African Journal of Biotechnology Vol. 9 (17), pp. 2499-2509, 26 April, 2010
Available online at http://www.academicjournals.org/AJB
ISSN 1684–5315 © 2009 Academic Journals

Full Length Research Paper

Mining olive genome through library sequencing and bioinformatics: Novel sequences and new microsatellites

Ekrem Dundar* and Oznur Suakar

Balikesir Universitesi, Fen Edebiyat Fakultesi, Biyoloji Bolumu, Balikesir, Turkey.

Accepted 27 April, 2010

As one of the initial steps of olive (*Olea europaea* L.) genome analysis, a small insert genomic DNA library was constructed (digesting olive genomic DNA with *Sma*I and cloning the digestion products into pUC19 vector) and randomly picked 83 colonies were sequenced. Analysis of the insert sequences revealed 12 clones that have no matches to previously characterized/ confirmed sequence records, and 5 insert sequences that are completely new to any nucleotide database available. The remaining sequences had homology to previously described protein coding genes (13%), ribosomal RNAs/tRNAs (24%), phage DNA (1%) and non-functional sequences (such as "chloroplast DNA", "*Lotus* chromosome 3" or "*Arabidopsis* chromosome 2") that are confirmed for accuracy but have not been assigned a function (22%). Analysis of the insert sequences employing multiple bioinformatics tools including a secondary structure prediction analysis revealed potential properties such as coding regions, regulatory sequences and microsatellites that helped to extract more information especially about insert sequences with no hits to any sequence record with a described function. Our results and analyses also suggested that olive di-nucleotide microsatellites with a repeat number of three [$(XY)_3$] could be informative and therefore should not be excluded from studies involving microsatellite analysis. Common insights extracted from multiple bioinformatics analyses suggested that the utilization of these tools can be useful for mining genomic sequences.

Key words: *Olea europaea* L., olive genome, simple sequence repeat, novel sequences, di-nucleotide simple sequence repeat content, secondary structure prediction, *Burkholderia* phage.

INTRODUCTION

Olive (*Olea europaea* L.) has been one of the most important cultivated fruit trees throughout history with various aspects including a very long life time in addition to its economic and health value. With the light of recent molecular genetic studies, another aspect of olive has become "rich genetic diversity" (Hatzopoulos et al., 2002). This genetic diversity at cultivar level is important due to significant economic aspects such as yield and chemical/aromatic composition of fruit and olive oil. To resolve the genetic complexity and to differentiate cultivars from one another a number of molecular systematic studies have been conducted such as microsatellite markers (simple sequence repeat, SSR) (Cipriani et al., 2002; Gil et al., 2006; Baldoni et al., 2009; Muzzalupo et al., 2009), random amplified polymorphic DNA (RAPD) (Khadari et al., 2003), amplified fragment length polymorphism (AFLP) (Angiolillo et al., 1999; Sensi et al., 2003) and restriction fragment length polymorphism (RFLP) (la Rosa et al., 2003). Although there are attempts toward olive genome sequencing (Cattonaro et

*Corresponding author. E-mail: dundar@balikesir.edu.tr, ekremdundar@gmail.com. Tel: +90- 537-624-4220. Fax: +90- 266-612-1215.

al., 2008), the complete sequence of olive genome has not been accomplished yet. Hence, the lack of olive genome sequence is a big handicap to resolve olive genetic complexity. Even combining all the molecular genetic studies on olive in the literature and accumulated nucleotide records are far from representing olive genome and unraveling olive gene composition.

Genomic survey sequences are important resources for molecular analysis of genomes with respect to genetic mapping, map-based gene cloning and contributing to initial steps of genome sequencing. When running an inquiry at GenBank nucleotide database (January, 2010) for *Zea*, *Arabidopsis*, *Glycine*, *Nicotiana*, *Vitis* and *Populus*; 4452798, 228949, 1849484, 1833749, 736685 and 577502 records return, respectively, while only 4907 total nucleotide records were found for *Olea*. When subtracting expressed sequence tags (EST) records (3758) from this number, and considering a large percentage of the remaining 1147 records as ribosomal RNA sequences and SSR records, it is clearly seen that the olive genome is very poorly known at sequence level. Only 2 genomic survey sequence (GSS) records of olive versus 229278 GSS records of *Vitis*, a comparable economic fruit tree for example, further displays the lack in sequence resources of olive genome.

To obtain a general idea about the structure and sequence composition of an organism, construction and sequencing of a small insert genomic DNA library is one of the fastest approaches (Frediani et al., 1999; Swan et al., 2002). For this purpose, a small insert genomic DNA library was constructed (digesting olive genomic DNA with *Sma*I and cloning the digestion products into pUC19 vector) and randomly picked 83 colonies were sequenced. Obtained sequences were analyzed with respect to sequence type, sequence origin and bioinformatic properties such as hairpin formation, functional potentials and SSR content.

MATERIALS AND METHODS

Plant material and genomic DNA isolation

Leaves of *O. europaea* cv. Ayvalik were collected from Gomec Olive Orchard of Edremit Zeytincilik Fidan Uretme İstasyonu (Edremit Olive Germplasm Station) and used for genomic DNA isolation. A slightly modified protocol of phenol/ chloroform DNA extraction (Dellaporta et al., 1983) was used for extracting genomic DNA. Briefly, the fresh leaves were first ground in liquid nitrogen using a mortar and pestle. Six hundred µl extraction buffer (33.6 g/ L urea, 0.5 M ethylenediaminetetraacetic acid (EDTA) pH 8, 1 M Tris-HCl pH 8, 5 M NaCl, 10% SDS) was added to 0.1 g of liquid nitrogen powdered leaves before they were thawed. After adding 500 µl phenol/chloroform/isoamyl alcohol (25: 24: 1), the ground tissue was shaked for 5 min by inverting up and down. The mixture was then centrifuged for 5 min at 13000 g and the supernatant was transferred into a clean microfuge tube where it was mixed with 1/10 supernatant volume of 3 M sodium acetate (pH 5.2). One supernatant volume of isopropanol (at room temperature) was then added and the mixture was centrifuged for 1 min at 13000 g after mixing again by inverting. The pellet was resuspended in 500 µl TE (10 mM Tris-HCl - 1mM EDTA, pH 8) and incubated with 5 µl RNase A (10 mg/ ml) for 30 min at 37℃ to remove RNA. The mixture was then mixed with 55 µl 3M sodium acetate (pH 5.2), 1 ml ethanol (95%), incubated for 30 min at - 20℃ (or for 10 min at - 80℃), and centrifuged for 10 min at 17000 g. The pellet was resuspended in 50 µl TE and it was further cleaned (for efficient restriction digestion) with columns of DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following manufacturer's instructions.

Library construction

For genomic DNA library construction, around 50 µg of genomic DNA was digested with *Sma*I, purified on column using QIAquick polymerase chain reaction (PCR) purification kit (Qiagen, Hilden, Germany), and ligated into the pUC19 (Fermentas, Vilnius, Lithuania) that was pre-digested with the same enzyme (*Sma*I) and phosphatase treated. 5 - 10 µl of the ligation reaction was used to transform chemically competent *Escherichia coli* strains DH5-α (Invitrogen, Carlsbad, CA) and GM2163 (Fermentas, Vilnius, Lithuania). Obtained white colonies on ampicillin/isopropyl-β-D-thiogalactopyranoside (IPTG)/ 5-bromo-4-chloro-indoly-β-d-galactoside (X-Gal) containing plates were inoculated into liquid LB (Luria-Bertani) medium containing 100 µg/ mL ampicillin, shake-incubated overnight and employed for plasmid isolation using GeneJET Plasmid Miniprep Kit (Fermentas, Vilnius, Lithuania). Isolated plasmids were digested with *Sma*I for insert control and sequenced at RefGen (Gen Arastirmalari ve Biyoteknoloji, Ankara) using an ABI 3130XL Genetic Anaylzer (Applied Biosystems, Fostercity, CA) with a BigDye Cycle Sequencing kit (Applied Biosystems, Fostercity, CA). Inserts were sequenced at least twice with a house primer (pUC47, forward) at RefGen, and confirmed lengths of the sequences were used for analysis. The clones were numbered based on sequence of isolation and named after Oe (*O. europaea*) as in Oe150. The GenBank accession numbers of the sequences are provided in Table 1.

Bioinformatics analysis

Insert sequences obtained were first analyzed using FincTV v1.4 (Geospiza, Seattle, WA) and BioEdit (Hall, 1999) for chromatogram quality and contig construction. After accurate sequences were constructed using at least 2 independent sequencing results, they were basic local alignment search tool (BLAST)-searched against non-redundant (nr) databases of BLASTn and BLASTx of GenBank (Altschul et al., 1990) and the homolog sequences from other plants (or from other organisms) were determined. In cases where no acceptable (expect value greater than 1e-5) matches returned for these databases, the insert sequences were further analyzed using other online bioinformatics tools [such as RepeatMasker (http://www.repeatmasker.org) for potential repetitive sequences, Web Promoter Scan (http://www-bimas.cit.nih.gov/ molbio/proscan) for potential promoter candidate sequences, and GENSCAN (http://genes.mit.edu/GENSCAN.html) and GenMark (http://exon.biology.gatech.edu/eukhmm.cgi) for potential gene finding)] in addition to all other relevant GenBank nucleotide databases (refseq_rna, refseq_genomic, chromosome, est, gss, HTGS, pat, dbsts, wgs, env_nt). These tools were mentioned in the

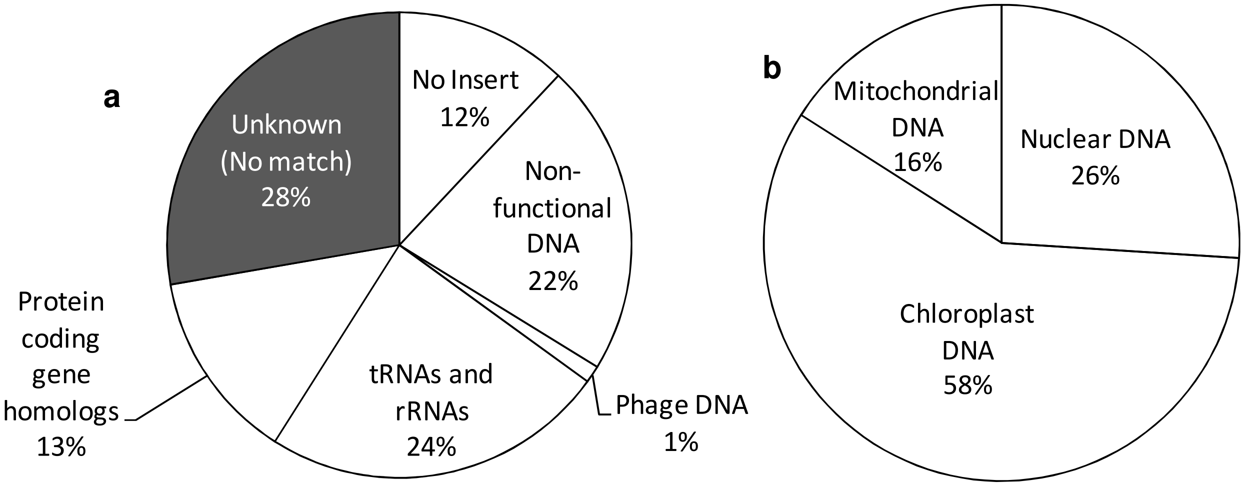

Figure 1. Insert profile of genomic DNA based on homology with GenBank sequences. (a) Overall profile of sequences with functional categories. (b) Percentages of sequences displaying homology with nuclear, chloroplast and mitochondrial sequence records.

results section if a significant result was obtained. An online (http://www.tbi.univie.ac.at/RNA) RNA secondary structure predicttion software for long sequences (Hofacker et al., 1994; Mathews et al., 1999) was utilized to predict hairpin formation potentials of insert sequences. To predict any potential promoters, a promoter prediction database (http://www-bimas.cit.nih.gov/ cgi-bin/molbio/proscan) was utilized. To determine the type and number of SSRs, tandem repeat finder (TRF) (Benson, 1999) was used. Since we were unable to detect di-nucleotide repeats with a repeat number of 3 [$(XY)_3$] using TRF, insert sequences were manually searched for these repeats simply using a word processor. When a single di-nucleotide SSR could be counted for both conjugate ones (as in AGAGAGA that could be counted both as $(AG)_3$ and $(GA)_3$), the number entered just 1 combining the cells for both (Table 4).

RESULTS

Overall insert profile

To conduct a simple molecular analysis of olive genome, recombinant plasmids from 83 randomly picked white colonies (Materials and Methods) were isolated and sequenced for their inserts. While a small amount (12%) of the colonies had insert-free plasmids (pUC19), a great majority (88%) of the colonies did yield inserts (Figure 1a) ranging from 50 nucleotides to 10000 nucleotides in size (restriction digestion results not shown). Based on homologous records of GenBank, a significant majority (58%) of all inserts had similarity to chloroplast genome while nuclear genome and mitochondrial genome were represented with 26 and 16%, respectively (Figure 1b). The largest percentage (28%) of all inserts had no similarity to any records of the non-redundant databases of BLASTn and BLASTx. Non- functional DNA (DNA sequences that are confirmed but not yet assigned any function such as sequence records annotated “chloroplast DNA”, “*Lotus* chromosome 3” or “*Arabidopsis* chromosome 2”) was represented with 22% while rRNAs and tRNAs (combined) constituted 24% of all inserts analyzed (Figure 1a). Insert sequences that have homology to protein coding genes represented 13% of the library. One interesting finding was to detect a *Burkholderia* phage (Summer et al., 2007) DNA (GenBank Accession: gi163716655) in olive leaves.

Insert sequences displaying homology to GenBank records with functional information

Insert sequences that had homology to GenBank records with a described function were categorized in RNA homologes (rRNAs and tRNAs) and protein coding gene homologes. Sequences displaying homology with previously described protein coding genes were further detailed in Table 1. Some sequences (Oe80-Oe160) that were categorized under RNA homologs were also listed in Table 1 since they also contain protein coding region (NADH dehydrogenase subunit 2, *ndhB* gene). Interestingly, *ndhB* homolog inserts constituted 8% of the library. Homologs of Ycf2, a hypothetical protein that is predicted to be an essential gene (Drescher et al., 2000), were the second most abundant (4%) inserts while all other inserts except RNA polymerase beta subunit homologs (2%) were detected once in the library (Table 1).

Table 1. NCBI BLASTn and BLASTx records (with functional information) displaying homology with insert sequences obtained from the genomic DNA library.

Clone name	GenBank Accn	Size (nt)	Matches from NCBI BLAST searches (BLASTn and BLAST x Databases)	E- Value	GenBank accession no (match)
Oe12	GS262902	372	*Pinus radiata* cytochrome *P450* (PRE74) mRNA	6e-04	gi2935524
Oe80 Oe113 Oe119 Oe128 Oe143 Oe155 Oe160	GS262903	735	*Lactuca sativa* NADH dehydrogenase subunit 2 (*ndhB*) gene, partial cds; ribosomal protein S7 (*rps7*) gene, complete cds; ribosomal protein S12 (*rps12*) gene, partial cds; tRNA-Val, 16S ribosomal RNA, tRNA-Ile, and tRNA-Ala genes, complete sequence; 23S ribosomal RNA gene, partial sequence; and unknown gene; chloroplast	0.0	gi66865837
Oe117 Oe145	GS262905	725	*Nicotiana tabacum Ycf2* gene (chloroplast gene)	3e-117	gi7564778
Oe125	GS262907	777	*Olea europaea matK* gene for maturase (chloroplast gene)	0.0	embAJ429335
Oe129 Oe164	GS262908	832	*Nicotiana tabacum* RNA polymerase beta subunit (chloroplast gene)	1e-127	gi3735092
Oe137	GS262909	225	*Anethum graveolens Ycf2* gene (chloroplast gene)	2e-89	gi156573684
Oe163	GS262911	801	*Sonneratia alba* PSII 10 kDa phosphoprotein (chloroplast gene)	4e-13	gbACS68679
Oe110	GS262904	715	*Solanum tuberosum* apocytochrome b (*cob*) gene, mitochondrial gene encoding mitochondrial protein	2e-118	gi800845
Oe123	GS262906	687	*Triticum aestivum* mitochondrial ATP synthase B chain precursor	1e-37	gi3800069
Oe141	GS262910	647	*Cucumis sativus* NADH dehydrogenase subunit 5 (*nad5*) gene, partial cds; mitochondrial gene for mitochondrial product	4e-23	gi31322694

Sequences were sorted according to organelle that have similarity. All inserts except Oe12 (which resembles to a nuclear gene) have homology to either chloroplast (rows 2 - 7, highlighted gray) or mitochondrial (rows 8 - 10) genes. Accession numbers in the second column from the left belong to the sequences obtained through this study while the ones in the last column belong to the homologous sequences in GenBank.

Novel sequences

Although almost all of the sequences are new records for nucleotide databanks (as olive sequences), a large amount of the inserts (28%) contained sequences that have no similarity to any sequence records described before (Table 2). In non-confirmed/non-characterized sequence containing databases (such as whole genome shotgun sequences, high throughput genomic sequences and expressed sequence tags) however, most of the novel sequences did have matching records which provided clues like “it is homologous to an expressed sequence or to a sequence of a different plant”. Most interestingly, 5 insert sequences (Oe101, Oe106, Oe130, Oe146 and Oe150) had no significant matches in any database available (Table 2). These sequences were further analyzed with various bioinformatics tools to gather any potential information leading to functional prediction. In Table 2, analysis results from all other databases of BLASTn (NCBI, National Center for Biotechnology Information), repetitive DNA database (Repeat Masker, http://www.repeatmasker.org), promoter prediction database (Web promoter scan service) and nucleotide compositions are listed. Promoter prediction analysis (http://www-bimas.cit.nih.gov/cgi-bin/molbio/proscan) for unknown sequences suggested putative promoter regions for only Oe121/Oe136/Oe158 (on reverse strand in 769 to 519, score: 79) and Oe130 (on reverse

Table 2. Insert sequences (obtained from the genomic DNA library) that display no homology with any record of the BLASTn and BLASTx databases.

Clone name	GenBank accession no	Size (nt)	Match in any database* (with accession no and E-Value)	Any information obtained from bioinformatics analysis
Oe3	GS262912	762	Database: Whole Genome Shotgun Sequences GenBank Accn: embCAAP02001845 *Vitis vinifera*, whole genome shotgun sequence. Expect = 2e-91	The largest ORF is 58 aa. AT rich (59 %). Contains 9 AAA repeats and 26 TTT repeats (both dispersed). G (20 %) and C (21 %) ratios are close to equal (A: 33%, T: 25).
Oe101	GS262914	210	None	The largest ORF is 69 aa. AT rich (67 %). G (20 %) and C (13 %) ratios are not close.
Oe131	GS262918	712	Database: High Throughput Genomic Sequences GenBank Accn: dbjAP009926 *Lotus japonicus* clone. Expect = 2e-65	The largest ORF is 50 aa. AT rich (57 %). G (21.90 %) and C (21.49 %) ratios are close to equal (A: 30 %, T: 26 %)
Oe138 Oe142	GS262919	727	Database: Repeat Masker** Matching Record: Simple repeat “TAAAAA)n#Sim”. Expect = N/A	The largest ORF is 59 aa. Constitutes 2 % of the library (2 clones). AT rich (70%). G (14.86 %) and C (14.72 %) ratios are close to equal (A: 32%, T: 38%).
Oe146	GS262921	153	None	The largest ORF is 46 aa. AT rich (66 %). G (12 %) and C (22 %) are not close.
Oe149 Oe162	GS262920	705	Database: Whole Genome Shotgun Sequences GenBank Accn: gbAASG02018402 *Ricinus communis* whole genome shotgun sequence. Expect = 5e-14	The largest ORF is 96 aa. Constitutes 2 % of the library (2 clones). AT rich (57 %). G (21.75 %) and C (21.33 %) ratios are close to equal (A: 30 %, T: 26 %).
Oe4	GS262913	147	Database: EST GenBank Accn: embAJ785235 *Lycopersicon esculentum* fruit 12 days post anthesis. Expect = 2e-25	The largest ORF size is 46 aa. Not AT rich (50.62 %). G (47 %) and C (32 %) ratios are not close to equal. (A: 31 %, T: 50 %)
Oe106	GS262915	702	None	The largest ORF size is 107 aa . Slightly AT rich (56 %). G (20 %) and C (24 %) ratios are not close to equal (A: 25 %, T: 31 %).
Oe121 Oe136 Oe158	GS262916	774	Database: EST GenBank Accn: gbFG447054 *Actinidia deliciosa* dormant buds before hydrogen cyanamide treatment Expect = 2e-68	The largest ORF size 122 aa. Not AT rich (53 %). G (21 %) and C (24 %) ratios are not close to equal (A: 25 %, T: 30 %). Constitutes 3 % of the library (3 clones).
Oe130	GS262917	753	None	The largest ORF size is 170 aa. GC rich (64 %). G (36 %) and C (28 %) ratios are not close to equal (A: 17 %, T: 17 %).
Oe150	GS262922	177	None	The largest ORF size is 58 aa. Slightly AT rich (54 %). G (27 %) and C (18 %) ratios are not close to equal (A: 28 %, T: 27)
Oe154	GS262923	737	Database: EST GenBank Accn: embCT983580 *Eucalyptus gunnii* differentiating xylem Expect = 6e-86	The largest ORF size is 141 aa. Not AT rich (54 %). G (21 %) and C (25 %) ratios are not close to equal (A: 24 %, T: 30 %).

Results from further bioinformatics analysis including BLAST searches from the other databases of NCBI are provided. Lines highlighted gray contain the sequences that share common features including hairpin formation. All other BLASTn and BLASTx databases were blast-searched if no hits were returned for nr (non-redundant) and EST (expressed sequence tags) database. Underlined sequences contain a predicted promoter region. ** http://www.repeatmasker.org

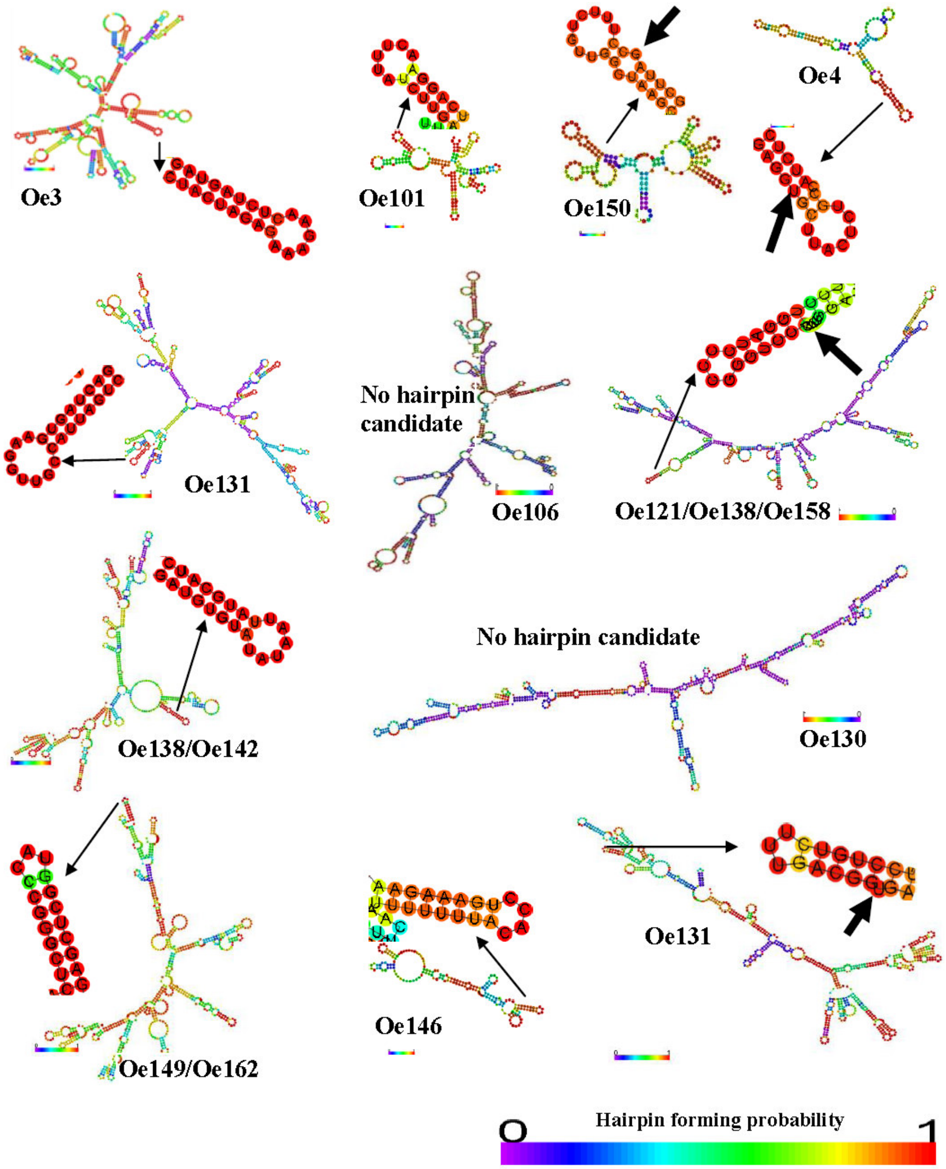

Figure 2. Analysis of sequences for hairpin formation potentials. An RNA secondary structure prediction software for long sequences (Hofacker et al., 1994; Mathews et al., 1999) was utilized to predict hairpin formation potentials. According to the prediction program, red (dark) hairpins have the highest probability to form. Potential hairpins that were also manually checked to confirm base pairs, are magnified and pointed with thin arrows. The thick arrows point out the false base pairings of the less probable hairpins. Most probable hairpins were detected in AT rich sequences that have a very close G and C amounts, contain no open reading frame and no EST matches.

reverse strand in 744 to 494, score: 89).

Hairpins are known to play important roles in various gene regulation processes including intron splicing (Perea and Jacq, 1985), mRNA stability (Klaff et al., 1996) and gene silencing (Helliwell et al., 2002). Of the 12 novel sequences analyzed using an RNA secondary structure prediction software (see Materials and Methods for details), 6 displayed one or more hairpins that were predicted to form (Figure 2). Interestingly, almost all of the sequences that were predicted to form strong hairpins

Table 3. Total SSR content and total SSR rate of the insert sequences.

Clone name	Sequence size (nt)	Number of SSRs	Repeating unit and frequency	SSR Rate
Protein Coding Gene Homologs				
Oe12	372	2	(GAAGATGAAC)$_2$, (TTA)$_3$	0.538%
Oe80- Oe160	735	0	-	0.000%
Oe117 Oe145	725	3	(AAAAAGTATCT)$_2$, (AATTGA)$_2$, (TGGAAA)$_2$	0.414%
Oe125	777	2	(AATATCCA)$_2$, (AAAG)$_2$	0.257%
Oe129 Oe164	832	2	(ACACGT)$_2$, (TCTAG)$_3$	0.240%
Oe137	225	0	-	0.000%
Oe163	801	3	(A)$_{17}$, (GCTTA)$_2$, (T)$_{10}$	0.375%
Oe110	715	2	(AGAT)$_3$, (CAGGATG)$_2$	0.280%
Oe123	687	2	(AATGA)$_2$, (TTCATG)$_2$	0.291%
Oe141	647	2	(GAAGC)$_2$, (GAGAA)$_2$	0.309%
Total	6516	18		0.276%
Unkown Sequences				
Oe3	762	4	(GATT)$_3$, (TTAAGGT)$_2$, (CAAAAGA)$_2$, (TCTTTC)$_2$	0.525%
Oe101	197	1	(TCTTTA)$_3$	0.508%
Oe131	712	3	(CTCTAG)$_3$, (GCAAGAA)$_2$, (GTTACAG)$_2$	0.421%
Oe138 Oe142	727	1	(ATTTTT)$_5$	0.138%
Oe146	153	1	(TTTTG)$_3$	0.654%
Oe149 Oe162	705	1	(AAAG)$_3$	0.142%
Oe4	147	0	-	0.000%
Oe106	702	1	(ATGA)$_3$	0.142%
Oe121 Oe136 Oe138	774	7	(AGTCA)$_2$, (GTGCATTA)$_2$, (TTAGT)$_2$, (CTTTA)$_2$, (CCTTT)$_3$, (AAGAA)$_2$, (AAAG)$_2$	0.904%
Oe130	753	5	(CGCC)$_3$, (AGG)$_4$, (CGAGCAGGAGGA)$_3$, (GCCGG)$_4$, (CGGCCCG)$_3$	0.664%
Oe150	177	0	-	0.000%
Oe154	737	1	(TCTCCA)$_2$	0.136%
Total	6546	25		0.382%

Mitochondrial gene homologs and unknown sequences with common features including hairpin formation and AT percent are shaded.

(Oe3, Oe101, Oe131, Oe138/Oe142, Oe146 and Oe149/Oe162) also had higher adenine-thymine (AT) percent, shorter open reading frame (ORFs), and very close to equal guanine (G) vs. cytosine (C) amounts. Sequences with no strong hairpins, on the other hand, shared features like lower AT percent, larger ORFs (and/

or EST matches) and unequal G vs. C amount.

Microsatellites

Types and numbers of SSRs in each insert sequence are shown in Tables 3 and 4. A control group of olive DNA (148 olive mRNAs totalling in 95552 nucleotides randomly picked from GenBank) used to compare the types and numbers of SSRs to that of the present study, yielded significant differences both in types and total numbers of SSRs. Overall SSR rate of insert sequences (0.329%) was 2.74 fold more than that (115 SSRs/ 95552 = 0.120%) of the control group while some unknown sequences (such as Oe138/Oe142, Oe149/Oe162, Oe106 and Oe154) contained SSRs at a similar rate with that of control and some insert sequences (Oe80-Oe160, Oe137, Oe4 and Oe150) contained no SSRs (Table 3). Detailed analysis of all SSRs in terms of existence rate in all insert sequences revealed that only di-nucleotide SSRs were widely represented in all sequences with mostly more than one (Table 4), while all other SSRs (tri-nucleotide SSRs, tetra-nucleotide SSRs, penta-nucleotide SSRs etc.) were not commonly detected (Table 3). Among di-nucleotide SSRs, $(AG)_3$/ $(GA)_3$ and $(AT)_3$/ $(TA)_3$ were the most abundant while $(GC)_3$/ $(CG)_3$ was the least abundant (Table 4). When looking at the total amount of di-nucleotide SSRs in each insert, some sequences (Oe12, Oe137, Oe146, Oe4 and Oe106, totalling in 1599 nucleotides) interestingly displayed no di-nucleotide SSRs while mitochondrial sequences and an unknown sequence (Oe149/Oe162) contained 2.5 to 4.9 times more di-nucleotide SSRs than that of the control sequences. Oe80 - Oe160, homolog of a chloroplast gene that is a complex sequence consisting of protein coding genes, rRNAs and tRNAs also contained significantly more (3.13 fold of control) di-nucleotide SSRs than that of the control (Table 4). Di-nucleotide SSR rate of all the inserts combined (0.329%) was 1.9 fold of the control group (0.174%).

DISCUSSION

Distribution profile of the insert sequences

An overview of the distribution of sequences obtained displays that olive homologs of chloroplast sequences constitute more than half of the genome (Figure 1). Although this is not unexpected since the total genomic DNA was extracted from green leaves that are known to contain more chloroplast DNA than nuclear DNA (Jope et al., 1978), it also prompts the possibility of nuclear DNA to contain some portion of cytoplasmic DNA as it has been reported for *Arabidopsis* genome where 75% of the mitochondrial DNA was found in chromosome 2 (Lin et al., 1999), and for various photosynthetic organisms including plants to contain nuclear genes with chloroplast origin (Martin et al., 1998). Genomic DNA libraries excluding organelle DNA still display significant amounts of chloroplast and mitochondrial DNA. A previous study covering 84% of *Nicotiana plumbaginifolia* genome (Chen et al., 1996) reported 6% of 22000 clones contained sequences derived from chloroplast DNA. Another study on rice genome (Baba et al., 2000) with 69276 clones revealed 11.8 and 0.9% chloroplast DNA and mitochondrial DNA, respectively. Considering these libraries were made from nuclear DNA (Chen et al., 1996; Baba et al., 2000), our results (that were generated using total genomic DNA) for the percentages of chloroplast and mitochondrial DNA in olive genome are in the range observed for other plants.

All of the insert sequences (except Oe12) having homology with previously characterized sequence records of GenBank (Table 1) are homologs of either chloroplast or mitochondrial DNA, while all of the nuclear sequences (except Oe12) fall in the non-functional category (Figure 1). Considering chloroplast genome is more conserved among plants (Palmer et al., 1988) than that of mitochondrial genome (Fauron et al., 2004), and the distribution of insert sequences among three organelles (Figure 1), it is possible to predict that most of the unknown sequences may be homologs of nuclear genome. Another interesting aspect of the insert profile is that all of the rRNA/tRNA sequences (except 2) are chloroplast sequences which both confirms the dominance of chloroplast DNA (in the starting total genomic DNA) and displays the abundance of RNA coding regions in chloroplast genome. Detection of a *Burkholderia* phage DNA suggests possible existence of *Burkholderia* and/ or the phage in olive leaves.

Bioinformatics properties of the sequences

Insert sequences bearing homology to protein coding genes constituted 13% of the library. For a genome with 46 chromosomes that potentially include numerous non coding DNA regions in addition to introns, it is quite efficient to get 13% gene information even if all except 1 belong to cytoplasmic organelles. On the other hand, the amount of unknown sequences was significant (28%), suggesting olive genome's difference from all other plants with a sequenced genome. Five insert sequences (Oe101, Oe106, Oe130, Oe146 and Oe150) ranging from 153 to 753 nucleotides had no significant hits from any

Table 4. Numbers of dinucleotide repeat microsatellites found in protein coding gene homologs and unknown insert sequences.

Clone name	Size (nt)	Microsatellite (SSR)													
		$(AG)_3$	$(GA)_3$	$(TC)_3$	$(CT)_3$	$(AC)_3$	$(CA)_3$	$(TG)_3$	$(GT)_3$	$(AT)_3$	$(TA)_3$	$(GC)_3$	$(CG)_3$	Total	Rate
Protein coding gene homologs															
Oe12	372	-	-	-	-	-	-	-	-	-	-	-	-	0	0.000%
Oe80 Oe160	735	1	-	1	-	1	-	-	-	-	-	1	-	4	0.544%
Oe117 Oe145	725	1		-	-	-	-	-	-	1	-	-	-	2	0.276%
Oe125	777	-	-	-	-	-	-	-	-	-	2	-	-	2	0.257%
Oe129 Oe164	832	-	-	-	-	-	-	-	-	$(AT)_4$, $(TA)_4$		-	-	1	0.120%
Oe137	225	-	-	-	-	-	-	-	-	-	-	-	-	0	0.000%
Oe163	801	-	-	-	-	-	-	1		1	-	-	-	2	0.250%
Oe110	715	1	-	2		1	1	-	-	$(AT)_4$		-	-	6	0.839%
Oe123	687	2		-	-	-	-	-	-	1		-	-	3	0.437%
Oe141	647	-	2	2		-	-	-	-	-	-	-	-	4	0.618%
Total	6516	7		5		2	1	1		7		1	0	24	0.368%
Unknown sequences															
Oe3	762	-	-	-	1	-	-	-	-	-	-	-	-	1	0.131%
Oe101	197	-	-	-	1	-	-	-	-	-	-	-	-	1	0.508%
Oe131	712	1	-	-	-	-	-	-	-	-	-		-	1	0.140%
Oe138 Oe142	727	-	-	-	-	-	-	1	-	$(TA)_3$, $(TA)_4$		-	-	3	0.413%
Oe146	153	-	-	-	-	-	-	-	-	-	-	-	-	0	0.000%
Oe149 Oe162	705	2		2		1	$(CA)_4$	-	-	-	-	-	-	6	0.851%
Oe4	147	-	-	-	-	-	-	-	-	-	-	-	-	0	0.000%
Oe106	702	-	-	-	-	-	-	-	-	-	-	-	-	0	0.000%
Oe121 Oe136 Oe158	774	-	-	-	-	-	-	-	-	1	-	-	1	2	0.258%
Oe130	753	-	-	1	-	-	-	-	-	-	-	1		2	0.266%
Oe150	177	-	1	-	-	-	-	-	-	-	-	-	-	1	0.565%
Oe154	737	-	-	$(TC)_3$, $(TC)_4$		-	-	-	-	-	-	-	-	2	0.271%
Total	6546	4		7		1	1	1		3		2		19	0.290%
Overall	13062	11		12		5		2		10		3		43	0.329%
Overall SSR Rate		0.084%		0.092%		0.038%		0.015%		0.077%		0.023%			
SSR Rates of control mRNAs*		0.0330%		0.0280%		0.0157%		0.0190%		0.0260%		0.0021%		166	0.174%

Mitochondrial gene homologs and unknown sequences with common features including hairpin formation and AT percent are shaded. Microsatellites with more than 3 di-nucleotide repeats were typed in. Three di-nucleotide microsatellites were also typed in when found together with a 4 di-nucleotide one. For GenBank_accn numbers (Tables 1 and 2). *Control mRNAs are 148 olive genes (95552 total nucleotides all together) randomly picked from GenBank.

nucleotide database available. With this respect, these sequences are completely novel to nucleotide databases. With the vast amount of bioinformatics tools available, however, it was possible to extract some information about these sequences. Oe121/Oe138/Oe158 and Oe130 were strongly predicted to be promoters (Table 2), they did not yield any strong hairpins (Figure 2), they both contained the largest amount (7 and 5 SSRs, respectively) of SSRs (Table 3) and they were 2 of only 3 inserts that contained $(GC)_3/(CG)_3$ repeats (Table 4). Another interesting observation of unknown sequences is that all of them having high AT% with G and C amounts close to equal (Table 2, shaded sequences), and having no EST matches, also contain strong hairpins while those that are not AT rich, G and C amounts are not close to equal, and mostly have EST matches, do not contain any strong hairpins. Considering hairpins' roles in intron splicing (Perea and Jacq, 1985) and in mRNA stability (Klaff et al., 1996), hairpin forming sequences can be predicted to be introns or to intervene the splicing process.

Di-nucleotide microsatellites with a repeat number of three [$(XY)_3$] can be informative

Microsatellites have become a popular tool for genetic analysis of plants. They have already been utilized in olive research such as cultivar identification and population genetics (Cipriani et al., 2002; Hanley et al., 2002; Khadari et al., 2003; Gil et al., 2006; Omrani-Sabbaghi et al., 2007; Doveri et al., 2008; Muzzalupo et al., 2009). Not all the microsatellites are polymorphic (Rallo et al., 2000) hence, new SSRs are always useful for genetic analysis of different populations. Therefore, microsatellites in all insert sequences have been identified (Tables 3 and 4) as potential SSR marker candidates in addition to informative properties of the sequences. Analysis of total SSRs revealed that only di-nucleotide SSRs were widely represented in all sequences with mostly more than one (Table 4), while all other SSRs were not commonly detected (Table 3). The reasons for larger repeating unit SSRs to be less widely present could be due to their larger sizes. With this respect, di-nucleotide SSRs were found more useful to extract information. $(AG)_3$/ $(GA)_3$ and $(TC)_3$/ $(CT)_3$ repeats were the most abundant in overall olive genomic library as in other plants (Wang et al., 1994; Guo et al., 2000; Suwabe et al., 2004) but mitochondrial sequences were significantly rich (0.44%) in these repeats compared to chloroplast sequences (0.05%) and unknown sequences (0.16%) (Table 3). $(AT)_4/(TA)_4$ repeats were found more (0.01%) in chloroplast and mitochondrial sequences than in unknown sequences (0.004%). Taking these results into account, it is possible to make predictions such as Oe149/Oe162 could be a mitochondrial sequence. These observations suggest di-nucleotide SSR content of unknown sequences can be informative about the organelle origin of the sequence.

In previous studies of olive and other plants, repeat numbers of di-nucleotide microsatellites ranged from 5 to 52 (Guo et al., 2000; Rallo et al., 2000; Cipriani et al., 2002; Suwabe et al., 2004; Gil et al., 2006; Tobias et al., 2005; Vogel et al., 2009) while we have detected 3 to 4 repeats of di-nucleotides (Table 4). A question may arise here whether $(XY)_3$ repeats may be at levels that should just be encountered randomly ($1/4^6$ = 0.024%). Rates of SSRs in control sequences (randomly picked olive sequences totaling in 95552 nucleotides), however, clearly display the significantly different levels of SSRs reported such as $(TC)_3/(CT)_3$ proportion (0.092%) being 3.28 fold more than that of control (0.028%), and $(GC)_3/(CG)_3$ proportion (0.023%) and being 11 fold more than the control (0.0021%) proportion (Table 4). With this respect, our results suggest that di-nucleotide SSR content of sequences can be informative for various aspects of DNA sequences such as the organelle origin of the sequence.

Overall results and analyses presented in this work make a contribution to olive genomic sequence resources in addition to providing five completely novel sequences for the nucleotide databases. Detailed analysis of the sequences using multiple bioinformatics tools revealed information content of the sequences that can be utilized to make predictions toward functional information. SSR content analysis suggests that di-nucleotide microsatellites repeated three times [$(XY)_3$] can be informative about sequences and hence they should not be excluded from SSR studies. Finally, the strategy devised with multiple bioinformatics tools proved useful to extract information from genomic sequences about which no other information is available.

ACKNOWLEDGEMENT

Authors would like to thank TUBİTAK-TOVAG for supporting this research with Grant number 106O616.

Abbreviations

SSR, Simple sequence repeat; **RAPD**, random amplified polymorphic DNA; **AFLP**, amplified fragment length polymorphism; **RFLP**, restriction fragment length polymorphism; **EST**, expressed sequence tags; **GSS**, genomic survey sequence; **PCR**, polymerase chain reaction; **IPTG**, isopropyl-β-D-thiogalactopyranoside; **X-Gal**, 5-bromo-4-chloro-indoly-β-d-galactoside.

REFERENCES

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Angiolillo A, Mencuccini M, Baldoni L (1999). Olive genetic diversity assessed using amplified fragment length polymorphism. Theor. Appl. Genet. 98: 411-421.

Baba T, Katagiri S, Tanoue H, Tanaka R, Ikeno M, Ohta T, Umehara Y, Matsumoto T, Sasaki T (2000). Construction and characterization of rice genomic libraries: PAC library of Japonica variety, Nipponbare and BAC library of Indica variety, Kasalath. Bull. Natl. Inst. Agrobiol. Res. 14: 41-52.

Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27: 573-580.

Cattonaro F, Felice N, Stefan A, Vendramin G, Baldoni L, Porceddu E, Porceddu A, Pe ME, Morgante M (2008). Toward olive genome sequencing: First insights into the genome organization. *In* Proceedings of the 52nd Italian Soceity of Agricultural Genetics Annual Congress, Padova, Italy, p. D.78

Chen C-M, Wang C-T, Lee F-M, Ho C-H (1996). Construction and characterization of a *Nicotiana plumbaginifolia* genomic library in a yeast artificial chromosome. Plant Sci. 114: 159-169.

Cipriani G, Marrazzo MT, Marconi R, Cimato A, Testolin R (2002). Microsatellite markers isolated in olive (*Olea europaea* L.) are suitable for individual fingerprinting and reveal polymorphism within ancient cultivars. Theor. Appl. Genet. 104: 223-228.

Dellaporta SL, Wood J, Hicks JB (1983). A plant DNA minipreparation: version II. Plant Mol. Biol. Rep. 1: 19-21.

Doveri S, Sabino Gil F, Díaz A, Reale S, Busconi M, da Câmara Machado A, Martín A, Fogher C, Donini P, Lee D (2008). Standardization of a set of microsatellite markers for use in cultivar identification studies in olive (*Olea europaea* L.). Sci. Hort. 116: 367-373.

Drescher A, Ruf S, Calsa TJ, Carrer H, Bock R (2000). The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. The Plant J. 22: 97-104.

Fauron C, Allen J, Clifton S, Newton K (2004). Plant mitochondrial genomes. *In* H Daniell, CD Chase, eds, Molecular Biology and Biotechnology of Plant Organelles. Netherlands: Springer, pp 151-177.

Frediani M, Gelati MT, Maggini F, Galasso I, Minelli S, Ceccarelli M, Cionini PG (1999). A family of dispersed repeats in the genome of *Vicia faba* : structure, chromosomal organization, redundancy modulation, and evolution. Chromosoma. 108: 317-324.

Gil FS, Busconi M, Machado ADC, Fogher C (2006). Development and characterization of microsatellite loci from *Olea europaea*. Mol. Ecol. Notes. 6: 1275-1277.

Guo JC, Hu XW, Yanagihara S, Yoshinobu E (2000). Isolation and characterization of microsatellites in snap bean. Acta Bot. Sin. 42: 1179-1183.

Hall TA (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In Nucleic Acids Symposium Series, 41: 95-98.

Hanley S, Barker JHA, Van Ooijen JW, Aldana C, Harris SI, Ahman I, Larsson S, Karp A (2002). A genetic linkage map of willow (*Salix viminalis*) based on AFLP and microsatellite markers. Theor. Appl. Genet. 105: 1087-1096.

Hatzopoulos P, Banilas G, Giannoulia K, Gazis F, Nikoloudakis N, Milloni D, Haralampidis K (2002). Breeding, molecular markers and molecular biology of the olive tree. Eur. J. Lipid Sci. Technol. 104: 574-586.

Helliwell CA, Wesley SV, Wielopolska AJ, Waterhouse PM (2002). High-throughput vectors for efficient gene silencing in plants. Funct. Plant Biol. 29: 1217-1225.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M and Schuster P (1994). Fast folding and comparison of RNA secondary structures. Monatsh. Chem. 125: 167-188.

Jope CA, Hira A, Wildman SG (1978). Evidence that the amount of chloroplast DNA exceeds that of nuclear DNA in mature leaves. J. Cell Biol. 79: 631-636.

Khadari B, Breton C, Moutier N, Roger JP, Besnard G, Bervillé A, Dosba F (2003). The use of molecular markers for germplasm management in a French olive collection. Theor. Appl. Genet. 106: 521-529.

Klaff P, Riesner D, Steger G (1996). RNA structure and the regulation of gene expression. Plant Mol. Biol. 32: 89-106.

la Rosa R, Angiolillo A, Guerrero C, Pellegrini M, Rallo L, Besnard G, Berville A, Martin A, Baldoni L (2003). A first linkage map of olive (*Olea europaea* L.) cultivars using RAPD, AFLP, RFLP and SSR markers. Theor. Appl. Genet. 106: 1273-1282.

Lin X KS, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, Feldblyum TV, Buell CR, Ketchum KA, Lee J, Ronning CM, Koo HL, Moffat KS, Cronin LA, Shen M, Pai G, Van Aken S, Umayam L, Tallon LJ, Gill JE, Adams MD, Carrera AJ, Creasy TH, Goodman HM, Somerville CR, Copenhaver GP, Preuss D, Nierman WC, White O, Eisen JA, Salzberg SL, Fraser CM, Venter JC (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature, 402: 731-732.

Martin W, Stoebe B, Goremykin V, Hapsmann S, Haseqawa M, Kwallik KV (1998). Gene transfer to the nucleus and the evolution of chloroplasts. Nature, 393: 162-165.

Mathews DH, Sabina J, Zucker M, Turner H (1999). Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J. Mol. Biol. 288: 911-940.

Muzzalupo I, Stefanizzi F, Perri E (2009). Evaluation of olives cultivated in Southern Italy by simple sequence repeat markers. HortScience. 44: 582-588.

Omrani-Sabbaghi A, Shahriari M, Falahati-Anbaran M, Mohammadi SA, Nankali A, Mardi M, Ghareyazie B (2007). Microsatellite markers based assessment of genetic diversity in Iranian olive (*Olea europaea* L.) collections. Sci. Hort. 112: 439-447.

Palmer JD, Jansen RK, Michaels HJ, Chase MW and Manhart JR (1988). Chloroplast DNA variation and plant phylogeny. Ann. Mo. Bot. Gard. 75: 1180-1206.

Perea J, Jacq C (1985). Role of the 5' hairpin structure in the splicing accuracy of the fourth intron of the yeast cob-box gene. EMBO J. 4: 3281-3288.

Rallo P, Dorado G and Martin A (2000). Development of simple sequence repeats (SSRs) in olive tree (*Olea europaea* L.). Theor. Appl. Genet. 101: 984-989.

Sensi E, Vignani R, Scali M, Masi E, Cresti M (2003). DNA fingerprinting and genetic relatedness among cultivated varieties of *Olea europaea* L. estimated by AFLP analysis. Sci. Hort. 97: 379-388.

Summer EJ, Gill JJ, Upton C, Gonzalez CF and Young R (2007). Role of phages in the pathogenesis of *Burkholderia*, or 'Where are the toxin genes in *Burkholderia* phages?'. Curr. Opin. Microbiol. 10: 410-407.

Suwabe K, Iketani H, Nunome T, Ohyama A, Hirai M, Fukuoka H (2004). Characteristics of microsatellites in *Brassica rapa* genome and their potential utilization for comparative genomics in Cruciferae. Breed. Sci. 54: 85-90.

Swan KA, Curtis DE, McKusick KB, Voinov AV, Mapa FA and Cancilla MR (2002). High-throughput gene mapping in *Caenorhabditis elegans*. Genome Res. 12: 1100-1105.

Tobias C, Twigg P, Hayden D, Vogel K, Mitchell R, Lazo G, Chow E , Sarath G (2005). Analysis of expressed sequence tags and the identification of associated short tandem repeats in switchgrass. Theor. Appl. Genet. 111: 956-964.

Vogel J, Tuna M, Budak H, Huo N, Gu Y, Steinwand M (2009). Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. BMC Plant Biol. 9: 88.

Wang Z, Weber JL, Zhong G, Tanksley SD (1994). Survey of plant short tandem repeats. Theor. Appl. Genet. 88: 1-6.