BIOTECHNOLOGICAL STUDIES ON EGYPTIAN DATE PALM

By

MORAD MOKHTAR MOKHTAR MOHAMED
B.Sc. Agric. Sci. (Biotechnology), Fac. Agric., Al-Azhar Univ., 2011

THESIS
Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

Agricultural Sciences
(Genetics)

Department of Genetics
Faculty of Agriculture
Cairo University
EGYPT

2016

APPROVAL SHEET

BIOTECHNOLOGICAL STUDIES ON EGYPTIAN DATE PALM

M.Sc. Thesis
In
Agric. Sci. (Genetics)

By

MORAD MOKHTAR MOKHTAR MOHAMED
B.Sc. Agric. Sci. (Biotechnology), Fac. Agric., Al-Azhar Univ., 2011

APPROVAL COMMITTEE

Dr. DINA AZIZ EL-KHISHIN...
Head Research of Genetics, Agricultural Research Center

Dr. MONA HASHEM AHMED HUSSEIN………..................................
Professor of Genetics, Faculty of Agriculture, Cairo University

Dr. SALAH EL-DIN SAYED MOHAMED EL-ASSAL …………
Professor of Genetics, Faculty of Agriculture, Cairo University

Dr. EBTISSAM HUSSEIN ALY HUSSEIN ……………………………
Professor of Genetics, Faculty of Agriculture, Cairo University

Date: / /

SUPERVISION SHEET

BIOTECHNOLOGICAL STUDIES ON EGYPTIAN DATE PALM

M.Sc. Thesis
In
Agricultural Sci. (Genetics)

By

MORAD MOKHTAR MOKHTAR MOHAMED
B.Sc. Agric. Sci. (Biotechnology), Fac. Agric., Al-Azhar Univ., 2011

SUPERVISION COMMITTEE

Dr. EBTISSAM HUSSEIN ALY HUSSEIN
Professor of Genetics, Faculty of Agriculture, Cairo University

Dr. SALAH EL-DIN SAYED MOHAMED EL-ASSAL
Professor of Genetics, Faculty of Agriculture, Cairo University

Dr. SAMI SAID ADAWY (Late)
Head Research of Genetics, Agricultural Genetic Engineering Research Institute, Agricultural Research Center

Name of Candidate: Morad Mokhtar Mokhtar Mohamed **Degree:** M.Sc.
Title of Thesis: Biotechnological Studies on Egyptian Date Palm.
Supervisors: Dr. Ebtissam Hussein Aly Hussein
Dr. Salah El-Din Sayed Mohamed EL-Assal
Dr. Sami Said Adawy
Department: Genetics Approval: / /

ABSTRACT

In recent years, date palm has been subjected to intensive genome sequencing studies. The advances in bioinformatics have provided the scientists with tools to develop useful SSR markers that help the breeder accelerating their breeding programs. A total of 172,075 SSR motifs was identified in date palm genome sequence with a frequency of 450.97 SSRs per Mbp. Out of these, 130,014 SSRs (75.6%) were located within the intergenic regions. While, only 42,061 SSRs (24.4%) were located within the genic regions. A number of 111,403 of SSR primer pairs were designed, that represent 292.39 SSR primers per Mb. Out of the 111,403 only 31,380 SSR primers were developed in the genic regions, while 80,023 primers were developed in the intergenic regions. A number of 250,507 SNPs were recognized in 84,172 SSR flanking regions, which represent 75.55% of the total SSR flanking regions. Out of 12,274 genes only 463 genes comprising 896 SSR primers were mapped onto 111 pathways using KEGG data base. The most abundant enzymes were identified in the pathway related to the biosynthesis of antibiotics. Validation of the designed SSR primers was conducted using *in silico* and *in vitro* PCR. We tested 1031 SSR primers using both publicly available date palm genome sequences as templates in the *in silico* PCR reactions. When using the date palm genome PDK30 sequence, all the 1031 SSR primer pairs successfully found complementary sequences. However, only 903 SSR primers could successfully hit within the ATBV01 genome. For *in vitro* validation, 31 SSR primers among those used in the *in silico* PCR were synthesized and tested for their ability to detect polymorphism among six Egyptian date palm cultivars. All tested primers have successfully amplified products, but only 16 primers detected polymorphic amplicons among the studied date palm cultivars.

Key words: Date palm, SSR, Intergenic regions, Genic regions, SNPs, pathways, KEGG, *In vitro* PCR, *In silico* PCR.

DEDICATION

I dedicate this work to whom my heart felt thanks to: the soul of Prof. Dr. Sami Adawy who participated in supervision of this study but passed away during preparing the paper and thesis, to the soul of my father, the most influential person in my life, who showed me what the attitude towards life's responsibilities should be, to my mother who brought me into this world and made me who I am with constant dedication and unconditional support, to my brothers and sisters for all the support, and especially to my wife for her support and patience, for all the support and encouragement they continually offered along the period of my post graduation.

ACKNOWLEDGEMENT

I wish to express my sincere thanks, deep gratitude and appreciation to ***Prof. Dr. EBTISSAM HUSSEIN ALI HUSSEIN*** *Professor of Genetics, Fac. Agric., Cairo Univ., for her, supervision, continued assistance, guidance, great interest, encouragement, following the progress of the work with great interest and continuous criticism through the course of the study and during writing the paper and the thesis..*

I would like to express my appreciation to ***Prof. Dr. Sami Said Adawy*** *Deputy Director of Agricultural Genetic Engineering Research Institute, ARC, Giza, Egypt (who passed away during preparing the paper and the thesis) for his supervision, continued assistance, guidance, great interest, encouragement, following the progress of the work with great interest, and continuous criticism through the course of the study.*

I would like to express my appreciation to ***Prof. Dr. Salah El-Din Sayed El-Assal*** *Professor of Genetics, Fac. Agric., Cairo Univ., for his, supervision, continued assistance, guidance, great interest, encouragement, following the progress of the work with great interest, continuous criticism through the course of the study and during writing the paper and the thesis. Many thanks are due to all the members of the Genetics Department, Fac. Agric., Cairo Univ., and to all members of (MGGM Lab.), Agricultural Genetic Engineering Research Institute, ARC, Egypt for their continuous help, facilities and moral support.*

CONTENTS

Page

INTRODUCTION……………………………………………………….. **1**

REVIEW OF LITERATURE……………………………………… **5**

1. Characteristics of date palm trees……………………………... **5**

2. Plant genomic studies……………………………………………….. **5**

3. Date palm genomic studies………………………………………. **6**

4. Simple sequence repeats in date palm………………………. **8**

5. Single nucleotide polymorphisms………………………………. **12**

a. SNP identification based on EST sequence data……………... **13**

b. SNP identification from sequenced genomes……………..... **13**

6. SNPs studies in date palm………………………………………… **14**

7. Resistance genes……………………………………………………… **16**

8. Detection of resistance genes in date palm…………………... **22**

9. Genome annotation…………………………………………………... **22**

MATERIALS AND METHODS……………………………….. **27**

1. Source of genomic sequences and SNPs data……………... **27**

2. SSR motifs mining and primers design……………………… **27**

3. Validation of the designed SSR primers…………………….. **28**

a. *In silico* validation……………………………………………... **28**

b. *In vitro* validation…………………………………………………… **58**

4. Sequence annotation and pathway analysis…………………... **63**

RESULTS AND DISCUSSION………………………………… **64**

1. Distribution of the different SSR motifs in the date palm genome…………………………………………………………………. **64**

2. SSR primers design………………………………………………… **75**

3. Determination of SNPs located within the SSRs flanking regions……………………………………………………….. **76**

4. Development of date palm SSR primers database…………. **77**

5. Detection of SSRs in disease resistance related genes…….. **78**

6. Validation of the designed primers……………………………. **79**

a. *In silico* validation…………………………………………………. **79**

b. *In vitro* validation of the designed SSR primers…………….. **79**

7. Functional classification by KEGG pathway analyses…… 236

CONCLUSION……………………………………………………………. 296

SUMMARY…………………………………………………………………. 297

REFERENCES……………………………………………………………. 302

ARABIC SUMMARY

LIST OF TABLES

No. **Title** **Page**

1. The 1,031 SSR primer pairs used in the *in silico* validation methods…………………………………………….. 29

2. The names, sequences and melting temperatures of the thirty one SSR primer pairs used in *in vitro* validation of the designed primers in the present investigation………… 59

3. Summary of the identified SSR motifs in the date palm genome sequences……………………………………………... 65

4. The details of the identified SSRs in date palm intergenic regions…………………………………………………………... 66

5. The details of the identified SSRs in date palm exon regions…………………………………………………………... 67

6. The details of the identified SSRs in date palm intron regions…………………………………………………………... 68

7. Comparison between non-triplet repeat types according to the number of repeats in both genic and intergenic regions…………………………………………………………… 70

8. Comparison between triplet repeat types according to the number of repeats in both genic and intergenic regions ... 70

9. Different substitution probabilities that may convert tri repeats in date palm exons to stop codons…………………… 75

10. SSR primers designed within disease resistance related genes……………………………………………………………… 80

11. *In silico* PCR results across PDK30 sequence, which represent primers Id in our data base, contig Id in date palm genome sequence (PDK30 sequence), the start position of primers on contig and the length of each product sequence when using these primers across the date palm genome sequence.. 81

12. *In silico* PCR results across ATB01 sequence, which represent primers Id in our data base, contig Id in date palm genome sequence (ATB01 sequence), the start position of primers on contig and the length of each product sequence when using these primers across the date palm genome sequence.. 148

13. The comparison between the summarized results of the *in silico* and *in vitro* PCR.. 237

14. The contig ID in date palm genome sequence (Contig ID), the gene start on contig (Start), the gene end on contig (End), primers names in our data base (Primers name) 241

15. The data for the pathways name, name of each enzyme, EC numbers and the position of their coding genes within the genome.. 256

LIST OF FIGURES

No.	Title	Page
1.	The distribution of the various SSR classes in genic (introns and exons) and intergenic regions in date palm genome…………………………………………………………...	69
2.	The comparison between mono, di and tri- repeats in both genic and intergenic regions………………………………..	72
3.	Comparison between tetra, penta and hexa- repeats in both genic and intergenic regions……………………………………….	73
4.	Comparison between primers product length within genic and intergenic regions on date palm genome…………………	76
5.	Comparison between the number of SSR primers and the number of detected SNPs within the genic and the intergenic regions in the date palm genome……………………..	77
6.	*In vitro* PCR analysis of the genomic DNA from six date palm cultivars………………………………………………………..	231
7.	*In vitro* PCR analysis of the genomic DNA from six date palm cultivars………………………………………………………..	233
8.	The annotation of the 63 enzymes assigned using our SSR primers in the biosynthesis of antibiotics pathway as revealed by the Blast2GO program……………………………...	290
9.	The annotation of the 23 enzymes assigned using our SSR primers in the starch and sucrose metabolism pathway as revealed by the Blast2GO program……….........................……	292

10. The annotation of the 23 enzymes assigned using our SSR primers in the starch and sucrose metabolism pathway as revealed by the Blast2GO program…………….. 294

11. The annotation of the 23 enzymes assigned using our SSR primers in the starch and sucrose metabolism pathway as revealed by the Blast2GO program…………….. 295

INTRODUCTION

Date palm (*Phoenix dactylifera* L.) is one of the most economically important fruit trees in the Middle East and North Africa. It is a dioecious perennial monocotyledon, cross pollinated plant belonging to the order Arecaceae (Barrow, 1998). It has been cultivated in the Middle East since about 4000 B.C. (Zohary and Spiegel-Roy, 1975). There are over 2000 varieties that vary in shape, color, size, and weight (Al-Farsi and Lee, 2008). Despite such a large number of varieties, for many years the detection of genetic variation in date palm relied mainly on the morphological variation between cultivars (Elhoumaizi *et al*., 2002). Morphological variations are uncertain and unstable because they could be affected by environmental factors, epistasis, pleiotropic effects and other factors. Therefore, as long as breeding programs depended on morphological markers alone, the pace of progress was bound to be slow.

During the last few decades molecular markers were employed to detect the genetic polymorphism more accurately, and as useful tools in crop improvement and breeding programs. Amplified fragment length polymorphism (AFLP) (Adawy *et al*., 2005), simple sequence repeats (SSRs) (Hamwieh *et al*., 2010), randomly amplified polymorphic DNA (RAPD) and inter simple sequence repeats (ISSR) (Hussein *et al*., 2005) are among the most widely used molecular markers. Compared with other types of molecular markers, SSRs have many advantages, such as simplicity, effectiveness, abundance, hyper variability, reproducibility, co-dominant inheritance and extensive genomic coverage (Powell *et al*., 1996 and Wei *et al*., 2011).

A growing number of genes that confer resistance to a diverse spectrum of pathogens have been isolated from a wide range of plant species (Richter and Ronald, 2000 and Hulbert *et al*., 2001). These ″R″ genes have been classified into several groups based on the structural similarities of their predicted protein products.

As R genes from different plant species share conserved domains, they can be used to screen plant genomes for R genes and putative R genes (e.g., resistance gene analogs, RGAs), and to create molecular markers (Takken *et al*., 2006).

In many plants, single nucleotide polymorphism (SNP) markers are increasingly becoming the marker system of choice. For many crop plants there are surprisingly low numbers of validated SNP markers available. However, These SNP markers are needed in large numbers for studies regarding genetic variation, linkage mapping, population structure analysis, association genetics, map-based gene isolation and plant breeding (Ganal *et al*., 2009).

One of the main benefits of sequencing genomes is the identification of genes involved in the biological pathways. Analysis of biological pathways in a genome is a complicated task since a number of biological entities are involved in pathways and biological pathways in different organisms are not identical. Computational pathway identification and analysis thus utilize a number of computational tools and databases and are typically done in comparison with pathways in other organisms. So, information systems for reconstructing, annotating, and analyzing biological pathways are much needed (Choi and Kim, 2008).

Functional annotation allows categorization of genes in functional classes, which can be very useful to understand the physiological meaning of large amounts of genes and to assess functional differences between subgroups of sequences. The Gene Ontology (GO) developed at the GO Consortium (Ashburner *et al*., 2000) provides a suitable framework for this kind of analysis, due to the wide scope of biology covered and its directed acyclic graph (DAG) structure that enables visualization in the context of biological dependences.

A variety of desktop and web applications are available to electronically assign GO terms to unknown sequences based on similarity (Khan *et al*., 2003, Zehetner, 2003, Groth *et al*., 2004 and Martin *et al*., 2004) or to analyze genomic data in the context of gene annotation (Doniger *et al*., 2003 and Al-Shahrour *et al*., 2004).

In recent years, date palm was subjected to intensive genome sequencing studies. Yang *et al*. (2010) reported the complete chloroplast genome sequence. The mitochondrial genome was assembled by Fang *et al.* (2012) with an approximate length of about 715,001 bp. While, the first nuclear genome sequence of date palm was published by Al-Dous *et al*. (2011), where it covered ~60% of the genome (~380 Mb). They recognized more than 3.5 million polymorphic sites among the nine investigated varieties of date palm. Also, Al-Mssallem *et al*. (2013) reported the second nuclear genome