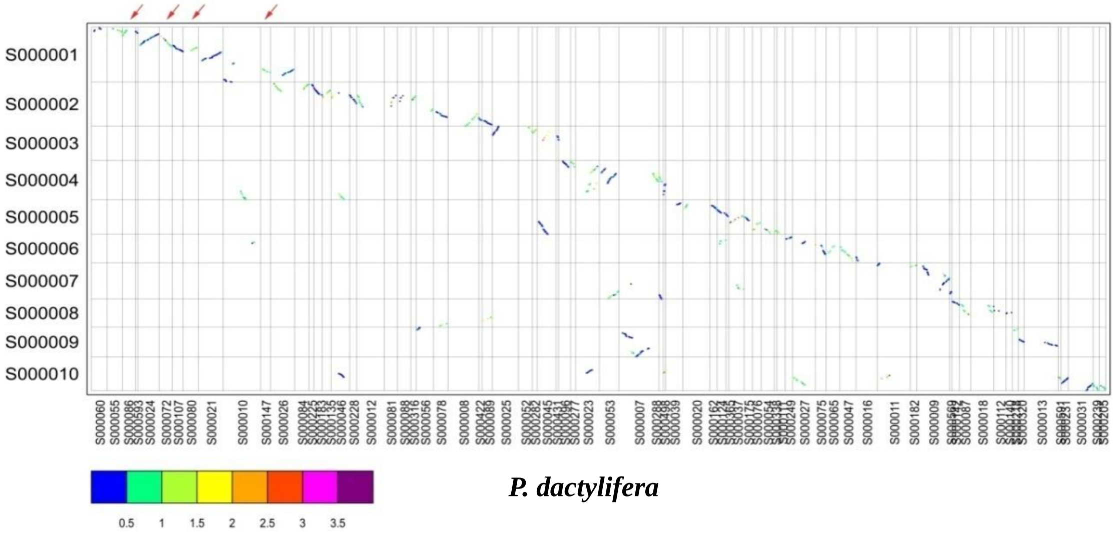
Supplementary Figure S2. Sequence alignments of 6 fosmid sequences (top blue lines) with *P. dactylifera* scaffolds (bottom blue lines). Unmapped repetitive sequences are often filled with unassembled Ns (black segments) in scaffolds. Repeats are highlighted in purple to differentiate them from the unique matches (red connecting lines). The fosmid sequences of variety Deglet Noor were downloaded from GenBank (GI number: 334263611, 334263613, 334263615, 270610521-270610523)]



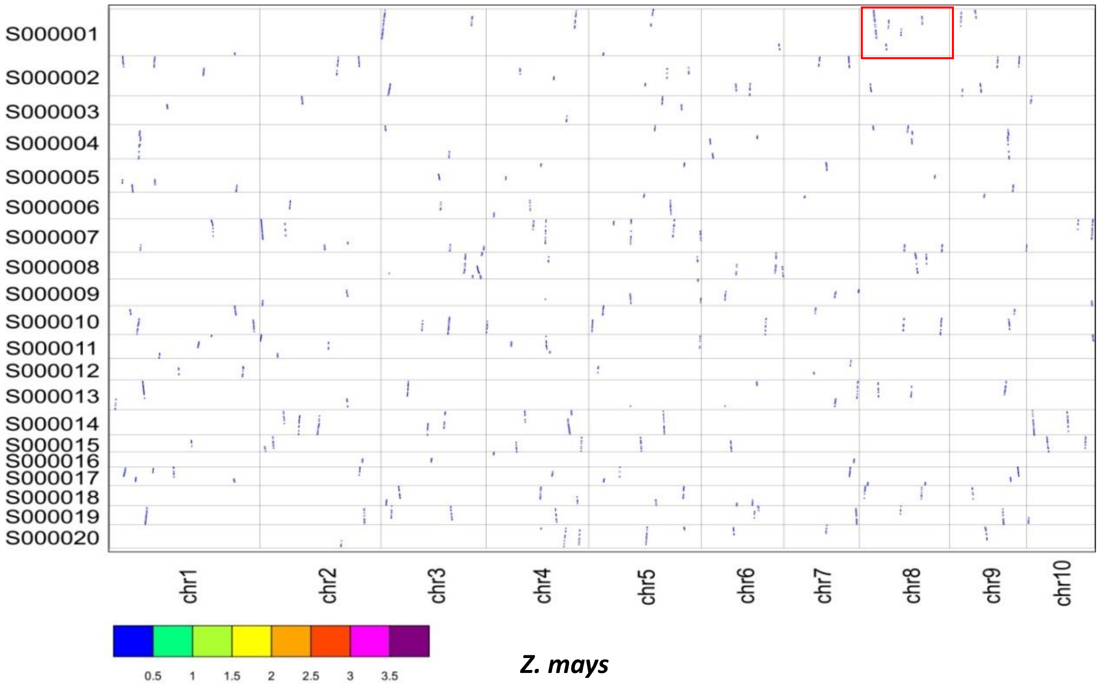
Supplementary Figure S3. Comparison between scaffolds with genes and without genes in the *P. dactylifera* genome assembly. (A). Correlation between gene density and scaffold size. (**B)**. Correlation between scaffold number and scaffold accumulative size.



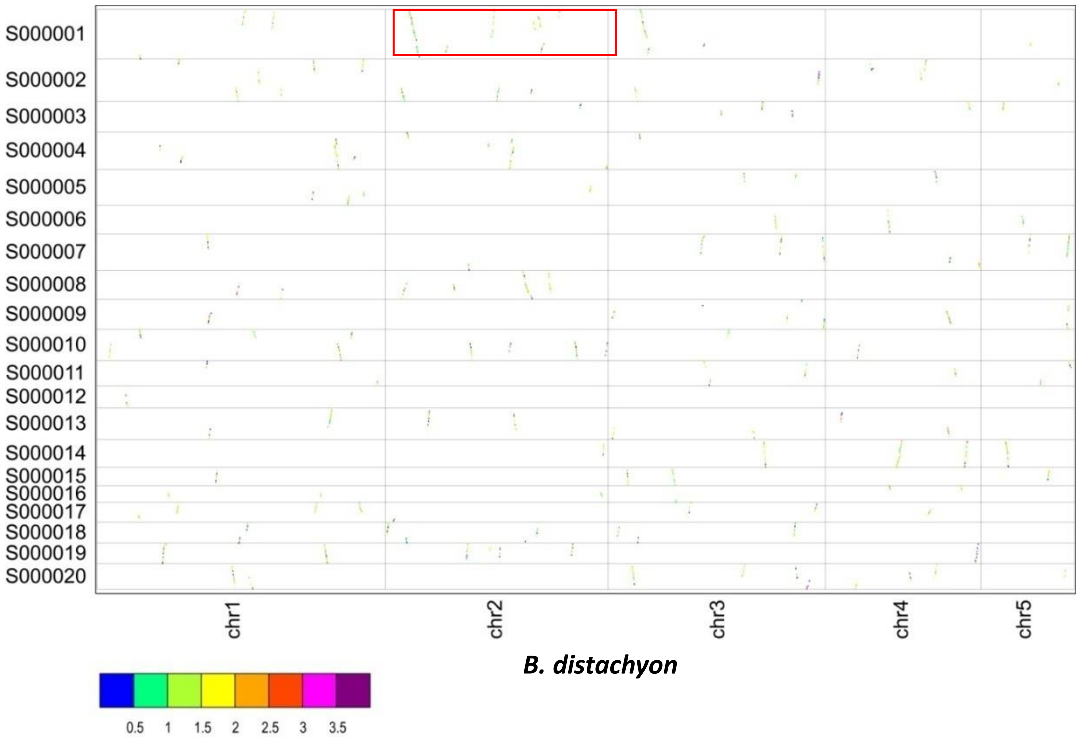
Supplementary Figure S4. The Ks (a) and 4DTv(b) distribution of gene pairs in the collinear regions of *P. dactylifera* genome assembly.

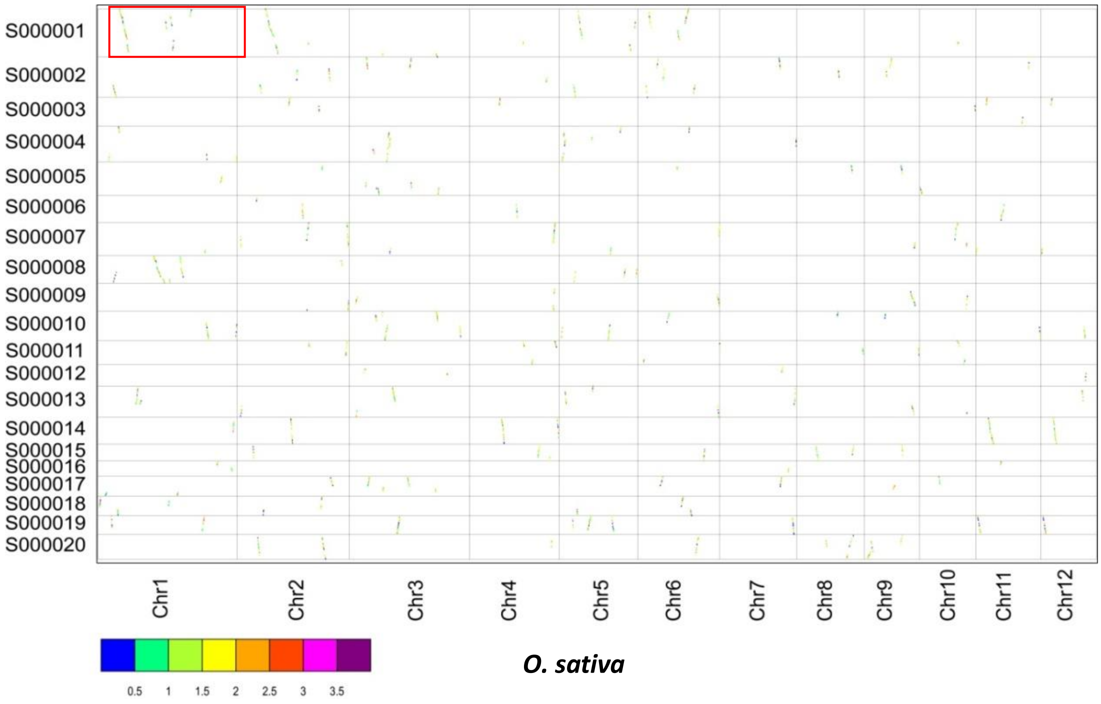


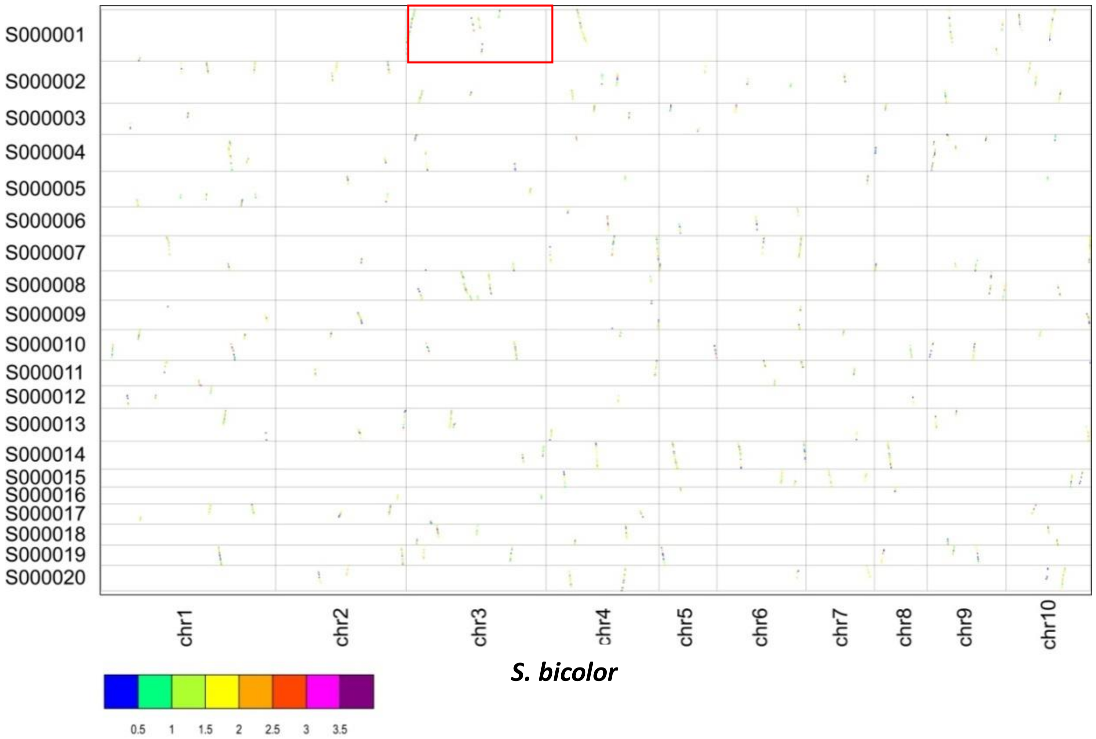
P. dactylifera



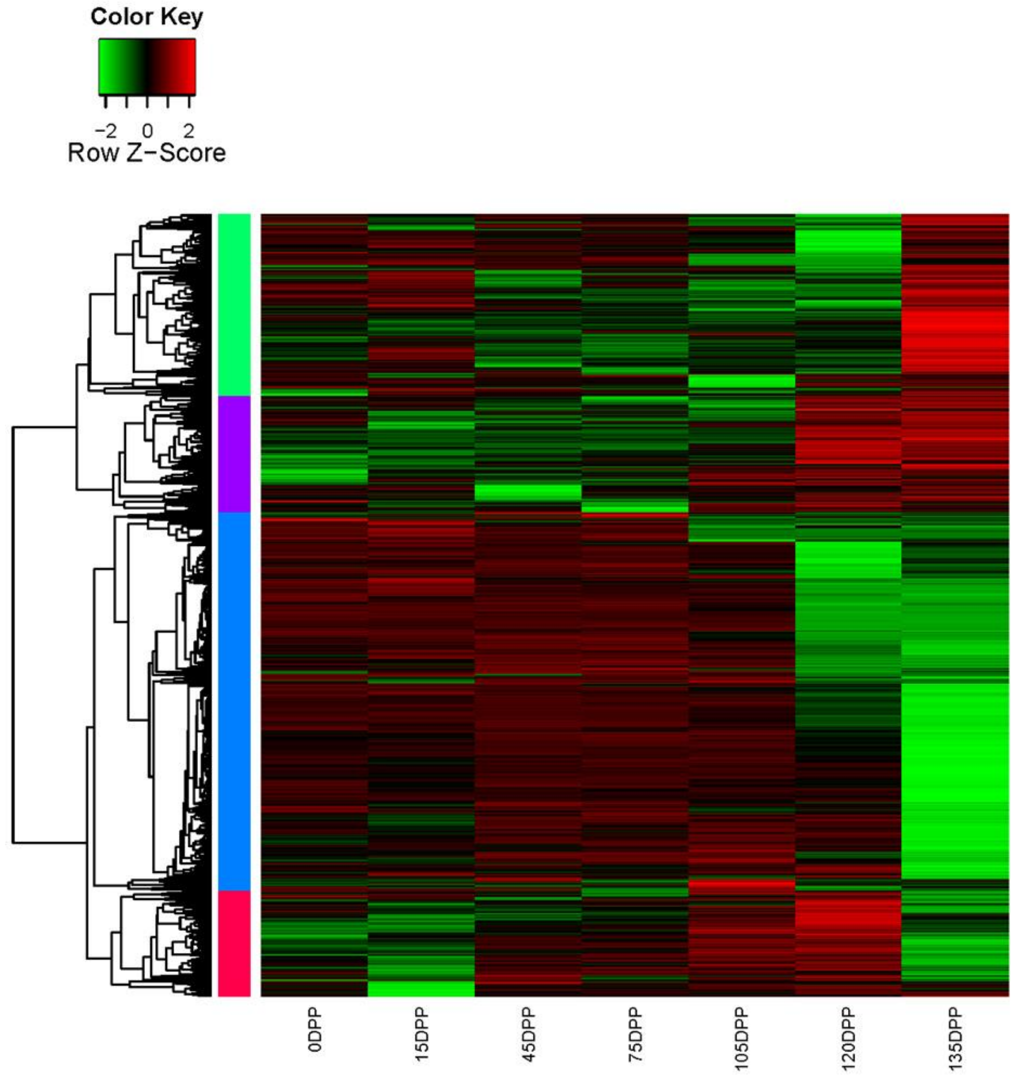
Z. mays


S000001
S000002
S000003
S000004
S000005
S000006
S000007
S000008
S000009
S000010
S000011
S000012
S000013
S000014
S000015
S000016
S000017
S000018
S000019
S000020
chr1
chr2
chr3
chr4
chr5
B. distachyon
0.5
1
1.5
2
2.5
3
3.5

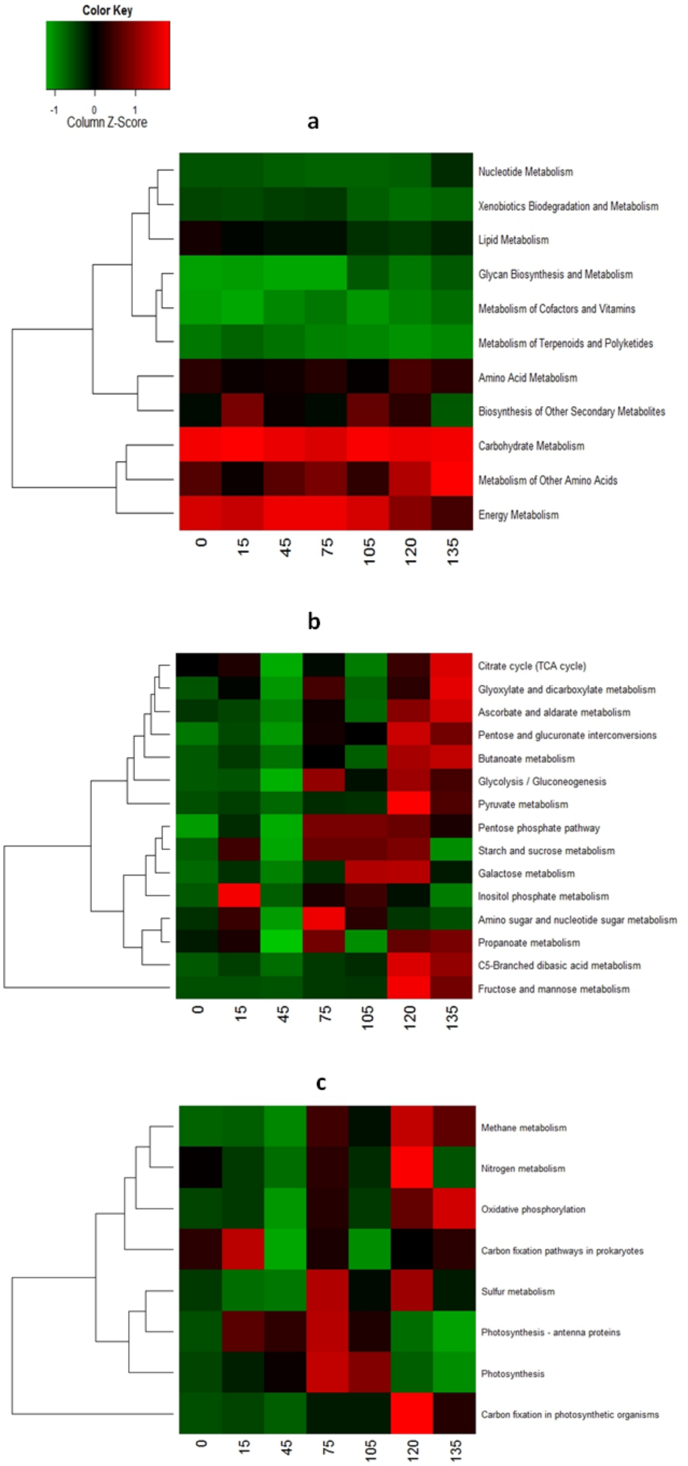

S000001
S000002
S000003
S000004
S000005
S000006
S000007
S000008
S000009
S000010
S000011
S000012
S000013
S000014
S000015
S000016
S000017
S000018
S000019
S000020
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
O. sativa
0.5
1
1.5
2
2.5
3
3.5



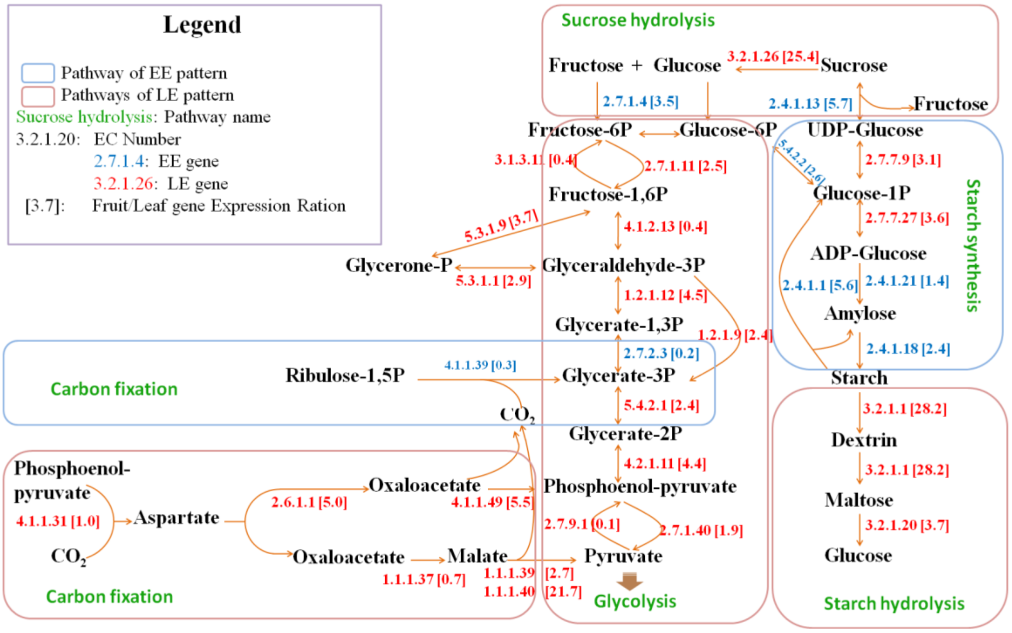
Supplementary Figure S5. Alignment of 20 largest scaffolds of the *P. dactylifera* assembly with other monocots chromosomes. The color bar indicates Ks values. The red arrow indicates the syntenies separated from S000001 and maintain the same pattern in other monocots. The red box indicates syntenies between monocot chromosome and the largest scaffolds in the *P. dactylifera* genome assembly.



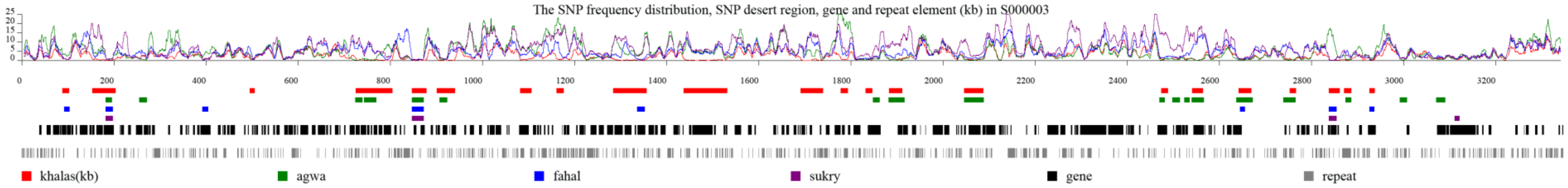
Supplementary Figure S6. Hierarchical cluster analysis of DEG gene expression based on log ratio RPKM data. Two-dimensional hierarchical clustering classifies 4,134 differential expression profiles into four expression cluster groups according to the similarity of their expression profiles.Purple bar represents for the up-regulated group; Blue represents for the down-regulated group; Green represents for the irregulated group 1; Red represents for the irregulated group 2.



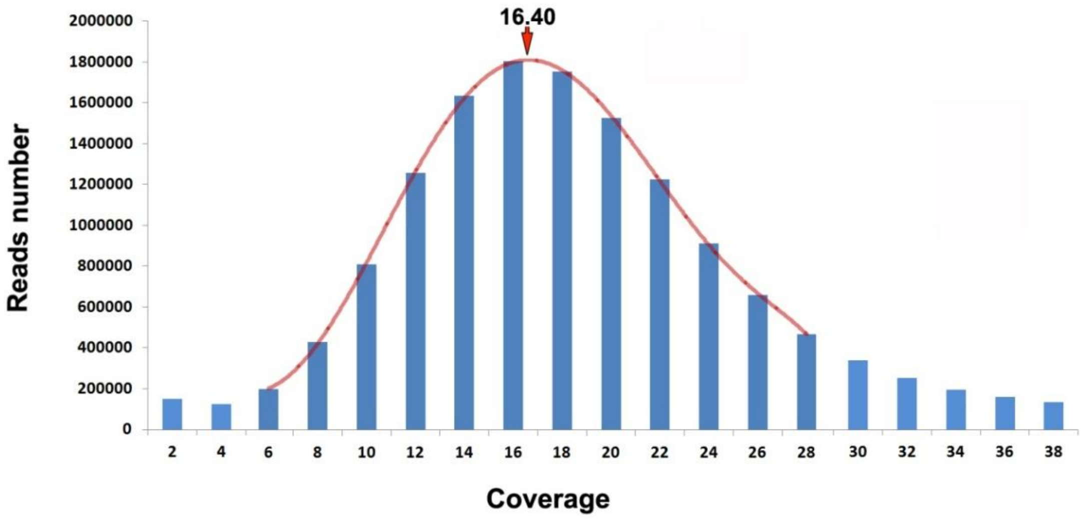
Supplementary Figure S7. Expression profiles of metabolic pathways. Important metabolism pathways based on KEGG annotation were selected and clustered based on scaled gene expression level. The date palm fruit samples of 0, 15, 45, 75, 105, 120, and 135 days post pollination were used for the transcriptomic study. We show here: **(a)** KEGG 3rd level pathways related to nutrition metabolism, (**b)** Carbon Carbohydrate metabolism pathways, and (**c)** Energy metabolism pathways.



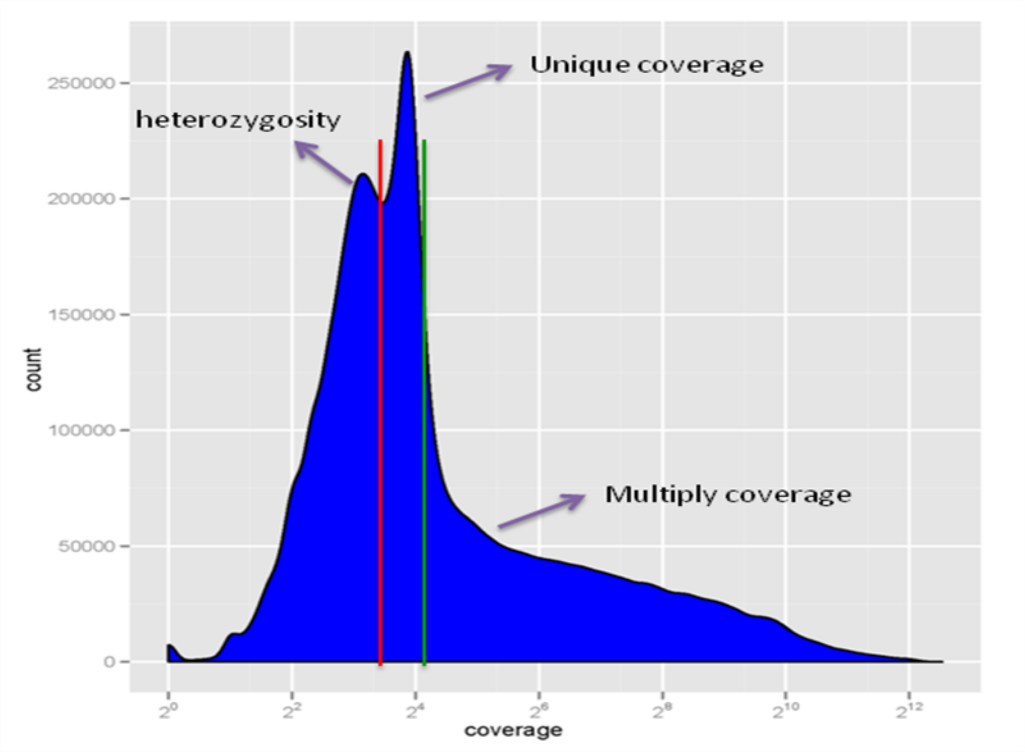
Supplementary Figure S8. Sugar metabolism pattern based on transcriptomic data in fruit developmental stages. Gene expressions at the early stages (EE) of fruit development and at the late stages (LE) of fruit development are summarized here. The fruit/leaf expression ratio is calculated using maximal expression level among 7 stages of the fruit development. More details are shown in Supplementary Data 2.



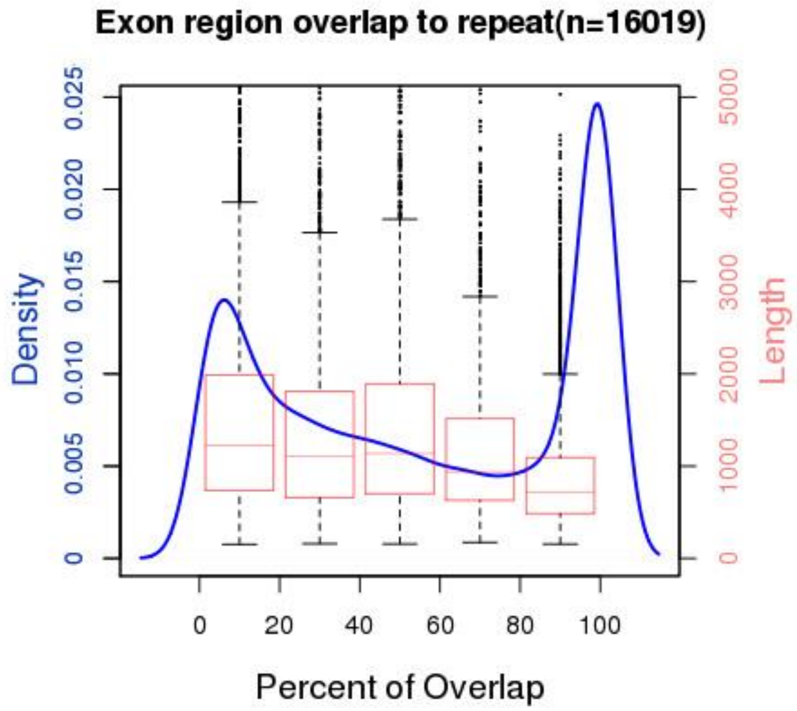
Supplementary Figure S9. An example of SNP desert regions found in the four *P. dactylifera* varieties. Horizontal colored bars below the plot of SNP rate indicate SNP desert regions.



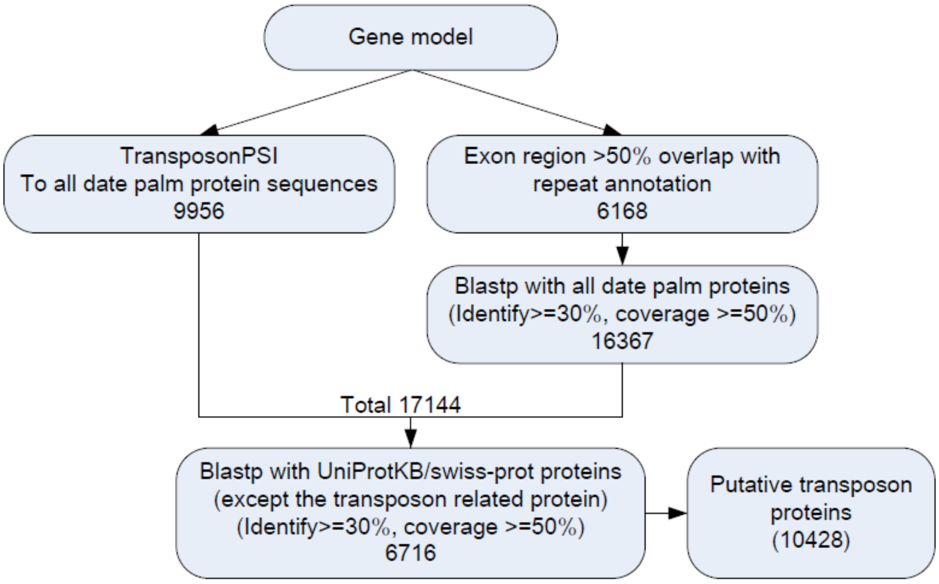
Supplementary Figure S10. Genome size estimation based on repeat values of Roche/454 sequence reads masked by 20-mers.



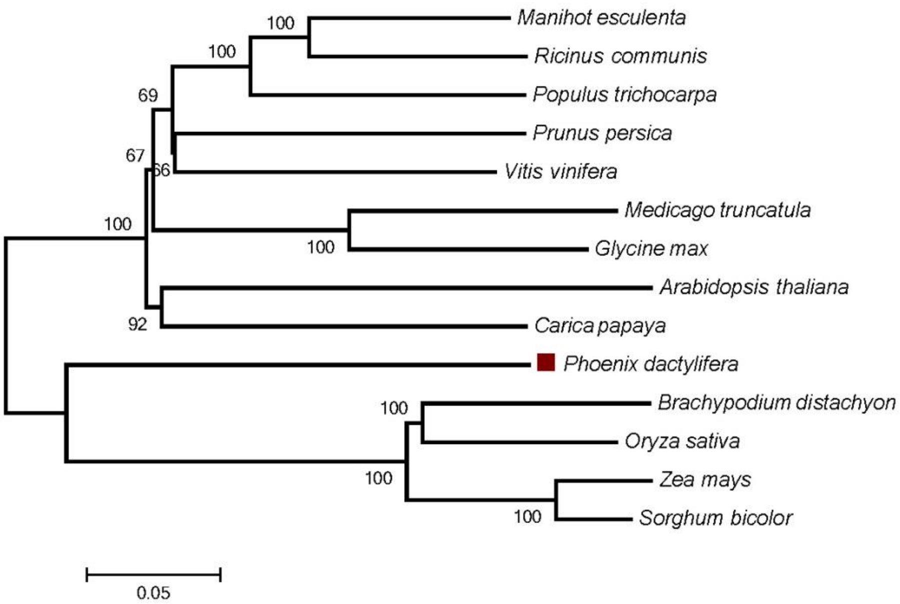
Supplementary Figure S11. Genome size estimation based on the coverage of Newbler-assembled contigs.



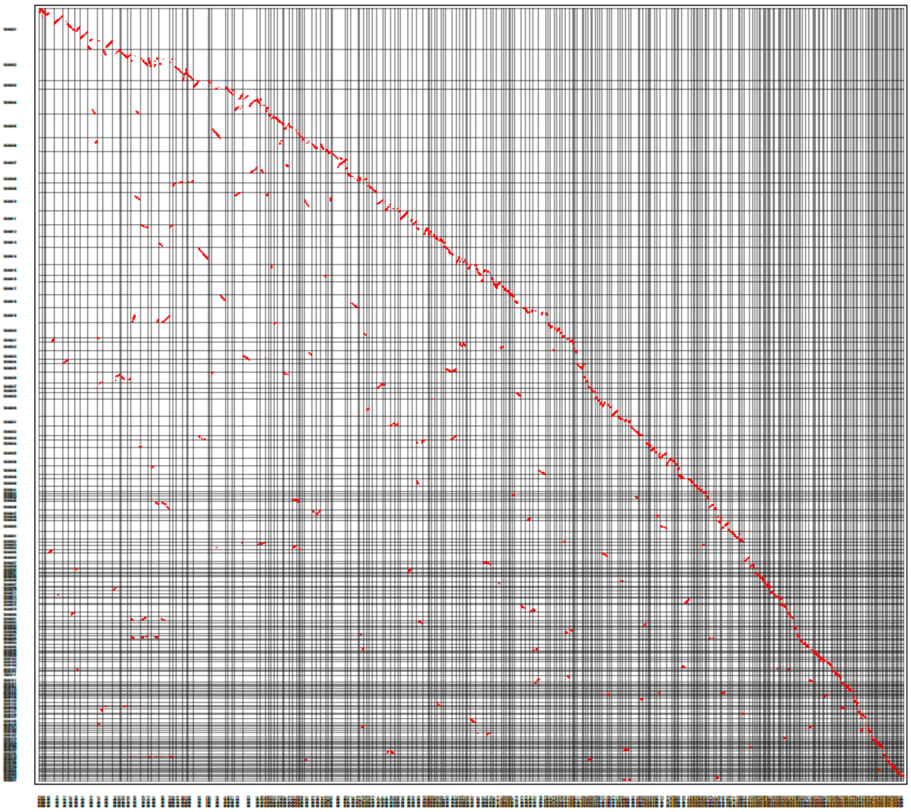
Supplementary Figure S12. Status of gene models that overlap with repeat annotation.


Gene model
TransposonPSI
To all date palm protein sequences
9956
Exon region >50% overlap with
repeat annotation
6168
Blastp with all date palm proteins
(Identify>=30%, coverage >=50%)
16367
Total 17144
Blastp with UniProtKB/swiss-prot proteins
(except the transposon related protein)
(Identify>=30%, coverage >=50%)
6716
Putative transposon
proteins
(10428)

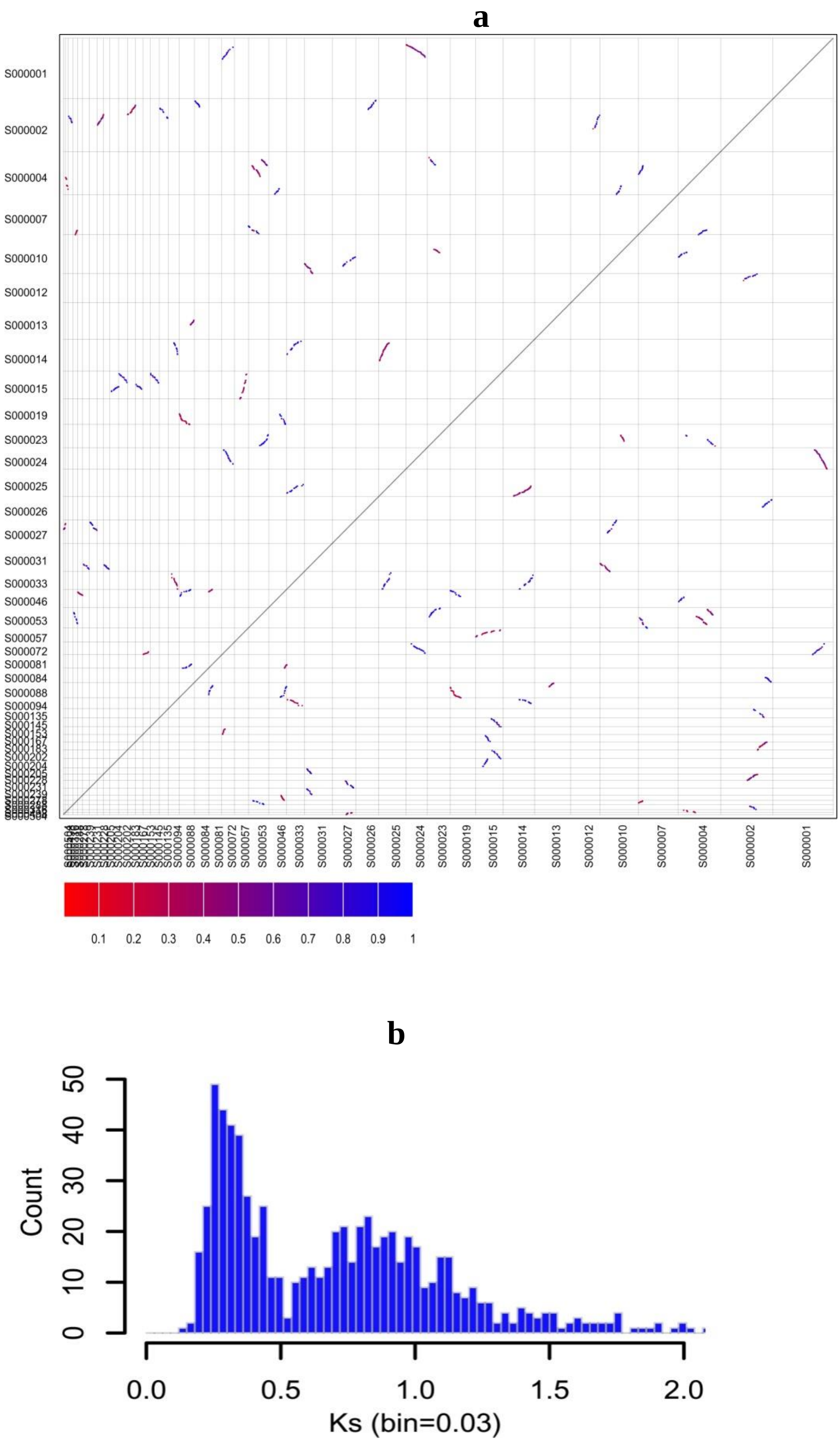
Supplementary Figure S13. Identification of putative TE-containing proteins.



Supplementary Figure S14. Phylogenetic tree of 13 sequenced plant genomes and *P. dactylifera* genome. We select the genes with only single copy in each orthology group. Amino acid sequence of genes were concatenated to structure the NJ tree by MEGA5.



Supplementary Figure S15. Collinear blocks in *P. dactylifera* genome assembly



Supplementary Figure S16. Orthologous relationships (a) and Ks distribution of genes (b) in multi-syntenic regions. We chose the blocks size with more than 10 gene pairs in this plot. Scaffolds length depends on all its gene numbers.

Supplementary Tables

Supplementary Table S1. An overview of sequence data for *P. dactylifera* genome assembly

Library Type	Variety	No.of Librariy.	Insert Size (kb)	Sequencing Platform	Sequencing Reads (M)*	Usable Data (Gb)**	Coverage
Fragment	Khalas	4	0.5-0.8	GS FLX	31.40	11.60	17.3
BAC	Khalas	2	100+	3730XL	0.05	NA	3.8***
Mate pair	Khalas	9	1-8	SOLiD 3(4)	1,931.98	73.94	122.1
Mate pair	Sukry	5	0.6-6	SOLiD 4	820.02	37.14	61.4
Mate pair	Agwa	8	0.4-6	SOLiD 4	1,168.00	53.38	88.2
Mate pair	Fahal	9	0.4-15	SOLiD 4	1,226.94	54.92	90.7
RNA-seq	Khalas	7	0.3	SOLiD 4	395.67	19.78	NA
cDNA	Khalas	14	0.5-0.8	GS FLX	14.44	4.98	NA

Note: * and ** indicate numbers of reads used for the assembly and uniquely mapped data, respectively. ***Indicates clone coverage, when the insert size of the BAC clones is assumed ~100 kb in average.

Supplementary Table S2. Genome assembly validation based on BAC and fosmid sequences

	BAC	**Fosmid**
Number	10	6
Total length (bp)	1,276,062	217,445
Matched (%)	1,257,235 (98.52%)	209,809 (96.49%)
Unique aligned (%)	1,007,504 (78.95%)	73,759 (33.92%)
Repeat aligned (%)	249,731 (19.57%)	136,050 (62.57%)
Aligned in single scaffolds (%)	1,190,835 (93.32%)	110,797 (50.95%)

Supplementary Table S3. Genome assembly validation based on EST data. ESTs which generated from 8 *P. dactylifera* tissues, the Qatar solexa transcriptome data, and GS FLX data generated by CNRS were mapped to our genome assembly, using BLAT with the parameters "alignment identity ≥ 95%", and "alignment coverage ≥ 90%, coverage ≥ 50%" , respectively.

Dataset	**Number**	**Total length (bp)**	**>90% of sequence covered by one scaffold**		**>50% of sequence covered by one scaffold**	
			Number	**Percent**	**Number**	**Percent**
Roche/454 Date Palm assembled ESTs:						
>100 bp	67,867	102,241,549	57,899	85.3%	65,328	96.3%
>200 bp	67,642	102,201,947	57,757	85.4%	65,141	96.3%
>500 bp	62,451	100,112,022	53,676	85.9%	60,432	96.8%
>1000 bp	41,824	84,967,502	35,871	85.8%	40,741	97.4%
Qatar Solexa Date Palm assembled ESTs:						
>100 bp	28,889	30,553,282	24,357	84.3%	26,236	90.8%
>200 bp	28,664	30,512,558	24,177	84.3%	26,047	90.9%
>500 bp	22,344	28,140,076	19,191	85.9%	20,707	92.7%
>1000 bp	11,730	20,391,890	10,236	87.3%	11,012	93.9%
CNRS Date Palm ESTs:						
>100 bp	37,048	37,848,924	28,704	77.5%	35,714	96.4%
>200 bp	37,048	37,845,043	28,691	77.4%	35,690	96.3%
>500 bp	34,615	36,790,333	26,814	77.5%	33,357	96.4%
>1000 bp	14,039	21,581,452	10,874	77.5%	13,545	96.5%

Supplementary Table S4. General statistics of *P. dactylifera* genome assembly and annotation.

	Name	Parameter
Genome	Genome size (bp)	558,022,834
	Total Contigs(>500 bp)	82,354
	Max Contig Length (bp)	4,533,682
	Genome GC%	36.4
Coding region	Gene number	41,660
	Gene Total Length (bp)	165,435,063 (29.6%)
	Gene Avg. Length (bp)	3,971 (Median: 2,232; Max: 73,150; Min:150)
	Gene Density	1.58/20 kb
	Gene GC%	39.3
	CDS Total Length (bp)	52,648,171 (9.4%)
	CDS Avg. Length (bp)	1,200 (Median: 960; Max: 15,312; Min: 69)
	CDS GC%	48.6
	Exon Total Length (bp)	51,580,262 (9.2%)
	Exon Avg. Length (bp)	273 (Median: 145)
	Exon per Gene	4.6
	Exon GC%	48
	Intron Total Length (bp)	110,854,801 (19.9%)
	Intron Avg. Length (bp)	749 (Median: 288)
	Intron GC%	35.2
Noncoding region	tRNA Number	414
	tNRA total length (bp)	31,133
	rRNA Number	677 (5S=219; 45S=458)
	rRNA total length (bp)	247,686
	snoRNA Number	62 (H/ACA 44;C/D 18)
	snoRNA length (bp)	7,199 (H/ACA 4,008;C/D 3,191)

Supplementary Table S5. Microsatellite annotation of *P. dactylifera* genome

	n=6-11		n>11	
	% GC	% genome	% GC	% genome
Mononuclieotides	7.88	1.76	13.23	0.09
Dinuclieotids	33.25	0.14	30.23	0.11
Trinuclieotides	35.67	0.04	13.18	0.01
Tetranulieotides	21.30	0.01	27.57	0.00
All periods		**1.94**		**0.21**

Supplementary Table S6. Repetitive sequences annotation of *P. dactylifera* genome assembly. We first sequenced 140 randomly selected BACs using Roche/GS FLX platform and generated 8,746 contigs with a total length of 17,631,122 bp (N50 = 10,345 bp). Combining with other reads from capillary sequenceing, we assembled repeat consensuses using RepeatScout [52], LTR-finder [53], and MITE-hunter [54].

Class	Sub-class	Content (%)*		Copy number	
		in assembly	in genome	in genome	STDEV
retrotransposon	Ty1/Copia	5.53	14.03	324,380	6,032
	Ty3/Gypsy	1.93	4.17	94,881	3,692
	LINE	0.78	0.46	12,445	314
	unknown	1.83	3.33	76,691	1,834
transposon	hAT/Ac	0.14	0.29	6,417	287
	CACTA/EnSpm	0.03	0.03	1,022	284
MITEs		0.3	0.18	6,617	230
centromere		0.01	0.01	585	64
teleomere		0.02	0.4	7,929	220
rRNA	5S	0.01	0.1	3,652	121
	45S	0.02	0.76	14,633	830
gene families		0.55	1.26	29,373	755
unknown repeats		10.16	13.38	512,638	18,289
Total		**21.31**	**38.41**		

* indicates repeat annotation for both assembled scaffolds (in assembly) and GS FLX reads (in genome). STDEV: standard deviation.

Supplementary Table S7. LTR retrotransposons Ty1/Copia are more than Ty3/Gypsy in the *P. dactylifera* assembly but not in other sequenced plant genomes (% of genome bp).

Sequenced plant genome	Ty1 (%)	Ty3 (%)	Ty1/Ty3
Selaginella moellendorffii[55]	2.7	21.1	0.13
Oryza sativa [13,14]	2.47	12.03	0.21
Sorghum bicolo [56]	5.18	19	0.27
Zea may [57]	21.75	37.73	0.58
Brachypodium distachyon [58]	4.9	16.1	0.30
*Phoenix dactylifera**	10.15	4.82	**2.11**
*Cocos nuncifera**[#]	16.53	5.33	**3.10**
*Areca catechu**[#]	9.6	7.74	**1.24**
Arabidopsis thaliana[59]	1.4	5.2	0.27
Malus domestica[60]	5.5	25.2	0.22
Populus trichocarpa[61]	1.6	4.9	0.33
Vitis vinifera[62]	4.8	14	0.34
Glycine max[63]	10.7	25.3	0.42
Ricinus communis[64]	4.77	11.45	0.42
Cajanus cajan[65]	6.22	11.79	0.53
Medicago truncatula[51]	4.1	5.7	0.72
Lotus japonicus[66]	7.16	8.81	0.81
Carica papaya[67]	5.5	27.8	0.20
Fragaria vesca[68]	4.58	5.99	0.76
Brassica rapa[69]	2.84	3.12	0.91

*Based on randomly selected Roche 454 reads scanned with Repbase using RepeatMasker (http://repeatmasker.org), we randomly sampled 2,760,440 sequences (Roche/454 reads) from date palm, 2,722,940 from coconut palm, and 1,239,603 from areca palm, and scanned them for copia and gypsy sequences based on information from Repbase library (Version 15.1)[70]. [#] indicates the unpublished in-house sequence data.

Supplementary Table S8. LEA gene family in the *P. dactylifera* and other sequenced genomes

Species	LEA_1	LEA_2	LEA_3	LEA_4	LEA_5	LEA_6	SMP	Dehydrin	Total	P-value* for LEA2
Phoenix dactylifera	3	62	5	3	2	2	4	3	84	NA
Brachypodium distachyon	5	45	5	7	2	1	6	7	78	0.015
Oryza sativa	4	52	4	4	1	2	6	7	80	0.11
Sorghum bicolor	4	46	7	6	2	1	7	3	76	0.036
Zea mays	3	60	7	5	2	1	4	5	87	0.24
Arabidopsis thaliana	3	40	4	9	2	3	6	7	74	0.0048
Carica papaya	2	19	3	3	2	2	3	3	37	0.0082
Glycine max	5	75	8	5	5	2	5	3	108	0.25
Medicago truncatula	4	21	3	1	2	3	5	2	41	0.0060
Populus trichocarpa	3	47	4	8	1	2	2	2	69	0.22
Ricinus communis	1	27	3	4	2	2	4	4	47	0.027
Theobroma cacao[71]	3	37	3	4	3	0	4	4	58	0.10
Vitis vinifera	3	20	1	5	3	0	3	2	37	0.016
Selaginella moellendorffii	0	7	0	7	2	0	4	0	20	0

* Comparison between other plants to *Phoenix dactylifera* by Z test.

Supplementary Table S9. An overview of fruit transcriptomic data

Libraries	Total reads	Mapped reads	Unique mapped reads	Expressed mRNAs
0DPP	76,515,395	53,825,507	11,396,571	17,464
15DPP	83,950,995	59,009,944	9,903,023	17,570
45DPP	77,132,786	53,999,668	12,355,849	15,920
75DPP	74,369,980	55,665,445	14,302,703	18,329
105DPP	71,372,371	52,916,490	12,403,501	16,349
120DPP	83,506,278	62,428,925	14,407,421	15,709
135DPP	98,713,383	70,080,905	10,247,674	15,978
Green leaf	94,042,176	45,735,383	14,697,605	16,022

Supplementary Table S10. Gene ontology (GO) enrichment analysis of up-regulated genes in date palm fruits

GO ID	GO Term	P-value	FDR*
GO:0006094	gluconeogenesis	2.6E-18	1.2E-16
GO:0044262	cellular carbohydrate metabolic process	1.7E-04	7.9E-03
GO:0044283	small molecule biosynthetic process	1.0E-03	4.8E-02

*FDR: False discovery rate

Supplementary Table S11. Gene ontology (GO) enrichment analysis of down-regulated genes in date palm fruits

GO ID	GO Term	P-value	FDR
GO:0060255	regulation of macromolecule metabolic process	1.9E-14	1.6E-12
GO:0050789	regulation of biological process	1.6E-10	1.4E-08
GO:0006350	transcription	1.8E-10	1.6E-08
GO:0065007	biological regulation	1.9E-10	1.7E-08
GO:0006108	malate metabolic process	2.0E-08	1.8E-06
GO:0007018	microtubule-based movement	1.3E-07	1.2E-05
GO:0043648	dicarboxylic acid metabolic process	2.4E-06	2.1E-04
GO:0051252	regulation of RNA metabolic process	1.4E-05	1.2E-03
GO:0016614	oxidoreductase activity, acting on CH-OH group of donors	2.4E-08	1.3E-06
GO:0016798	hydrolase activity, acting on glycosyl bonds	4.3E-04	2.3E-02
GO:0005576	extracellular region	2.2E-07	1.1E-06

Supplementary Table S12. SNP density (per kb) among varieties (both intervariety and introvariety in a 10-kb window and 1-kb step).

Varieties	Origin	Collection site	Coverage	SNP density (SNP/kb)										
				Khalas	Sukry	Fahal	Agwa	Khalt	AlrijalF	KhalsFx	Medjool	Medjool BC4	Deglet Noor	Deglet Noor BC5
Khalas	KSA	Al-Hssa, KSA	122	**2.57**	3.01	2.89	3.05	2.09	2.46	1.7	2.8	2.55	3.03	2.55
Sukry	KSA	Al-Qasim, KSA	61		**6.24**	3.11	3.65	3.59	3.66	3.37	4.29	4.11	4.37	4.06
Fahal (M)	KSA	Al-Hssa, KSA	91			**5.51**	3.59	3.49	3.46	3.26	4.2	4.02	4.28	3.96
Agwa	KSA	Al-Medina, KSA	88				**6.1**	3.64	3.77	3.42	4.31	4.14	4.41	4.08
Khalt (M) *	Qatar	Qatar	12					**4.25**	2.71	2.23	3.12	2.89	3.24	2.87
AlrijalF*	Qatar	Qatar	15						**6.63**	2.58	3.44	3.22	3.48	3.17
KhalsFx *	KSA/USA	California, USA	39							**3.85**	2.88	2.66	3.13	2.71
Medjool *	North Africa	California, USA	28								**5.45**	2.71	3.56	3.31
Medjool BC4(M)*	California	California, USA	27									**4.44**	3.4	3.08
Deglet Noor *	North Africa	California, USA	24										**6.55**	2.7
Deglet Noor BC5(M)*	California	California, USA	24											**5.07**

*These raw data were downloaded from the SRA of GenBank. M indicates male variety. Only major alleles were considered as intervariety SNPs.

Supplementary Table S13. Indel density of four varieties

Varieties	Indels	Indel density (indels/kb)
Khalas	56,463	0.100257
Agwa	138,270	0.245515
Fahal (M)	86,495	0.153582
Sukry	113,057	0.200746

Supplementary Table S14. SNP desert size with different sliding windows.

The minimum SNP desert size	Total SNP desert length	Percent of Khalas genome	N50 size
1 kb	131,465,000 bp	21.7%	32,000 bp
5 kb	121,052,000 bp	20.0%	37,000 bp
10 kb	108,259,000 bp	17.9%	43,000 bp
20 kb	86,267,000 bp	14.3%	56,000 bp

Supplementary Table S15. GO cluster for genes in *P. dactylifera* unique families.

GO id	Gene no.	GO items	P value
0004540; 0004518; 0016891; 0004523; 0004519; 0016893; 0004521; 0016788; 0016787; 0008784	435	MF: Catalytic activity	0
0006139; 0006278; 0044238; 0015074; 0044267; 0019538; 0006259; 0044260; 0043283; 0006260; 0006333; 0006325; 0006323; 0051276; 0006996	2782	BP: DNA/RNA Metabolic process	0
0043170	2701	BP: Macromolecule metabolic process;	0
0046872; 0043167; 0008270; 0046914	1429	MF: Ion binding	0
0044464; 0005623	2003	CC: Cell part	0

Supplementary Table S16. Reference summary of *P. dactylifera* genome size estimation

Prime Estimate	Genus	Species	Chr No.	Ploidy	Estimation method	1C (Mb)	Original Reference
	Phoenix	*dactylifera*	36	2	FC:PI	680	Al-Dous et al., 2011[6]
Prime	*Phoenix*	*canariensis*	36	2	FC:PI	880	Suda et al., 2005[72]
Prime	*Phoenix*	*dactylifera*	36	2	Fe	929	Olszewska and Osiecka, 1982[73]
	Phoenix	*dactylifera*	36	2	FC:PI	1,296	Zonneveld et al., 2005[74]
Prime	*Phoenix*	*theophrasti*	36	2	Fe	1,296	Röser et al., 1997[75]
Prime	*Phoenix*	*rupicola*	36	2	Fe	1,467	Röser et al., 1997
Prime	*Phoenix*	*roebelenii*	36	2	Fe	1,491	Röser et al., 1997

Supplementary Table S17. Clustered top30 GO enrichment results (p<0.0001) for the genes in multi-synteny regions. BP: biological_process; MF: molecular_function; CC: cellular_component.

GO id	Gene no.	GO items	P value
0065007; 0050794;0050789; 0045449; 0019219;0010468; 0031323; 0019222;0006350; 0006355; 0050790;0006351; 0065009; 0032774	153	BP: Regulation of cellular process/Transcription/ Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process/ Gene expression/ Transcription, DNA-dependent/RNA biosynthetic process	1.37e-33
0006259; 0003677	84	BP: DNA metabolic process MF: DNA binding;	5.53e-30
0010467	159	BP: Gene expression	2.27e-23
0015074; 0043234; 0032991	26	BP: Protein complex; DNA integration CC: Macromolecular complex	7.48e-18
0006260; 0042578	19	BP: DNA replication MF: Phosphoric ester hydrolase activity	4.84e-14

Supplementary Table S18. Numbers of NBS genes found in *P. dactylifera* and other plant genomes. Pd: *Phoenix dactylifera*; Zm: *Zea mays*; Sb: *Sorghum bicolor*; Os: *Oryza sativa*; Tc: *Theobroma cacao*; Pt: *Populus trichocarpa*; Vv: *Vitis vinifera*; Mt: *Medicago truncatula*; At: *Arabidopsis thaliana*.

Gene family	Pd	Zm	Sb	Os	Tc	Pt	Vv	Mt	At
TIR-NBS-LRR	NA	NA	NA	NA	8	78	97	118	93
CC-NBS-LRR	19	72	132	276	82	120	203	152	51
NBS-LRR	16	23	52	182	104	132	159	NA	3
NBS	69	10	34	27	53	62	36	328	1
CC-NBS	40	24	27	23	46	14	26	25	5
TIR-NBS	NA	NA	NA	NA	4	10	14	38	21
Total NBS-LRR genes	35	95	184	458	194	330	459	NA	147
Total NBS genes	144	129	245	508	297	416	535	661	174

Supplementary Table S19. Energy and sugar metabolism related genes in plant genomes. We identified metabolism pathways based on *P. dactylifera* (Pd) gene models as well as information from other representative species, including *O. sativa* (Os), *S. bicolor* (Sb), *A. thaliana* (At), *P. trichocarpa* (Pt), and *V. vinifera* (Vv), using the KAAS online search service (http://www.genome.ad.jp/tools/kaas/). The orthologs related to energy and sugar metabolisms were extrtacted

Energy and sugar related pathways	Gene number					
	Pd	Os	Sb	At	Pt	Vv
Amino sugar and nucleotide sugar metabolism	31	34	32	29	43	33
Carbon fixation in photosynthetic organisms	43	45	41	43	51	31
Fructose and mannose metabolism	17	19	16	16	19	13
Galactose metabolism	18	16	18	19	24	15
Glycolysis/Gluconeogenesis	90	97	96	89	110	76
Pentose and glucuronate interconversions	17	16	13	35	36	22
Photosynthesis	50	55	41	51	56	32
Photosynthesis - antenna proteins	19	15	16	20	22	11
Starch and sucrose metabolism	105	111	100	112	145	102

Supplementary Notes

Supplementary Note 1. Genome size estimation

Although there are several studies that estimate the genome size of *P. dactylifera* as ranging from 680 Mb to 1,491 Mb, all are based on results from flow cytometry (Supplementary Table S16), and the physical size of the *P. dactylifera* genome remains elusive. We estimated the genome size in three different ways. First, based on all of the Roche/454 GS FLX reads, we calculated the frequency of every 20-mer sequence for all reads in a hash table and estimated the coverage based on the frequencies of median repeat values (Supplementary Fig. S10). The estimated genome size based on this procedure was ~638.2 Mb. Second, we estimated the genome size based on the coverage distribution of the Newbler contigs (Supplementary Supplementary Fig. S11). We classified the contigs into heterozygous, unique, and multiple, based on coverage. The heterozygous contigs, whose collective size is half of the heterozygous contigs, are 104,212,972 bp. The unique regions cover 438,982,154 bp. The length of the regions covered by multiple contigs was calculated based on the total read lengths mapped in the contigs and the average coverage of the unique regions. We estimated that the average collapsed ratio in these regions is approximately 1:5.56 (approximately 20.18 Mb in our assembly). The actual length of the multi-covered regions is 171,206,668 bp (“the pyrosequencing reads in the multi-covered regions + the unassembled repeat reads” / the average_unique_coverage). The genome size was thus calculated as follows: 104,212,972*0.5 + 438,982,154 + 171,206,668 = ~662.3 Mb. Third, we estimated the repetitive sequence content. The total repeat content was estimated to be 38.41% by the genome average, but we only have 21.31% repeats in the 558.02 Mb genome assembly. If the missing part of the repeat content is factored in, we obtain a genome size of ~713.0 Mb. Taking all estimates into consideration, we believe that the *P. dactylifera* genome is ~671.2 Mb, which is close to the estimate reported by Al-Dous et al[6].

Supplementary Note 2. Gene prediction and annotation

We predicted 50,132 *ab initio* gene models using Fgenesh++[12] with monocot parameters (contig size > 500 bp). A total of 40,588 of the gene models contain polyA signals, and 42,080 of the gene models have both start and stop codons. We used all plant protein sequences from the NCBI Refseq databases (release 44)[76] and Swiss-Prot (2010_11)[77,78] for splicing variant alignments, using Spaln[79,80] with default parameters. We also constructed a series of EST libraries (sequences generated from both the SOLiD and pyrosequencing platforms) from different tissues (male and female flowers, flower buds, green and yellow leaves, roots, and fruits) and assembled them into 67,651 transcription units with an N50 size of 1,911 bp[11]. PASA[81] was used for EST alignments. Transcripts were assembled with a maximal intron length of

3,000 bp, yielding 66,738 (98%) transcription units with their spliced variants aligned to the genome assembly. We selected 671 intact and highly credible transcripts for EVM[82] training by manual inspection (using Blastp with the rice and Arabidopsis records in UniProt with E < 1e-5 and 90% identity for confirmation). EVM combined all predictions, and the output was patched with two runs of PASA for the small exon and potential UTR regions.

The output models with premature termination were discarded, although some of them have supported expression tag reads. Overlapping models with repeat annotations at exon regions were selected and show a different pattern in the length and percent of the overlap (Supplementary Figure S12). Gene models with smaller degrees of overlap to repeat annotations were preferred to numbers of gene families, and the parts that were fully contained in the repeat regions were annotated as TE proteins or ORFs in TE regions with a smaller length distribution. We used a 50% exon overlap as a cutoff to distinguish the models with TE proteins. Some gene models with a 50% repeat overlap were also discarded if there were no homologous proteins in NCBI or functional domains. The remaining repeat related proteins were approximately 6,843 and more than 78% (5356) of the models had functional domains related to transposition. We used these protein sequences to with all known date palm proteins, and combined the outputs of TransposonPSI to blast against all UniProtKB/Swiss-Prot proteins to retrieve the functional genes. There were also some unique date palm non-TE protein sequences that could not be identified. We estimate that the total number of TE proteins in our annotation is no more than 10,428 (Supplementary Fig. S13). Finally, we identified 41,660 gene models (42,957 isoforms) in 10,363 scaffolds of the date palm genome. Approximately 84% of these gene models (35,106) have cDNA support (31,493, > 80% coverage) or SOLiD RNA-seq reads support (24,501, > 5 unique-mapped reads in at least one tissue). The predicted gene models were investigated (Supplementary Fig. S3); however, we did not observe any gene models in contigs shorter than 500 bp. We thus excluded sequences smaller than 500 bp from further analysis.

In the annotation effort, we transferred GO function IDs using InterProScan v4.7[83] (50% exon overlap, Blastp E < 1e-10, identity > 50%, and coverage > 50%) and enzyme EC IDs using the KEGG annotation system[43] (default parameters), and finally annotated 31,943 gene models. We also identified orthologous genes in *P. dactylifera*, *A. thaliana*, *O. sativa*, *S. bicolor*, and *V. vinifera* using amino acid sequences from Phytozome v7.0 and only kept the longest sequences. All-to-all Blastp was performed with E < 1e-5. OrthoMCL 2.0.2[84] was used to construct the orthologous groups using the best reciprocal hit approach. We selected the genes with single copies in each orthology group. The amino acid sequences of genes were concatenated to obtain the NJ (neighbor-joining) tree using the MEGA5 program. *P. dactylifera* is a prior speciation to the grass family after the divergence between monocots and dicots (Supplementary Fig. S14) .

We also predicted RNA genes including 414 tRNA (38 tRNA pseudogenes), 219 5S rRNA, 458 45S rRNA, 44 box H/ACA snoRNA, and 18 box C/D snoRNA genes (Supplementary Data 3).

Supplementary Note 3. Genome-wide duplication

We used MCscan[85] to define syntenic regions between *P. dactylifera* and other plants. One isoform for each gene was selected for this exercise. The best five mutual hits of the Blastp results were used as MCscan inputs. Only the syntenic segments that had more than five gene pairs were considered for 4DTv calculation. Raw 4DTv values were corrected for possible multiple transversions at the same site. Ks values were calculated using the KaKs_Calculator[86] with the NG (Nei-Gojobori) model. The Kernel density estimation of 4DTv distance, or Ks, was performed using a Perl module (Supplementary Fig. S4). There are 15,202 (36% of the total) genes located in the synteny blocks. Considering the scattered assembling, it is quite a large fraction of the *P. dactylifera* genome. If we plot all the contigs/scaffolds that contain collinear blocks, we see that almost all of the collinear blocks are in one-vs.-one matches (Supplementary Fig. S15). We also assessed co-linearities with other monocot plants, and a few examples are shown in Supplementary Fig. S5.

The regions with multi-syntenic regions were also selected to trace the second peak in the Ks plot (Supplementary Fig. S16). The 745 gene pairs in these regions as well as GO cluster results are consistent with previous results. Most retained genes after GWD (genome-wide duplication) or segmental duplications, have the function of transcription regulation (Supplementary Table S17).

Supplementary Note 4. Identification of NBS family in the *P. dactylifera* genome

The genes that encode NBS (nucleotide-binding site) proteins play a key role in plant pathogen sensing, host defence, and cell cycle progression[87]. The NBS-LRR gene family is rather abundant in plant genomes, ranging from 0.6% to approximately 2% of the total genes[87-89]. The resistance protein (R protein) gene family can be subdivided into different groups based on the structure of N-terminal and C-terminal domains. The N-terminal domain either has a CC (coiled-coil) /TIR (Toll-interleukin receptor) motif or not. While the C-terminal domain can either contain a LRR (leucine-rich repeat) motif or not[90].

The *P. dactylifera* protein sequences were screened using Hidden Markov Models (HMMs) with Pfam NBS (NB-ARC) family PF00931 domains (E-value cutoff of 1.0)[91] using hmmsearch v3 software[92]. Sequences of non-conserved domains were manually removed, and 144 sequences were retained. The 144 predicted NBS-encoding amino acid sequences were used to detect TIR domains using an HMM model with Pfam TIR PF01582 (E-value cutoff 1.0) domains and LRR motifs in the C-terminal domains. The HMM models Pfam LRR_1 (PF00560), LRR_2 (PF07723) and LRR_3 (PF07725) (E-value cutoff 1.0)[91] were employed to screen the predicted NBS-encoding amino acid sequences. CC (coiled-coil) motifs were screened using Paircoil2[93] (P-score cutoff of 0.025). Totally, 144 non-redundant NBS-encoding genes were identified and manually validated, which account for approximately 0.35% of the gene models in the *P. dactylifera* genome. In other monocot plants such as *Z. may, S. bicolor,* and *O. sativa***,** the percentage is 0.4%, 0.68% and 1.35%, respectively[94]. In some eudicot plants, for instance, *T. cacao***,** *P. trichocarpa*, *V. vinifera*, *M. truncatula*, and *A. thaliana*, the percentage of NBS-encoding genes is 0.9%, 1%, 1.8%, 1.2% and 0.7%, respectively (Supplementary Table S18)[88,89,95]. The TIR-NBS-LRR and TIR-NBS orthologous genes are absent in the *P. dactylifera* genome as well as in the other three monocot plants, and the CC-NBS-LRR gene is comparatively less frequent in the *P. dactylifera* genome than in other monocot plants.

Supplementary Note 5. Mining genes related to energy and sugar metabolism in representative plant genomes

We identified metabolic pathways based on well-defined gene models of *P. dactylifera* and information from other representative species, including *O. sativa, S. bicolor, A. thaliana, P. trichocarpa,* and *V. vinifera*, using the KAAS online search service (http://www.genome.ad.jp/tools/kaas/). The orthologs related to energy and sugar metabolism were extracted (Supplementary Table S19). We identified 148 orthologs representing 390 genes in the *P. dactylifera* genome assembly. The duplication and expansion of carbon and sugar metabolism-related genes appears to be correlated with unique sugar metabolism pathways. For instance, ribose-5-phosphate isomerase A has five copies in *P. dactylifera*, but only two and three orthologs are found in *O. sativa* and *S. bicolor,* respectively. We also found 14 pyruvate kinase genes in *P. dactylifera*, but detected only 8 in rice and 10 in *S. bicolor*. Another example is *P. dactylifera* pyruvate orthophosphate dikinase, which has two tandem copies that are 92.4% identical in amino acid sequence, corresponding to a single copy in maize and sorghum. Similarly, two pairs of adjacent phosphoglycerate kinase genes recruited new paralogs via fragmental duplication.

Supplementary References

52 Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358 (2005).

53 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265-268 (2007).

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

55 Banks, J. A. *et al.* The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960-963 (2011).

56 Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).

57 Vielle-Calzada, J. P. *et al.* The *Palomero* genome suggests metal effects on domestication. *Science* **326**, 1078 (2009).

58 The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768 (2010).

59 The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).

60 Velasco, R. *et al.* The genome of the domesticated apple (*Malus* x *domestica Borkh.*). *Nat. Genet.* **42**, 833-839 (2010).

61 Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (*Torr. & Gray*). *Science* **313**, 1596-1604 (2006).

62 Velasco, R. *et al.* A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).

63 Kim, M. Y. *et al.* Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja Sieb.* and *Zucc.*) genome. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 22032-22037 (2010).

64 Chan, A. P. *et al.* Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* **28**, 951-956 (2010).

65 Varshney, R. K. *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **6**, 83-89 (2011).

66 Sato, S. *et al.* Genome structure of the legume, *Lotus japonicus*. *DNA Res.* **15**, 227-239 (2008).

67 Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). *Nature* **452**, 991-996 (2008).

68 Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, 109-116 (2011).

69 Wang, X. *et al.* The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **28**, 1035-1039 (2011).

70 Jurka, J. *et al.* Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462-467 (2005).

71 Argout, X. *et al.* The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101-108 (2011).

72 Suda, J., Kyncl, T. & Jarolimova, V. Genome size variation in Macaronesian angiosperms: forty percent of the Canarian endemic flora completed. *Plant Sys. Evol.* **252**, 215 (2005).

73 Olszewska, M. J. & Osiecka, R. Relationship between 2C DNA content, life-cycle, systematic

position & level of DNA endoreplication in parenchyma cell nuclei during root growth and differentiation in some monocots. *Biochem. Physiol. Pflanz.* **177**, 319-336 (1982).

74 Zonneveld, B. J., Leitch, I. J. & Bennett, M. D. First nuclear DNA amounts in more than 300 angiosperms. *Ann. Bot.* **96**, 229-244 (2005).

75 Röser, M., Johnson, M. & Hanson, L. Nuclear DNA amounts in palms (*Arecaceae*). *Bot. Acta.* **110**, 79-89 (1997).

76 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61-65 (2007).

77 UniProt Consortium. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142-148 (2010).

78 Jain, E. *et al.* Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**, 136 (2009).

79 Gotoh, O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res.* **36**, 2630-2638 (2008).

80 Gotoh, O. Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics* **24**, 2438-2444 (2008).

81 Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).

82 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).

83 Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848 (2001).

84 Qin, E. *et al.* A genome sequence of novel SARS-CoV isolates: the genotype, GD-Ins29, leads to a hypothesis of viral transmission in South China. *Genomics Proteomics Bioinformatics* **1**, 101-107 (2003).

85 Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944-1954 (2008).

86 Zhang, Z. *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259-263 (2006).

87 DeYoung, B. J. & Innes, R. W. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* **7**, 1243–1249 (2006).

88 Mun, J. H., Yu, H. J., Park, S. & Park, B. S. Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa*. *Mol. Genet. Genomics* **282**, 617-631 (2009).

89 Ameline-Torregrosa, C. *et al.* Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. *Plant Physiol.* **146**, 5-21 (2008).

90 McHale, L., Tan, X. P., Koehl, P. & Michelmore, R. W. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7**, 212 (2006).

91 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211-D222 (2010).

92 Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205-211 (2009).

93 McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**, 356-358 (2006).

94 Li, J. *et al.* Unique evolutionary pattern of numbers of gramineous NBS–LRR genes. *Mol. Genet. Genomics* **283**, 427–438 (2010).

95 Porter, B. W. *et al.* Genome-wide analysis of *Carica papaya* reveals a small NBS resistance gene family. *Mol. Genet. Genomics* **281**, 609-626 (2009).