Chapter 7

Olive Tree Genomic

Rosario Muleo, Michele Morgante, Riccardo Velasco,
Andrea Cavallini, Gaetano Perrotta and Luciana Baldoni

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51720

1. Introduction

The cultivation of olive trees dates back to ancient time. Mythology ascribes to the divine will the domestication of this species; the goddess Athena taught the people of the city of Athens, as a gift, the cultivation of the tree and the treatment of the drupe (Kakridis, 1986). Despite the economic, cultural and ecological importance of olive groves in the Mediterranean area, now extending to other regions, olive has been a poorly characterized species at genetic and genomic level among other fruit tree crops. Therefore, still remains unknown the inheritance of most genes controlling the agronomical performance and quality traits, even though in the last thirty years a wide molecular survey has been performed on the olive germplasm (Rugini et al., 2011). In the Mediterranean Basin, in fact, is conserved the majority of a large number of olive cultivars estimated in more than 1,200 (Bartolini et al., 2004).

Olea europaea subsp. *europaea* is present in two forms, namely wild (*Olea europaea* subsp. *europaea* var. *sylvestris*) and cultivated (*Olea europaea* subsp. *europaea* var. *europaea*); it is a diploid species ($2n = 2x = 46$), and the genome size range between 2.90 pg/2C and 3.07 pg/2C, with 1C = 1,400-1,500 Mbp (Loureiro et al., 2007). Crosses with other subspecies and with the wild plants are possible and may produce fertile offsprings, providing access to an enormous pool of genetic variability. Over the last two decades, new knowledge on olive genetics has been produced, with the development of nuclear and plastidial molecular markers and linkage maps.

The long generation time of the species has severely restricted breeding strategies to clonal or varietal selection and, in a very few cases, to inter-varietal crosses. Approaches of marker assisted selection could speed up the cross breeding programs but QTL markers are not yet available. The first linkage map of *Olea europaea* was constructed by de La Rosa and co-workers (2003), through the use of dominant PCR markers, such as RAPDs and AFLPs, and codominant marker as RFLPs and SSRs on a cross progeny between two highly heterozygous cultivars. Other maps have been constructed by the use of RAPDs,

microsatellites and SCAR markers on a Frantoio x Kalamata progeny (Wu et al., 2004) and, more recently, a new maps has been derived through SSR, AFLP, ISSR, RAPD and SCAR marker, scored on a 140 F1 progeny from a Picholine Marocaine x Picholine du Languedoc cultivars cross (El Aabidine et al., 2010). In any case, no QTLs of agronomical interest for olive breeding have been detected.

The Italian project, OLEA, is an initiative, mainly supported by Italian Minister of Agricultural, Food and Forestry Policies, dedicated toward the development genomic resources of olive, and it aims to identify, isolate and determine the function of genes that are associated with both vegetative and reproductive phenotype. Therefore, the knowledge of the genetic structural basis is the first step to identify the relevant differences in the control of gene expression of the same sets of genes that exist among different genotypes. The development of new molecular tools through approaches of structural and functional genomics, together with those from proteomics, metabolomics, mapping and genotyping, will allow to advance in molecular breeding of olive, pull out under-exploited natural diversity that is present in the *Olea* complex and in olive germplasm, dissect the molecular mechanisms underlying traits related to high valued compounds and those involved in plant-environment interactions, establish a platform for a rapid and cost-effective transfer of knowledge and technologies.

2. Genome sequencing and assembly

The olive genome is being sequenced using a combination of Next Generation Sequencing (NGS) technologies and a combination of assembly approaches, using the cultivar Leccino as the genotype to be sequenced. The Whole Genome Shotgun approach to assembling 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 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 clone sequenced on average at a 50X coverage. The advantages of the BAC approach are of two types: on one hand each BAC pool is much smaller in size than 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 maternal and paternal-derived genomes that strongly affects WGS approaches. The advantage of the WGS approach is the much more complete and homogeneous coverage of the entire genome. The two assemblies derived, 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 we have produced all of the data needed for the Whole Genome Shotgun component. We have produced approximately 90 Gbp of Illumina sequence data, corresponding to a nominal coverage of 60X of the olive genome. The Illumina sequences were obtained from two paired-end libraries with 500-600 bp inserts that were sequenced on the Illumina Genome Analyser 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 HiSeq2000 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 HiSeq2000 to produce 100 bp reads and reach a coverage of 4 genome equivalents (6 Gbp).

We have produced approximately 18 Gbp of Roche-454 sequence data, corresponding to 12X coverage approximately. 12 Gbp were obtained as long single reads of which approximately one third were 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 were obtained as paired end reads from 3 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 sequenced, i.e. all those sequences that have been obtained from inserts of larger size, will be utilised in order to scaffold into larger assemblies the contigs obtained from the assembly of the reads from the shorter inserts and 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.

We have performed a number of assemblies 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 data sets have been initially filtered for low quality sequences and for chloroplast DNA contamination and then 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. We expect that the addition of the remaining 454 sequenced from the large insert libraries (3 and 8 Kbp inserts) should greatly improve the assembly by increasing considerably the N50 size of the scaffolds. However, due to the problems posed by the high levels of sequence heterozygosity present in the olive genome of cultivar Leccino, we consider the sequencing of the pools of BACs a necessary component of our strategy in order to obtain a satisfactory assembly. The problems here are represented by the difficulties in obtaining BAC libraries with large insert sizes (>100 Kbp) from cultivar Leccino. Should this not prove feasible we will anyhow resort to using a fosmid library (40 Kbp inserts).

3. Analysis of the repetitive component of the genome

3.1. Assembly of olive repetitive sequences

Some of the biggest technical challenges in sequencing eukaryotic genomes are caused by repetitive DNA (Alkan et al., 2011): that is, sequences that are similar or identical to sequences elsewhere in the genome.

The first step in characterizing and sequencing large genomes has to be a genome survey, from which important information about common repeat sequences can be obtained. NGS data are particularly suitable to identify sequences present in many copies per genome, by assembling reads according to their sequence.

The olive genome is largely uncharacterized, despite the growing importance of this tree as oil crop. Concerning repeated sequences, the most characterized are tandem repeats belonging to 4 families, 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). Also putative retrotransposon fragments have been isolated and sequenced (Stergiou et al., 2002; Natali et al., 2007), but a comprehensive picture of RE landscape in the olive genome is still lacking.

We have performed a deep analysis of the repetitive component of olive genome, using NGS techniques (454 and Illumina). We have used around 25 million Illumina paired-end reads of 75 nt, corresponding to 1.8 billion nt and a 1.3 x coverage, and around 8 million 454 single reads, with mean read length of 407 nt, corresponding to a total of 3.3 billion nt and a 2.3 x coverage.

This large amount of sequencing data cannot be sufficient for whole genome assembly, but it enables representative sampling of elements present in a genome in multiple copies. Moreover, the proportion of individual sequences in the reads reflects their genomic abundance, thus providing a simple and reliable means for quantification of repetitive elements (Macas et al., 2007).

In our experiments, we performed de novo repeat identification and reconstruction by direct assembly of the reads. Due to the relatively low genome coverage of the sequencing, most of the contigs that are obtained 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 (Novak et al., 2010). Even though 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 full length repetitive elements using PCR (Swaminathan et al., 2007).

We assembled Illumina and 454 sequence reads by overlapping DNA sequence fragments using CLC-BIO and CAP3 as aligners. In spite of recent progresses, a major challenge remains when reads map to multiple locations, i.e. with multi-reads. The occurrence of multi-reads is strongly dependent on the read length: they are most common in the Illumina sequence packages, and less common in 454 sequence packages, in which sequence length is rapidly growing to lengths similar to those achieved by classical Sanger sequencing, though at higher costs than Illumina.

The sequencing coverage affects heavily the possibility to recover repeated sequences. Obviously, the larger is the coverage, the higher is the possibility that multi-reads are not resolved and discarded. For example, it has been demonstrated, in pea, that a very low coverage (0.008 ×) of the genome allows to obtain repetitive sequences present with at least

1000 copies (Macas et al., 2007). Hence, we decided to proceed to the final assembly of Illumina reads after having splitted the sequence read datasets into subpackages of different genome coverages.

In a first assembly, we assembled the complete pool of Illumina reads using CLC-BIO and subsequently CAP3 assembler. In other experiments, the pool of Illumina reads was splitted into 8, 16, 32, 250, or 500 subpackages and assembled separately (indicated as split 1, 2, 3, 4, and 5, respectively); for each splitting, the resulting contigs were assembled on their turn using CAP3 assembler obtaining 210,063 supercontigs.

All supercontigs were then mapped with all Illumina 75 nt long reads (Table 1). It can be observed that major splittings allow to recover the most redundant supercontigs, that are not found in the lower splittings, because of their too large coverage and, hence, the occurrence of multi-reads. Due to the different redundancy observed in the different subpackages, we decided to use all supercontigs in the final assembly.

Split	Nr. of sub-packages	Subpackage coverage	Nr. of assembled supercontigs	Mean length	Mean nr. of mapped reads	Average coverage	N_{50}
0	0	1.309 x	44336	235.6	19.61	6.07	243
1	4	0.327 x	78983	200.5	19.96	6.01	201
2	8	0.163 x	50698	204.2	31.72	9.56	204
3	32	0.041 x	22749	252.3	68.58	14.35	265
4	244	0.005 x	14748	240.6	218.57	74.42	258
5	489	0.003 x	11819	223.6	212.77	67.99	239

Table 1. Characteristics of supercontig sets obtained by CLC Bio Workbench and CAP3 assembly after different splitting of Illumina reads.

Concerning 454 sequence reads, we did not proceed to such a subdivision, estimating that the superior length of reads compared to that of Illumina ones allowed to recover also highly repeated sequences. In fact, in longer sequences, the occurrence of multi-reads is naturally reduced.

All Illumina- and 454-derived supercontigs and contigs longer than 80 nt were masked against an in-house made database of chloroplast and mitochondrial sequences using RepeatMasker, and organellar sequences were removed. Then, a final assembly was performed, using CAP3, among all datasets, i.e. six Illumina datasets (split 0-5) and one 454 dataset. The resulting whole genome dataset included 238,914 supercontigs, with mean length of 667.9 nt and N_{50} = 1.331.

3.2. Estimation of copy number of assembled sequences

Assuming that Illumina sequence reads in our experiments are sampled without bias for particular sequence types, mapping the whole genomic dataset with Illumina sequence reads provides a method of estimating the copy number of any genomic sequence in the dataset (Swaminathan et al., 2007).

Data in the literature and slot blot experiments previously performed in our lab (Giordani, personal communications) allowed estimation of the copy number per haploid genome of 16 sequences. The 16 sequences with known redundancy were inserted in the whole genomic database and used as reference for the estimation of copy number by mapping on them a pool of around 270 million Illumina 75 nt reads (coverage 14.4 x). We adopted a classification commonly used in biochemical experiments (Britten & Kohne, 1968) and defined supercontigs as highly repeated (HR, redundancy > 10.000 copies per genome, 3,619 supercontigs), medium repeated (MR, redundancy ranging between 100 and 10,000 copies per genome, 67,045 supercontigs) and "unique" (U, redundancy < 100 copies per genome, 168,250 supercontigs).

3.3. Olive genome composition

HR and MR supercontig datasets were annotated to produce the OLEAREP 1.0 database. The annotation pipeline is reported in Figure 1.

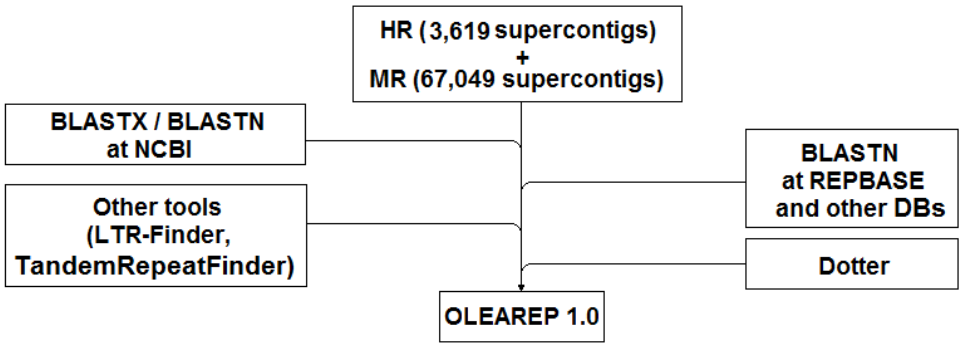


Figure 1. The annotation pipeline for the production of OLEAREP database.

The distribution of sequence type in the HR and MR datasets and in the whole OLEAREP 1.0 database is reported in Table 2.

The average coverage of each HR and MR sequence was used to estimate the redundancy of the various types of repeat classes. Concerning the whole olive genome, around 50% appears to be made of highly repeated sequences. Of these, around 2/3 are tandem repeats belonging to five major families and other minor families (Figure 2). Such extreme redundancy of tandem repeats appears a peculiar feature of olive genome, not found in the

other plant species whose genome has been sequenced. On the contrary, medium repeated component is mainly composed of LTR-retrotransposons, while tandem repeats are much less represented in this genome portion.

Sequence type		Nr. of sequences (%)			
		HR		MR	
DNA transposons		31	(0.86)	2,183	(3.26)
Retrotransposons	LTR-*Copia*	134	(3.70)	7,569	(11.29)
	LTR-*Gypsy*	258	(7.13)	8,066	(12.03)
	Non-LTR	29	(0.80)	949	(1.42)
Tandem repeats		1,535	(42.42)	6,718	(10.02)
rDNA		29	(0.80)	555	(0.83)
Putative genes		46	(1.27)	2,729	(4.07)
Unknown repeats		317	(8.76)	1,795	(2.68)
No hits found		1,240	(34.26)	36,481	(54.41)
Total		3,619		67,045	

Table 2. Functional percentage distribution of the supercontigs in OLEAREP 1.0.

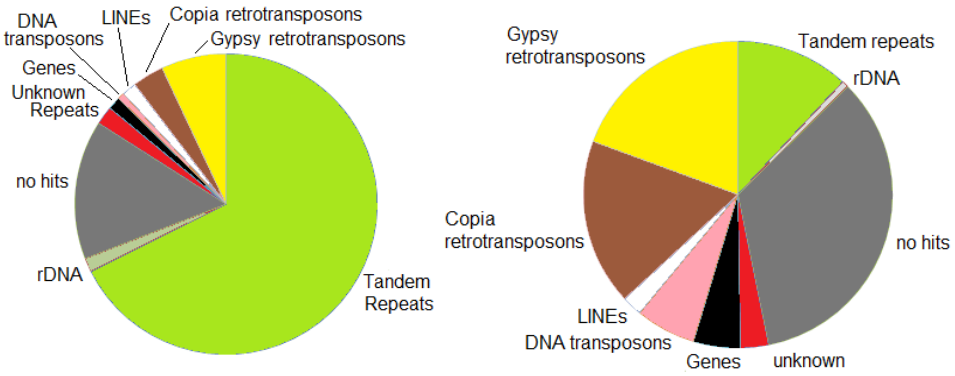


Figure 2. HR (left) and MR (right) fraction composition.

4. Olive chloroplast genome

The chloroplast genome of the olive has an organisation and gene order that is 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). 40 polymorphisms have been identified in the plastome sequence, poorly able to differentiate among olive cultivars.

5. miRNA

A first inventory of sRNAs in olive has been obtained from juvenile and adult shoots, revealing that the 24-nt class dominates the sRNA transcriptome and atypically accumulates to levels never seen in other plant species, suggesting an active role of heterochromatin silencing in the maintenance and integrity of its large genome (Donaire et al., 2011). A total of 18 known miRNA families were identified in the libraries.

6. Analysis of transcriptome

Despite its global importance, genomic sequence resources available for olive are still scarce, though an increasing number of expressed gene functions are being described in the last few years through limited NGS approaches. Recently, many EST sequences from large scale transcriptomic analyses of different organs, such as fruits and leaves have been released (Alagna et al., 2009; Galla et al., 2009; Ozgenturk et al., 2009).

While these studies have highlighted the utility of cDNA sequencing for candidate gene discovery and gene function, a comprehensive description of genes expressed in *Olea europaea* remains unavailable.

Over the past several years, the NGS technology has emerged as a cutting edge approach for high-throughput sequence determination and this has dramatically improved the efficiency and speed of gene discovery (Ansorge, 2009). Furthermore, NGS has also significantly accelerated and improved the sensitivity of gene-expression profiling and, is expected to boost collaborative and comparative genomics studies (Strickler et al., 2012).

In this study, we generated over one million sequence reads with 454 FLX technology (Roche Diagnostics Corporation, Basel, Switzerland) and identified a number of gene functions potentially involved in the expression of major traits that control productivity and quality of olive and oil production.

The starting materials used to explore the olive transcriptome were flower and fruit samples from five different genotypes. Flower tissues at different developmental stages were sampled from Leccino, Dolce Agogia and Frantoio varieties. Two 454 sequencing libraries were obtained from retro-transcribed pooled RNA samples, extracted from flower buds at all stages of development until anthesis of Leccino and Dolce Agogia genotypes, respectively. Furthermore, pooled flower samples of Leccino and Frantoio genotypes, collected after anthesis, were used for synthesis of cDNAs and for the subsequent preparation of two additional 454 sequencing libraries.

In order to gain information on genes expressed in the drupe, with particular regard to those involved in response to pathogen infections, another set of four 454 sequencing libraries was obtained from fruit samples (at about 17 weeks after flowering) of Ortice and Ruveia genotypes collected before and after the infection caused by the olive fruit fly (*Bactrocera oleae*).

The eight 454 cDNA libraries, four from flower and four from fruit tissues, were sequenced in two separate runs by using 454 GS FLX Titanium Sequencer (Roche Diagnostics Corporation, Basel, Switzerland); each library was loaded on ¼ sector of a picotiter sequencing plate.

We identified a total of more than 1 million sequence reads with an average length of 356 bp, corresponding to a little less of half billion bases, about 60% of them are from fruit and 40% from flower samples (Table 3).

Sample	Reads	Total Bases	Length Average (bp)
Flower_1	113,134	42,568,083	376.55
Flower_2	146,576	54,807,276	374.23
Flower_3	67,797	21,939,320	323.92
Flower_4	137,824	48,459,361	351.82
Flower_total	465,331	167,774,040	360.55
Fruit_1	173,118	63,611,956	367.45
Fruit_2	197,782	73,464,937	371.44
Fruit_3	146,765	52,792,933	359.71
Fruit_4	177,846	61,220,898	344.23
Fruit_total	695,611	251,090,724	361.02
Olea_total	1,160,942	418,864,764	356.63

Table 3. Raw sequencing data

Assembling of adaptor-trimmed 454 sequence data was performed using GSAssembler Software (Roche Diagnostics Corporation, Basel, Switzerland). To build a compilation of gene structures and functions expressed in *Olea*, we first assembled row data from all the eight libraries together (Table 4).

			Contigs		Singletons	
Samples in Assembly	Reads in Assembly	%	Number	Length Average (bp)	Number	Length Average (bp)
Flower + Fruit samples	964.266	83,05	25.342	892	112.717	323

Table 4. Total assembling

More than 83% of raw sequences were included in the assembly with 112,717 remaining as singletons. This produced a set of 25,342 contigs with an average length of 892 bp (Table 4). As expected, when sequences from flower and fruit samples are assembled separately, the

number of EST sequences assembled in contigs are significantly lower; however the average length of contigs and singletons remains similar (Table 5).

			Contigs		Singletons	
Samples in Assembly	**Reads in Assembly**	**%**	**Number**	**Length Average (bp)**	**Number**	**Length Average (bp)**
Flower	338.853	72,82	14.599	804	91.999	345
Fruit	570.878	82,01	15.058	884	72.662	333

Table 5. Flower and Fruit Assembling

To assess the representativeness and the overall quality of the assembling, three randomly chosen gene sequences, among those already characterized in *Olea*, were used as a reference to map contig and singleton sequences produced by the assembling (Figure 3).

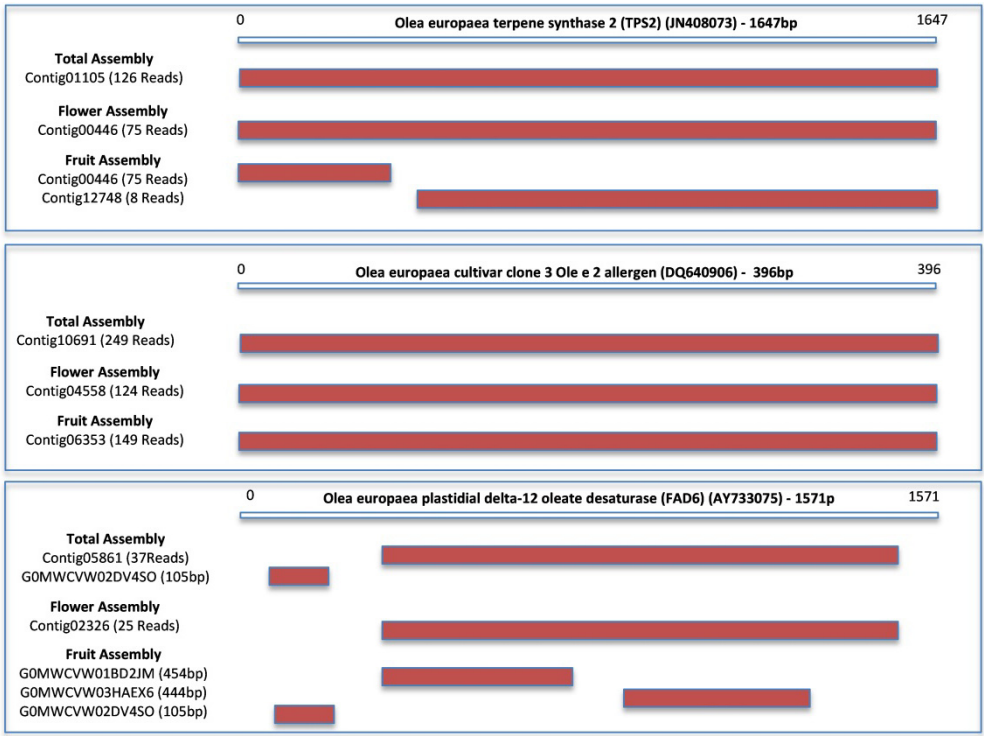


Figure 3. Overview of assembling procedure

The fact that two out of three selected genes are 100% covered by the total assembly with a single contig composed by a great number of EST's, indicates the coverage of the assembly is sufficient to characterize the full coding sequence of high-medium expressed transcripts. Only FAD 6 shows partial coverage, especially in the fruit assembly where only three

matching singleton sequences were found (Figure 3). This is most probably due to the sharp decrease of FAD 6 transcript abundance in fruits sampled at late developing stages, from 15 to 20 weeks after flowering (Matteucci et al., 2011).

To predict gene functions, we used a BlastX-based annotation (E-value ≤ 1-e^{-5}) of unigenes comparing them to NCBI non-redundant (nr) database (http://www.ncbi.nlm.nih.gov/). About 52% of the unigenes match to known functional genes; while the remaining 48% has no function assigned (Figure 4).

The majority of the BlastX annotated unigenes matches most to *Vitis vinifera*, *Populus trichocarpa* and *Ricinus communis* counterpart sequences, in decreasing order (Figure 4).

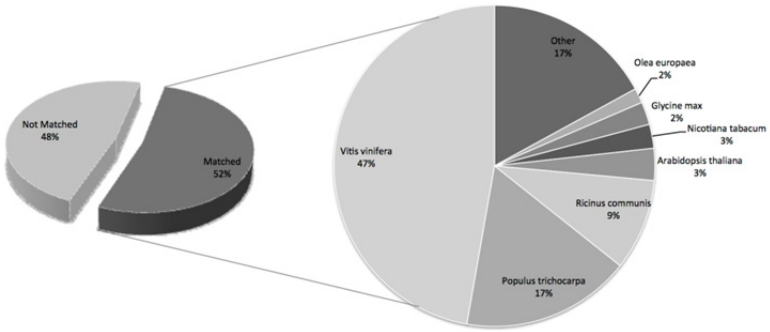
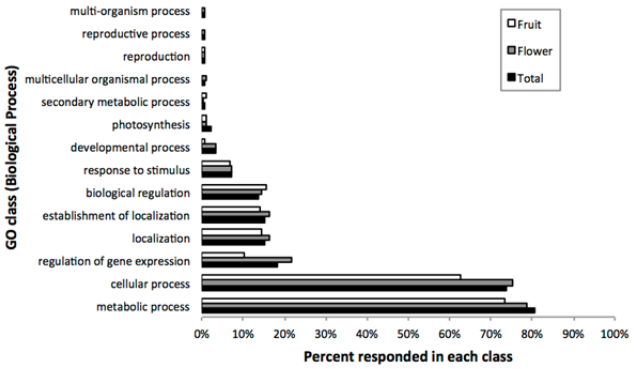


Figure 4. Overall profile of unigenes based on homology with GenBank sequences

We also mapped the GI identifiers (http://www.ncbi.nlm.nih.gov/) of the best BlastX hits to UniprotKB protein database (http://www.uniprot.org/) in order to extract Gene Ontology (GO, http://www.geneontology.org/). Approximately one-fourth of the unigene set was assigned to GO terms. This allowed us to group unigenes in 14 sub-categories of biological processes, 9 sub-categories of cellular components and 11 sub-categories of molecular functions (Figure 5).



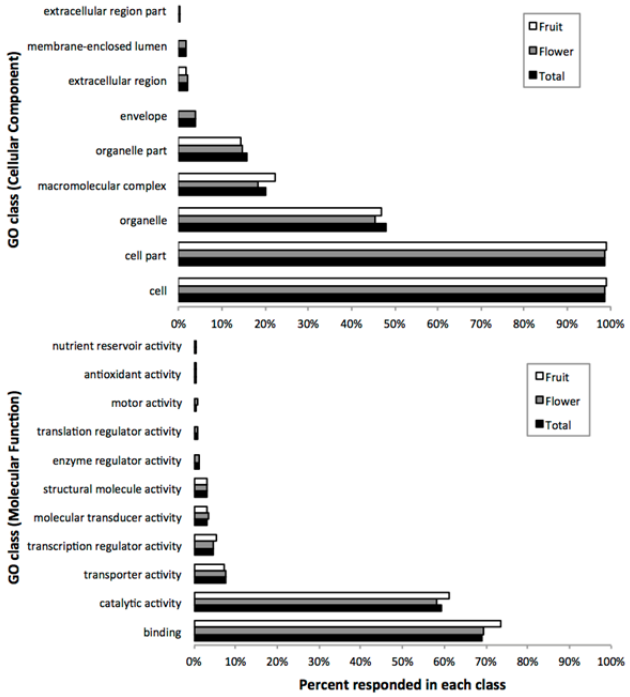


Figure 5. GO terms distribution in the cellular components, biological processes and molecular functions vocabularies.

Metabolic process sub-category, consisting of more than 11,000 genes, is dominant in biological process. While, binding and cell part subcategories, consisting of about 13,0000 and 8,000 genes, are dominant in molecular function and cellular component, respectively. We also noticed an appreciable number of genes included in cellular process, catalytic activity and organelle sub-categories (Figure 5). However, at this level of detail no dramatic differences are evident between flower and fruit tissue transcriptomes.

Next generation RNA sequence from additional organs/tissues/genotypes of *Olea europaea*, as well as, full comparative data analysis between and within the sequenced samples are currently in progress. These studies will certainly provide valuable information about gene functions that trigger key metabolic pathways for the expression of desired traits.

7. Conclusion

Genomic sequences for olive will enable researchers to explore the breadth of genetic diversity present within the species and within the breeding germplasm using high throughput methods of resequencing based on NGS technologies. This will give access to all types of variations, namely Single Nucleotide Polymorphisms (SNPs), small insertion/deletions and structural variants (large insertion/deletions). It will allow a better

assessment of the relationships among the different accessions, of the geographical patterns of distribution of genetic variation and of the genetic consequences of olive trees domestication. It will finally form the basis for the development of novel molecular marker assays.

They will also allow the analysis of global gene expression and specific gene expression of olive tissues in diverse developmental stages and conditions. The identification and characterization of expression of important genes involved in agronomic and productive traits affecting fruit production and quality, biotic and abiotic stress resistance, important development characters (e.g., juvenility, self-incompatibility, ovary abortion, chill response), it may offer a significant amount of tools and open new opportunities for improvement either through molecular breeding and/or genetic engineering.

Many researches will be focused on gene network activities, using olive microarray and/or qPCR to address the expression patterns of genes, during plant and fruit development and ripening of drupe fruits, fatty acid metabolism as well as phenylpropanoid metabolism. Moreover, new generation of molecular markers will be developed, helpful to localize genes involved in both monogenic and polygenic agronomic traits, to construct genetic fine-maps; these markers will be also used for marker-assisted selection (MAS) to obtain elite genotypes by allowing the analysis of cross progenies at earlier stages.

Author details

Rosario Muleo
University of Tuscia, Dept. DAFNE, Viterbo, Italy

Michele Morgante
IGA-Institute of Applied Genomics, Udine, Italy

Riccardo Velasco
Edmund Mach Foundation, IASMA, San Michele all'Adige, Italy

Andrea Cavallini
Dept. Crop Species Biology, Pisa, Italy

Gaetano Perrotta
ENEA, Trisaia, Rotondella (MT), Italy

Luciana Baldoni
CNR- Institute of Plant Genetics, Perugia, Italy

Acknowledgement

This research was partially supported by Progetto Strategico MIPAF "OLEA - *Genomica e Miglioramento genetico dell'olivo*", D.M. 27011/7643/10, and by the Province of Trento and Edmund Mach Foundation. We thank the Roche Diagnostic Spa, Applied Science to support the OLEA Italian Project.

8. References

Alagna, F.; D'Agostino, N.; Torchia, L.; Servili, M.; Rao, R.; Pietrella, M.; Giuliano, G.; Chiusano, M.L.; Baldoni, L.; Perrotta, G. (2009). Comparative 454 Pyrosequencing of Transcripts from Two Olive Genotypes During Fruit Development. *BMC Genomics*, Vol. 10, pp. 1-15, ISSN 1471-2164

Alkan, C.; Coe, B.P. & Eichler, E.E. (2011). Genome Structural Variation Discovery and Genotyping. *Nature Reviews Genetic,* Vol. 12, No. 5, pp. 363-376, ISSN 1471-0056

Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, Vol. 25, No. 4, pp. 195-203, ISSN 1871-6784

Bartolini, G.; Prevost, G. & Messeri, C. (1994). Olive tree germplasm: descriptor lists of cultivated varieties in the world. *Acta Horticulturae*, Vol. 365, pp. 116-118, ISSN 0567-7572

Britten, R.J. & Kohne, D.E. (1968). Repeated sequences in DNA. *Science*, Vol. 161, pp. 529-540, ISSN 0036-8075

Contento, A.; Ceccarelli, M.; Gelati, M.T.; Maggini, F.; Baldoni, L.; Cionini, P.G. (2002). Diversity of *Olea* genotypes and the origin of cultivated olives. *Theoretical and Applied Genetics*, Vol. 104, No., 8, pp. 1229-38. ISSN 1432-2242

De La Rosa, R.; Angiolillo, A.; Guerrero, C.; Pellegrini, M.; Rallo, L.; Besnard, G.; Bervillé, A.; Martin, A. & Baldoni, L. (2003). A first linkage map of olive (*Olea europaea* L.) cultivars using RAPD, AFLP, RFLP and SSR markers. *Theoretical and Applied Genetics* Vol. 106, No. 7, pp. 1273-1282. ISSN 1432-2242

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* Vol. 6, No. 11, pp. e27916. ISSN 1832-6203

El Aabidine, A.Z.; Charafi, J.; Grout, C.; Doligez, A.; Santoni, S.; Moukhli, A.; Jay-Allemand, C.; El Modafar, C.; Khadari, B. (2010). Construction of a genetic linkage map for the olive based on AFLP and SSR markers, Crop Science Vol. 50, No. 6, pp. 2291-2302. ISSN 1435-0645

Galla, G.; Barcaccia, G.; Ramina, A.; Collani, S.; Alagna, F.; Baldoni, L.; Cultrera, N.G.M.; Martinelli, F.; Sebastiani, L.; Tonutti, P. (2009). Computational annotation of genes differentially expressed along olive fruit development. *BMC Plant Biology* Vol. 9, pp. 128. ISSN 1471-2229

Kakridis, I.Th. (1986). *Greek Mythology*, Tome II. Athens: Editons Ekdotiki of Athens, Greece.

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* Vol. 41, No.4, pp. 527-534. ISSN 1480-3321

Lorite, P.; Garcia, M.F.; Carrillo, J. A.; Palomeque, T. (2001). A new repetitive DNA sequence family in the olive (*Olea europaea* L.), *Hereditas* Vol. 134, No.1, pp. 73-78. ISSN 1601-5223

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, *Genetic Resources and Crop Evolution*, Vol. 54, No.1, pp. 21–25. ISSN 1573-5109

Macas, J.; Neumann, P. & Navratilova, A. (2007). Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*, *BMC Genomics*, Vol. 8, pp. 427. ISSN 1471-2164

Mariotti, R.; Cultrera, N.G.M.; Muñoz Díez, C.; Baldoni, L.; Rubin,i A. (2010). Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison, *BMC Plant Biology*, Vol. 10, pp. 211. ISSN 1471-2229

Matteucci, M.; D'Angeli, S.; Errico, S.; Lamanna, R.; Perrotta, G.; Altamura MM. (2011). Cold affects the transcription of fatty acid desaturases and oil quality in the fruit of *Olea europaea* L. genotypes with different cold hardiness. *Journal of Experimental Botany*, Vol. 62, No. 10, pp. 3403-20. ISSN 1460-2431

Minelli, S.; Maggini, F.; Gelati, M.T.; Angiolillo, A.; Cionini, P.G. (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 Research*, Vol. 8, No.7, pp. 615-619. ISSN 1573-6849

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*. *Molecular Breeding*, Vol. 19, No.3, pp. 255-65, ISSN 1572-9788

Novak, P.; Neumann, P. & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, Vol. 11, pp. 378, ISSN 1471-2105

Ozgenturk, N.O.; Oruc, F.; Sezerman, U.; Kucukural, A.; Korkut, S.V.; Toksoz, F., Un, C. (2010). Generation and analysis of Expressed Sequence Tags from *Olea europaea* L., *Comparative and Functional Genomics*, Article ID 757512, 9 pages doi:10.1155/2010/757512, ISSN 1532-6268

Rugini, E.; De Pace, C.; Gutiérrez-Pesce P.; Muleo R. (2011). *Olea*. In: *Wild Crop Relatives: Genomic and Breeding Resources, Temperate Fruits*, Chittaranjan Kole (Ed.), 79-114, Springer-Verlag, ISBN 978-3-642-16057-8, BERLIN-HEIDELBERG, HEIDELBERG, DORDRECHT, LONDON

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. *Theoretical and Applied Genetics*, Vol. 104, No.6-7, pp. 926-33, ISSN 0040-5752

Strickler, S.R.; Bombarely, A. & Mueller, L.A. (2012). Designing a transcriptome next generation sequencing project for a nonmodel plant species. *American Journal of Botany*, Vol. 99, No.2, pp. 257-266, ISSN 1537-2197.

Swaminathan, K.; Varala, K. & Hudson, M.E. (2007). Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics*, Vol. 8, pp. 132, ISSN

Wu, S.B.; Collins, G.; Sedgley, M. (2004). A molecular linkage map of olive (*Olea europaea* L.) based on RAPD, microsatellite, and SCAR markers. *Genome,* Vol. 47, No.1, pp. 26-35, ISSN 1480-3321