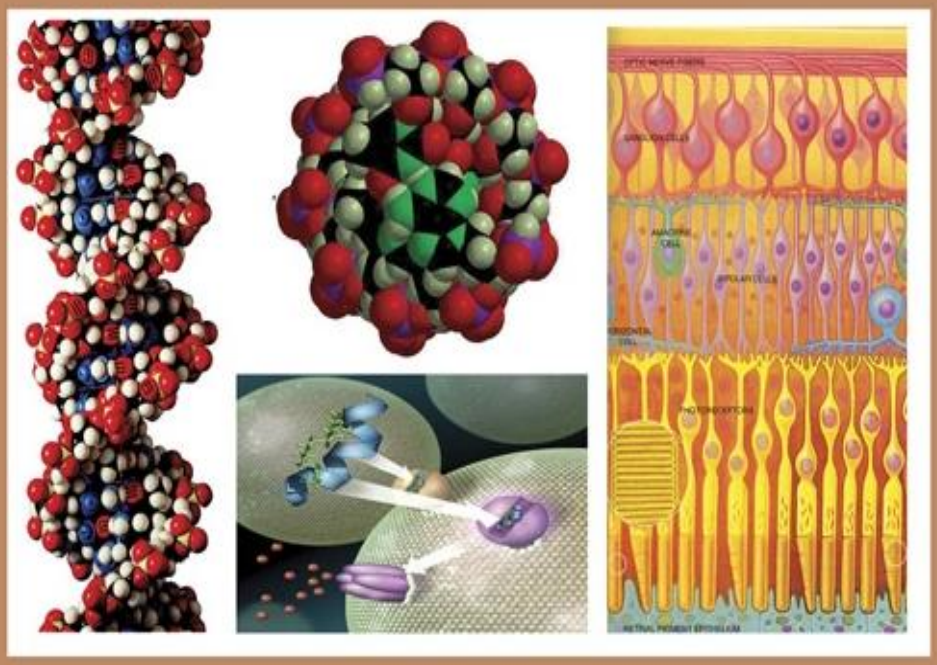
EGYPTIAN ACADEMIC JOURNAL OF

BIOLOGICAL SCIENCES

PHYSIOLOGY & MOLECULARBIOLOGY

C



ISSN
2090-0767

WWW.EAJBS.EG.NET

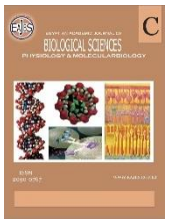
Vol. 14 No. 2 (2022)

.

Citation: *Egypt.Acad.J.Biolog.Sci. (C.Physiology and Molecular biology) Vol. 14(2) pp311-330 (2022)*
DOI: 10.21608/EAJBSC.2022.272232

Egypt. Acad. J. Biolog. Sci., 14(2):311-330 (2022)

Egyptian Academic Journal of Biological Sciences
C. Physiology & Molecular Biology
ISSN 2090-0767
www.eajbsc.journals.ekb.eg



Whole Genome Sequencing of Date Palm (*Phoenix dactylifera* L.) Cultivars Using NGS.

Sara Aly[1], Mohamed E. Saad[2,3], Ashraf Hendam[4], Amina Abdel-Hamied[2], Ahmed Farouk Al-Sadek[4], E. A. Madboly[5], Hoda S. Barakat[1] and H. El-Atroush[1*]

1- Department of Botany, Faculty of Science, Ain Shams University, Cairo, Egypt.
2- Department of Nucleic acid, proteomics and Genomics, Genomics facility central lab, Agricultural Genetic Engineering Research Institute (AGERI), Agricultural Research Center ARC, Egypt.
3- Department of Biology, Faculty of Science, Taibah University, Almadinah Almonawarah, Kingdom of Saudi Arabia.
4- Bioinformatics Unit, Central Lab for Agricultural Experts Systems, Agricultural Research Center (ARC), Giza, Egypt.
5- The Central Laboratory for Date Palm Research and Development (CLDPRD), Agricultural Research Center (ARC), Giza, Egypt.
*E. Mail: hala.elatroush1st@gmail.com

ARTICLE INFO
Article History
Received:1/10/2022
Accepted:17/11/2022
Available:21/11/2022

Keywords:
Whole Genome (WGS) Sequencing -Next Generation Sequencing (NGS), SNP, Indel, DNA Markrs, Date Palm.

ABSTRACT

Date palm (*Phoenix dactylifera* L.) is related to the family Arecaceae, which is considered one of the most ancient economically cultivated crops. Mainly, it is grown in the arid regions of the Middle East and North Africa. The crucial matter in maintaining the diverse number of date palm cultivars in Egypt is its biodiversity conservation of it. In order to progress programs and cultivar characterization and conservation to combat genetic erosion, we must estimate the genetic variability and right date palm cultivar identification this is the important point to present a comprehensive investigation for Egyptian date palm genome variations and develop novel DNA markers (SNPs and indels) in four date palm cultivars using SOLiD sequencing.

INTRODUCTION

Date palm (*Phoenix dactylifera* L.) is one of the most greatest economically important plants which possess a lot of useful products (Ahmed *et al.*, 1995). It plays an important role in the desert ecological system. It acts as a shelter for ground crops, and also, plays a great role in land reclamation and desertification control (Gotch *et al.*, 2006).Date palm (2n = 36) is dioecious, from the family Arecaceae (Palmae) monocotyledonous, perennial fruit tree (Barrow, 1998). It is widely cultivated in the arid regions of the Arabian Peninsula, the Middle East and North Africa (Chao and Krueger, 2007).

In Egypt, date palm is considered as very important crop. There is a great national interest in increasing date palm production area to cover the local market and for exportation. In 2014, Egypt has reached a maximum production according to the last FAO statistics (Bekheet and El-Sharabasy, 2015). Although there are about 52 identified cultivars, there is only 20 date palm cultivars are commercially distributed in the Egyptian market. However, the precise number of Egyptian date palm cultivars is argued.

This may be due to the difference in identification criteria and the dependence on local experience (Rizk *et al.,* 2004; Bekheet and El-Sharabasy, 2015). In 2008Sharma et al. found that DNA sequencing is a clear-cut approach for identifying genomic variations. Now, there are improvement techniques that led to the enhancement of molecular marker systems which contributed to increasing the sensitivity and resolution of these techniques in genetic diversity studies and cultivar identification purposes (Agarwal *et al.,* 2008). Earlier studies usually paid more attention only to markers in crucial genomic regions to follow important traits. However, with the advent of NGS, markers are used to review as many loci as possible throughout the entire genome and with nucleotide-level precision (Varshney *et al.,* 2014). The development of DNA markers using next-generation sequencing (NGS) technology became the most effective way in improving date palm valuable resources (Cullis, 2011; Faqir *et al.,* 2017). In addition, a major step in understanding the genetic basis of an organism is the complete sequences studies (Sterky and Lundeberg, 2000). Finally, we found that, NGS application is very important in many fields related to agriculture in order to find targets for genetic manipulation, evolution studies, exploring the bionetworks, and understanding the fundamental principles of functional genomics (Ashraf *et al.,* 2022)

Only, in the last decade deeply studies of the genome sequence of date palm in a much more comprehensive view. Different research teams tried to discover date palm genome sequencing. In 2009, Qatar,s team from Weil Cornell Medical College sequenced the entire date palm genome using an illumina sequencer. They announced the first draft genome map in 2009 for a Khalas variety female date palm. They estimated genome size of 658Mb, assembled 58% of the genome (382Mb) and predicted 25,059 gene models. They sequenced eight other cultivars, including females of the Medjool and Deglet Noor varieties and their backcrossed males. Finally, they postulated a sex-linked region of the date palm genome (Al-Dous *et al.,* 2011).

King Abdulaziz City for Science and Technology (KACST) in collaboration with the Beijing Institute of Genomics, Chinese Academy of Science (BIG/CAS) estabilished a date palm genome project by using both Roche 454 and ABI SOLiD. They presented a more complete genome assembly for the cultivar Khalas and compared it with three other cultivars (Agwa, Sukry and Fahal) to determine sequence variations among them. They estimated a genome size of 605.4 Mb, covered approximately 90% of the genome (671Mb) and predicted 41,660 gene models. Also, they applied transcriptomic studies on genes related to fruit development and abiotic resistance. (Zhang *et al.,* 2011; Al-Mssallem *et al.,* 2013). In A whole genome re-sequencing study took place at the Center for Genomics and Systems Biology, New York University, Abu Dhabi using an illumina sequencer in 2015. They wanted to assume the origin of date palms and determine the genomic diversity in 62 different date palm cultivars collected from various areas,Finally, they discovered genes controlling traits of interest. In this study, four-Egyptian cultivars were presented Zaidi, Samany, Zagloul and Hayani (Hazzouri *et al.,* 2015).

Since the diversity within a species can't be studied by the sequencing of a single reference genome, re-sequencing of different cultivars plays a key role in any study (Bolger *et al.,* 2014).

In the present study, whole genome sequencing for date palm cultivars was carried out using NGS. A comparative genomic investigation for Egyptian date palm genome variations among four female cultivars which include, Amhat, Sewi, Zagloul and Hayani and develop novel DNA markers (SNPs/indels) for date palms using next-generation sequencing.

MATERIALS AND METHODS

Plant Material:

Four female date palm (*Phoenix dactylifera* L.) cultivars were shown in Table (1).

Table 1: Show Four Female Date Palm Cultivars**.**

Serial	Cultivar	Type of fruit	Sex	Collection region	Coordinates
1	Amhat	Fresh	Female	Sakara – Giza, Egypt	**N 29º51'51.8'' E 31º13'49.7''**
2	Sewi	Semi dry	Female	Sakara - Giza, Egypt	**N 29º51'51.8'' E 31º13'49.7''**
3	Zagloul	Fresh	Female	Sakara - Giza, Egypt	**N 29º51'51.8'' E 31º13'49.7''**
4	Hyani	Fresh	Female	Sakara - Giza, Egypt	**N 29º51'51.8'' E 31º13'49.7''**

DNA Extraction and Quantification:

Healthy green young leaves were chosen from mature palm trees, thoroughly rinsed with tap water, washed with 70% ethyl alcohol and dried with clean tissues. Then the leaves were cut into small pieces to be preserved at -80 ^{O}C or directly applied to the DNA isolation protocol. Genomic DNA was extracted for the four cultivars using DNeasy™ Plant Mini Kit (Qiagen Inc., cat. no. 69104) and quantified using a Qubit 2.0 fluorometer (supplied by Invitrogen).

The whole genome sequencing workflow includes 4 steps: (1) Fragment library preparation. Throughout library preparation, forward and reverse adaptors are attached to the ends of sheared DNA fragments. The four-date palm cultivars were multiplexed as each one is characterized by a specific barcode. (2) Templated bead preparation (Emulsion, amplification and enrichment). In the SOLiD system, the DNA library to be sequenced needs amplification on P1 beads during emulsion PCR (ePCR), then enrichment of the templated beads takes place. (3) 5500xl SOLiD sequencing. The beads are then deposited and attached to a sequencing flow chip. The flow chip was mounted on the 5500xl genetic analyzer. The clonally amplified DNA fragments linked to beads were sequenced parallelly using sequential ligation of dye-labeled, two-base encoded, oligonucleotide probes. The four libraries were sequenced utilizing the FWD1 sequencing chemistry. (4) Data analysis

Bioinformatics tools were used to analyze the raw data of the four date palm cultivars from SOLiD 5500xl to 1) to calculate the genome coverage. 2) align the sequenced genome to a reference one. 3) compare the variations found among cultivars through indels and SNPs polymorphism.

After run completion, the system exported the data in the eXtensible SeQuence (XSQ) file format. There is one XSQ file for each lane, this file was then converted to a color space fasta (csfasta) and a quality file that contains the Phred quality scores (.qual) using NGS plumbing tool. The theoretical redundancy of coverage (c) was calculated for each of the four cultivars separately as LN/G, where L is the read length, N is the number of reads and G is the haploid genome length (Lander and Waterman, 1988). The reads passed the quality control of SOLiD were aligned by Bfast tool (Version bfast-0.7.0a) using the reference genome of cv. Khalas (GCF_000413155.1_DPV01) with a total length of 556.481 Mb and 80,317 scaffolds. The reference genome was converted to BRG format by the fasta2brg command then an index of the reference genome was created using the index of Bfast. The match command of the Bfast was used for seeding alignment locations, local align was used for local alignments of the reads and finally post process command was used for filtering the alignments.The Samtools was used to convert the same files from the Bfast to bam files using the view command. The bam files were merged using the Samtools merge command. The bam file was sorted by the

sort command and an index of the aligned reads was created by the index command. Finally, an index of the reference genome was created by faidx command. The picard-tools were used for marking and removing duplicate reads. The insertion/deletion (indel) and SNP calling were conducted by SNVer-0.4.1 using the SNVerIndividual command which creates files with vcf format for both indels and SNPs. SNPs and indels were visualized using Integrative Genomics Viewer (IGV) program. Dendrograms were drawn using the SNP variations using the program MegaX (Kumar *et al.,* 2018).

RESULTS AND DISCUSSION

Whole Genome Sequencing and Sequence-Based Markers:

1. Library Preparation Results:

1.a Electrophoresis of the Sheared DNA Library.

Sheared genomic DNA was run on a 2% agarose gel to ensure the success of the shearing process. Figure (1) shows that the four date palm cultivars were successfully sheared as the majority of the sheared DNA produced lies between sizes of 100- 250 bp.

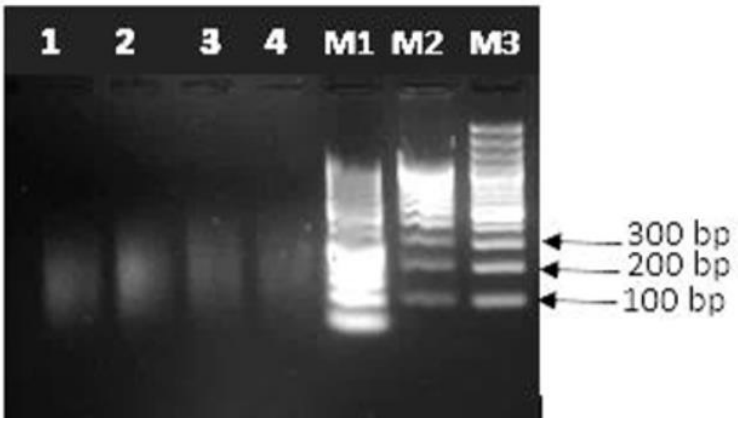


Fig. 1: Date palm genomic DNA after shearing, 1-4 are Amhat, Sewi, Zaghloul and Hyani cultivars respectively. M refers to the DNA ladder; M1: 50bp (gene ruler), M2: 100 bp (sib enzyme), M3: 100 bp plus (gene ruler).

1. b Quantization of the Size-Selected DNA:

After the size selection of the sheared, end-polished DNA fragments by Agencourt AMPure XP reagent beads, the quantity of the size selected DNA yield was measured by the Qubit 2.0 fluorometer using the Qubit dsDNA HS (High Sensitivity) assay kit (Table 2). The average yield of size-selected DNA is about 12% of the input quantity among the four libraries.

1.c Quantization of the Ligated DNA:

The size-selected DNA was subjected to dA tailing, adaptor ligation and purification steps, then the quantity of the ligated DNA was measured by the Qubit 2.0 fluorometer using the Qubit® dsDNA HS (High Sensitivity) assay kit (Table 2). The quantity of the ligated DNA after purification for the four libraries was good enough to proceed into further steps.

Table 2: Size selected DNA concentration of sheared date palm libraries.

Sample	DNA conc. (ng/µl)	Total DNA conc. (µg)	Yield relative to DNA input quantity
1)Amhat	11.3	**0.38**	**12.8%**
2) Sewi	14.1	**0.48**	**16%**
3) Zagloul	8.98	**0.3**	**12 %**
4) Hayani	9.62	**0.32**	**10.9%**

1.d Quantization of the Amplified Libraries:

Using Qubit After libraries PCR amplification and purification, the quantity of the amplified DNA was measured by the Qubit 2.0 fluorometer using the Qubit dsDNA HS (High Sensitivity) assay kit (Table 3).

Table 3: Ligated DNA concentration of date palm fragment, barcoded libraries.

Sample	DNA conc. (ng/µl)	Total DNA conc. (µg)
1)Amhat	18.6	**0.39**
2) Sewi	15.5	**0.32**
3) Zagloul	12.4	**0.26**
4) Hayani	8.22	**0.17**

Using Quantitative real-time PCR (qPCR) was used to determine the amount of amplifiable template in a SOLiD library Samples were diluted to very low concentrations for quantitation using TaqMan method (Table 4)

Table 4: Amplified DNA libraries concentration of date palm cultivars according to Qubit readings**.**

Sample	DNA conc. (ng/µl)
1)Amhat	11.8
2) Sewi	11.5
3) Zagloul	15.4
4) Hayani	7.51

1. e Checking the Size Distribution of The Libraries:

Using gel electrophoresis, the amplified libraries were run on 2% agarose gel to ensure size distribution. Figure (2) shows the 4 fragment libraries after amplification with an average size of 250 bp.

The aamplified DNA libraries concentration of date palm cultivars according to quantitative real-time PCR readings were shown in Table 5.

Table 5: Amplified DNA libraries concentration of date palm cultivars according to quantitative real-time PCR readings.

Sample	DNA conc. (pM)
1)Amhat	0.4
2) Sewi	0.4
3) Zagloul	0.73
4) Hayani	0.68

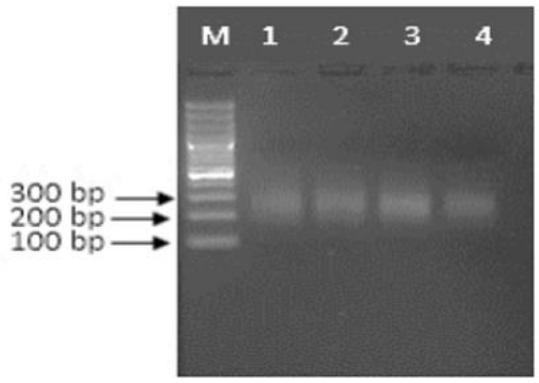


Fig. 2: The date palm fragment barcoded libraries after amplification. M refers to DNA ladder, 100 bp plus (gene ruler). 1-4 is Amhat, Sewi, Zaghloul and Hyani cultivars respectively.

Using QIAxcel Advanced:

The four amplified libraries were run on the multi-capillary QIAxcel Advanced machine to determine size distribution and concentration. Figure (3) shows a simulated gel image produced for the four fragment libraries. An electropherogram was drawn for each lane and converted to a gel image with annotated bands for each lane

.

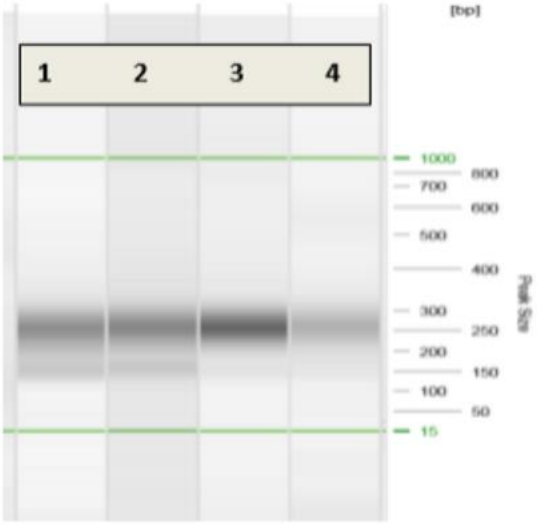


Fig. 3: A simulated gel image for the four fragment libraries drawn using QIAxcel Advanced machine. 1-4 is Amhat, Sewi, Zaghloul and Hyani cultivars respectively.

Figure (4) shows gel images produced for each of the four fragment libraries in a separate lane with annotated bands. The analysis software exports also a table with the size and concentration of each band. The maximum peak was estimated for each cultivar to ensure size distribution. It was the 239 bp band for cultivar ‘Amhat’ library in the gel image (a), 280 bp in (b) for ‘Sewi’, 250 bp in (c) for ‘Zagloul’ and 280 bp in (d) for ‘Hyani’ cultivar. In Table (6), the average concentration for each library was documented**.**

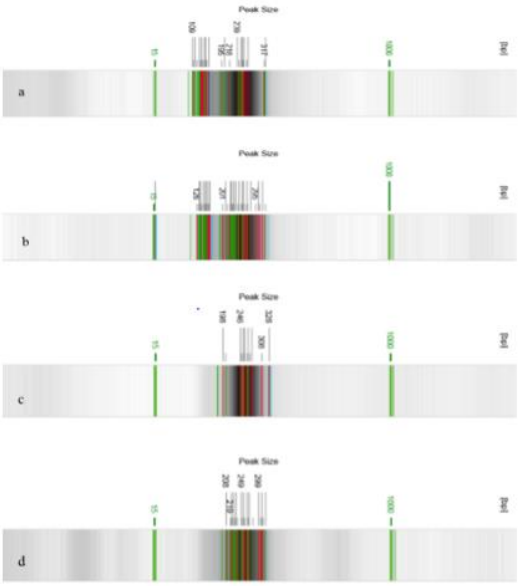


Fig. 4: Gel images for each lane with annotated bands for the four fragment libraries drawn using QIAxcel Advanced machine. a-d are Amhat, Sewi, Zaghloul and Hyani cultivars respectively.

Table 6: Amplified DNA libraries concentration of date palm cultivars according to QIAxcel Advanced machine**.**

Sample	DNA conc. (ng/µl)
1)Amhat	11.40
2) Sewi	11.55
3) Zagloul	17.58
4) Hayani	6.27

2. Templated Beads Quantitation Using Nanodrop:

Three readings for optical densities were recorded at A600 nm for the P2 enriched beads sample using the nanodrop spectrophotometer. Then the concentration of the beads was calculated according to a previously drawn standard curve. A total of 4.5 million beads were obtained after the enrichment process representing 16% of the input beads (Table 7). After considering dilutions, it was found that the sample has 11.5 million beads in each µL. This was deposited on 6 lanes of the 5500 SOLiD flow chip. Thus, 26 µL of the enriched beads needed to be deposited in each lane.

Table 7: Nanodrop results for the templated beads and total bead concentration**.**

A600 (1)	A600 (2)	A600 (3)	Average	Total no. of beads in 1 ml (million beads)	% Of P2 enriched beads relative to the total input
0.823	**1.078**	**0.982**	**0.961**	**4.548**	**16%**

3. Monitoring the Run:

The sequencing run was monitored to assess the quantity and quality of beads deposited in each lane. The run was monitored for each separate lane in real-time. Run data is shown for one lane for demonstration. Primarily, the 5500xl genetic analyzer records a focal map for the beads on the flow chip, this is done only at the start of the run to establish the position of each covalently attached bead (X, Y).

Bead Assessment:

After the focal map and before the first sequencing ligation, the analyzer generates a bead assessment report for each lane.

Figure (5) shows the bead assessment results for one lane as an example, in which P2 records 142.25 million beads. P2 value gives an indication of the total number of enriched beads deposited in a lane. An optimum P2 value spans around 177M beads (every lane has 708 imaged panels and the bead density target is 250,000 beads/panel). One panel is one image area in a lane. N2S (noise to signal) number gives an indication of the sample's clonality which is recorded here as 17.8%. A low number means that the beads are likely monoclonal (having a single template on each bead), and a higher number reflects the beads' polyclonality (beads with multiple templates, so multiple colors). An optimum N2S value shouldn't exceed 15%. The example results also showed 53% as the percentage of beads on the axis. The percent on the axis shows the frequency with which enriched templated beads lie within 10 degrees of each color channel. Spectrally purer beads (closer to the axis) indicate more monoclonality. A higher number of beads on the axis is better. The Median P2/panel value indicates how many beads on average per panel were deposited and detected when the probe annealing to the P2 tag was interrogated. Results showed 201K in this example lane. 250 K beads/panel is the optimum bead density target.

a

P2:	142.25M
N2S:	17.87%
Percent on-axis:	53.00%
Median P2/panel:	201K

b

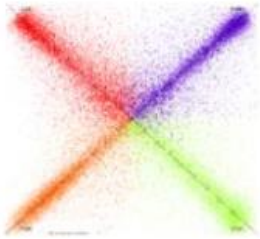


Fig. 5: Bead assessment results for one lane in the 5500 xl SOLiD genetic analyzer run. a) summary report b) satay plot.

The color balance or satay plot (shown in figure 5-b) is an indicator of the spectral purity and signal intensity of the beads.

Each quadrant represents a nucleotide and the color noise associated with it. The figure shows dense colors on axes referring to high spectral purity.

A satay report is shown for each of the five sequencing primers used during the 15 ligation cycles. Figure (6) demonstrates the difference in bead distribution around axes in the satay plots among the 5 sequencing primers in the first and last ligation cycles in one lane as an example. Figure (7) shows the fraction of Good/Best beads with an average of 50%. Best beads are those beads present on the axis in the satay plot (monoclonal ones), while good beads are beads that lie within 10 degrees of each color channel. This value decreases every ligation cycle because a number of beads are blocked every cycle to avoid dephasing.

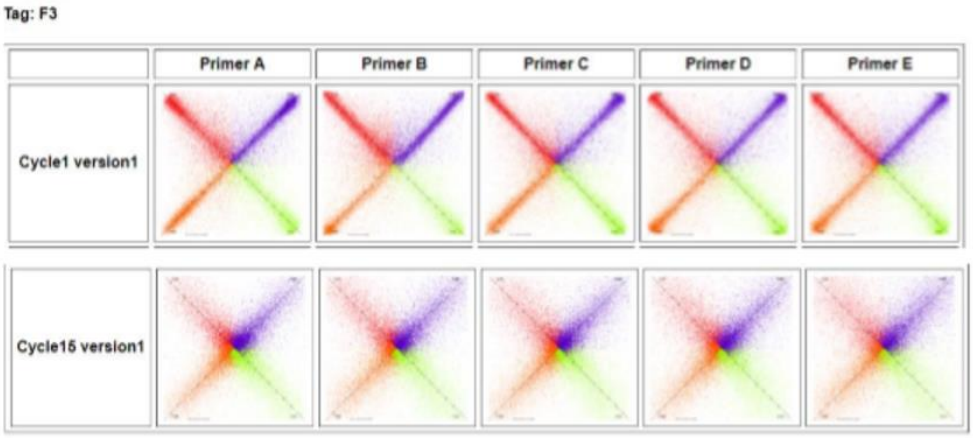


Fig. 6: Satay report for the 5 sequencing primers in the first and last ligation cycles.

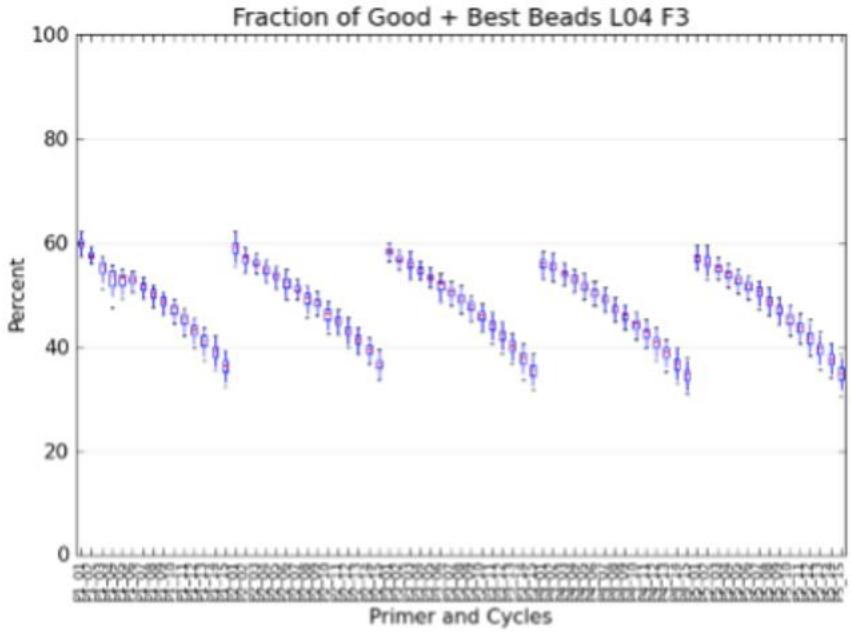


Fig. 7: The fraction of Good/Best beads with an average of 50%.

4. Bioinformatic Results:

4.1 Genome Coverage Analysis:

The average sequencing coverage slightly varied across the four date palm cultivars (Table 8). The redundancy or depth of coverage is the average number of times that a base in the reference is covered by randomly distributed raw sequencing reads across a genome (Sims *et al.,* 2014). 'Zagloul' had the highest depth at 25x and 'Hyani' recorded the lowest (18x) while 'Amhat' and 'Sewi' reported 22.6x and 21x respectively.

Higher coverage values were indicated in other date palm sequencing studies concerning different cultivars; where Al-Dous *et al.* (2011) reported a sequence redundancy of 53.4× for postassembly reads of the genome for 'Khalas' cultivar. Also, for the same cultivar, Al-Mssallem *et al.* (2013) revealed a depth of 17.3x generated from four fragment libraries and the depth was further increased to 122.1x through 9 mate-paired libraries, while Hazzouri *et al.* (2015) obtained a mean sequencing depth of 20.8x per sample from paired-end libraries during their whole genome resequencing study on 62 cultivars. These elevated coverage estimates were because of the dependence on more sequencing runs of different types.

Table 8: Depth of coverage of the four samples**.**

Sample	Depth or redundancy of coverage
1)Amhat	22.6 x
2) Sewi	21 x
3) Zagloul	25 x
4) Hayani	18.2 x

4.2 Single Nucleotide Polymorphisms and Insertion/Deletion Mutations:

Some genomic variations from the genome sequence were documented to understand the genetic diversity within the date palms. SNPs and indels were obtained for each cultivar compared to the genome of the Saudi cultivar 'Khalas' as a reference. The Integrative Genomics Viewer (IGV) program was used to illustrate the position of SNPs and indels in the sequencing reads relative to the reference genome. Full data analysis obtained with the aligned sequence results for the four date palm cultivars will be uploaded to the NCBI database. Figure (8) demonstrates an SNP example in cultivar 'Amhat' that was assigned by the reference as 'A' (at the bottom of the figure) and altered in the Egyptian cultivar 'Amhat' to 'T'. Figures (9 and10) (4.20 and 21) show examples of insertion and deletion mutations respectively in the 'Amhat' cultivar. These variations are listed in excel sheets for each cultivar (Tables 9&10).

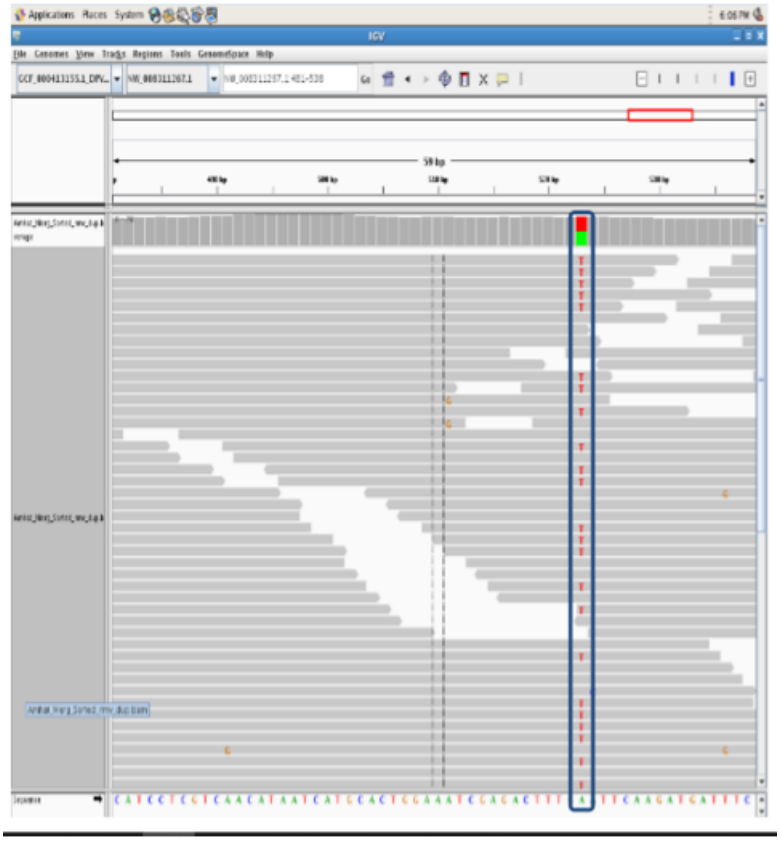


Fig. 8: Example of SNP variation in cultivar 'Amhat'. Grey bars represent the aligned sequencing reads. The 'A' base in the reference genome was altered to 'T' (scaffold NW_008311267.1 and position 523 bp).

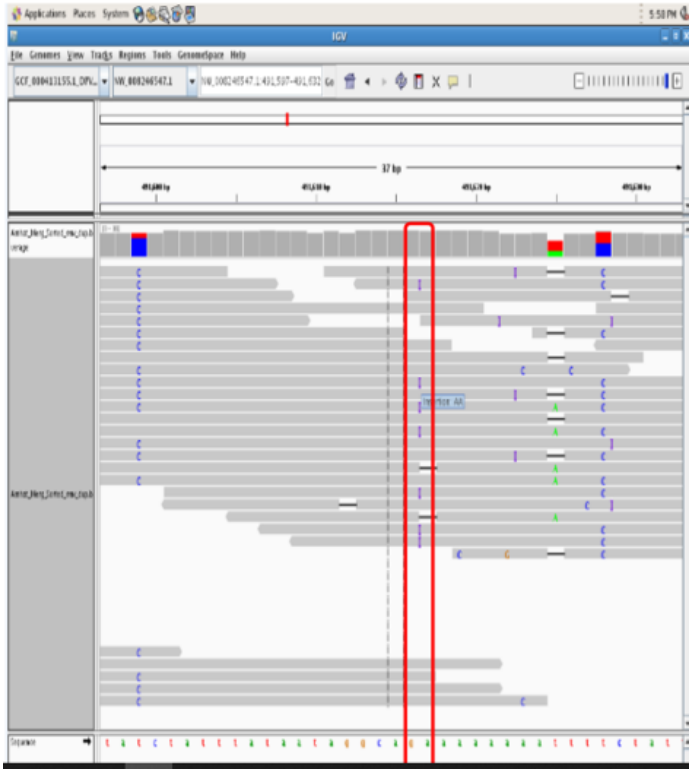


Fig. 9: Example of insertion of two bases 'AA' in 'Amhat' cultivar in (scaffold NW_008246547.1 and position 491,617 bp.

Parts of the SNP and indel output data sheets for cultivar 'Hayani' as an example are shown in Tables (9 and 10) for SNPs, insertions and deletions respectively as the whole data were represented in over 1000 pages. Thus, various numbers for SNPs and indels were estimated from raw data (Table 11). However, variations at a specific level of coverage are only accepted. An SNP was called when it has at least 20x and not more than 70x depth of coverage and indel depth ranged from 10x to 70x. Unusually high coverage sequences (produced from the incorrect mapping of reads from duplicated regions) and low coverage (due to low-quality sequences) are avoided in calling SNPs and indels.

Finally, after filtration 'Zagloul' recorded the highest number of SNPs (1,101,303), insertions (63,648) and deletions (75,709) while 'Amhat' had the lowest SNP (10,427), insertion (7257) and deletion (8782) values among the four cultivars (Table 11 and figure 10). The produced SNPs and indels could be used in genotyping assays to clearly identify date palm cultivars. In this respect, 1,748,109 SNPs were discovered in the Saudi cultivar 'Khalas' indicating a heterozygosity rate of 0.46% (Al-Dous *et al.*, 2011). A number of 7,176,238 SNPs across 62 date palm cultivars were reported by Hazzouri *et al.* (2015), this elevated value may be because of the increased numbers of tested cultivars or higher coverage estimates.

Table 9: Part of the SNP output data sheet for 'Hayani cultivar**.**

Scaffold	Position	Reference	Alteration	Depth
NW_008246507.1	16730	C	T	23
NW_008246507.1	18799	T	C	25
NW_008246507.1	19385	G	C	22
NW_008246507.1	21958	C	A	22
NW_008246507.1	24920	C	T	25
NW_008246507.1	24929	T	A	24
NW_008246507.1	24954	A	T	24
NW_008246507.1	24957	G	A	24
NW_008246507.1	24982	G	T	23
NW_008246507.1	25019	T	A	25
NW_008246507.1	175341	G	T	31
NW_008246507.1	175351	C	T	30
NW_008246507.1	175363	C	T	33
NW_008246507.1	791214	A	G	36
NW_008246507.1	791220	T	C	40
NW_008246507.1	791257	G	A	39
NW_008246507.1	791292	T	C	28
NW_008246507.1	175370	A	G	28

Table 10: Part of the deletion output data sheet for 'Hayani' cultivar. (REF refers to the base where the deletion starts in the reference genome).

1	Scaffold	Position	REF	Insertion	Depth
2	NW_008246704.1	405014	T	+AA	57
3	NW_008288390.1	210	A	+TGCTAT	57
4	NW_008247948.1	21165	A	+G	56
5	NW_008249240.1	3623	T	+G	56
6	NW_008251476.1	5135	T	+G	56
7	NW_008275961.1	874	A	+CTG	56
8	NW_008280741.1	148	A	+TC	56
9	NW_008280741.1	149	T	+CC	56
10	NW_008246541.1	1394271	T	+AAAC	55
11	NW_008251663.1	5082	A	+C	50
12	NW_008254034.1	2591	G	+A	50
13	NW_008255812.1	1699	A	+G	50
14	NW_008258753.1	374	C	+G	50
15	NW_008259074.1	361	A	+T	50
16	NW_008305052.1	364	A	+TTG	50
17	NW_008246557.1	455415	a	+AG	49
18	NW_008246559.1	593735	T	+A	49
19	NW_008246576.1	54223	G	+A	49

Table 11: SNP, insertion and deletion numbers among the four samples before and after filtration**.**

Sample	SNP		Insertion		Deletion	
	Raw	After filtration 70≥DP≥20	Raw	After filtration 70≥DP≥10	Raw	After filtration 70≥DP≥10
1) Amhat	647517	10427	13422	7257	18358	8782
2) Sewi	1009726	45073	25665	19030	32867	22748
3) Zagloul	1883318	1101303	67899	63648	79825	75709
4)Hayani	1512401	531221	48588	43909	57054	51810

By examining indel loci, it was found that the numbers of indels greatly varied across four cultivars as well as in relation to their length (Tables 11, 12 and Fig. 10). The more the length of the indel is, the fewer indels are called among the four cultivars. Insertions started from a single base to a fragment of 46 bases in length, while deletions ranged from one to 23 bases. These variations may have great effects on each cultivar's identity. This is because insertions and deletions are directly related to significant genomic variations that may reach 10–20% among plant cultivars (Britten *et al.*, 2003; Morgante *et al.*, 2005; Ding *et al.*, 2007).

Table 12: Insertion and deletion numbers against their lengths among the four samples

Length	Number of insertions				Number of Deletions			
	Amhat	Sewi	Zagloul	Hayani	Amhat	Sewi	Zagloul	Hayani
1	5088	13468	43656	31014	6162	15663	49702	34994
2	864	2288	7556	5028	992	2662	9406	6188
3	347	959	3359	2185	462	1253	4371	2918
4	296	733	2807	1770	444	1161	3997	2568
5	132	308	1051	681	136	380	1482	910
6	99	283	916	572	120	322	1312	876
7	78	206	967	568	148	381	1365	873
8	76	207	796	530	122	308	1270	823
9	56	108	479	273	62	149	646	398
10	24	83	322	215	35	107	491	295
11	23	71	270	162	26	98	379	226
12	16	52	219	125	21	70	333	185
13	14	39	152	92	16	48	215	131
14	13	28	121	64	10	44	181	102
15	7	21	79	54	9	34	168	80
16	11	20	101	47	4	19	106	57
17	8	10	67	53	5	21	100	70
18	5	19	67	46	5	14	87	45
19	7	12	51	36	0	6	58	38
20	4	7	41	33	3	8	40	31
21	5	6	34	24	0	0	0	1
22	10	7	25	14	0	0	0	0
23	2	4	41	18	0	0	0	1
24	4	9	18	17				
25	5	5	31	21				
26	3	8	25	13				
27	5	3	18	11				
28	7	3	20	18				
29	4	6	25	18				
30	5	4	28	16				
31	4	4	20	15				
32	5	7	28	18				
33	3	4	24	16				
34	4	4	31	17				
35	5	3	19	13				
36	5	3	19	17				
37	1	2	20	15				
38	0	7	18	12				
39	0	3	18	11				

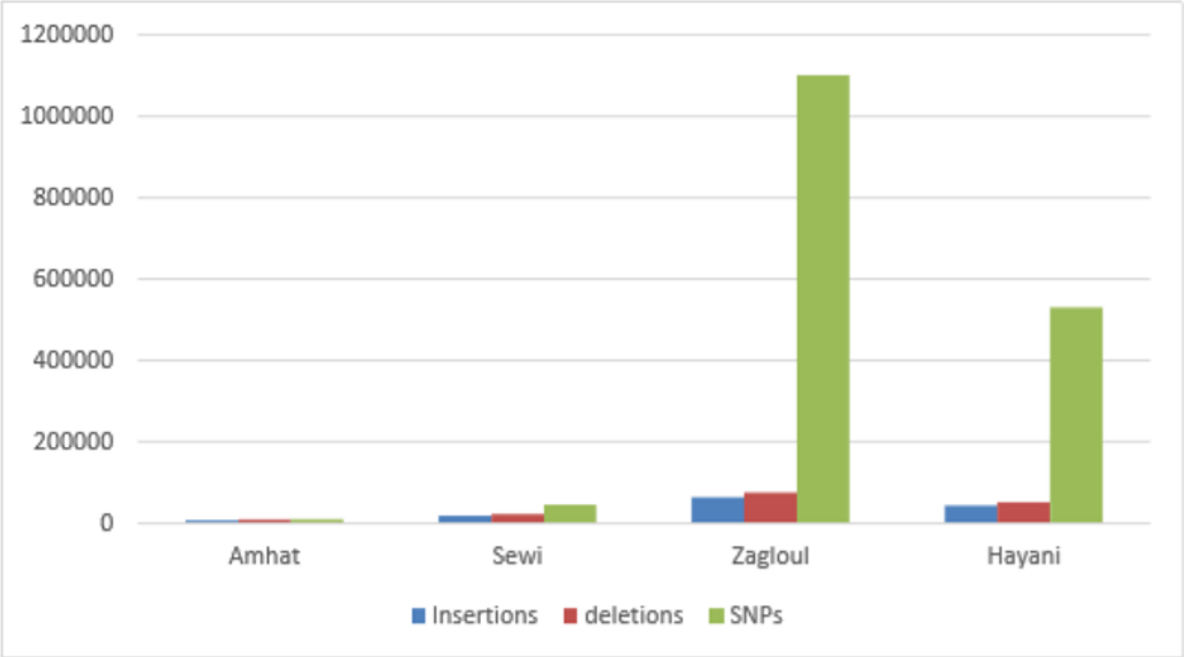


Fig. 10: Histogram of insertion, deletion and SNPs after filtration among the four date palm cultivars. Cultivar names are written on the x-axis and numbers of different parameters are represented on the y-axis.

4.3 Comparative Genomic Analyses Using Single Nucleotide Polymorphisms:

Common SNPs among the four date palm cultivars were collected in a comparative way to show genomic variation at a single nucleotide precision (Table13).

Table (13): Part of common SNPs output data sheet among the four cultivars.

1	Scaffold	Position	Reference	Amhat	Sewi	Zagloul	Hayani
2	NW_008246509.1	1211980	T	A	A	A	A
3	NW_008246509.1	1212911	T	G	G	G	G
4	NW_008246509.1	1212918	T	G	G	G	G
5	NW_008246508.1	3189927	A	G	G	G	G
6	NW_008246508.1	3525020	T	G	G	G	G
7	NW_008246509.1	91746	A	T	T	T	T
8	NW_008246510.1	2122242	A	G	G	G	G
9	NW_008246510.1	2155770	T	C	C	C	C
10	NW_008246510.1	2155800	C	T	T	T	T
11	NW_008246510.1	2155836	A	T	T	T	T
12	NW_008246519.1	1612012	G	A	A	A	A
13	NW_008246527.1	26900	C	T	T	T	T
14	NW_008246527.1	26934	C	A	A	A	A
15	NW_008246527.1	1033426	A	G	G	G	G
16	NW_008246533.1	875453	C	A	A	A	A
17	NW_008247030.1	83213	T	A	A	A	A
18	NW_008246543.1	1165815	C	T	T	T	T
19	NW_008246543.1	1273709	C	T	T	T	T
20	NW_008246562.1	768057	C	A	A	A	A
21	NW_008246562.1	853030	A	G	G	G	G
22	NW_008246562.1	462296	C	A	A	A	A
23	NW_008246575.1	94947	G	A	A	A	A
24	NW_008246575.1	94956	T	C	C	C	C
25	NW_008246595.1	94009	T	C	C	C	C

A total of 2848 SNP loci were examined among the four cultivars, 2823 of which were monomorphic (calling for the same base) while only 25 loci were polymorphic representing a low percentage of polymorphism (0.88%) (Table 14 and fig11). Polymorphic SNPs across the four cultivars are listed in table (15) with their specific scaffold ID and position. These novel DNA markers (SNPs) allow much more precise genotyping for cultivars.

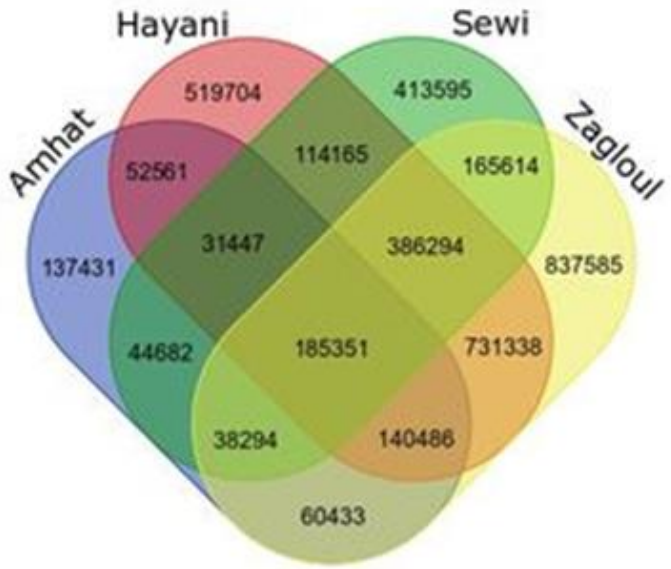


Fig. 11: Show monomorphic and polymorphic SNPs across the four cultivars relative to the reference genome of the Saudi cultivar 'Khalas'

Table14: Common SNP data among the four cultivars.

SNP parameter	Value
Total common SNPs among the 4 cultivars	2848
Monomorphic SNPs	2823
Polymorphic SNPs	25
Percentage of polymorphism	0.88%

By inspecting the 2848 common SNP loci, it was found that transition and transversion SNPs are represented in the four date palm genomes in different quantities, however, the transition/transversion ratio of the SNPs didn't much vary between cultivars recording an average of 1.65 (Table 16). Base substitution mutation is the base of single nucleotide polymorphism which involves either a transition (purine/purine or pyrimidine/pyrimidine) or transversion (purine against pyrimidine or vice versa) exchange. However, transitions are more frequent in nature than transversions (Weising *et al.*, 2005) as reported in this study (Table 16).

Table 15: List of polymorphic SNPs among the four cultivars relative to the reference genome (cv Khalas) and their scaffold ID and position.

S.	Scaffold	Position	Reference	Amhat	Sewi	Zagloul	Hayani
1	NW_008246541.1	1393586	A	C	T	C	C
2	NW_008246566.1	194039	C	G	A	A	A
3	NW_008246532.1	434717	T	G	A	A	A
4	NW_008246552.1	469706	A	T	G	T	T
5	NW_008246962.1	124509	G	A	A	C	C
6	NW_008247024.1	52103	T	G	C	G	C
7	NW_008261704.1	741	A	G	G	G	C
8	NW_008246608.1	419266	G	T	A	A	A
9	NW_008246532.1	434702	A	C	G	G	G
10	NW_008246562.1	926441	T	A	A	C	C
11	NW_008246622.1	434273	A	G	G	G	T
12	NW_008246776.1	77929	C	G	G	G	T
13	NW_008246649.1	31784	C	A	T	T	T
14	NW_008252287.1	43	C	T	A	T	T
15	NW_008251753.1	1955	C	A	T	A	A
16	NW_008255739.1	734	C	A	G	T	A
17	NW_008248344.1	14283	G	C	A	A	A
18	NW_008309191.1	311	A	G	T	T	G
19	NW_008263565.1	126	G	T	A	T	T
20	NW_008251753.1	1544	G	A	T	A	T
21	NW_008255739.1	640	T	C	G	C	C
22	NW_008305030.1	475	G	A	T	T	T
23	NW_008252087.1	782	A	C	C	T	T
24	NW_008308248.1	323	G	C	C	C	A
25	NW_008293249.1	41	C	A	A	T	A

Table 16: SNP numbers, types and transition/transversion ratios among the four cultivars

Cultivar	Number of SNPs	SNP type						Transition/ Trans-version ratio
		Transition		Transversion				
		A/G	C/T	A/C	A/T	C/G	G/T	
Amhat	10427	1574	1583	516	677	389	532	1.5
Sewi	45073	6698	7186	2072	2878	1602	2154	1.6
Zagloul	1101303	169647	182504	47475	61538	40390	49805	1.76
Hayani	531221	80962	87316	23121	30515	19362	24215	1.73
Average								1.65

4.4 Genetic Relationships and Cluster Analysis Among the Four Date Palm Cultivars Using Different Marker Types:

SNPs gained great interest as they are the smallest element of genetic variation and the origin of most genetic variation among individuals (Johnson *et al.*, 2001). The dendrogram constructed for the four Egyptian date palm cultivars from SNP data in comparison with the Saudi cultivar Khalas (reference genome) emphasized the generally narrow genetic background among the Egyptian cultivars (Fig. 12) as concluded earlier from PCR-based and morphological markers (Aly *et al.*,2019); while showing divergence between the Egyptian date palms and the Saudi cultivar.

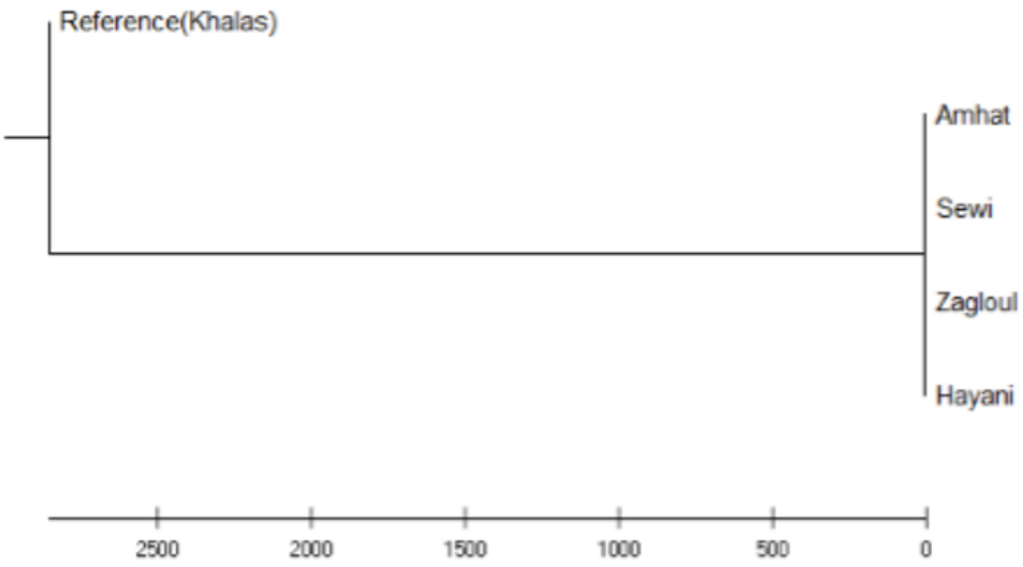


Fig. 12: Dendrogram for the four date palm cultivars constructed from SNP data in comparison with cv. Khalas (reference genome) using Un-weighed pair-group Arithmetic Average (UPGMA).

Two dendrograms were constructed for the four date palm cultivars from PCR-based markers (SCoT and SSR) collectively (Fig. 13) (Aly *et al.*,2019) and SNP data (Fig. 14). Results clearly showed the same pattern of relations in both trees, although, PCR-based markers recorded higher PIC values (Table 17). This reflects the power of SNP markers in revealing genomic variation precisely in fewer DNA loci. Similarly, a study aimed to detect the genetic diversity of wild almonds using traditional and new-generation markers, reported higher PIC values for SSR markers over SNPs (Sorkheh *et al.*, 2017).

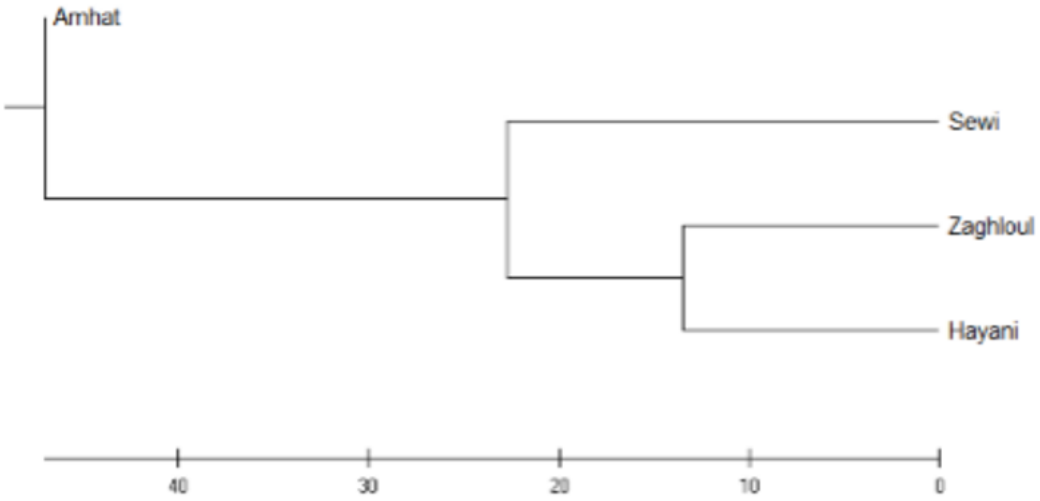


Fig. 13: Dendrogram for the four date palm cultivars constructed from collective SCoT and SSR data using Un-weighed pair group Arithmetic Average (UPGMA).

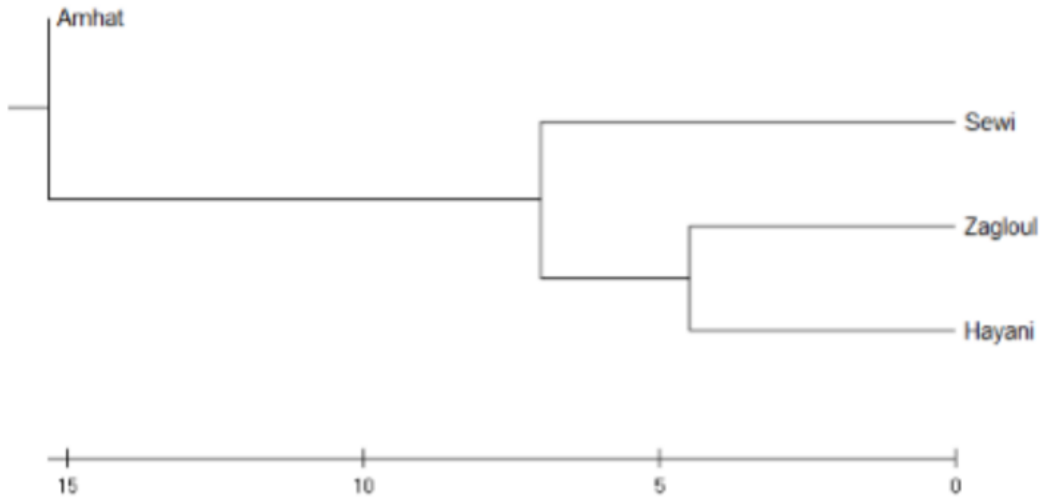


Fig. 14: Dendrogram for the four date palm cultivars constructed from SNP data using Un-weighed pair-group Arithmetic Average (UPGMA).

Table 17: Comparison between PCR-based and SNP markers in detecting genomic variations.

Parameter	Value	
	PCR-based markers	SNP
Total number of tested loci	209	2848
Number of polymorphic loci	120	25
Percentage of polymorphism	57.4%	0.88%
Nei's gene diversity (h)	0.155	0.0036
PIC	0.12	0.003

In this study, the whole genome sequence has increased our knowledge about the palm genome and opened new opportunities to further expand the identification of genetic variation. The advancement of new sequencing technologies had made it possible to highlight SNPs/indel that could differentially distinguish genotypes with distinct traits. However, many other biological questions can be answered through polymorphisms that are assayed in a subset of genomic regions. A massive flow of data from the entire genome sequence of the date palm is to be expected and available, which will help in understanding the genetic bases of this crop and its interaction with the environment, as well as for identification of a molecular marker linked to sex and insights into improving yield, quality and disease resistance. The entire genome will yield the discovery of functions for various genes and establish strategies to manipulate them in order to improve cultivars of interest and create new ones. Therefore, a lot of information still needs to be collected and the genome of the date palm must be studied in more detail.

REFERENCES

Agarwal M, Shrivastava N, Padh H. 2008. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Reproduction,* 27(4): 617-31.doi: 10.1007/s00299-008-0507-z.

Ahmed IA, Ahmed AWK, Robinson RK. 1995. Chemical composition of date varieties as influenced by the stage of ripening. *Food chemistry,* 54:305-309.

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh

YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J. 2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature biotechnology*, 29:521-527.

Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, Yu X, Liu J, Pan L, Zhang T, Yin Y, Xin C, Wu H, Zhang G, Ba Abdullah MM, Huang D, Fang Y, Alnakhli YO, Jia S, Yin A, Alhuzimi EM, Alsaihati BA, Al-Owayyed SA, Zhao D, Zhang S, Al-Otaibi NA, Sun G, Majrashi MA, Li F, Tala, Wang J, Yun Q, Alnassar NA, Wang L, Yang M, Al-Jelaify RF, Liu K, Gao S, Chen K, Alkhaldi SR, Liu G, Zhang M, Guo H, Yu J. 2013. Genome sequence of the date palm *Phoenix dactylifera* L. *Nature Communications*, 4:2274. DOI: 10.1038/ncomms3274

Aly,S., Saad,M.E., Madboly,E.A., Barakat,H.S. and El-Atroush,H. 2019 Genetic Diversity of Egyptian Date Palms (Phoenix dactylifera L.) Using Morphological and Molecular Markers. *Egyptian Academic Journalof Biological Science (C.Physiology and Molecular Biology)*, 10(2):55-69

Ashraf, M.F., Hou, D., Hussain Q., Imran M., Pei, J., Ali,M. , Shehzad, A. , Anwar ,M ,, Noman, A ,Waseem, M. and Lin,X. 2022 .Entailing the Next-Generation Sequencing and Metabolome for Sustainable Agriculture by Improving Plant Tolerance *Int. J. Mol. Sci.* , 23, 651. https://doi.org/10.3390/ijms23020651

Barrow SC. 1998. A monograph of *phoenix* L. (palmae: Coryphoideae). *Kew bulletin*,:513-575.

Bekheet SA, El-Sharabasy SF. 2015. Date Palm Status and Perspective in Egypt. In: Al-Khayri JM, Jain SM, Johnson DV, editors. Date Palm Genetic Resources and Utilization: Volume 1: Africa and the Americas. Dordrecht: Springer Netherlands. p 75-123.

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF. 2014. Plant genome sequencing—applications for crop improvement. *Current opinion in biotechnology*, 26:31-37.

Chao CT, Krueger RR. 2007. The date palm (*Phoenix dactylifera* L.): overview of biology, uses, and cultivation. *Horticultural Science*, 42:1077-1082.

Cullis C. 2011. Molecular markers in date palm. In: Date Palm Biotechnology: Springer. p 361-370.

Faqir N, Muhammad A, Hyder MZ. 2017. Diversity assessment and cultivar identification in date palm through molecular markers-a review. *Turkish Journal of Agriculture-Food Science and Technology*, 5:1516-1523.

Gotch T, Noack D, Axford G. 2006. Feral tree invasions of desert springs. In: Third International Date Palm Conference. Abu Dhabi, United Arab Emirates. p 40.

Hazzouri KM, Flowers JM, Visser HJ, Khierallah HS, Rosas U, Pham GM, Meyer RS, Johansen CK, Fresquez ZA, Masmoudi K. 2015. Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. *Nature communications*, 9;6:8824. doi: 10.1038/ncomms9824.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular biology and evolution*,:msy096.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2:231-239.

Rizk R, El Sharabasy S, El Bana A. 2004. Morphological Diversity of Date palm (*Phoenix dactylifera* L.) in Egypt. I. Dry date cultivars.

Egyptian Journal of Biotechnology, 16:482-500.

Sharma A, Namdeo AG, Mahadik K. 2008. Molecular Markers: New Prospects in Plant Genome Analysis. *Pharmacognosy reviews*, 2.(3). Available online: http://www.phcogrev.com

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15:121-132.

Sterky F., Lundeberg J. 2000. Sequence analysis of genes and genomes. *Journal of Biotechnology*, 76:1-31.

Varshney RK, Terauchi R, McCouch SR. 2014. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS biology*, 12:e1001883.

Zhang X, Tan J, Yang M, Yin Y, Al-Mssallem I, Yu J. 2011. Date palm genome project at the Kingdom of Saudi Arabia. In: Date Palm Biotechnology: Springer. p 427-448.