Transcriptome characterization and expression profiling in chestnut cultivars resistant or susceptible to the gall wasp Dryocosmus kuriphilus

Alberto Acquadro, Daniela Torello Marinoni*, Chiara Sartor, Francesca Dini, Matteo Macchio, and Roberto Botta

Transcriptome characterization and expression profiling in chestnut cultivars resistant or susceptible to the gall wasp *Dryocosmus kuriphilus*

DISAFA, Dipartimento di Scienze Agrarie, Forestali e Alimentari, Università di Torino, Largo Paolo Braccini 2, 10095 Grugliasco (Italy)

***Corresponding author:**
Daniela Torello Marinoni
e-mail address: daniela.marinoni@unito.it
telephone: + 39 11 6708816
fax: +39 11 6708658

ORCID Alberto Acquadro: 0000-0002-5322-9701
ORCID Daniela Torello Marinoni: 0000-0002-3679-4813
ORCID Roberto Botta: 0000-0002-1952-8775

Acknowledgments

Research founded by Regione Piemonte Administration and by the cooperation program Italy-France Alcotra Project: *Salvaguardia dell'ecosistema castagno.*

Abstract –. The oriental gall wasp *Dryocosmus kuriphilus* represents a limiting pest for the European Chestnut (*Castanea sativa*, Fagaceae) as it creates severe yield losses. The European Chestnut is a deciduous tree, having major social, economic and environmental importance in Southern Europe, covering an area of 2.53 million hectares, including 75,000 ha devoted to fruit production. Cultivars show different susceptibility and very few are resistant to gall wasp. To deeply investigate the plant response and understand which factors can lead the plant to develop or not the gall, the study of transcriptome is basic (fundamental). To date little transcriptomic information are available for *C. sativa* species. Hence, we present a *de novo* assembly of the chestnut transcriptome of the resistant Euro-Japanese hybrid 'Bouche de Bétizac' (BB) and the susceptible cultivar 'Madonna' (M), collecting RNA from buds at different stages of budburst. The two transcriptomes were assembled into 34,081 (BB) and 30,605 (M) unigenes, respectively. The former was used as a reference sequence for further characterization analyses, highlighting the presence of 1,444 putative Resistance Gene Analogues (RGAs) and about 1,135 unigenes, as putative MiRNA targets. A global quantitative transcriptome profiling comparing the resistant and the susceptible cultivars, in the presence or not of the gall wasp, revealed some GO enrichments as "response to stimulus" (GO:0050896), and "developmental processes" (e.g.: post embryonic development, GO:0009791). Many up-regulated genes appeared to be transcription factors (e.g.: RAV1, AP2/ERF, WRKY33) or protein regulators (e.g.: RAPTOR1B) and storage proteins (e.g.: LEA D29) involved in "post-embryonic development". Our analysis was able to provide a large amount of information, including 7k simple sequence repeat (SSR) and 335k single-nucleotide polymorphism (SNP)/INDEL markers, and generated the first reference unigene catalogue for the European Chestnut. The transcriptome data for *C. sativa* will contribute to understand the genetic basis of the resistance to gall wasp and will provide useful information for next molecular genetic studies of this species and its relatives.

Keywords: RNA-Seq, *Castanea*, resistance, assembly, Gene Ontology

Introduction

The oriental gall wasp *Dryocosmus kuriphilus* Yasumatsu (Hymenoptera: Cynipidae) is considered the most invasive alien pest for the European chestnut (*Castanea sativa* Miller) currently reported in almost the whole Europe. The purpose of this study was a deeply investigation of the chestnut transcriptome to understand the genetic basis of the plant response to gall wasp infestation and to provide a large amount of information for molecular genetic studies on this species.

The European sweet chestnut (*Castanea sativa* Miller) is a multipurpose tree species mostly distributed across Southern Europe, from Turkey to Portugal, but found also in Northern countries such as UK. Its wide distribution and presence in mountain and high hill areas makes the tree an important resource as provider of ecosystemic services in these sites. Besides its value as a forest species and its importance for the landscape and environment, *Castanea sativa* still represents a relevant economic resource for the nut production, being Italy, Turkey and Portugal the major producing countries in Europe. Despite the progressive decline of the last 50 years of production in Europe due to a series of factors including diseases and pest, climatic change, aging and urbanization of mountain people, recently *C. sativa* production is showing a slow production recovery (FAOSTAT 2018).

The diffusion of the gall wasp *D. kuriphilus* in Europe, represented a major threat for chestnut; the pest, native of China, was reported for the first time in Piemonte (NW Italy) in 2002 (Brussino et al. 2002) and now spread in most of the European Countries where chestnut is present. The pest lays eggs into the buds in early summer of the first year; larvae and eggs are found in the buds at the end of winter of the following year, but there is no outer symptoms of the presence of the wasp until after budburst. The evidence of the infestation is the formation of galls, round green and reddish structures that develop on the young shoots in spring, due to the reaction of the plant to the presence of the feeding larvae. Following pupation, adults fly out of the gall in early summer and lay eggs into the new formed dormant buds of the chestnut tree. The thelytokous parthenogenetic reproduction system of the wasp causes an exponential population increase in a short time, while dispersal through propagation material is favoured by the absence of external symptoms in buds during winter.

The damage that galls can cause involve directly leaves, shoots and inflorescences, and indirectly the whole biomass; leaf surface is reduced, and the amount of vegetative buds is decreased, year-by-year (Kato and Hijii 1997). Although plant death is rare and usually associated with other factors such as diseases, the interruption of plant growth and the reduction of fruiting results in yield losses of up to 50-70% in the Chinese chestnut (*C. mollissima*) and Japanese chestnut (*C. crenata*) (Dixon et al. 1986). The assessment of yield loss in *Castanea sativa* (Sartor et al. 2015) showed similar data indicating that infestation values, determined as No. galls/bud, above 0.6 cause a drastic decrease of productivity (60% on average).

In Japan, after the accidental introduction of the gall wasp in 1941, breeding programs were carried out to obtain resistant cultivars starting from resistant genotypes found in *C. crenata*. More recently, the trait was found to be present in other *Castanea* species (*C. mollissima*, *C. pumila*) and in *C. sativa* (Sartor et al. 2015). Studies on the genetic bases of resistance, agree on the hypothesis that more mechanisms may be responsible of the trait in the different chestnut genotypes (Shimura 1972b; Anagnostakis et al. 2009).

Following the introduction of the wasp, in Italy several strategies of control were tested, the most successful being the biological control by *Torymus sinensis* (Kamijio) (Quacchia et al. 2008; Picciau et al. 2017; Ferracini et al., 2018). In parallel, studies were conducted on the susceptibility and resistance to the pest in the cultivated and wild

C. sativa germplasm (Sartor et al. 2015). In fact, there were reports of resistance in *C. crenata* and a large variation in susceptibility observed across genotypes (Shimura 1972a). Among cultivars, the interspecific hybrid 'Bouche de Bétizac' (*C. sativa* 'Bouche Rouge' x *C. crenata* 'CA04') was found to be asymptomatic in spring (no galls), although buds were oviposited and contained larvae in winter (Sartor et al. 2009; Dini et al. 2012). In this case, the occurrence of a hypersensitive response at budburst was postulated to explain larvae death and regular shoot development (Dini et al. 2012). Following these advancements, a segregating progeny accounting 250 individuals was obtained from 'Bouche de Bétizac' X 'Madonna' (*C. sativa*, highly susceptible cultivar) in order to map the trait (Torello Marinoni et al. 2017) and a transcriptome analysis, described in this paper, was carried out. The purpose was to create a catalogue of *C. sativa* unigenes, likely including genes involved in plant-insect interaction, and to isolate molecular markers for the mapping of traits of interest.

With a similar approach, a highly informative genetic map of Chinese chestnut was constructed to extend genomic studies in the Fagaceae and to aid the introgression of Chinese chestnut blight resistance genes into American chestnut (Kubisiak et al. 2013). The transcriptome-based genetic map was created with 329 simple sequence repeat and 1,064 single nucleotide polymorphism markers all derived from expressed sequence tag sequences. Genetic maps for each parent were developed and combined to establish 12 consensus linkage groups spanning 742 cM. Another paper compared the root transcriptome of the susceptible species *C. sativa* and the resistant species *C. crenata* after *P. cinnamomi* inoculation to elucidate chestnut defense mechanisms to ink disease (Serrazina et al. 2015); results of RNA-seq enabled the selection of candidate genes for ink disease resistance in *Castanea.*

In this paper, we sequenced, assembled and functionally characterized the transcriptome of two chestnut cultivars (a cynipid-resistant and a cynipid-susceptible), during the early stages of the interaction plant-pest, generating an extraordinary amount of information. A set of genes regulated in both the susceptible and the resistant cultivars was highlighted. The functional annotation of RGAs and miRNA target genes was attempted and SSR and SNP markers were identified/classified to populate a catalogue suitable for genetic trait dissection. The chestnut transcriptome assembly will open the possibility to deeply study the plant response and understand which factors can lead the plant to develop or not the gall.

Materials and Methods

Chestnut Material

Buds from cultivar 'Madonna' (*C. sativa*) and the Eurojapanese hybrid 'Bouche de Bétizac' (*C. sativa* 'Bouche Rouge' X *C. crenata* 'CA04'), were harvested from single plants at different times of budburst from April 21st to May 12th. The cultivar 'Madonna' buds were harvested in areas highly infested by the cynipid, while the 'Bouche de Bétizac' buds were collected from plants infested by the cynipid, maintained in screenhouses set up in the forest nursery of Chiusa Pesio (CN, Piedmont, Italy) as described in Sartor et al. 2015. The collection was carried out once a week in order to gather material representative of four different stages of bud sprouting (1-closed bud; 2-bud that initiates to swell; 3-end of bud swelling: scales separated; 4-brown scales fallen, bud enclosed by green scales), and be able to sample tissues during the defensive response. Buds were immediately frozen in liquid nitrogen and stored at -80°C until use.

The identity of the cultivars was checked by SSR analysis (CsCAT1, CsCAT3, CsCAT6, CsCAT16, CsCAT17, QpZAG110; Steinkellner et al. 1997; Marinoni et al. 2003) according to the protocol by Torello Marinoni et al. (2013).

RNA and DNA extraction

Buds were disrupted in liquid nitrogen using a baked mortar and pestle treated with DEPC water. Nucleic acids were extracted using a buffer containing: 2% CTAB, 2% polyvinylpyrrolidone (PVP) K-30 (soluble), 100 mM Tris HCl (pH 8.0), 25 mM EDTA, 2.0 M NaCl, 0.5 g/L spermidine, 2% b-mercaptoethanol. After two buffered chloroform extractions, the upper phase containing the nucleic acids was divided in two parts for RNA and DNA extraction. DNA was precipitated with 0.7 volumes of isopropanol and washed in 70% ethanol; it was dried and resuspended in 50 ml of sterile water.

Total RNA was precipitated overnight with 8 M LiCl at 4°C. The next day RNA was added of an SSTE buffer (5.0 M NaCl, 0.5% SDS, 10 mM Tris HCl, pH 8.0, 1 mM EDTA), and treated at 65°C for 10 min. Following two chloroform purifications, RNA was precipitated and then washed in ethanol 100% and 70%. Total RNA was purified with RNeasy Mini Kit (Qiagen). RNA yield and quality were evaluated using spectrophotometric determinations (Dini et al. 2012).

DNA samples were used to check the cynipid presence/absence by diagnostic PCR following the protocol by Sartor et al. (2012), choosing the 320bp amplicon of 28S Ribosomal DNA sequence as marker to detect the larva presence.

RNA sequencing

For RNA sequencing we considered 4 thesis: ‘Bouche de Bétizac’ infested, ‘Bouche de Bétizac’ not infested, ‘Madonna’ infested, ‘Madonna’ not infested. To have a representative sample, the total RNA extracted from single buds, belonging to the same thesis, but collected at four different stages of sprouting, was pooled together..

From these materials, in collaboration with Evrogen (Moscow, Russia), 4 tagged cDNA libraries were obtained: ‘Bouche de Bétizac’ infested (BI), ‘Bouche de Bétizac’ uninfested (BNI), ‘Madonna’ infested (MI), ‘Madonna’ uninfested (MNI). The sequencing was commissioned to BMR Genomics (Padova, Italy), preparing a unique pool of the four tagged samples in equimolar concentration. For the sequencing a single Hiseq 1000 protocol (2PE x 100bp) was used, following the TruSeq DNA protocol (Illumina).

Transcriptome assembly and completeness

The Illumina reads were adapter trimmed (Scythe, https://github.com/vsbuffalo/scythe) and quality filtered (Sickle, https://github.com/najoshi/sickle). Reads were then separated according to their sequence tag using a custom Python script. Reads sizing less than 15 bases were deleted. Quality check of the raw/filtered reads was carried out using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Filtered reads were assembled using ABySS v.2.0 (Simpson et al. 2009) with default parameters; a preliminary assembly optimization was performed by varying *k* values in the range 25-60 and the best assembly was picked using some metrics adopted in Assemblathon2 (https://github.com/ucdavis-bioinformatics/assemblathon2-analysis). A final filtering cutoff criterium (500 bp) was applied and shorter contigs were filtered out.

The ‘Bouche de Bétizac’ unigenes set was selected for transcriptome fine characterization procedures. Its initial gene set was further filtered, for contaminants deriving from pests and fungi, using Blastn against three databases: 1) *Biorhiza pallida* transcriptome (gently provided by Prof. Graham Stone, Univ. of Edinburgh); 2) *Synergus japonicus* draft genome (gently provided by Prof. Graham Stone, Univ. of Edinburgh); 3) *Aureobasidium pullulans*. High ranked contigs having high identity with insects and the fungus sequences were removed. Transcriptome completeness was then assessed by means of the CEGMA pipeline (Parra et al. 2007) measuring the percentages of 248 different Core Eukaryotic Genes (CEGs) mapped in the chestnut assembly.

Transcriptome functional annotation

Functional annotation of the unigenes was locally performed with Blast2GO (Conesa et al. 2005) and gene ontology (GO) terms were predicted by assigning functional classifications (Gene Ontology Consortium, 2000) as well as potential properties of gene products. The blast cut-off E-value was 10^{-5}. The GO terms were assigned to the representative transcripts for each sample through an enrichment analysis using Fisher's exact test (p-value <0.01), with a false discovery rate (FDR) correction in terms of biological processes and molecular functions. The unigenes were submitted to ORF predictor (Min et al. 2005; http://bioinformatics.ysu.edu/tools/OrfPredictor.html) and compared with data from the *C. mollissima* genome (www.hardwoodgenomics.org/organism/Castanea/mollissima).

Resistance genes analogs (RGA) analysis **-** candidates genes were identified by means of a Blastp analysis against the Plant Resistance Genes database (http://prgdb.crg.eu; Sanseverino et al. 2012). Positive hits were validated via HMMERv3 (hmmer.janelia.org/software) software, searching against PFAM hidden Markov models for NB-ARC, TIR and several leucine-rich repeat motifs (Finn et al. 2016).

Mirna target analysis - Transcribed sequences were subjected to psRNATarget (Dai et al. 2011; http://plantgrn.noble.org/psRNATarget/) analysis against miRBase Release 21 (Griffiths-Jones et al. 2008). A maximum expectation of 2.5 was adopted, allowing a maximum energy to unpair the target site of 25, and considering 17 bp upstream and 13 bp downstream of the target site sequence. Inhibition of translation was considered for mismatches in the 9th to 11th mature miRNA nucleotides. Any enrichment of GO terms was verified by comparing the putative miRNA targets against the whole transcript dataset by means of the Gossip package implemented in the Blast2Go suite: a Fisher’s exact test was applied collecting terms with *P* values $<e^{-4}$ and false discovery rate <0.01.

Differentially expressed genes (DEGs)

The produced clean reads were mapped to the reference transcriptome (B) using BWA software with default parameters. Only the reads that could be uniquely mapped to the transcriptome were used for subsequent processing. The retained reads were quantified using the “count” function implemented in GFOLD algorithm (Feng et al. 2012; https://bitbucket.org/feeldead/gfold), which considered the expression level of each gene by normalizing to the read per kilobase of exon per million mapped reads (RPKM) value. Differentially expressed genes were identified using the “diff” function of GFOLD algorithm, which was biologically meaningful for single replicate experiments. Genes with four fold change (GFOLD > 1 or < -1) were considered differentially expressed between two samples. The transcriptome profile of the two varieties was analysed considering four pairwise comparisons (Figure 2A; BI vs BNI; BI vs MI; MI vs MNI; BNI vs MNI).

SSR identification and primer design

SSR motifs were identified with the suite SciRoKo (Kofler et al. 2007; http://kofler.or.at/bioinformatics). Perfect and imperfect mono, di-, tri-, tetra-, penta- and hexanucleotide motifs were targeted with default parameters. Primer pairs were designed from the flanking sequences using Primer3 software (Rozen and Skaletsky 2000) in batch mode, as implemented in the SciRoKo package. The target amplicon size range was set as 125-450 bp. The optimal annealing temperature was 60° C, and the optimal primer length 20 bp.

SNP mining

Two transcriptomic read sets were constructed by pooling filtered reads from infested and not infested 'Bouche de Bétizac' buds (BI + BNI), and reads from infested and not infested 'Madonna' buds (MI + MNI). The two pools were independently back-aligned to the reference transcriptome ('Bouche de Bétizac') with the Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net) using *mem* as algorithm with default parameters. Automated SNP calling on was carried out on sorted bam alignment files by SAMtools mpileup and vcftools in a multi-sample call pipeline. To populate the starting SNP table, a minimum mapping quality of 25 was required, with a minimum SNP quality of 20. SNP characterization (test-cross/intercross; homozygous/heterozygous) was addressed through custom bash scripts. The full SNP data set was organized into a relational database, available upon request.

Results

Transcriptome sequencing and assembly

The whole Illumina sequencing experiment resulted in 361M raw pair ended reads with an average length of 100 bp. Filtering/trimming operations reduced the reads to 298M (83%, Table 1). The total amount of high quality sequence was 59.7 Gb ('Bouche de Bétizac', 28 Gb; 'Madonna', 31.7 Gb).

'Bouche de Bétizac' assembly metrics - The best draft assembly was established using a k value of 47 and the initial reference resulted in about 1 million contigs. By applying a cutoff of 500 bp, the assembled transcriptome was reduced to 39,365 scaffolds of average length of 1,032 bp. The 50% of the *de novo* assembly (N_{50}) was included in 11.440 scaffolds of 1,142 bp or larger, with 41.76% G+C bases (Table 2). About 40.6 Mb represented the final assembly span. The longest scaffold was 13.7 kb. Contigs in the 500-1000 bp range were 24,880 (63.2%, Fig. 1). The contigs exceeding 1kb were 14,481 (36.8%), of these 13,876 (35.3%) were in the length range 1000-3000 bp.

'Madonna' assembly metrics - The best draft assembly was established using a k value of 45 and consisted of 30,605 scaffolds of average length of 1,018 bp. The 50% of the *de novo* assembly (N_{50}) was included in 8,886 scaffolds of 1,120 bp or larger, with 41.44% G+C bases (Table 2). About 31.2 Mb represented the final assembly span. The longest scaffold was 5.4 kb. Contigs in the 500-1000 bp range were 19,809 (64.7%, Fig. 1). The contigs exceeding 1kb were 10,796 (35.3%), of these 10,362 (33.9%) were in the length range 1,000-3,000 bp.

Transcriptome filtering and completeness

The *de novo* assembled transcriptomes of both cultivars were produced and, through a reciprocal blast analysis, 20,950 common unigenes were detected (in addition, 18,415 were specific for 'Bouche de Bétizac', 9,655 for Madonna). High ranked contigs having high identity with insects and the fungus sequences were removed, Overall, 1,039 cynipid-like and 4,245 fungi-like contaminant sequences were filtered out. The resulting 'Bouche de Bétizac' filtered transcriptome contained 34,081 contigs. The assembled transcriptomes of the two cultivars are provided in Online Resource file 1 ('Bouche de Bétizac') and in Online Resource file 2 ('Madonna'). The CEGMA (Core Eukaryotic Genes Mapping Approach), pipeline was adopted to assess the completeness of the transcriptome of 'Bouche de Bétizac' (both unfiltered and insect-filtered sets). The transcriptome draft (unfiltered) was surveyed for the presence of 248 conserved eukaryotic genes (CEGs). More than 69% of the 248 full-length CEGs were mapped (Online Resource file 3**),** some of the missing CEGs were present as partial matches and when included the mapped CEGs rose to 90%. The same pipeline was used to analyse a filtered transcriptome (Online Resource file1), where some insect contaminant transcripts were removed (about 1039 sequences). Results resembled the unfiltered data in both complete and partial alignments. ORF predictor detected about 33,000 sequences with ORF, which were used for further analyses (RGA and miRNA target mining).

Transcriptome functional annotation and differential gene expression

Blast2GO analysis of the 34,081 contigs produced the following results: 21,926 contigs were fully annotated, 2,202 contigs received just a Blast annotation, 3,683 received a GO Mapping, 573 received an "InterPro Scan" annotation **(**Online Resource Fig. 1a). About 7,000 sequences were not effectively annotated. Overall, the Blast2GO annotation permitted the functional annotation of 27,345 (71.3%) unigenes. The annotation was mainly referred to UniProtKB (UniProt Consortium) and TAIR 10 (www.arabidopsis.org/) databases (Online Resource Fig. 1b); four species provided a total of 16,391 "useful" annotations. The higher number of homologies with the chestnut transcriptome reported in Online Resource Fig. 1c was found in *Arabidopsis thaliana* (6,310 sequences), *Vitis vinifera* (5,565 sequences), *Populus trichocarpa* (2,339 sequences), and *Ricinus communis* (2,177 sequences).

The transcriptome profile of the two cultivars was analysed and four comparisons were carried out (Fig. 2a; BI vs BNI; BI vs MI; MI vs MNI; BNI vs MNI). Infested 'Bouche de Betizac' buds (cynipid-resistant) compared to the healthy buds (BI vs BNI) showed a relative low regulation (389 genes up and 168 down-regulated). Infested 'Madonna' buds (cynipid-susceptible), compared to the healthy buds (MI vs MNI), showed a major down-regulation (706 genes up and 2,108 down-regulated). Infested 'Bouche de Betizac' buds (cynipid-resistant) compared to infested 'Madonna' buds (cynipid-susceptible; BI vs MI) showed a high level of regulation (2,488 genes up and 2,178 down-regulated). Indeed a very high variation in transcriptome was observed between the two healthy cultivars (BNI vs MNI; 2,5k genes upregulated and 4,9k genes down-regulated, Fig. 2a). The genes regulated in infested 'Bouche de Bétizac' and 'Madonna transcriptomes are shown in Online Resource table 1 and were further analyzed. Considering the biological processes (Fig. 3), almost 300 "response to stimulus" (GO:0050896) related genes appeared up regulated. Up to 70 genes involved in "post embryo development" (GO:0009791) appeared regulated and most of them were transcription factors involved in the plant development (Fig. 3). Many up and down regulated proteins appeared associated to "death" (GO:0016265) processes and "apoptosis" (GO:006915) and some of them were involved in the hyper-sensitive response. In the molecular functions, the "transcriptional regulator activity" (GO:0030528) term was enriched only in the up-regulated genes.

In the cellular components, the “vacuolar part” (GO:0044437) term appeared significantly enriched only in up-regulated genes (Fig. 3). Considering only the genes that were commonly regulated “BI vs BNI” and “BI vs MI” (Venn diagram intersection, Fig. 2b) and not regulated in “BNI vs MNI” and “MI vs MNI”, ~100 genes were highlighted (Fig. 2b**;** Online Resource table 2). Some GO enrichments were still highlighted for specific GO terms (Fig. 3), such as some biological processes involved in response to stimulus (GO:0050896), and developmental processes (e.g.: post embryonic development, GO:0009791). Many up-regulated genes appeared to be transcription factors (e.g.: RAV1, AP2/ERF, WRKY33) or protein regulators (e.g.: RAPTOR1B) and storage proteins (e.g.: LEA D29) involved in post-embryonic development.

Resistance genes and miRNA target

Protein sequences of 112 reference RGAs (from RGDB) were used to perform BLASTp searches against chestnut unigenes. A total of 1,444 unigenes (Online Resource file 4), showing homology to 82 univoque proteins out of the 112 encoded reference RGAs, were identified. Their putative functions (Fig. 4) were identified and 32 (39.0%) belonged to the CC-NB-LRR type (CNL), 9 (10.9%) belonged to the TIR-NB-LRR type (TNL), 4 (4.9%) belonged to the NB-LRR type (NL), 12 (14.6%) belonged to the receptor-like protein type (RLP), 10 (12.2%) belonged to the receptor-like kinase class (RLK), and 1 (1.2%) belonged to the kinase-resistance related type. Some of these proteins appeared as regulated by the presence of the cynipid (Fig. 2).

The assembled transcriptomes of both cultivars were scanned for the presence of recognition sites for known plant miRNAs (miRNA targets) and the results are provided in Table 3 as well as in Online Resource file 5 **(‘**Bouche de Bétizac’) and in Online Resource file 6 (‘Madonna’). The contigs/scaffolds of ‘Bouche de Bétizac’ showed, in total, target annealing sites for 249 miRNAs, located in 1135 transcripts (Online Resource table 3). A total of 185 targets belonged to the ath-miR5021 family; a total of 146 targets belonged to the ath-miR5658; a total of 139 targets belonged to the ath-miR414 family. ReviGO analysis (Fig. 5) showed some GO enrichments for miRNA targets transcripts (Online Resource table 4), particularly for the categories: nitrogen compound metabolic process (GO:0006807), nucleobase-containing compound metabolic process (GO:0006139), developmental process (GO:0032502), shoot system development (GO:0048367), phyllome development (GO:0048827) and response to stimulus (GO:50896). They included 21 genes: three were involved in the trichome development, four were genes involved in stress related phenomena and fourteen were miRNA target related to tissue development (leaf, shoot, inflorescence). The latter group contained three structural proteins, five enzymes and six transcriptional factors. Some of these proteins **(**Online Resource table 1) were significantly regulated in the resistant cultivar in the presence of the cynipid.

SSR identification and primer design

A screening of the reference transcriptome resulted in the identification of 5,713 scaffolds containing 11,364 putative SSRs. About 14.9% of the unigenes contained an SSR (one SSR per 3.5 Kb) and the most abundant repeat motifs were mono-nucleotides (4,519; 39.8%), followed by tri-nucleotides (2,296; 20.2%), di-nucleotides (1,908; 16.8%), hexanucleotides (1,282; 11.3%), penta-nucleotides (719; 6.3%) and tetra-nucleotides (640; 5.6%). The most common di- and tri-nucleotide motifs were AG (1,449, 12.8%) and AAG (629; 5.5%). Frequencies and repeat numbers for the 20 most present SSR motifs are reported in Figure 6 and complete statistics are presented in Online Resource file 7). A batch analysis permitted the design of PCR primers for all the loci, leading to the generation of 7,176 putative markers (Online Resource file 8).

SNP mining

Considering the two cultivars ‘Bouche de Bétizac’ (as reference) and ‘Madonna’, 335,468 reliable SNPs/Indels (DP>10), across the two accessions, were detected. On the whole, 321,939 were SNPs and 13,529 were Indels and among SNPs, 206,015 were transitions, 115,515 are transversions (ratio=1.78). Since the assembled transcriptome is 36,094,445 bp long in 34,081 contigs, the average SNP frequency was calculated at 1⁄124 bp with a mean of 8.4 SNP/INDEL per contig.

SNP (inter/intra genotype) - The number of SNPs/Indels between the two accessions was 335,468. The number of SNPs in homozygous state between them was 25,154. The number of heterozygous loci was 232,578 in the ‘Bouche de Bétizac’ genotype, and 159,742 in the ‘Madonna’ genotype.

SNP markers - Considering only SNP variants, loci were classified into those expected to segregate in a 1:1 ratio ("testcross markers" AA x AB or AB x BB), and those in a 1:2:1 ratio ("intercross markers"; AB x AB). Testcross markers were 76,764, considering the ‘Bouche de Bétizac’ genotype over the ‘Madonna’ one and 149,600 considering ‘Madonna’ genotype over ‘Bouche de Bétizac’ one (Table 4). Overall testcross markers were 226,364. Intercross markers were 82,978 (24.7% of the total). Testcross and intercross SNPs were in all 309,343 (92.2% of the total).

Discussion

The first fully resistant to gall wasp cultivar found was the hybrid ‘Bouche de Bétizac’ that showed no symptoms both in orchard and under controlled conditions (Sartor et al., 2009). For this reason this cultivar was used for a transcriptomic approach, starting from the hypothesis that resistance may be due to a hypersensitive reaction in the bud tissues (Dini et al. 2012).

To conduct genetic dissection of traits involved in the insect-plant interaction and proceed to breeding practices, there is a need for a transcriptome/genome reference sequence. High-throughput RNA sequencing is a useful approach to obtaining a complete set of transcripts from species of interest. Because of the potential advantages of these technologies (high-throughput vs low costs), many transcriptomes from model/non-model species have been sequenced and assembled in the last years (over 2,100 papers in the period 2000-2018, ISI - Web of Science survey). This consolidated approach was used, in the present study, to reconstruct the transcriptome of *C. sativa* buds under biotic stress (i.e.: in the presence/absence of the chestnut gall wasp). In ‘Bouche de Bétizac’, a total of 34,081 unigenes were obtained by optimizing assembly procedures. The transcriptomes of both cultivars were properly assembled, and while the ‘Bouche de Bétizac’ unigenes set (belonging to the resistant cultivar) was selected for the functional characterization, the ‘Madonna’ one was just used for RNAseq data analysis, and provided as supplementary materials. The ‘Bouche de Bétizac’ assembly was evaluated for its completeness with CEGMA, Parra et al. 2007). This pipeline uses 248 Core Eukaryotic Genes (CEGs), which are highly conserved, present in low copy numbers in higher eukaryotes, to describe the gene space. Based on the average degree of conservation observed from each CEG, the CEGMA pipeline divides the CEGs into four groups (group 1 has the least conserved CEGs while group 4 has the most conserved CEGs). The alignment trends (Online Resource file 1) using the new *C. sativa* reference transcriptome (‘Bouche de Bétizac’, filtered) showed how the lack of complete alignments for less conserved ortholog groups (65%) is due to divergence, while partial alignments confirm an even representation of different ortholog groups, indicating a gene space coverage of about 90% (range 86-95 %). Blast2GO analysis annotated 27,345 (71.3%) unigenes containing a wide range of biological, cellular and molecular functions typical of a plant transcriptome resembling similar transcriptome assemblies (Garcia-Seco et al. 2015; Cardoso-Silva et al. 2014), involving all the physiological processes (i.e.: Biological Process), the majority of cellular compartments (i.e.: Cellular Component), and the functions of the proteins produced (i.e.: Molecular Function).

A differential gene expression analysis was conducted to highlight regulated genes emerging in two genetic contexts (cultivars susceptible/resistant to the cynipid) following the infestation with the wasp. The RNAseq analyses were conducted using bulks of buds to catch macroscopic variations between resistant and sensitive varieties in the presence of the pest. To the scope, genes were analyzed with the GFOLD suite, specifically implemented as a tool for studies with few or no replicates (Feng et al. 2012), as it generalizes the fold change by considering the posterior distribution of log fold change, such that each gene is assigned a reliable fold change. Considering a 1-fold variation, up to 557 genes showed to be regulated in the infested 'Bouche de Bétizac' transcriptome, while 2814 in the infested cultivar 'Madonna', most of which (75%) were down-regulated genes. This high difference in number of regulated genes was mostly expected since the phenotype of the sensitive buds in respect of the resistant ones appeared very different (Dini et al. 2012). As reviewed elsewhere (Schuman et al. 2016) plant respond to pests with a multilayer approach. Here, we observed, as expected, many and different "response to stimulus" up-regulated genes. Among them, some were involved in the likely recognition of specific elicitors and patterns of damage. Indeed, as many as sixty LRR proteins were observed to be regulated during the interaction between chestnut bud and cynipid; some other genes were implied with a "transcriptional regulator activity" role. Intriguingly, 16 WRKY and 6 ERF/AP2 genes were here observed as up regulated. Recently, those categories of transcription factors are reported to be involved in response to both aphid attack and *P. syringae* infection regulation, but it is known that some of them are up-regulated in an insect-specific manner (e.g.: WRKY-33, Barah et al. 2013). Dini et al. (2012) highlighted the occurrence of an HR in the resistant cultivar 'Bouche de Bétizac' as response to the cynipid infestation, resulting in cell and larvae death. This fact was here confirmed, since more than 100 genes appeared associated to "death" and "apoptosis" processes, including genes for HR response. Also the "vacuolar part" term appeared significantly enriched in up-regulated genes; they were prominently positively regulated membrane proteins coding for ionic channels and pumps, consistently with the role attributed to vacuolar proteins in plant immunity (Hatsugai 2015; Zhang et al. 2010). Some other genes involved in "post embryo development" were over-represented. For example, Contig_32550 is a putative homologue to RAV1 transcription factor, which has been suggested to be a negative regulator of growth and development. The regulation of RAV1 (and RAV2) may serve not only for immediate physiological responses, but also for developmental adaptation in response to the environmental stimuli, such as response to touch stimulus, happening during the cynipid growth in the gall infection.

During gall formation, many biochemical, physiological, and molecular changes occur in plant tissues requiring continuous activity of unknown stimuli (Harper et al. 2004). Studies on the insect-plant interaction carried out on cynipids of oak and rose showed three stages of gall formation: initiation, growth and maturation (Harper et al. 2004). The growth stage starts with cell proliferation, differentiation and hypertrophy due to stimuli from the cynipid gall wasps that are able to redirect host-plant development to form novel structures to protect and nourish the developing larvae. Comparisons between inner-gall and non-gall tissue protein signatures by Schönrogge et al. (2000) have identified a number of inner-gall proteins, such as a NAD-dependent formate dehydrogenase (NFD) and a biotin carboxyl carrier protein (BCCP), the latter being a subunit of a class II acetyl CoA-carboxylase (ACCase), involved in the production of triacylglycerol lipids. Further studies (Harper et al. 2004) showed that genes for this inner-gall putative BCCP reveal differential expression throughout gall development. Fluorescent in situ hybridization demonstrated many of the inner-gall cells to be polytenized. The expression of putative BCCP and the polytenization of the nuclei in gall cells are typical of nutritive, secretory cells, found in seeds and also in tapetal cells in pollen. Recent studies (Pawłowski et al. 2017) on protein patterns in healthy and gall tissues showed changes in abundance of 21 proteins. Interestingly, some functions (Online Resource Table 1) appeared here to be up-regulated in infested 'Bouche de Bétizac', such as some subunit of ATP synthase and many HSPs, while others showed to be down-regulated in Madonna infested galls, such as ascorbate peroxidase, actin and many stress-related and pathogenesis-related proteins (PRP).

Each contig was scanned for the presence of recognition sites for known plant miRNAs (miRNA targets). The latter analysis highlighted some findings, which deserves to be deeply discussed. microRNAs (miRNAs) were discovered in '90s and are now recognized as one of the major regulatory gene families in plants and in eukaryotes in general. They play important roles in a variety of biological phenomena, such as development and responses to abiotic and biotic stresses, by regulating complementary target transcripts. In particular, it was primarily recognized that miRNA activity results in gene expression repression impairing mRNA stability, by guiding mRNA degradation, at a protein synthesis initiation level, by its inhibition, or through degradation of a protein, via the binding of the 3′UTR of a target transcript. We highlighted target annealing sites for 249 miRNAs located in 1,135 transcripts and, interestingly, we observed enrichments for certain specific GO categories. The more intriguing were the ones involved in some developmental processes (e.g.: shoot/phyllome development), as well as the ones involved to the response to stimulus (Online Resource Table 3). The gFOLD analysis showed a trend of up-regulation in infested buds belonging to the susceptible cultivar ('Madonna') and thus, as expected, we spotted some enriched GO terms likely related to plant/insect interaction, which deserve to be discussed. A first example of regulated genes is referred to some miRNA targets (in contig_4038, contig_37578, contig_2740 and contig_15619), regulated by a unique miRNA (miR5658), which are involved in the energetic regulation/reprogramming of the cell in condition of stress and involved in defense responses, as well as in growth and development (Online Resource Table 4). Some other miRNA target genes were observed (Online Resource Table 4) and most of them were transcription factors (GRF7, ZFP8, DOT2, TCP3, ATHB-15) or generally involved in gene regulation (LHP1, YUC4, ARF) playing a role in root, shoot and flower development as factors most likely involved in the reshaping of the tissue towards the formation of a suitable gall. These enrichments are intriguing and will be the target for future analyses aimed at understanding the functions of the identified genes and thus providing useful tools to molecular breeders. Indeed, RNA interference technology involving siRNA and miRNA have emerged as an attractive tool used by plant biologists not only to decipher the plant function, but also to develop plants with improved and novel traits by the manipulation of both desirable and undesirable genes.

One of our target was the identification of SNP markers; however, transcriptomic data from NGS sequencing made it also possible the mining of microsatellite motifs. SSR markers are multi-allelic and are widely applied for genetic analyses, regardless of their cost for development and for implementation in throughput facilities. During the last years the exploitation of publicly available EST database, at first, and the explosion of NGS data production, secondarily, leaded to the identification of several thousands of new markers in virtually almost every plant species (Portis et al. 2016). The most abundant repeat motif were mononucleotides, followed by trinucleotides, dinucleotides and hexanucleotides. This seems not consistent with the observations in Poplar and Arabidopsis (Morgante et al. 2002), where tri-nucleotides are the most represented in transcriptomes; from our analyses most of the mononucleotide motifs could be represented by terminal polyA (this happens for over 3,500 sequences). A more realistic number of mononucleotidic repetition could be thus calculated to be about one thousand (8.8%). The most common di- and tri-nucleotide motifs were AG (1,449, 12.8%) and AAG (629; 5.5%) in accordance to the observation in a previous study, where AG and AAG were the predominant motifs (Stagel et al. 2008; Zeng et al. 2010; Portis et al. 2007; Morgante et al. 2002). About 14.9% of the unigenes contained an SSR (one SSR per 3.5 Kb), which is a value comparable to the success rate recorded from other fruit crops (Rai et al. 2015).

SNP frequencies in the *Castanea sativa* transcriptome appear to be comparable to that found in the outbred, highly heterozygous *Cynara cardunculus* transcriptome (Scaglione et al. 2012) and among *Citrus* species ESTs (Jiang et al. 2010). Overall, 335k SNP/indels were identified and the number of loci in heterozygous state was very high (232.578 in 'Bouche de Bétizac', and 159.742 in 'Madonna'). This is expected in *Castanea sativa* that being an outbreeding crop, presents a high level of heterozygosis. Genetics of *C. sativa* has been limited studied so far, and very few mapping populations are available. The testcross and intercross SNP markers identified (in all 309.343 markers) will be suitable information for mapping purposes using F_1 progeny in a 2 way pseudo-test cross approach. The core set of SNPs will be pivotal to setup SNP arrays or to design custom SPET (Single Primer Enrichment Technology, Nugen) assays, as example of targeted resequencing approach, through the selection of known polymorphic regions in gene space.

Overall, the *de novo* assembly of the transcriptome chestnut, combined with extensive homology analyses, yielded a number of contigs comparable to those found in literature for similar RNAseq experiments. The work is still at a preliminary stage, but it was successful in classifying contigs, and once data will be readily available to the international scientific community, they will guide a better understanding of the interaction chestnut-gall wasp. Altogether, the corpus of produced information will lead to investigate the different mechanisms of resistance against cynipid, and to address breeding strategies towards resistant cultivars. To the scope, the bioinformatics pipeline adopted enabled identification of a large set of SSR/SNP markers for practical applications in breeding programs and provenance/pedigree tracking. We believe that the availability of these transcriptome data for *C. sativa* will contribute to understand the genetic basis of the resistance to gall wasp and meet the informational needs for molecular genetic studies of this species and its relatives.

Data Archiving Statement

Sequences have been submitted to NCBI's Short Read Archive (SRA).
SRA accession: PRJNA509688

Compliance with Ethical Standards

This article does not contain any studied with human participants or animals performed by any of the authors.

Conflict of Interest

The authors declare that they have no conflict of interest and in particular:

Author 1 (Alberto Acquadro) declare that he has no conflict of interest

Author 2 (Daniela Torello Marinoni) declare that she has no conflict of interest

Author 3 (Chiara Sartor) declare that she has no conflict of interest

Author 4 (Francesca Dini) declare that she has no conflict of interest

Author 5 (Matteo Macchio) declare that he has no conflict of interest

Author 6 (Roberto Botta) declare that he has no conflict of interest

References

Anagnostakis S, Clark S, McNab H (2009) Preliminary report on the segregation of resistance in Chestnut to infestation by oriental Chestnut Gall Wasp. Acta Hort 815:33-35

Barah P, Winge P, Kusnierczyk A, Tran DH, Bones AM. (2013) Molecular Signatures in *Arabidopsis thaliana* in Response to Insect Attack and Bacterial Infection. PLoS ONE 8(3):e58987. https://doi.org/10.1371/journal.pone.0058987

Bradnam KR, Fass JN, Alexandrov A et al (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species, GigaScience 2(1):1–31. https://doi.org/10.1186/2047-217X-2-10

Brussino G, Bosio G, Baudino M, Giordano R, Ramello F, Melika G (2002) Pericoloso insetto esotico per il castagno europeo. L'informatore Agrario 37:59-61

Cardoso-Silva CB, Costa EA, Mancini MC et al (2014) *De Novo* assembly and transcriptome analysis of contrasting sugarcane varieties. PLoS ONE 9(2): e88462. https://doi.org/10.1371/journal.pone.0088462

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Research 39 (Issue suppl_2):W155–W159. https://doi.org/10.1093/nar/gkr319

Dini F, Sartor C, Botta R (2012) Detection of a hypersensitive reaction in the chestnut hybrid 'Bouche de Bétizac' infested by *Dryocosmus kuriphilus* Yasumatsu. Plant Physiology and Biochemistry 60:67-73

Dixon WN, Burns RE, Stange LA (1986) Oriental chestnut gall wasp. *Dryocosmus kuriphilus.* Plant Industry. Florida Department of Agriculture and Consumer Service, Gainsville (US). Div. Entomol. Circ., 1-2, n°287.

FAOSTAT (2018) Food and Agriculture Organization of the United Nations statistics database, Rome.

Feng J1, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y (2012) GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics 28(21):2782-8. https;//doi.org/ 10.1093/bioinformatics/bts515.

Ferracini C, Ferrari E, Pontini M, Saladini MA, Alma A (2018) Effectiveness of *Torymus sinensis*: a successful long-term control of the Asian chestnut gall wasp in Italy. Journal of Pest Science. https://doi.org/10.1007/s10340-018-0989-6

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J., Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44 (D1):D279-D285. https://doi.org/10.1093/nar/gkv1344

Garcia-Seco D, Zhang Y, Gutierrez-Mañero FJ, Martin C, Ramos-Solano B (2015) RNA-Seq analysis and transcriptome assembly for blackberry (*Rubus* sp. Var. Lochness) fruit. BMC genomics 16:5 https://doi.org/10.1186/s12864-014-1198-1

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Research. 36 (suppl_1): D154–D158. http://dx.doi.org/10.1093/nar/gkm952

Gross P, Price PW (1988) Plant Influences on Parasitism of Two Leafminers: A Test of Enemy-Free Space. Ecology 69(5):1506-1516. https://doi.org/10.2307/1941648

Harper LJ, Schönrogge K, Lim KY, Francis P, Lichtenstein CP (2004) Cynipid galls: insect-induced modifications of plant development create novel plant organs. Plant, Cell and Environment 27 (3):327-335

Hatsugai N, Yamada K, Goto-Yamada S, Hara-Nishimura I (2015) Vacuolar processing enzyme in plant programmed cell death. Frontiers in Plant Science 6:234. https://doi.org/10.3389/fpls.2015.00234

Huang MY, Huang WD, Chou HM, Chen CC, Chen PJ, Chang YT, and Yang CM (2015) Structural, biochemical, and physiological characterization of photosynthesis in leaf-derived cup-shaped galls on *Litsea acuminata*. BMC Plant Biol 15:61. https://doi.org/10.1186/s12870-015-0446-0

Inbar M, Izhaki I, Koplovich A, Lupo I, Silanikove N, Glasser T, Gerchman Y, Perevolotsky A, Lev-Yadun S (2010) Why do many galls have conspicuous colors? A new hypothesis. Arthropod-Plant Interactions 4:1-6 https://doi.org 10.1007/s11829-009-9082-7

Jiang D, Ye QL, Wang F, Cao L (2010) The mining of citrus EST-SNP and its application in cultivar discrimination. Agricultural Sciences in China 9(2):179-190. https://doi.org/10.1016/S1671-2927(09)60082-1

Kato K, Hijii N (1997) Effects of gall formation by *Dryocosmus kuriphilus* Yasumatsu (Hym., Cynipidae) on the growth of chestnut trees. J Appl Entomol 121:9-15

Kofler R, Schlötterer C, Lelley T (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23(13):1683–1685

Manoj KR, Shekhawat NS (2015) Genomic resources in fruit plants: an assessment of current status Critical Reviews in Biotechnology 35(4):438-447. https://doi.org/10.3109/07388551.2014.898127

Marinoni D, Akkak A, Bounous G, Edwards KJ, Botta R (2003) Development and characterization of microsatellite markers in *Castanea sativa* (Mill.). Molecular Breeding 11:127-136

Min XJ, Butler G, Storms R, Tsang A (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Research 33(suppl_2):W677–W680. https://doi.org/10.1093/nar/gki394

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with non repetitive DNA in plant genomes. Nature Genetics. 30:194–200

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, Bioinformatics 23(9):1061–1067

Pawłowski TA, Staszak AM, Karolewski P, Giertych MJ (2017) Plant development eprogramming by cynipid gall wasp: proteomic analysis. Acta Physiol Plant 39:114. https://doi.org/10.1007/s11738-017-2414-9

Petricka JJ, Clay NK, Nelson TM (2008) Vein patterning screens and the defectively organized tributaries mutants in *Arabidopsis thaliana*. https://doi.org/10.1111/j.1365-313X.2008.03595.x

Picciau L, Ferracini C, Alma A (2017) Reproductive traits in *Torymus sinensis*, biocontrol agent of the asian chestnut gall wasp: implications for biological control success. Bulletin of Insectology 70(1):49-56

Portis E, Nagy I, Sasvari Z, Stagel A, Barchi L, Lanteri S (2007) The design of Capsicum spp. SSR assays via analysis of in silico DNA sequence, and their potential utility for genetic mapping. Plant Science 172:640–648. https://doi.org/10.1016/j.plantsci.2006.11.016

Portis E, Portis F, Valente L, Moglia A, Barchi L, Lanteri S, Acquadro A (2016) A Genome-Wide Survey of the Microsatellite Content of the Globe Artichoke Genome and the Development of a Web-Based Database. PLoS ONE 11(9):e016284. http://dx.doi.org/10.1371/journal.pone.0162841

Quacchia A, Moriya S, Bosio G, Scapin G, Alma A (2008) Rearing, release and settlement prospect in Italy of *Torymus sinensis*, the biological control agent of the chestnut gall wasp *Dryocosmus kurip*hilus. BioControl 53: 829–839

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology 132:365-386

Sanseverino W, Hermoso A, D'Alessandro R, Vlasova A, Andolfo G, Frusciante L, Lowy E, Roma G, Ercolano MR (2012) PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. Nucleic Acids Research 41(D1):D1167–D1171. https://doi.org/10.1093/nar/gks1183

Sartor C, Torello Marinoni D, Quacchia A, Botta R (2012) Quick detection of *Dryocosmus kuriphilus* Yasumatsu (Hymenoptera: Cynipidae) in chestnut dormant buds by nested PCR. Bullettin of Entomological Research 102 (3):367-371

Sartor C, Dini F, Torello Marinoni D, Mellano MG, Beccaro GL, Alma A, Quacchia A, Botta R (2015) Impact of the Asian wasp *Dryocosmus kuriphilus* (Yasumatsu) on cultivated chestnut: Yield loss and cultivar susceptibility. Scientia Horticulturae 197:454-460. https://doi.org/10.1016/j.scienta.2015.10.004

Sartor C, Botta R, Mellano MG, Beccaro GL, Bounous G, Torello Marinoni D, Quacchia A, Alma A (2009) Evaluation of susceptibility to *Dryocosmus kuriphilus* Yasumatsu (Hymenoptera: Cynipidae) in *Castanea sativa* Miller and in hybrid cultivars. Acta Hort 815:289-298

Scaglione D, Lanteri S, Acquadro A, Lai Z, Knapp SJ, Rieseberg L, Portis E (2012) Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. Plant biotechnology journal 10(8):956-969

Schönrogge K, Harper LJ, Lichtenstein CP (2000) The protein content of tissues in cynipid galls (hymenoptera: Cynipidae): similarities between cynipid galls and seeds. Plant Cell and Environment 23:215-222

Schuman MC, Baldwin IT (2016) The layers of plant responses to insect herbivores. Annual Review of Entomology 61:373-394. https://doi.org/10.1146/annurev-ento-010715-023851

Serrazina S, Santos C, Machado H, Pesquita C, Vicentini R, Pais MS et al (2015) *Castanea* root transcriptome in response to *Phytophthora cinnamomi* challenge. Tree Genet Genomes 11:1–19.

Shimura I (1972a) Breeding of chestnut varieties resistant to chestnut gall wasp, *Dryocosmus kuriphilus* Yasumatsu. Japan Agricultural Research Quarterly 6:224-230

Shimura I (1972b) Studies on the breeding of chestnut, *Castanea* spp. II. Parasitic variation in the chestnut gall wasp, *Dryocosmus kuriphilus* Yasumatsu. Bulletin of the Horticultural Research, Station A11:1-13

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: A parallel assembler for short read sequence data. Genome Research 19(6):1117–1123

Singh N, Srivastava S, Shasany AK, Sharma A (2016) Identification of miRNAs and their targets involved in the secondary metabolic pathways of *Mentha* spp. Comput Biol Chem 64:154-162. https://doi.org/10.1016/j.compbiolchem.2016.06.004

Stàgel A, Portis E, Toppino L, Rotino GL, Lanteri S (2008) Gene-based microsatellite development for mapping and phylogeny studies in eggplant. BMC Genomics 9:357. https://doi.org/10.1186/1471-2164-9-357

Steinkellner H, Fluch S, Turetschek E, Lexer C, Streiff R, Kremer A, Burg K, Glossl J (1997) Identification and characterization of (GA/GT)n microsatellite loci from *Quercus petraea*. Plant Mol Biol 33:1093-1096

Torello Marinoni D, Akkak A, Beltramo C, Guaraldo P, Boccacci P, Bounous G, Ferrara AM, Ebone A, Viotto E, Botta R (2013) Genetic and morphological characterization of chestnut (*Castanea sativa* Mill.) germplasm in Piedmont (north-western Italy). Tree Genetics & Genomes 9(4):1017-1030. https://doi.org/10.1007/s11295-013-0613-0

Torello Marinoni D, Nishio, Portis E, Valentini N, Sartor C, Dini F, Ruffa P, Oglietti S, Martino G, Akkak A, Botta R (2017) Development of a genetic linkage map for molecular breeding of chestnut. Acta Hort https://doi.org/10.17660/ActaHortic.2018.1220.4

Zeng S, Xiao G, Guo J, Fei Z, Xu Y, Roe BA, Wang Y (2010) Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. BMC Genomics 11:94. https://doi.org/10.1186/1471-2164-11-94

Zhang H, Zheng X, Zhang Z (2010) The role of vacuolar processing enzymes in plant immunity. Plant Signaling & Behavior 5(12):1565-1567. https://doi.org/10.4161/psb.5.12.13809

Wen C, Cheng Q, Zhao L, Mao A, Yang J, Yu S, Weng Y, Xu Y (2016) Identification and characterisation of Dof transcription factors in the cucumber genome Sci Reports 6: 23072. https://doi.org/10.1038/srep23072

TABLES

Table 1 Sequencing results and statistics after polishing procedures.

Cultivar	Raw reads (M)	filtered/trimmed reads (M)	Amount of sequence (Gb)
‘Bouche de Bétizac’, infested	59.77	49.02	9.8
‘Bouche de Bétizac’, not infested	110.15	90.93	18.2
‘Bouche de Bétizac’ (total)	169.92	139.95	28.0
‘Madonna’ infested	75.21	62.74	12.6
‘Madonna’ not infested	115.81	95.55	19.1
‘Madonna’ (total)	191.02	158.29	31.7
Total	**360.94**	**298.24**	**59.7**

Table 2 Statistics for the *de novo* assembled transcriptomes. Values were calculated using the Perl script assemblathon.pl (Bradnam et al. 2013).

Characteristics	‘Bouche de Bétizac’	‘Bouche de Bétizac’ filtered	‘Madonna’
Number of scaffolds	39,365	34,081	30,605
Total size of scaffolds	40,621,562 (39	36,094,445	31,159,190
Longest scaffold	13,723	13,723	5,405
Shortest scaffold	500	500	500
Number of scaffolds > 1K nt	14,481 (36.8%)	13,112 (38.5%)	10,782 (35.2%)
Number of scaffolds > 10K nt	4 (0.0%)	4 (0.0%)	0
Number of scaffolds > 100K nt	0	0	0
Mean scaffold size	1,032	1,059	1,018
Median scaffold size	818	835	801
N50 scaffold length	1,142	1,191	1,12
L50 scaffold count	11,44	9,747	8,886
% GC	41.76	41.50	41.44

Table 3 Abundance of putative miRNA annealing sites in the *Castanea sativa* transcriptome. miRNA families occurring fewer than four times were incorporated into the category labelled ‘other’.

miRNA family	No. of targets	miRNA family	No. of targets
ath-miR5021	185	ath-miR1886.1	5
ath-miR5658	146	ath-miR1886.2	5
ath-miR414	139	ath-miR4243	5
ath-miR838	15	ath-miR5641	5
ath-miR854a	13	ath-miR156i	5
ath-miR854b	13	ath-miR156a	4
ath-miR854c	13	ath-miR156b	4
ath-miR854d	13	ath-miR156c	4
ath-miR854e	13	ath-miR156d	4
ath-miR5653	13	ath-miR156e	4
ath-miR865-3p	11	ath-miR156f	4
ath-miR834	10	ath-miR171b	4
ath-miR400	9	ath-miR171c	4
ath-miR396a	8	ath-miR395a	4
ath-miR396b	8	ath-miR395d	4
ath-miR5648-5p	8	ath-miR395e	4
ath-miR4221	7	ath-miR397a	4
ath-miR157d	6	ath-miR397b	4
ath-miR163	6	ath-miR407	4
ath-miR837-5p	6	ath-miR156h	4
ath-miR156j	6	ath-miR415	4
ath-miR5654-3p	6	ath-miR447a.2-3p	4
ath-miR157a	5	ath-miR472	4
ath-miR157b	5	ath-miR831	4
ath-miR157c	5	ath-miR861-5p	4
ath-miR395b	5	ath-miR773b-3p	4
ath-miR395c	5	ath-miR5016	4
ath-miR395f	5	ath-miR5652	4
ath-miR773a	5	ath-miR5998a	4
ath-miR830-3p	5	ath-miR5998b	4
ath-miR835-5p	5	Others	305
ath-miR866-3p	5		

Table 4 | Testcross and intercross markers evaluation. Data represents SNP sites having sequence information for each of the two samples analyzed.

SNP markers	‘B. de Bétizac’ vs ‘Madonna’	‘Madonna’ vs ‘B. de Bétizac’	In common
Putative testcross	76,764	149,600	-
Common intercross	-	-	82,978

FIGURE CAPTIONS

Fig. 1 The distribution of the scaffold length in the chestnut transcriptome Assembled scaffold size: the length interval measured were set to 500 bp. Blue bars represent 'Madonna' scaffold; grey bars represent 'Bouche de Betizac' scaffold.

Fig. 2 Differential gene expression in chestnut buds in the presence of the cynipid in resistant/susceptible cultivars. (a) Histogram representing variation of genes (> 1-fold) after inoculation of the cynipid in the two cultivars, considering four possible comparisons (BI vs BNI; BI vs MI; MI vs MNI; BNI vs MNI). **(b)** Venn diagram intersection of the four transcriptomes comparison. A white square highlights the genes commonly regulated in "BI vs BNI" and "BI vs MI" and not regulated in "BNI vs MNI" and "MI vs MNI".

Fig. 3. Enriched GO terms in terms of biological processes, cellular components and molecular functions in up-regulated and down-regulated genes of 'Bouche de Bétizac' over 'Madonna' transcriptomes; X-axis and Y-axis are expressed as semantic space, using color (log10 p-value) and size variation (log size).

Fig. 4 RGA analysis in the chestnut transcriptome (a) Representation of putative and univocal RGAs in the 34,081 analysed unigenes. **(b)** Analysis of the different categories of RGAs in the 82 univocal chestnut RGAs. The CNL class comprises resistance genes encoding proteins with at least a coiled-coil domain, a nucleotide binding site and a leucine-rich repeat (CC-NB-LRR); the TNL class includes those with a Toll-interleukin receptor-like domain, a nucleotide binding site and a leucine-rich repeat (TIR-NB-LRR); the RLP class, acronym for receptor-like protein, groups those with a receptor serine– threonine kinase-like domain, and an extracellular leucine- rich repeat (ser/thr-LRR); the RLK class contains those with a kinase domain, and an extracellular leucine-rich repeat (Kin-LRR); the 'Others' class includes all other genes which have been described as conferring resistance through different molecular mechanisms, e.g. Mlo and Asc-1; the kinase contain a kinase domain involved in resistance process.

Fig. 5 miRNA target enrichment analysis in the chestnut transcriptome (a) Representation GO category enriched in the miRNA target transcriptome subset. X axis is expressed in log10 p-value, Y axis is expressed as semantic space scale.

Fig. 6 SSR most represented motifs

SUPPLEMENTARY MATERIAL CAPTIONS

ONLINE RESOURCE FILES

Online Resource file 1 The assembled transcriptome of the cultivar ‘Bouche de Bétizac’, filtered for contaminants deriving from pests and fungi

Online Resource file 2 The assembled transcriptome of the cultivar ‘Madonna’

Online Resource file 3 CEGMA pipeline results on the *C. sativa* transcriptome Prots = number of 248 ultra-conserved CEGs present in genome; %Completeness = percentage of 248 ultra-conserved CEGs present; Total = total number of CEGs present including putative orthologs; Average = average number of orthologs per CEG; %Ortho = percentage of detected CEGS that have more than 1 ortholog

Online Resource file 4 Unigenes, showing homology to 82 univoque proteins out of the 112 encoded reference RGAs, identified

Online Resources file 5 miRNA targets in cultivar ‘Bouche de Bétizac’

Online Resources file 6 miRNA targets in cultivar ‘Madonna’

Online Resource file 7 Complete statistics for the identified SSR loci

Online Resource file 8 primers designed for the selected SSR loci

ONLINE RESOURCE TABLE

Online Resource table 1 List of genes regulated in infested ‘Bouche de Bétizac’ and ‘Madonna transcriptomes

Online Resource table 2 List of the genes commonly regulated in “BI vs BNI” and “BI vs MI” (Venn intersection, Fig. 2C) and not regulated in “BNI vs MNI” and “MI vs MNI”

Online Resource table 3 Genes showing enriched GO-terms among the putative miRNA target transcripts

Online Resource table 4 The over-representation of GO-terms among the putative miRNA target transcripts. P-value (<0.01) was used to assess statistical significance. AgriGO was used to obtain GO terms from presumptive miRNA target and ReviGO was used to evaluate enriched GO terms

ONLINE RESOURCE FIGURE

Online Resource Figure 1 Annotation and categorization analysis of the chestnut transcriptome. (a) Blast2GO results. **(b)** Species more represented in the blast analysis and top blast hits. **(c)** Chart giving the distribution of the number of annotations (GO-terms) retrieved from the different source databases (e.g. UniProt, PDB, TAIR)