Genome Sequencing, Transcriptomics, and Proteomics

9

Rosario Muleo, Michele Morgante, Federica Cattonaro, Simone Scalabrin, Andrea Cavallini, Lucia Natali, Gaetano Perrotta, Loredana Lopez, Riccardo Velasco and Panagiotis Kalaitzis

Abstract

This review encompasses the current status of major areas of progress in olive tree genome sequencing, including insights into genome function derived from large-scale gene expressing profiling, and studies on genomic architecture of repetitive sequences, smaller RNA, and proteomics. Olive tree genomics, as well as other omics, is progressing owing to recent developments in next-generation sequencing (NGS) technologies. Biological insights, therefore, are not only resulted from the sequencing initiative, since from genetic mapping, gene expression profiling, gene discovery research, and proteomics over nearly last seven years a large amount of information has been provided by different laboratories. The availability of high-quality genome assembly provides olive biologists with valuable new tools to improve and develop new varieties more efficiently, enabling the implementation of marker-assisted selection and genomic selection, and contributing to the comprehension of the

R. Muleo (✉)
Laboratory of Molecular Ecophysiology and Biotechnology of Woody Plants, Department of Agricultural and Forestry Science, University of Tuscia, S.N.C., Via S.C. De Lellis, s.n.c., 01100 Viterbo, Italy
e-mail: muleo@unitus.it

M. Morgante
Department of Crop and Environmental Sciences, University of Udine, Via delle Scienze, 208, loc. Rizzi, 33100 Udine, Italy

M. Morgante · F. Cattonaro · S. Scalabrin
IGA Technology Service, Via J. Linussio, 51 Z.I.U, 33100 Udine, Italy

A. Cavallini · L. Natali
Department of Agricultural, Food and Environmental Sciences, University of Pisa, Via del Borghetto 80, 56124 Pisa, Italy

G. Perrotta · L. Lopez
Italian National Agency for New Technologies, Energy and Sustainable Economic Development, TRISAIA Research Center, S. S. 106 Jonica Km 419 +500, 75026 Rotondella, Matera, Italy

R. Velasco
Research and Innovation Centre, Fondazione Edmund Mach, via Mach 1, 38010 San Michele All’Adige, Trento, Italy

P. Kalaitzis
Department of Horticultural Genetics and Biotechnology, Mediterranean Agronomic Institute of Chania (MAICH), Alsyllio Agrokepio, 1 Makedonias str., 73100 Chania, Crete, Greece

E. Rugini et al. (eds.), *The Olive Tree Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-319-48887-5_9

molecular determinants of key traits peculiar to the species of olive tree and giving important clues concerning the evolution of its complex genome.

1 Introduction

Unraveling the information through studies of omics can be equated to the discovery of the whole experience, which has been biologically accumulated during the evolutionary history of an organism, and it is the baseline which characterizes the set of adaptive events switched on in response to environmental factors and crop management choices that occur during the developmental stages of olive tree. The olive tree-cultivated form (*Olea europaea* L. subsp. *Europa*ea var. europaea) and the wild form (*Olea europaea* subsp. *Europaea* var. *sylvestris*) are mainly grown in the Mediterranean basin and Near Eastern, where the majority of a large number of olive cultivars estimated in more than 1200 (Bartolini et al. 1994) are also conserved. Olive is a diploid species ($2n = 2x = 46$) and the genome size ranges between 2.90 pg/2C and 3.07 pg/2C, with 1C = 1400–1500 Mbp (Loureiro et al. 2007).

Olive species is becoming one of the most economically important evergreen fruit crops in all Mediterranean climate types around the world. Despite its global importance and its metabolic peculiarities, available information on genomic and transcriotomic sequences for olive are still scarces, recently, an increasing number of expressed gene functions are being described. Besides, the Olive Genome Project (OLEA) (http://www.oleagenome.org) and the International Olive (*O. europaea*) Genome Consortium (IOGC) (http://olivegenome.karatekin.edu.tr) are expected to provide high-resolution information for functional studies and for discovery of new molecular markers. A striking feature coming from the analysis of studies conducted until now on olive genome indicates the presence of a greater number of repeated elements and among them the tandem repeat sequences (excluding rDNA) accounted for 31.16 % of the reads, the LTR-REs (*Gypsy* plus *Copia* elements) accounted for 38.84 % of the reads matching the whole genome set of assembled sequences (WGSAS), while low percentages of the presence were accounted for DNA transposons and non-LTR-REs. In this chapter, more accurate details are reported. From these studies, the peculiarity of genome evolution in this species has been evidenced with a very large fraction of the genome produced by tandem repeats amplification and LTR-RES. The role of this very large fraction of genome still remains unknown, but it can negatively affect the assembly of genome since olive is a highly heterozygous species.

The occurrence of a large and highly variable germplasm for this species, and for the related species, will allow to explore genetic variability concerning this genome fraction, possibly enabling to clarify the mechanisms by which such sequences have been produced and maintained during evolution and their function. The acquired knowledge will identify the relevant differences in the control of gene expression of the same sets of genes that exist among different genotypes. Following the genome assembly, the considerable task of annotating the genome remains. This includes predicting key features such as polymorphisms, GC content, repeated sequences, and genes. A suite of bioinformatics tools is available for predicting protein-coding genes and repeated sequences, based on sequence homology with other sequenced genomes and alignment of RNA sequences onto the assembly pseudo-chromosomes.

High-density genetic marker screen technology has been developed for olive, including single-nucleotide polymorphism (SNP) arrays (Kaya et al. 2013) and genotyping-by-sequencing (GBS) (İpek et al. 2016; Marchese et al. 2016). These technologies will be helpful for developing high-density genetic maps, fine mapping of major loci, genome-wide association studies (GWAS), genomic selection, and accelerate plant breeding (He et al. 2014). The data of structural and functional genomics, together with those from proteomics, metabolomics, mapping and genotyping, will be extremely useful for linking genotype to phenotype and pull out under-exploited natural diversity that is present in the *Olea* complex and in olive germplasm, enabling olive tree scientists to develop an understanding of the genetic regulatory mechanisms of key traits of high-quality production, synthesis of functional compounds, and those involved in plant–environment interactions and improved yield, and will provide fascinating opportunity in olive breeding programs, reducing the length and number of breeding cycles, labor, and cost (van Nocker and Gardiner 2014).

2 Genome Sequencing and Assembly

Two independent projects are focused on sequencing the olive genome. The Italian OLEA project is focused on the Leccino variety (Muleo et al. 2012), while the IOGC International Consortium has sequenced and assembled the genome of wild olive tree (*O. europaea*, var. *sylvestris*) with a coverage of 246X (Unver et al. 2016).

The *O. europaea* var. *sylvestris* genome was assembled with SOAPdenovo that produced a draft genome of 1.48 Gb, with the quality of genome assembled (N50) of 228 kb into twenty-three linkage groups that were anchored 50 % of the sequences, as resulted from the association with a newly constructed genetic map. Moreover, about 50 % of the total genome assembly was composed of repetitive DNA. The number of predicted gene models is 60,214, and 36,381 of them were anchored to chromosomes. Phylogenetic studies have highlighted that the genome underwent whole genome duplication event, before speciation from sesame. The olive genome with the last species shares a high degree of synteny for a large number of blocks.

The first draft of the olive genome sequence has been recently released by researchers of another independent project focused on sequencing the genome of almost 1200-year-old olive tree of Spanish cv. Farga (Cruz et al. 2016). The authors have assembled sequence data of 155,000 fosmid clones and 543 GB of raw DNA sequence from whole genome shotgun (WGS) that were generated by a combination of illumina sequencers run on short-insert paired-end (PE) libraries. Half of the 13,038 scaffolds (N50) were larger than 443.1 kb, and the final genome assembly of scaffolds indicated a total length of 1.31 Gb (95 % of 1.38 Gb estimated genome size), and the C-value with a median at 1.59 pg. These results confirm the existence of notable variation in the repetitive fraction of the genome for the species. The pipeline CEGMA estimated a genome completeness of about 98.79 %, and the heterozygous ratio identified by kmer individuals' analysis was 0.054. The number of gene-coding sequences with 56.339 predicted unique proteins generated from genome annotation was also supported by RNA sequencing from leaf, root, and fruit tissues at various stages. The higher number of proteins compared to closely related *Erynthranthe guttata* (24,373 predicted proteins) is consistent with the putative event of genome duplication in *O. europaea*. In this species, the chromosomal number is almost the double ($2n = 46$) than that found in *Sesamum indicum* ($2n = 26$) by Zhang and co-workers (2013), and *E. guttata* ($2n = 28$) by Fishman and co-workers (2014).

The olive genome of the cultivated variety Leccino is being sequenced, by the Olive Genome Project (OLEA) (http://www.oleagenome.org), using a combination of NGS Technologies and a combination of assembly approaches. The WGS approach to assemble the genome is being pursued using Illumina and 454-sequencing with a combination of long single reads, paired-end reads, and mate pairs until a coverage of at least 40 genome equivalents is reached. The assembly is being performed using Abyss and CLC assemblers. A bacterial artificial chromosome (BAC) pooling approach is being used to sequence random pools of 384 BACs using Illumina paired-end reads. A BAC coverage of approximately 3–4 genome equivalents is going to be sequenced, with each BAC pool sequenced at least at a 50X coverage. The advantages of the BAC approach are of two types: on the one hand, each BAC pool is much smaller in size than in the total genome size, reducing the assembly complexity. On the other hand, within each BAC pool we should not face the problem posed by sequence heterozygosity among maternally and paternally derived genomes that strongly affects WGS approaches and that is particularly challenging in the olive genome. The advantage of the WGS approach is the much more complete and homogeneous coverage of the entire genome. The two assemblies produced, the WGS and the pooled BAC assembly, will therefore be combined using a proprietary algorithm (GAM) to produce a consensus assembly. The consensus assembly will finally be anchored to the genetic map through the use of high-throughput genotyping technologies.

As of today, all the data needed for the WGS component have been produced. Gbp of Illumina sequence data was approximately produced, corresponding to a nominal coverage of 60X of the genome of cv Leccino. The Illumina sequences were obtained from two paired-end libraries with 500–600-bp inserts that were sequenced on the Illumina Genome Analyzer IIx producing 150-bp reads for a total coverage of 43X (65 Gbp) and from one paired-end library with 1000-bp inserts that was sequenced on the Illumina HiSeq 2000 system producing 100-bp reads for the remaining 17X coverage (25 Gbp). Finally, two mate-pair libraries with 3-kbp inserts were constructed and sequenced on the HiSeq 2000 to produce 100-bp reads and to reach a coverage of 4 genome equivalents (6 Gbp).

Eighteen Gbp of Roche-454 sequence data was approximately produced, corresponding to 12X coverage approximately. Twelve Gbp was obtained as long single reads of which approximately one-third was 400-bp-long reads (FLX TITANIUM technology) and two-thirds were 700-bp-long reads (FLX XL PLUS technology). Additionally, 6.2 Gbp of sequence data was obtained as paired-end reads from three libraries with 3-kbp inserts (3.8 Gbp) and 10 libraries with 8-kbp inserts (4.4 Gbp).

The 454 single reds and the Illumina paired-end reads are being used in a traditional WGS assembly. The Illumina mate-pair and the 454 paired-end sequences (i.e., all those sequences that have been obtained from inserts of larger size) will be utilized in order to scaffold into larger assemblies than the contigs obtained from the assembly of reads from the shorter inserts, with the aim to try to overcome the assembly problems posed by the occurrence of repetitive elements. Since many of the transposable elements in plant genomes are larger than 3 kbp, the larger inserts are going to be of crucial importance.

A number of assemblies were performed to test different strategies and to obtain a first rough draft of the olive genome. We tested assemblies both using the Illumina data only, as well as using Illumina and 454 data. All datasets have been initially filtered for low-quality sequences and for chloroplast DNA contamination and then were subject to assembly using the CLCBio assembler. When only the Illumina data were used (53X coverage after filtering), we produced an assembly of total size of 1.1 Gbp and N50 size of 1.7 kbp. The scaffolding using the mate-pair and paired-end information on the same assembly using the SSPACE tool increased the N50 size to 2.3 kbp. The addition of an initial set of 454 data (3.5 genome equivalents after filtering, single reads only) increased the total assembly size to 1.5 Gbp and the N50 size of contigs and scaffolds to 2.8 and 3.7 kbp, respectively. Finally, the addition of the remaining 454 sequences from the large insert libraries (3- and 8-kbp inserts) greatly improved the assembly, increasing considerably the N50 size of the scaffolds. The restriction to scaffolds of minimum 500 bp long the final assembly is 1.4 Gb long, and the N50 size of scaffolds is increased up to 10 kbp. However, due to the problems posed by the high levels of sequence heterozygosity present in the genome of cultivar Leccino, we consider the sequencing of the BAC pools a necessary component of our strategy in order to obtain a satisfactory assembly.

A large insert library (>100 kbp) of BAC clones was obtained from cultivar Leccino. 43,008 BAC clones were pooled into 112 plates of 384 BAC clones each. Eleven pools were initially sequenced with both Illumina Hiseq 2000 and Illumina Miseq, 100-bp and 250-bp paired-end, respectively, for a total of 60 Gbp. Reads were de novo assembled, and a total of 350 Mbp with N50 ranging in the 11 pools from 10 kbp to 21 kbp was produced, and some BACs were fully reconstructed (>100 kbp).

In order to evaluate the level of polymorphism in the *Olea* genome, we aligned the reads produced within the WGS approach on the assembled BAC pools. For each of the 11 pools, on single-copy regions, we detected SNPs. An high-frequency of SNP was found, detecting one SNP every 30 to 40 bp, proving a very high level of heterozygosity in this species. The degree of heterozygous in olive is comparable to that of the most heterozygous species, classifying it among complex genomes, such as Ciona savignyi (Small et al. 2007), Branchiostoma floridae (Putnam et al. 2008), and Strongylocentrotus purpuratus (Sea Urchin Genome Sequencing Consortium 2006). Further resequencing of different varieties is in progress and might reveal an even higher level of polymorphism within the *Olea* genome.

The International Consortium IOGC has used SOAPdenon method to assembly the genome, which generated a draft genome of 1.48 Gb. The dimension of the assembled genome resulted to be near to the estimated dimension of ~1.46 Gb. The researchers were able to anchor 50 % of sequences into 23 linkage groups, by using a constructed genetic map; the sequences have included 572 Mb. About 50 % of the total genome assembly was found to be composed of repetitive DNA. Transposable elements and interspersed repeats occupied 47 % of the genome. Phylogenetic and synteny analysis, and whole genome duplication analyses highlighted that the olive genome underwent duplication, before the event of speciation from sesame.

3 Analysis of the Repetitive Component and Olive Genome Composition

Some of the biggest technical challenges in sequencing eukaryotic genomes are caused by repetitive DNA (Faino and Thomma 2014): that is, sequences that are similar or identical to sequences elsewhere in the genome. An initial assembly of olive Illumina and 454 reads using RepeatExplorer (Novák et al. 2010) clearly showed five major clusters corresponding to five repeat families containing tandem repeats (Barghini et al. 2014). The repeat unit of four of these families (Oe80, Oe86, Oe178, and Oe218) were already identified as tandem repeats, isolated from genomic libraries, and, in some instances, localized by cytological hybridization on olive chromosomes (Katsiotis et al. 1998; Minelli et al. 2000; Lorite et al. 2001; Contento et al. 2002). The remaining family (Oe179) and a sixth minor family (Oe51) were also identified as tandem repeats. Besides clusters of tandem repeats, a number of minor clusters related mostly to *Gypsy* and *Copia* long-terminal-repeat (LTR) retrotransposons (REs) were identified (Fig. 1).

Then, a de novo assembly procedure was used to produce a large set of genomic sequences from Illumina and 454 reads (Barghini et al. 2014). The resulting whole genome set of assembled sequences (WGSAS) was composed of 210,068 sequences. Because of the relatively low genome coverage of the sequencing, most of the contigs that were obtained by both methods do not represent specific genomic loci; instead, they are probably composed of reads derived from multiple copies of repetitive elements, thus representing consensus sequences of genomic repeats (Novák et al. 2010). Although the exact form of this consensus does not necessarily occur in the genome, this representation of repetitive elements has been shown to be sufficiently accurate to enable amplification of the whole-length repetitive elements using PCR (Swaminathan et al. 2007). Moreover, the comparison with an available sequence library obtained by Sanger sequencing indicated a good correspondence

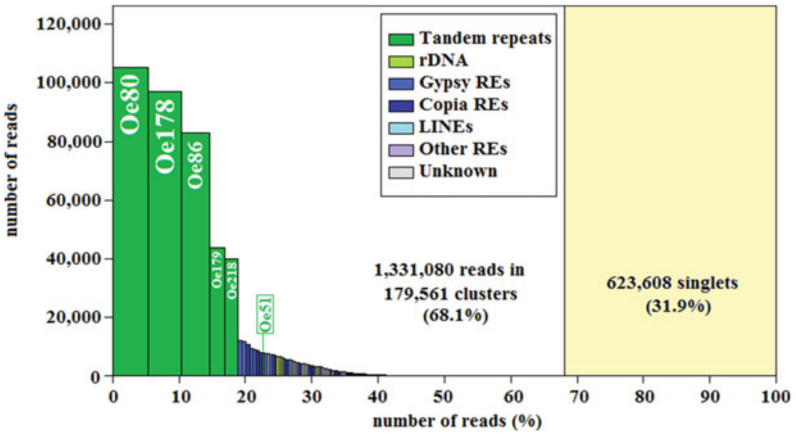

Fig. 1 Repeat abundance based on one genome equivalent of Illumina reads clustered using RepeatExplorer (Novák et al. 2010). Each bar in the histograms shows the individual size (height) of each cluster and the size relative to the total (width). The composition of each cluster is indicated by color, and single-copy, unclustered sequences are reflected to the right of the vertical bar. For the most redundant clusters, the annotation is reported within the bar (Barghini et al. 2014)

between virtual and real sequences (Barghini et al. 2014).

Assuming that Illumina sequence reads were sampled without bias for particular sequence types, mapping Illumina reads onto the WGSAS provided a method for estimating the redundancy of any genomic sequence in the dataset (Swaminathan et al. 2007; Tenaillon et al. 2011; Natali et al. 2013). All contigs with estimated redundancy higher than 100X (83,324 sequences) were selected and annotated to produce a collection of olive repeated sequences, hereafter called OLEAREP (Barghini et al. 2014).

The frequency distribution of different sequence types in OLEAREP is reported in Fig. 2, in which the dataset was further subdivided into two fractions, according to their average coverage, highly repeated (HR, average coverage >16,200), and medium repeated (MR, average coverage ranging between 16.2 and 16,200). Concerning the HR fraction, tandem repeats were the largest component, accounting around 2/3 of these contigs (Fig. 2). LTR-REs were also represented in the HR fraction, with *Gypsy* REs being more abundant in this fraction than *Copia* ones. Other classes of repeats (DNA transposons, rDNA, and putative genes) accounted only for minimal portions of HR set.

By converse, the MR fraction was mainly composed of LTR-REs (66.1 %), with *Gypsy* and *Copia* REs showing similar percentages (Fig. 2). Non-LTR-REs were poorly represented, as frequently observed in plant genomes. Putative DNA transposons accounted for 9.65 % of the MR fraction. All types of plant DNA transposons were found. Putative hAT and Mutator elements were by far the most redundant in this class, followed by putative Helitrons and CACTA elements. Tandem repeats were much less represented in this genome fraction than in HR.

Olive genome composition was estimated by counting the number of reads that mapped to each sequence. The percentage of HR sequences in the *Olea* genome was very high, amounting to 38.62 % at least. MR sequences accounted at least for 34.16 % of the genome, and low- or single-copy sequences represented only 16.92 % of the olive genome.

Olive genome composition was estimated also in terms of repeat types. The frequencies of each repeat type are reported in Table 1. Tandem repeat sequences (excluding rDNA) accounted for 31.16 % of the reads matching the WGSAS. LTR-REs amounted to 38.84 %, with *Gypsy* elements prevailing over *Copia* ones. DNA

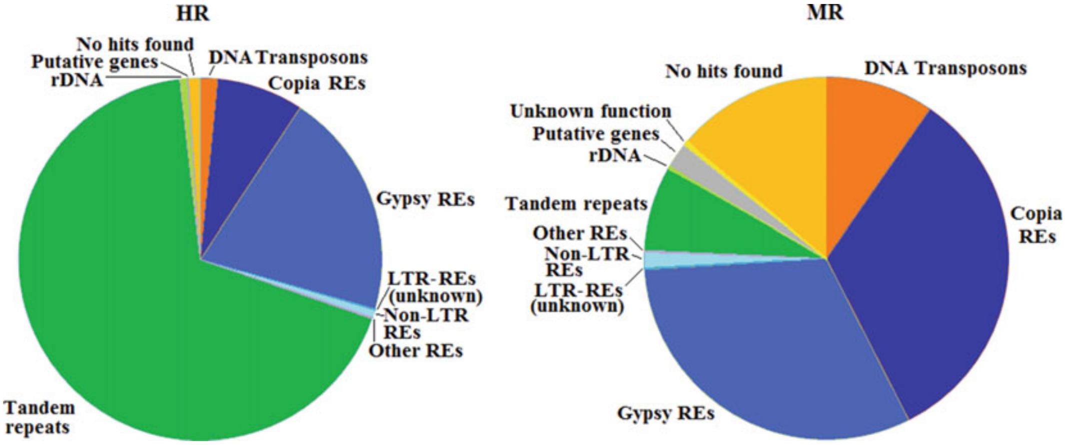

Fig. 2 Sequence composition of the OLEAREP database (HR and MR sequences) (Barghini et al. 2014)

transposons and non-LTR-REs showed low percentages.

On the whole, the OLEAREP database gives a precise characterization of the repetitive component of the *O. europaea* genome. It includes all already known olive repetitive sequences but also new, unknown sequences with high redundancy, which might represent new repeats to be still identified and characterized.

3.1 Analysis of Tandem Repeats

The large fraction of genome formed by tandem repeats is a peculiar feature of the olive genome. In many studies on genome assembly, tandem repeats are preliminarily removed, representing a negligible fraction of the genome and facilitating the assembly procedure (see e.g., for the sunflower genome, Staton et al. 2012). Until today, the largest fraction of tandem repeats found in a plant genome was estimated at around 23 % in the genome of cucumber (Huang et al. 2009).

Olive tandem repeats belong to six major families, defined according to their sequence and length. The first three families (Oe80, Oe178, and Oe86) correspond to the OeTaq80, OeTaq178, and OeGEM86 families described by Bitonti et al. (1999) and by Minelli and co-workers (2000) and account for about 72 % of tandem repeats. The fourth family (Oe179) was for the first time identified in this survey: It represents 12.6 % of the tandem repeats and the most common repeat unit is 179 bp in length; within this family, a number of repeats were truncated, with a variable length. In some cases, truncated elements were also arranged in repeat arrays, suggesting that the truncation has occurred while Oe179 was still replicating, with the truncated units that have continued their amplification.

The fifth family is Oe218, already described by Katsiotis and co-workers (1998), and accounting for 12.3 % of tandem repeats. The sixth major family was observed for the first time in this survey, representing only 2.2 % of the tandem repeats; the repeat unit is 51 bp.

Oe80, Oe178, and Oe218 constitute GC-rich, heavy satellites, having a GC content of 45.4, 43.2, and 41.8 %, respectively. By converse, Oe51 has a GC content of 33.5 %, constituting a light satellite. The GC contents of Oe86 and Oe179 (36.0 for each type) are similar to the mean GC content. All repeat families are present in multiple distinct contigs, indicating that distinct subtypes and higher-order structures of these sequences are present in the olive genome.

A distance tree, constructed using 100 sequences for each of the six repeat types, showed low sequence similarity among major tandem repeat families, suggesting an independent origin from each other (Fig. 3).

Table 1 Percentage distribution of repeat classes in the olive genome

Sequence type	Order	Superfamily	Number of contigs	Number of matched reads	Percentage
Retrotransposons	Unclassified		42	34,017	0.025
(Class I)	LTR	*Copia*	54,110	24,725,640	17.821
		Gypsy	47,920	28,884,342	20.819
		Retrovirus	101	74,960	0.054
		Endogenous retrovirus	4	6314	0.005
		Solo-LTR	52	18,355	0.013
		Unknown	189	174,016	0.125
	LINE	L1	2384	1,739,119	1.253
		RTE	453	123,845	0.089
		Unknown	38	20,591	0.015
	SINE	tRNA	268	64,093	0.046
	Total				40.265
DNA transposons (Class II)	Unclassified		67	32,668	0.024
	Subclass I	Tc1-Mariner	217	74,711	0.054
		hAT	7187	2,784,674	2.007
		Mutator	5790	3,335,678	2.404
		PiggyBac	1	34	0.000
		PIF-Harbinger	754	250,771	0.181
		CACTA	1212	496,957	0.358
		Crypton	7	2054	0.001
	Subclass II	Helitron	1297	672,682	0.485
	Total				5.514
Tandem repeats			11,260	43,233,770	31.161
rDNA			356	1,932,081	1.393
Unknown			308	179,225	0.129
No hits found			74,292	14,584,090	10.512
Total reads excluding organellar ones		138,741,954			

The measurement of the nucleotide diversity (the number of nucleotide substitutions per site), of each tandem repeat family, has shown that Oe218 is the most variable, followed by Oe178, and Oe80. Minor variations were observed within the families Oe179, Oe86, and Oe51. Actually, it is known that tandem repeats are characterized by large instability, depending on the repeat unit length, on the purity (i.e., similarity) of repeats, on the base composition, on external factors such as biotic and abiotic stresses (Gemayel et al. 2012). Moreover, the mutation rate in tandem repeats is estimated between 10^{-3} and 10^{-6} per cellular generation (Verstrepen et al. 2005). Such a high mutation rate should be related to the hypermethylation of these sequences (Hu et al. 2012).

It is hypothesized that the tandem repeats have a role in the genome. Beside their structural role in participating in centromeres and telomeres (Gemayel et al. 2012), tandem repeats can accumulate and generate intercalary heterochromatic regions. For example, in maize, tandem repeats form chromosomal knobs that reduce

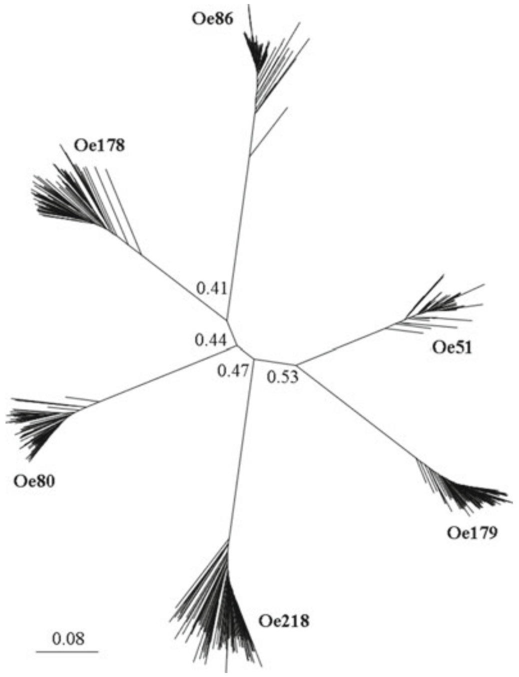

Fig. 3 Distance tree of olive tandem repeats (100 sequences per family); bootstrap values higher than 0.4 are shown. *Bar* represents the nucleotide distance (Barghini et al. 2014)

recombination rate in adjacent regions (Ghaffari et al. 2013).

In conclusion, our findings on olive genome evidenced the peculiarity of genome evolution in this species, with a very large fraction of the genome produced by tandem repeats amplification. The occurrence of a large and highly variable germplasm for this species will allow to explore genetic variability concerning this genome fraction, possibly enabling to clarify the mechanisms by which such sequences have been produced and maintained during evolution and their function.

3.2 Analysis of LTR Retrotransposons

Olive retrotransposon fragments were isolated and sequenced (Stergiou et al. 2002; Natali et al. 2007). However, the identification and accurate characterization of LTR-REs require the availability of sequences that span element length. In the frame of the project aimed to sequence the olive genome, a number of BAC clones were sequenced. These sequences were scored to identify full-length LTR-REs, searching for structural features and sequence similarities, i.e., the occurrence of two relatively intact LTRs, of identified polypurine tracts and primer-binding-sites, and of flanking tandem-site-duplications, and allowing the first characterization of intact elements in olive.

A set of 254 putative full-length REs was isolated (Barghini et al. 2015). The majority of isolated full-length REs belonged to the *Copia* superfamily (166), followed by the *Gypsy* superfamily (81, of which 36 contained an integrase chromodomain). Only seven REs remained unclassified.

In angiosperms, *Gypsy* and *Copia* superfamilies are differently represented in the genomes. Different ratios between *Gypsy* and *Copia* RE frequencies were reported ranging from 5:1 in papaya to 1:2 in grapevine (Vitte et al. 2014). Analysis of the whole olive genome showed a ratio of 1.17:1 (Barghini et al. 2014). The isolated olive full-length REs showed on the contrary a prevalence of *Copia* over *Gypsy* elements, indicating that the number of *Gypsy* families is lower than that of *Copia*, but *Gypsy* REs are more abundant than *Copia* REs (Barghini et al. 2015).

The relatively low frequency of REs in the olive genome could be related to a low rate of retrotransposition, but also to RE loss (Ma et al. 2004). RE DNA removal is driven in plants by a number of mechanisms, including DNA rearrangements and unequal homologous recombination; solo-LTRs are the main products of such processes (Ma and Bennetzen 2004).

Analyzing the relative redundancy of LTRs and inter-LTR regions in one and the same full-length RE was performed for evaluating the occurrence of solo-LTRs related (i.e, belonging to the same family) to that RE. Solo-LTRs related to the isolated full-length REs were rare: only 16 out of 254 REs showed a ratio between the number of mapped reads per Kb of LTR and inter-LTR >2.5. These ratios were especially high for two *Gypsy* and two *Copia* elements, indicating the occurrence of a large number of solo-LTRs for RE families that are related to these full-length elements (Barghini et al. 2015).

Concerning the amplification of REs, the identification of sister LTRs allowed us to date the insertion of REs in the olive genome, using the method established by San Miguel and co-workers (1998) in maize. Intact retroelements have a built-in molecular clock that is useful for estimating their insertion times, based on sister LTR divergence. In fact, when an RE inserts into the genome, its LTRs are usually 100 % identical (Kumar and Bennetzen 1999). Mutations then occur within the two LTRs, and as more time passes since the insertion, the larger the genetic distance between LTRs becomes. Hence, the RE insertion time can be estimated using a nucleotide substitution rate suitable for such elements (Ma and Bennetzen 2004).

Using a substitution rate per year of 3.6×10^{-9}, calculated comparing orthologous genes of olive and ash trees, the putative insertion times were calculated for each full-length LTR-RE. The putative mean age of analyzed LTR-REs was 17.94 MY (Barghini et al. 2015). Analysis of sister LTR similarity indicates that, in olive, both *Gypsy* and *Copia* REs have been active in the same period. Nearly all the identified full-length elements appear to be mobilized in a time-span of 40 MY (Fig. 4).

The mean insertion date of olive *Copia* full-length REs is lower than that of *Gypsy*. The insertion date profiles indicate that, during the last 40 MY, *Copia* and *Gypsy* REs have both been active, but with different time-courses. For example, only one isolated *Gypsy* full-length RE inserted between 1 and 8 MY ago. Moreover, the percentage of *Gypsy* REs inserted between 10 and 25 MY ago; hence, presumably, their retrotransposition activity is by far larger than that of *Copia* elements.

In contrast to other species, such as maize (Brunner et al. 2005) and sunflower (Buti et al. 2011) in which the retrotransposon burst is very recent and probably still occurring, in the olive genome the insertion of new REs for both *Gypsy* and *Copia* REs appears to be decreasing in frequency in the last 8 MY. A similar time-course of the RE amplification wave was reported in the genome of a gymnosperm, the Norway spruce (Nystedt et al. 2013). The observation of a considerable number of elements that have inserted more than 10 MY ago represents a clear

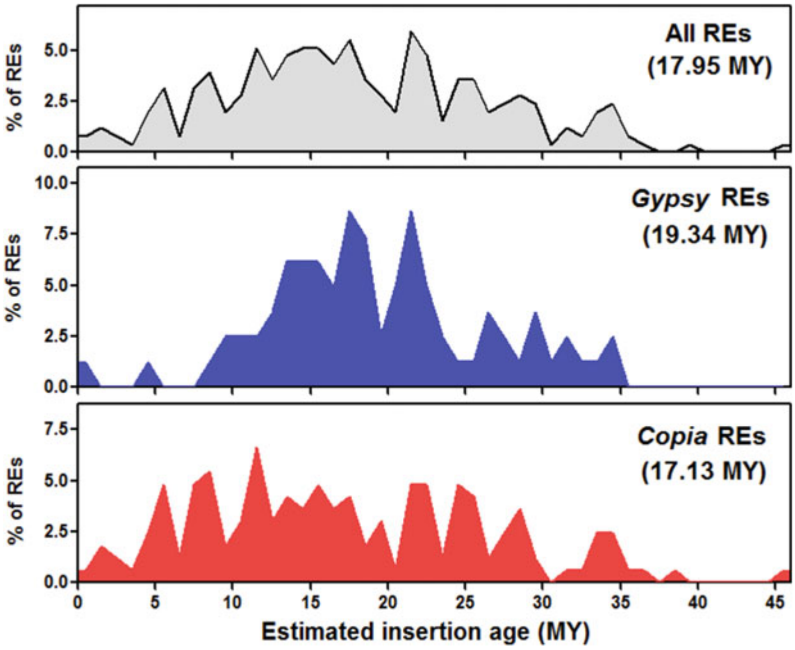

Fig. 4 Distributions of full-length REs of *Olea europaea*, according to their estimated insertion ages (MY). Mean insertion dates are reported in parentheses (Barghini et al. 2015)

distinctive feature of the olive genome in comparison with other angiosperm genomes analyzed so far.

4 Olive Chloroplast Genome

The chloroplast genome of the olive is already available from two independent groups (Mariotti et al. 2010; Besnard et al. 2011), with 155,889 bp in length. This genome has an organization and gene order conserved among numerous Angiosperm species and do not contain any of the inversions, gene duplications, insertions, inverted repeat expansions and gene/intron losses that have been found in the chloroplast genomes of the genera *Jasminum* and *Menodora*, from the same family as *Olea* (Mariotti et al. 2010). Forty polymorphisms have been identified in the plastome sequence, leading to a low number of chlorotypes distinguishing the olive cultivars.

5 Transcriptomics and Proteomics in Olive

Over the past several years, in olive, classical breeding programs have been focused on selecting for traits as short juvenile period, plant architecture suited for mechanical harvest, or oil characteristics, including fatty acid composition, phenolic, and volatile compounds to suit new markets. However, a better understanding of the genomic organization and the development of suitable molecular tools are mandatory steps to improve the efficiency of such breeding programmes. Nonetheless, transcriptomic data are already available for some of the olive genes involved in specific traits, such as fruit ripening, growth, and juvenile-phase transition. In the following paragraphs, we will focus on the gene expression studies performed so far by using high-throughput transcriptome sequencing technologies. Such transcriptomic approaches allowed to select a number of candidate genes that affect olive biology.

5.1 Flower and Fruit Development

In olive, floral biology has important practical implications, in addition to its scientific relevance, given that flower features and bloom affect fruits and yield. Furthermore, fruit development is the result of genetically programmed processes and environmental cues. High-throughput transcriptomics represent a key step for understanding the regulatory networks underlying plant reproduction and fruit growth and ripening.

Olive is a wind-pollinated, andromonoecious species whose cluster inflorescences are paniculate and whose flower position on the inflorescence may affect its development and fertility (Ben et al. 2013). Complex molecular and cellular processes are required for the development of reproductive tissues and the sculpting of the final form of the different organs (Irish 2010). A single tree may produce as many as 500,000 flowers, but only a small percentage of them (1–2 %) may set fruits due to several factors, such as wind pollination syndrome, flower development defects (i.e., ovary abortion), plant reproductive barriers (i.e, self-incompatibility and male sterility), and competition for maternal nutritional resources (Lavee et al. 1999; Rosati et al. 2011). The physiological changes that occur throughout flower development in olive have been investigated in the cultivar (cv.) Leccino at pre-anthesis and anthesis stages (Alagna et al. 2016). Analysis of the transcriptomic data generated by 454 sequencing (Table 2) revealed that among the flower transcripts, a large number of genes involved in the modification and degradation of proteins represented the substantial changes that occur during the development of flower verticils, including reproductive organs. Transcripts involved in cell-wall remodeling and polyamine biosynthesis were found upregulated at anthesis, while phenylpropanoid and flavonoid biosynthesis-related transcripts were downregulated during flower development. Finally, several genes involved in carbohydrates, lipid metabolism, transport, and cellular component organization were found more expressed at later stages of flower development (Alagna et al. 2016).

Some olive cv are characterized by a high number of male flowers, due to a high rate of ovary abortion and pistil desiccation (Reale et al. 2009; Rosati et al. 2011; Rapoport et al. 2012). The incidence of pistil abortion is influenced by nutritional or stress conditions (Bouranis et al. 1999; Fernández-Escobar et al. 2008), and it has been proposed that starch and sucrose metabolism might have a role in this process (Reale et al. 2009). In fact, flowers need carbohydrates to complete their differentiation, and changes in starch synthesis, degradation, and mobilization might affect the correct balance of nutrients in flower organs with consequences on the regular development of the pistils and ovary. Male sterility may also occur in olive (Cavallotti et al. 2003).

Transcripts potentially involved in the ovary abortion process were investigated by comparing cv. Leccino (low-ovary aborted flowers) and cv. Dolce Agogia (high-ovary aborted flowers) (Alagna et al. 2016). The analysis of the transcriptomic data (Table 2) identified several olive homologs of the genes involved in starch and sucrose metabolism, polyamine biosynthesis, cell-wall metabolism, programmed cell death (PCD), regulation of flavonoid biosynthesis, and MYB and MADS transcription factors (Alagna et al. 2016).

Self-incompatibility and interincompatibility represent the most important reproductive barriers in olive. In self-incompatible plants, the main recognition step is accomplished by the interaction between female and male determinants, which are usually encoded at a single polymorphic locus (*S*-locus) (Iwano and Takayama 2012). Subsequently, apposite cell signaling and cell–cell communication are fundamental for pollen acceptance and growth, requiring intricate intercellular communication between male and female cells. Pollen tube growth occurs directionally through the stigma and style to enter the ovary and is influenced by chemotropic agents, as well as a variety of lipids, ions, proteins, and metabolites that are produced by the pistil (Chapman and Goring 2010). A representative transcriptome of pollen and pistils has been just released (Carmona et al. 2015) (Table 2); cDNA libraries from pollen and pistil at different maturing and developing stages (with leaf and root as vegetative control) were 454 sequenced to provide a reproductive transcriptome and a user-friendly database (http://reprolive.eez.csic.es).

Candidate genes potentially involved in pollen–pistil interactions were identified by comparing the transcriptomes of cv. Frantoio (self-compatible) and cv. Leccino (self-incompatible) at anthesis (when the pollen grains reach the stigma upon self-pollination and potentially trigger the incompatibility reaction) (Alagna et al. 2016) (Table 2). The authors found more than 26 % of the upregulated genes in the self-incompatible cv. were involved in cell-wall degradation, whereas the upregulated genes in self-compatible cv. were mainly related to catabolic metabolism. Noteworthy, a significant number of altered transcripts belonged to hydrolase family (Alagna et al. 2016). It has been proposed that these enzymes may be involved in the remodeling of the pollen tube cell wall during its growth along the stylar transmitting tissues (Mollet et al. 2000), and in olive, they might also have a specific role in the self-incompatible interactions.

Fruit development and ripening takes place in about 4–5 months and includes the following phases: fruit set after fertilization, seed development, pit hardening, mesocarp development, and ripening. During the ripening process, fruit tissues undergo physiological and biochemical changes that include cell division and expansion, oil accumulation, metabolite storage, softening, phenol degradation, color change (due to anthocyanin accumulation in outer mesocarp cells). Several candidate genes putatively involved in olive fruit development were identified by comparative large-scale transcriptome analysis performed on fruits of cv. Coratina (high phenolic content) and cv. Tendellone (oleuropein-lacking

Table 2 An overview of the main set of transcriptome data generated from different cultivars, plant organs, and adaptive responses to stresses

Biological process	Publication	Cultivar	Tissue	Sequencing technology	Reads
Flower and fruit development	Alagna et al. (2016)	Leccino; Dolce Agogia; Frantoio	Flower (pre-anthesis, anthesis)	Pyrosequencing	465,000
	Carmona et al. (2015)		Pollen (mature, 1–5 h germination); pistil (stages 2–3–4); leaf (mature); root (mature, radicle)	Sanger	1549
				Pyrosequencing	2,077,309
	Alagna et al. 2009	Coratina; Tendellone	Fruit (45–135 days after flowering)	Pyrosequencing	261,485
	Iaria et al. (2016)	Leucocarpa; Cassanese	Fruit (100–130 days after flowering)	Illumina	103,359
	Muñoz-Merida et al. (2013)	Lechin de Sevilla; Picual; Arbequina; Picual*x*Arbequina	Fruit mesocarp (green, turning, purple); shoot (juvenile, adult, dormant); root (juvenile, adult); seed (green fruits); leaf (young)	Sanger	38,183
				Pyrosequencing	1,742,850
Fruit abscission	Gil-Amado and Gomez-Jimenez (2012)	Picual	Fruit AZ (54–217 days post-anthesis)	Pyrosequencing	443,811
	Parra et al. (2013)	Picual	Fruit pericarp and AZ tissues (217 days post-anthesis)	Pyrosequencing	397,457
Abiotic stress responses	Bazakos et al. (2015)	Kalamon	Leaf; root	Pyrosequencing	291,958
	Leyva-Perez et al. (2015)	Picual	Leaf	Illumina	149,638,888
	Guerra et al. (2015)	Leccino			75,645,229
miRNA	Donaire et al. (2011)	Picual; Arbequina	Shoot (juvenile, adult)	Pyrosequencing	169,699
	Yanik et al. (2013)	Ayvalik	Fruit (ripe, unripe); leaf (“on-year” and “off-year”)	Illumina	92,823,293

natural variant), at two developmental stages: 45 and 135 days after flowering (DAFs) (Alagna et al. 2009) (Table 2). About 25 % of the annotated enzyme-coding transcripts were involved in biosynthesis of lipids and fruit metabolites. Transcript fluctuations were consistent with the physiological status of the fruit. The higher expression of transcripts related to the biosynthesis of structural proteins at 45 DAF may be correlated with the intense and rapid cell divisions during fruit growth, while the higher expression of transcripts putatively associated with fatty acid biosynthesis and with the assembly of storage triacylglycerols at 135 DAF is in agreement with fatty acid accumulation pattern in olive fruits, starting at about 90 DAF until the end of fruit maturation (Conde et al. 2008). Among genotype-specific transcripts, several ones putatively involved in the biosynthesis of steroids with nutritional and health benefits were reported exclusively in cv. Coratina (Alagna et al. 2009).

To investigate whether these changes at the mRNA level correspond to variations at the protein level, a comparison between transcript and protein profiles was performed in Bianco and co-authors (Bianco et al. 2013). So far, this paper is the only one in the literature that monitors the proteome variations associated with olive fruit development by using comparative proteomics based on 2-DE coupled to MALDI-TOF mass spectrometry, providing new and important insights into fruit metabolism and oil accumulation process (Bianco et al. 2013).

Interestingly, comparison between transcriptomic and proteomic datasets revealed that most of the proteins and their putative transcripts associated with fatty acids biosynthesis and metabolism (enoyl ACP reductase and lipoxygenase), as well as transcripts and proteins linked to cell cycle, biosynthesis of structural proteins involved in cell expansion showed a similar increased pattern during drupe development (Bianco et al. 2013). Transcript and protein profile comparison also revealed some divergent patterns, indicative of possible post- transcriptional events in RuBisCO large subunit-binding protein subunit alpha and proteins associated with detoxification and oxidation-reduction processes (Bianco et al. 2013).

Transcripts involved in flavonoid and anthocyanin metabolism during drupe development were identified by comparing different Illumina RNA-seq libraries generated from drupes of cv. Leucocarpa (characterized by a switch-off in skin color at full ripeness) and cv. Cassanese (as control), sampled at 100 and 130 DAF (Iaria et al. 2016) (Table 2). The cv. Leucocarpa was characterized by a broad downregulation of chalcone synthase, dihydroflavonol 4-reductase, and anthocyanidin synthase transcripts compared to cv. Cassanese. Moreover, several members of MYB, MYC, and WD transcription factors related to the regulatory complexes that control anthocyanin structural genes at the transcriptional level were identified as differentially expressed (Iaria et al. 2016).

Oil synthesis starts after pit hardening, reaching a plateau after 75–90 days, while the phenolic fraction is maximum at fruit set and decreases rapidly along fruit development. To get information about genes involved in determining oil content and composition, mesocarp and seed transcriptomes from fruits of different cv (Picual and Arbequina with different characteristics regarding fruit and oil organoleptic properties) were investigated (Muñoz-Merida et al. 2013) (Table 2). To date, this paper is the largest contribution to transcript information in *Olea* (about 2 M reads) (Table 2). The assembly has rendered over 81,020 unigenes that have been functionally annotated. Interestingly, numerous transcripts are involved in lipid metabolic/biosynthetic process or lipid fatty acid metabolic/biosynthetic process and then associated with oil characteristics and production (Muñoz-Merida et al. 2013).

5.2 Fruit Abscission

Abscission and senescence are key physiological events that occur during the growth and development of fruits in higher plants. These bear commercial implications both for the plant yield and the harvest. In agricultural research, the manipulation of genes governing these phenomena is crucial to develop varieties that can produce fruits with longer shelf-life as well as crops that tolerate greater environmental stress. After fruit ripening, many fruit tree species undergo massive natural fruit abscission. Abscission occurs in an anatomically distinct layer of cells known as the abscission zone (AZ) (Gonzalez-Carranza and Roberts 2012) located between the pedicel and fruit, and the patterns of mature fruit abscission differ between cultivars (Gomez-Jimenez et al. 2010). Olive fruit has several AZs in the pedicel, but only one AZ at a time is selectively activated per specific developmental stage (Parra-Lobato and Gomez-Jimenez 2011). Probably, the induction of abscission depends on a complex interplay of plant hormone concentrations in addition to factors that alter the responsiveness and sensitivity of the tissues (Gonzalez-Carranza and Roberts 2012). To identify differences in transcript abundance related to the mature fruit abscission in olive, 454 pyrosequencing technology was used in cv. Picual comparing AZ transcripts at two different stages: pre-abscission vs. abscission (Gil-Amado and Gomez-Jimenez 2013) (Table 2). The authors identified 70 transcription factor genes induced during mature fruit abscission in AZ. Among them, the classes that are well represented included bZIP proteins, MYB proteins, and homeobox domain proteins (Gil-Amado and Gomez-Jimenez 2013). To significantly expand the olive transcript catalog, 454 pyrosequencing technology was also used to sequence two cDNA samples from fruit pericarp and AZ, which were collected from ripe fruits, when abscission occurs (Parra et al. 2013) (Table 2). Functional categorization of the differentially expressed genes showed that AZ tissues have an apparently higher response to external stimuli than that of ripe fruit, revealing a higher expression of genes involved in auxin-signaling, lignin, aromatic amino acid, isoprenoid, amino acid dephosphorylation-transport, and photosynthesis pathways (Parra et al. 2013). By contrast, fruit-enriched transcripts are involved in ATP synthesis coupled with proton transport, glycolysis, and cell-wall organization. Regarding the cross-talk between fruit and AZ, several transcription factors were identified, especially MADS-box, ZF, homeobox domain proteins, bHLH, and bZIP families (Parra et al. 2013). This represents the first effort to elucidate the molecular bases related to the mature fruit abscission in olive, as a model to study fleshy fruit abscission. In fact, most studies identifying transcriptional regulators during organ abscission have used *Arabidopsis* (Nath et al. 2007), while regarding fruit abscission transcriptomic data are available only in apple (Botton et al. 2011).

5.3 Abiotic Stress Responses

Abiotic stresses such as salinity, drought, and cold cause a plethora of responses at the morphological, physiological, biochemical, and molecular levels which reduce yield and plant productivity. All three abiotic stresses cause a primary loss of cell water and as a result a decrease of cell osmotic potential (Duque et al. 2013). Plants have evolved highly complex mechanisms to respond and tolerate such stresses which are partly coordinated by intricate gene regulatory networks.

Although olive is a tree species well adapted to xerothermic conditions and, therefore, to environments of high temperature and long drought; the rapid expansion of olive cultivation increases the need for use of low-quality saline water for irrigation. Such water causes salt stress which negatively affects shoots growth and fruit productivity. In olive, there are salt-tolerant and salt-sensitive genotypes which differ in their ability to exclude toxic ions and to control the net salt import to the shoot.

The molecular basis of this tolerance was investigated by comparative transcriptome analysis of two olive cultivars using microarrays (Bazakos et al. 2012). Despite the limited number of probe sets, transcriptional regulatory networks were constructed for both, cv. Kalamon and cv. Chondrolia Chalkidikis, while several hierarchically clustered interacting transcription factor regulators such as JERF and bZIP were identified (Bazakos et al. 2012). The higher complexity of the cv. Kalamon transcription factor network compared to the cv. Chondrolia Chalkidikis network might be indicative of a more coordinated effort to adapt to salinity. Moreover, the comparison of the interactions among transcription factors in olive with those reported for *Arabidopsis* indicates similarities in the response of a tree species with *Arabidopsis* at the transcriptional level under salinity stress (Bazakos et al. 2012).

A 454 pyrosequencing approach was also employed to characterize the transcriptome of leaves and roots of cv Kalamon in response to salinity (Bazakos et al. 2015) (Table 2). In roots, 24 differentially expressed clusters were identified comprising 9 down- and 15 upregulated genes, while 14 down- and 56 upregulated clusters of differentially expressed genes were identified in leaves (Bazakos et al. 2015). In addition, 433 unique transcripts encoding transcription factors were determined while the most abundant among them appeared to be senescence-associated as well as NAC domain family transcription factors which are known to be involved in salt stress responses (Bazakos et al. 2015). Transcripts implicated in salt tolerance, such as glutathione reductase, superoxide dismutase, and proline dehydrogenase, were also identified in the leaf transcriptome exposed to salinity (Bazakos et al. 2015).

In another report, transcriptome analysis of olive leaves of cv Picual during cold acclimation conditions resulted in the identification of 6309 differentially expressed transcripts (Leyva-Perez et al. 2015) (Table 2). Among them, the early response genes comprised of C-repeat binding factor transcription factors, fatty acid desaturases, wax synthesis and oligosaccharide metabolism (Leyva-Perez et al. 2015).

A RNA-Seq analysis was performed to identify in olive genes associated to cold stress response, studying short- and long-term transcriptional changes occurring in leaves of cv. Leccino exposed to a progressive lowering of temperatures until −4 °C (Guerra et al. 2015). The Illumina (Illumina Genome Analyzer IIx) sequencing approach generated 93, 927,355 pair-end reads for a total of 27.24 Gb, reduced to 75, 645,229 pair-end high-quality reads and 20.33 Gb after the trimming and filtering process. A total number of 85,752 contigs resulted and 44,332 leaf transcripts has been de novo assembled. Among them, 5464 differentially expressed genes (DEGs) were identified. Most of the typical components of the known and conserved molecular repertoire of the plant cold response have been found into the set of transcriptomic data, as transcriptions factors of cold signaling, induction of coldregulated genes (*cor*), genes involved in changes of membrane composition, and downregulation of photosynthesis-related genes (Guerra et al. 2015). Specific cold response genes of olive tree leaves, induced during cold acclimation, were identified, including genes of the glutathione cycle, polyamine and flavonoid pathways, likely to support reactive oxygen species (ROS) scavenging, as well as genes of the raffinose and trehalose carbohydrate biosynthetic pathways to sustain the accumulation of osmolytes. Moreover, genes involved in the signaling pathway of abscisic acid (ABA), synthesis of callose and lignins, indicated changes in composition of cell wall, were also strongly present (Guerra et al. 2015). The RNA-Seq data about CBF-like transcript has been confirmed trough expression profile studied by RT-PCR trials conducted in Leccino and in seven other cultivars differing for cold tolerance (Guerra et al. 2015).

The high-throughput transcriptome analyses of olive trees under abiotic stress resulted in the identification of a large number of genes involved in adaptation and tolerance, but future functional characterization will determine their physiological significance in these conditions.

5.4 Small RNAs

The microRNAs (miRNAs) are noncoding small RNA found in diverse eukaryotes, negatively regulating specific target messenger RNA (Reinhart et al. 2002). The plant miRNAs range in size from 20 to 24 bases (Dugas and Bartel 2004). They act as key regulators controlling the gene expression in a multitude of developmental and physiological processes (Pulido and Laufs 2010; Sunkar et al. 2012). So far, their involvement in developmental regulation and flowering processes has been extensively studied in a wide variety of herbaceous plant species (http://www.mirbase.org/) while the list of miRNAs from woody plants is scarce and restricted to conifers, poplar, grapevine, and citrus (Lu et al. 2008;

Morin et al. 2008; Song et al. 2009; Pantaleo et al. 2010). Recently, the first inventory of miRNAs in olive was reported (Donaire et al. 2011) (Table 2). Two distinct miRNA cDNA libraries were prepared from juvenile and adult shoots from the progeny of a genetic cross between the cv. Picual and cv. Arbequina and sequenced by deep pyrosequencing (Table 2). The vast majority of sequences (80 %) were singletons suggesting that the miRNA libraries were far from saturated and that, consequently, olive contained a large and diverse miRNA population. A hallmark signature of the olive miRNA population is the vast presence of the 24-nt species at a higher level with respect to many other plant species. Donaire and colleagues suggest an active role of heterochromatin silencing in the maintenance and integrity of the olive large genome (Donaire et al. 2011). Currently, miRNAs from about 24 broadly conserved families have been identified from eudicots to basal plants and deposited in the public miRNA database miRBase (Griffiths-Jones et al. 2008). In the olive miRNA dataset were identified 18 out of the 24 known miRNA families (Donaire et al. 2011).

Regulation of miRNA has a significant impact on the olive tree alternate bearing (Yanik et al. 2013). Alternate bearing is a common phenomenon among crop plants, defined as the tendency of certain fruit trees to produce a high-yield crop one year ("on-year"), followed by a low yield or even no crop the following year ("off-year"). Thus, this phenomenon may severely affect the olive fruit yield. Several miRNAs related to the alternate bearing were identified in a study performed in Yanik et al. 2013. In this work, six miRNA libraries were constructed from fruits (ripe and unripe) and leaves ("on-year" and "off-year" in July and in November, respectively) (Table 2). About 15,587,819 reads from each library were generated with the high-throughput Illumina sequencing system (Yanik et al. 2013). Predicted targets of miRNA were categorized into 108 process ontology groups with significant abundance. Among those, several alternate bearing-associated processes were found, such as development, hormone-mediated signaling, and organ morphogenesis.

Deeper sequencing or even alternative sequencing platforms could give better resolution in the olive small RNA population, therefore unraveling more miRNA.

6 Conclusions

With the availability of highly assembled genome sequence, the research activities will be focused on the challenge to translate the decoded genome into new tools that can be implemented by olive biologists and tree breeders for variety improvement. High-quality genome assembly greatly facilitates this task by enabling the complete inventory of DNA variation in olive species, including copy number variations, single-nucleotide mutations (insertions and deletions), epigenetic variations, such as DNA methylation, smaller RNA, and mobile elements. The next-generation sequencing (NGS) technology (Metzker 2010), which enables the rapid generation of a massive amount of sequencing data with a limited low cost, gives the opportunity to sequence whole genome and DNA variant identification of cultivated varieties and wild species. The capacity of sequencing large DNA fragments of several kbps, of increasing the proportion of the assembly anchored to genetic maps, assembling larger haplotyped scaffolds by new molecular and bioinformatics procedures will improve the current genome assembly in the near future.

The transcriptomic approaches discussed clearly demonstrate that the catalog of olive transcripts has been significantly expanded in recent years (Table 2), providing several answers to the various biological questions affecting the olive biology. Large datasets of transcriptomic sequences including miRNAs, mainly generated by pyrosequencing, have recently been reported for several tissues (Fig. 5a, b). However, genomic information in olive is well behind other species of woody plants, such as grape (Velasco et al. 2007) or poplar (Tuskan et al. 2006) whose complete genome sequences are already available. The lack of a complete and annotated

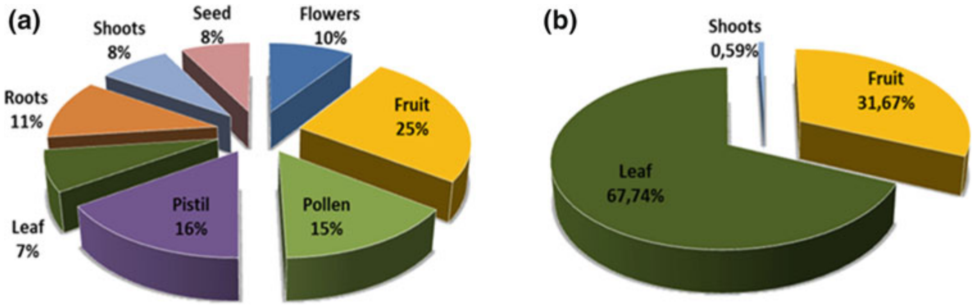

Fig. 5 **a** Pie-sectioned representation of transcriptomic reads generated for olive tissues (Illumina reads of leaf are not included). **b** Pie-sectioned representation of miRNA reads generated for olive tissues

genome makes impossible a deep and detailed interpretation of data already available in olive.

Recently, the advancements on sequencing technology have been impressive. The newer generation of sequencing methods based on single-molecule sequencing and in situ sequencing (to read nucleic acid composition directly in fixed cells and tissues), allows to obtain a great number of reads of several Mb in length, no GC bias, and high read accuracy at lower costs (Buermans and den Dunnen 2014). By applying these emerging sequencing technologies, the amount of genomic information will become accessible in a shorter time, allowing the easier sequencing and resequencing of the olive genome. All this will make easier to identify new gene functions and new molecular markers involved in the expression of fundamental agronomic and productive traits affecting olive biology, opening the possibility of developing molecular tools to the level currently available for other model plant species.

Acknowledgments This research was partially supported by Project MIPAF "OLEA—*Genomica e Miglioramento genetico dell'olivo*," D.M. 27011/7643/10. We thank the Roche Diagnostic Spa, Applied Science to support the OLEA Italian Project.

References

Alagna F, D'Agostino N, Torchia L, Servili M, Rao R et al (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genom 10:399

Alagna F, Cirilli M, Galla G, Carbone F, Daddiego L et al (2016) Transcript analysis and regulative events during flower development in olive (*Olea europaea* L.). PLoS ONE 11(4):e0152943

Barghini E, Natali L, Cossu RM, Giordani T, Pindo M et al (2014) The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. Genome Biol Evol 6(4):776–791

Barghini E, Natali L, Giordani T, Cossu RM, Scalabrin S et al (2015) LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. DNA Res 22:91–100

Bartolini G, Prevost G, Messeri C (1994) Olive tree germplasm: descriptor lists of cultivated varieties in the world. Acta Hortic 365:116–118

Bazakos C, Manioudaki ME, Therios I, Voyiatzis D, Kafetzopoulos D et al (2012) Comparative transcriptome analysis of two olive cultivars in response to NaCl-stress. PLoS ONE 7:e42931

Bazakos C, Manioudaki ME, Sarropoulou E, Spano T, Kalaitzis P (2015) 454 pyrosequencing of olive (*Olea europaea* L.) transcriptome in response to salinity. PLoS ONE 10:e0143000

Besnard G, Hernández P, Khadari B, Dorado G, Savolainen V (2011) Genomic profiling of plastid DNA variation in the Mediterranean olive tree. BMC Plant Biology 11:80

Ben Sadok I, Celton JM, Essalouh L, El Aabidine AZ, Garcia G et al (2013) QTL mapping of flowering and fruiting traits in olive. PLoS ONE 8:e62831

Bianco L, Alagna F, Baldoni L, Finnie C, Svensson B et al (2013) Proteome regulation during *Olea europaea* fruit development. PLoS ONE 8:e53563

Bitonti MB, Cozza R, Chiappetta A, Contento A, Minelli S et al (1999) Amount and organization of the heterochromatin in *Olea europaea* and related species. Heredity 83:188–195

Botton A, Eccher G, Forcato C, Ferrarini A, Begheldo M et al (2011) Signaling pathways mediating the induction of apple fruitlet abscission. Plant Physiol 155:185–208

Bouranis DL, Kitsaki CK, Chorianopoulou SN, Aivalakis G, Drossopoulos JB (1999) Nutritional dynamics of olive tree flowers. J Plant Nutr 22:245–257

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17:343–360

Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. Biochim Biophys Acta 1842:1932–1941

Buti M, Giordani T, Cattonaro F, Cossu RM, Pistelli L et al (2011) Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. Theor Appl Genet 123:779–791

Carmona RM, Zafra A, Seoane P, Castro AJ, Guerrero-Fernández D et al (2015) ReprOlive: a database with linked data for the olive tree (*Olea europaea* L.) reproductive transcriptome. Front Plant Sci 6:625

Cavallotti A, Regina TMR, Quagliariello C (2003) New sources 898 of cytoplasmic diversity in the Italian population of *Olea europaea* L. as revealed by RFLP analysis of mitochondrial DNA: characterization of the cox3 locus and possible relationship with cytoplasmic male sterility. Plant Sci 164:241–252

Chapman LA, Goring DR (2010) Pollen-pistil interactions regulating successful fertilization in the Brassicaceae. J Exp Bot 61:1987–1999

Conde C, Delrot S, Geros H (2008) Physiological, biochemical and molecular changes occurring during olive development and ripening. Plant Physiol 165:1545–1562

Contento A, Ceccarelli M, Gelati MT, Maggini F, Baldoni L et al (2002) Diversity of *Olea* genotypes and the origin of cultivated olives. Theor Appl Genet 104:1229–1238

Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M et al (2016) Genome sequence of the olive tree, Olea europaea. GigaScience 5:29

Donaire L, Pedrola L, de la Rosa R, Llave C (2011) High-throughput sequencing of RNA silencing-associated small RNAs in olive (*Olea europaea L.*). PLOS One 6(11):e27916

Dugas DV, Bartel B (2004) MicroRNA regulation of gene expression in plants. Curr Opin Plant Biol 7:512–520

Duque AS, De Almeida AM, Da Silva AB, Da Silva JM, Farinha AP et al (2013) Abiotic stress responses in plants: unraveling the complexity of genes and networks to survive. In: Vahdati K, Leslie C (eds) Abiotic stress—plant responses and application in agriculture. InTech Publisher, Rijeka, Croatia, pp 49–101

Faino L, Thomma BPHJ (2014) Get your high-quality low-cost genome sequence. Trends Plant Sci 19:288–291

Fernández-Escobar R, Ortiz-Urquiza A, Prado M, Rapoport HF (2008) Nitrogen status influence on olive tree flower quality and ovule longevity. Environ Exp Bot 64:113–119

Fishman L, Willis JH, Wu CA, Lee Y-W (2014) Comparative linkage maps suggest that fission, not polyploidy, underlies near-doubling of chromosome number within monkey flowers (*Mimulus*; Phrymaceae). Heredity 112:562–568

Gemayel R, Cho J, Boeynaems S, Verstrepen KJ (2012) Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes 3:461–480

Ghaffari R, Cannon EK, Kanizay LB, Lawrence CJ, Dawe RK (2013) Maize chromosomal knobs are located in gene-dense areas and suppress local recombination. Chromosoma 122:67–75

Gil-Amado JA, Gomez-Jimenez MC (2012) Regulation of polyamine metabolism and biosynthetic gene expression during olive MFA. Planta 235:1221–1237

Gil-Amado JA, Gomez-Jimenez MC (2013) Transcriptome analysis of mature fruit abscission control in olive. Plant Cell Physiol 54:244–269

Gomez-Jimenez MC, Paredes MA, Gallardo M, Sanchez-Calle IM (2010) Mature fruit abscission is associated with up-regulation of polyamine metabolism in the olive abscission zone. J Plant Physiol 167 (1):432–441

González-Carranza ZH, Shahid AA, Zhang L, Liu Y, Ninsuwan U et al (2012) A novel approach to dissect the abscission process in Arabidopsis. Plant Physiol 160(3):1342–1356

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158

Guerra D, Lamontanara A, Bagnaresi P, Orrù L, Rizza F et al (2015) Transcriptome changes associated with cold acclimation in leaves of olive tree (*Olea europaea* L.). Tree Genet Genomes 11:113

He J, Zhao X, Laroche A, Lu Z-X, Liu H-K et al (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted (MAS) tool to accelerate plant breeding. Front Plant Sci 5:484

Hu Y, Zhang L, He S, Huang M, Tan J et al (2012) Cold stress selectively unsilences tandem repeats in heterochromatin associated with accumulation of H3K9ac. Plant, Cell Environ 35:2130–2142

Huang S, Li R, Zhang Z, Li L, Gu X et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

Iaria DL, Chiappetta A, Muzzalupo I (2016) A De novo transcriptomic approach to identify flavonoids and anthocyanins “Switch-Off” in olive (*Olea europaea* L.) drupes at different stages of maturation. Front. Plant Sci 6:1246

İpek A, Yilmaz K, Sikici P, Tangu NA, Öz AT et al (2016) SNP discovery by GBS in olive and the construction of a high-density genetic linkage map. Biochem Genet 54(3):313–325

Irish VF (2010) The flowering of Arabidopsis flower development. Plant J 61:1014–1028

Iwano M, Takayama S (2012) Self/non-self-discrimination in angiosperm self-incompatibility. Curr Opin Plant Biol 15:78–83

Katsiotis A, Hagidimitriou M, Douka A, Hatzopoulos P (1998) Genomic organization, sequence interrelationship, and physical localization using in situ hybridization of two tandemly repeated DNA sequences in the genus *Olea*. Genome 41:527–534

Kaya HB, Cetin O, Kaya H, Sahin M, Sefer F et al (2013) SNP Discovery by Illumina-based transcriptome sequencing of the olive and the genetic characterization of Turkish Olive Genotypes revealed by AFLP, SSR and SNP markers. PLOS One 8(9):e73674

Kumar A, Bennetzen JB (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Lavee S, Rallo L, Rapoport HF, Troncoso A (1999) The floral biology of the olive—II. The effect of inflorescence load and distribution per shoot on fruit set and load. Sci Hortic 82:181–192

Leyva-Pérez Mde L, Valverde-Corredor A, Valderrama R, Jiménez-Ruiz J, Muñoz-Merida A et al (2015) Early and delayed long-term transcriptional changes and short-term transient responses during cold acclimation in olive leaves. DNA Res 22:1–11

Lorite P, Garcia MF, Carrillo JA, Palomeque T (2001) A new repetitive DNA sequence family in the olive (*Olea europaea* L.). Hereditas 134:73–78

Loureiro J, Rodriguez E, Costa A, Santos C (2007) Nuclear DNA content estimations in wild olive (*Olea europaea* L. ssp. *europaea* var. *sylvestris* Brot.) and Portuguese cultivars of *O. europaea* using flow cytometry. Genet Resour Crop Evol 54(1):21–25

Lu S, Sun YH, Chiang VL (2008) Stress-responsive microRNAs in *Populus*. Plant J 55:131–151

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. P Natl Acad Sci-Biol 101:12404–12410

Ma J, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14:860–869

Marchese A, Marra FP, Caruso T, Mhelembe K, Costa F et al (2016) The first high-density sequence characterized SNP-based linkage map of olive (*Olea europaea* L. subsp. *europaea*) developed using genotyping by sequencing. Aust. J Crop Sci 10(6):857–863

Mariotti R, Cultrera NGM, Muñoz Diez C, Baldini L, Rubini A (2010) Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison. BMC Plant Biology 10:211

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46

Minelli S, Maggini F, Gelati MT, Angiolillo A, Cionini PG (2000) The chromosome complement of *Olea europaea* L.: characterization by differential staining of the chromatin and in-situ hybridization of highly repeated DNA sequences. Chromosome Res 8:615–619

Mollet JC, Park SY, Nothnagel EA, Lord EM (2000) A lily stylar pectin is necessary for pollen tube adhesion to an in vitro stylar matrix. Plant Cell 12:1737–1750

Morin RD, Aksay G, Dolgosheina E, Ebhardt HA, Magrini V et al (2008) Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. Genome Res 18:571–584

Muleo R, Morgante M, Velasco R, Cavallini A, Perrotta G, Baldoni L (2012) Olive tree genomic. In: Mazzalupo I (ed) Olive germplasm. The olive cultivation, table and olive oil industry in Italy. InTech Publisher, Rijeka, Croatia, pp 133–148

Muñoz-Merida A, Gonzalez-Plaza JJ, Canada A, Blanco AM, Garcia-Lopez Mdel C et al (2013) De novo assembly and functional annotation of the olive (*Olea europaea*) transcriptome. DNA Res 20:93–108

Natali L, Giordani T, Buti M, Cavallini A (2007) Isolation of Ty1-*Copia* putative LTR sequences and their use as a tool to analyse genetic diversity in *Olea europaea*. Mol Breed 19:255–265

Natali L, Cossu RM, Barghini E, Giordani T, Buti M et al (2013) The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. BMC Genom 14:686

Nath P, Sane AP, Trivedi PK, Sane VA, Asif MH (2007) Role of transcription factors in regulating ripening, senescence and organ abscission in plants. Stewart Postharvest Rev 3:1–14

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584

Pantaleo V, Szittya G, Moxon S, Miozzi L, Moulton V et al (2010) Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. Plant J 62:960–976

Parra R, Paredes MA, Sanchez-Calle IM, Gomez-Jimenez MC (2013) Comparative transcriptional profiling analysis of olive ripe-fruit pericarp and abscission zone tissues shows expression differences and distinct patterns of transcriptional regulation. BMC Genom 14:866

Parra-Lobato MC, Gomez-Jimenez MC (2011) Polyamine-induced modulation of genes involved in ethylene biosynthesis and signalling pathways and nitric oxide production during olive mature fruit abscission. J Exp Bot 62:4447–4465

Pulido A, Laufs P (2010) Co-ordination of developmental processes by small RNAs during leaf development. J Exp Bot 61:1277–1291

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U et al (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198):1064–1071

Rapoport HF, Hammami SBM, Martins P, Pérez-Priego O, Orgaz F (2012) Influence of water deficits at different times during olive tree inflorescence and flower development. Environ Exp Bot 77:227–233

Reale L, Sgromo C, Ederli L, Pasqualini S, Orlandi F et al (2009) Morphological and cytological development and starch accumulation in hermaphrodite and staminate flowers of olive (*Olea europaea L.*). Sex Plant Reprod 22:109–119

Reinhart BJ1, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev 16:1616–1626

Rosati A, Caporali S, Paoletti A, Famiani F (2011) Pistil abortion is related to ovary mass in olive (Olea europaea L.). Sci Hortic-Amsterdam 127:515–519

San Miguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Sea Urchin Genome Sequencing Consortium, Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA et al (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314(5801):941–52 (Erratum in: Science. 2007 Feb 9;315(5813):766)

Small KS, Brudno M, Hill MM, Sidow A (2007) Extreme genomic variation in a natural population. P Natl Acad Sci-Biol 104(13):5698–5703

Song C, Fang J, Li X, Liu H, Thomas Chao C (2009) Identification and characterization of 27 conserved microRNAs in citrus. Planta 230:671–685

Staton SE, Bakken BH, Blackman BK, Chapman MA, Kane NC et al (2012) The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J 72:142–153

Stergiou G, Katsiotis A, Hagidimitriou M, Loukas M (2002) Genomic and chromosomal organization of Ty1-*Copia*-like sequences in *Olea europaea* and evolutionary relationships of *Olea* retroelements. Theor Appl Genet 104:926–933

Sunkar R, Li Y-F, Jagadeesvaran G (2012) Functions of miRNA in plant stress responses. Trends Plant Sci 17:196–203

Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. BMC Genom 8:132

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biol Evol 3:219–229

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Unver T, Turktas M, Dorado G, Hernandez P, Van de Peer Y (2016) De novo whole genome sequencing of olive tree (*Olea europaea* L.). In: Abstract of the XXIV international plant & animal genome, San Diego, 8–13 Jan 2016, P1164

van Nocker S, Gardiner SE (2014) Breeding better cultivars, faster: applications of new technologies for the rapid deployment of superior horticultural tree crops. Hort Res 1:14022

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2:e1326

Verstrepen KJ, Jansen A, Lewitter F, Fink GR (2005) Intragenic tandem repeats generate functional variability. Nat Genet 37:986–990

Vitte C, Fustier MA, Alix K, Tenaillon MI (2014) The bright side of transposons in crop evolution. Brief Funct Genomics 13:276–295

Yanik H, Turktas M, Dundar E, Hernandez P, Dorado G, Unver T (2013) Genome-wide identification of alternate bearing-associated microRNAs (miRNAs) in olive (*Olea europaea* L.). BMC Plant Biol 13:10

Zhang H, Miao H, Wang L, Qu L, Liu H, Wang Q, Yue M (2013) Genome sequencing of the important oilseed crop *Sesamum indicum* L. Genome Biol 4:401