Chestnut Transcriptome NG Sequencing: a New Tool to Investigate Gall Wasp Response

(Article begins on next page)

Chestnut Transcriptome NG Sequencing: a New Tool to Investigate Gall Wasp Response

F. Dini, C. Sartor, D. Torello Marinoni and R. Botta

DISAFA, Dipartimento di Scienze Agrarie, Forestali e Alimentari, Università degli Studi di Torino, Via Leonardo da Vinci 44, 10095 Grugliasco (Italy)

Keywords: RNA-Seq, Chestnut, Resistance, Gall wasp

Abstact

The *Dryocosmus kuriphilus* susceptibility in almost all chestnut cultivars and the absence of infestation symptoms in the cultivar 'Bouche de Bétizac' led us to investigate in depth plant-pest molecular interaction. For this reason, in 2012, the study of the gall wasp response in chestnut shifted from classical methodologies to an innovative approach potentially able to provide a large amount of information on a species poorly known from a genetic point of view.

Next Generation Sequencing platforms are able to realize a massive sequencing of DNA molecules spatially separated in a flow cell. This strategy represents a radical change in comparison with method described by Sanger sequencing, and allows to sequence hundreds of Mbp (million base pairs) to Gbp (billion base pairs) of DNA in a single analytical run. RNA-Seq allows to map and quantify the transcripts present in biological samples. Our approach was to use the Illumina technique. The experimental design was set to compare the genes activated at sprouting in susceptible and resistant chestnut cultivars in the presence and absence of infestation. We collected buds at different stages in order to identify the genes involved in plant-insect interaction at the onset of gall formation (susceptible genotype) or at the time of resistant response causing larva death. The sequencing of the cDNA libraries was carried out by BMR genomics (Padova, Italy). The reference chestnut transcriptome was assembled and included about 1,000,000 contig sequences. Since this is an extremely large amount of data, a cutoff at 500 bp was applied to the library obtaining 40,000 contigs that were used as a reference in the RNASeq analysis. The reads produced by sequencing the 4 treatments were mapped against the reference library. Reference contig identification, currently in progress, is being carried out by the program Blast2GO. This approach yielded an initial transcriptome assembly and will provide transcript sequences to the research community to facilitate further studies.

INTRODUCTION

The experiments with NGS platforms generate an unprecedented amount of information and this is a big challenge for data management, storage and analysis (Pop et al 2008).

Alignment and assembly, however, are more difficult for data obtained through NGS platforms compared to those obtained by Sanger method and require an adequate number of reads overlapping each other. A not adequate coverage may lead to a failure detection of a nucleotide variation, giving rise to false negatives for heterozygous samples (Bentley *et al.*, 2008). The process of de novo assembly leads to the formation of sets of contiguous reads (contigs), shorter and more numerous.

The platform specialized for sequencing short fragments ("short reads") of nucleic acids is the Illumina/Solexa: the major advantage of this technology is the paired ends sequencing of the cDNA molecules. This possibility provides information that facilitates the alignment and the "the novo" assembly, especially of "short reads" (Campbell *et al.*, 2008).

The analysis of mRNA with the approach of "RNA-Seq", based on NGS technology, allows to map and quantify all transcripts present in biological samples. The RNA-Seq shows several advantages compared to gene expression arrays (Wang *et al.*, 2009) that allow to analyze only

known genomic sequences; RNA-Seq let to sequence totally unknown transcripts and to derive the expression levels from the total number of reads showing a performance improvement in the quantitative identification of transcribed products both at high and low levels.

Since 2004, Piedmont Region (North Western Italy) has been fighting against *Dryocosmus kuriphilus* (Yasumatsu), a dangerous wasp that can affect the whole chestnut ecosystem. The interruption of the plant growth and the reduction of fruiting for chestnuts growers can be translated in yield losses of up to 50-70%. High infestations can cause, although rarely, plant death.

The wasp dangerousness is related to its ability to hide for a long time because infested buds, during vegetative rest, show no symptoms that could lead to suppose the presence of the early larval stages inside them. Moreover, its thelytokous parthenogenesis cause an exponential population increase in a short time. The typical symptoms of the presence of *D. kuriphilus* are represented by the galls, round green and reddish formations which are formed within 2-3 weeks on the shoot in spring, following a reaction of the plant to the wasp larvae presence. The damage that galls can cause involve directly leaves and shoots and indirectly the whole biomass; leaf surface is reduced, yellowing occurs in advance and the amount of vegetative buds is, every year, lower and lower (Kato and Hijii, 1997).

The plant resistance was studied in 'Bouche de Bétizac' (Sartor *et al.*, 2009; Dini *et al.*, 2012) demonstrating by histological observations the occurrence of a hypersensitive reaction that kills larvae in the early stages of budburst.

To study the plant response and understand which factors can lead the plant to develop or not the gall, we sequenced the mRNAs that the plant and the wasp transcribed during the early stages of the interaction. In this paper the first results are presented.

MATERIALS AND METHODS
Chestnut Samples Collection and RNA Preparation

In 2011, infested and not infested buds from cultivar 'Madonna' (*C. sativa*) and the hybrid 'Bouche de Bétizac' (*C. sativa* 'Bouche Rouge' X *C. crenata* CA04), were collected from single plants at different times of budburst from April 21st to May 12th. The harvest was carried out once a week in order to gather material representative of the different stages of bud sprouting and be able to observe the onset of the defensive response. The areas of sampling were highly infested by the cynipid and plant identity was checked by DNA analysis. The buds for the RNA extraction were frozen in liquid nitrogen, immediately after being detached from the branch, and kept at -80°C until use.

Nucleic acids extraction was performed from single buds following Dini et al. (2012).

After the validation of cynipid presence/absence by diagnostic PCR on bud DNA (Sartor *et al*. 2012), the total RNA extracted from single buds of the 4 different stages of bud development was pooled to have a representative sample. The resulting four samples (2 cultivars, infested and not infested) of total RNA were then shipped to Evrogen (Moscow, Russia) which processed them in order to isolate the mRNA and obtain 4 cDNA library. The sequencing was performed by BMR genomics (Padova, Italy) on a pool of cDNAs obtained by mixing equimolar amounts of the differently labeled cDNA libraries from the 4 treatments.

RNA-Seq

The protocol for Illumina sequencing was carried out by BMR genomics (Padova, Italy) by processing the cDNA samples according to the protocol TruSeq DNA.

The software chosen for the assembly of the reads was Newbler while CLC Genomics Workbench was used for analysis RNAseq. The Illumina reads were filtered and separated on the basis of the sequence tag (INDEX). Reads sizing less than 15 bases were deleted.

The package Blast2GO (Conesa *et al.*, 2005) was used to predict gene ontology (GO) terms for the contigs by assigning functional classifications (Gene Ontology Consortium, 2000) and potential properties of gene products to the contigs.

RESULTS AND DISCUSSION

The Illumina sequencing has resulted in 360,952,440 reads with an average length of 101 bp. Table 1 lists all the cleaning operations made on reads and its data output.

This material was then assembled for the realization of the reference library and resulting in total of about 1,000,000 contigs (exons).

We choose to apply to this library a cutoff of 500 bp both because it is an extremely large amount of data and in order to have reliable blast results. The 40,000 contigs of length equal to or greater than 500 were used as a reference in the analysis RNASeq. The results (not shown) are then displayed in a table reporting: value of expression (FPKM), length reference sequence, number of unique reads aligned on the reference, and total number of unique reads aligned and not aligned.

Functional classifications were performed on these 40,000 contigs using GO analysis to survey, categorize, and define the potential properties of gene products with respect to their predicted biological contexts. We used the program Blast2GO, a tool for assigning GO terms to unknown sequences. This resulted in GO functional classifications for 22.692 contigs (56,73 %), considering the scarsity of chestnut data in genebanks it was definitely a good result (Fig. 1). Among the transcript contigs classified by Blast2GO, 28.48% were assigned to "cellular component" ontology, 34.09% were assigned to "biological process" ontology, and 34.15% to "molecular function" ontology, with only 3,29% lacking assignment to GO classifications.

A first analysis conducted on all annotated sequences revealed the presence of an interesting group involved in the defense against pathogens shown in the combined graph of biological processes in Figure 3. The relationships between the categories are very complex and the numbers of sequences to be analyzed are quite high, for example, in the sole category "regulation of immune response" 304 contigs were identified.

CONCLUSIONS

The study carried out yielded a number of contigs comparable to those found in literature for similar RNAseq experiments and was successfull in classifying contigs according to GO. The work is still at a preliminary stage and it will be interesting to continue the analysis of the sequences to identify differentially expressed genes for better understanding the interaction chestnut-gall wasp. Moreover, these data of great interest will soon be available to the international scientific community.

Obtained sequences will also be useful to investigate the different mechanisms of resistance against cynipid in resistant *C. sativa* genotypes recently selected, and to identify genes involved in agronomical traits.

Literature Cited

Bentley D.R., Balasubramanian S. and Swerdlow H.P. 2008. Accurate whole human genome sequencing using reversible. Nature 456:53-59

Campbell P.J., Stephens P.J. and Pleasance E.D. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. 40:722-729.

Conesa A., Gotz S., Garcia-Gomez J.M., Terol J., Talon M. and Robles M. 2005. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676

Dini F., Sartor C. and 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, ISSN: 0981-9428, doi: 10.1016/j.plaphy.2012.07.023

Kato K. and 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.

Sartor C., Botta R., Mellano M.G., Beccaro G.L., Bounous G., Torello Marinoni D., Quacchia A. and Alma A. 2009. Evaluation of susceptibility to *Dryocosmus kuriphilus* Yasumatsu (Hymenoptera: Cynipidae) in Castanea sativa Miller and in hybrid cultivars. In: International Workshop on Chestnut Management in Mediterranean Countries - Problems and Prospects. Acta Horticulturae, 815:289-297, ISSN: 0567-7572

Sartor C., Torello Marinoni D., Quacchia A. and Botta R. 2012. Quick detection of *Dryocosmus kuriphilus* Yasumatsu (Hymenoptera: Cynipidae) in chestnut dormant buds by nested PCR. Bulletin Of Entomological Research, 102:367-371, ISSN: 0007-4853, doi: 10.1017/S0007485311000812

Wang Z., Gerstein M. and Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10:57-63.

Tables

Table 1. Report of the Illumina sequencing for the four treatment: Bouche de Bétizac infested and not Madonna infested and not; the table shows the output from the three different stages of raw data cleaning.

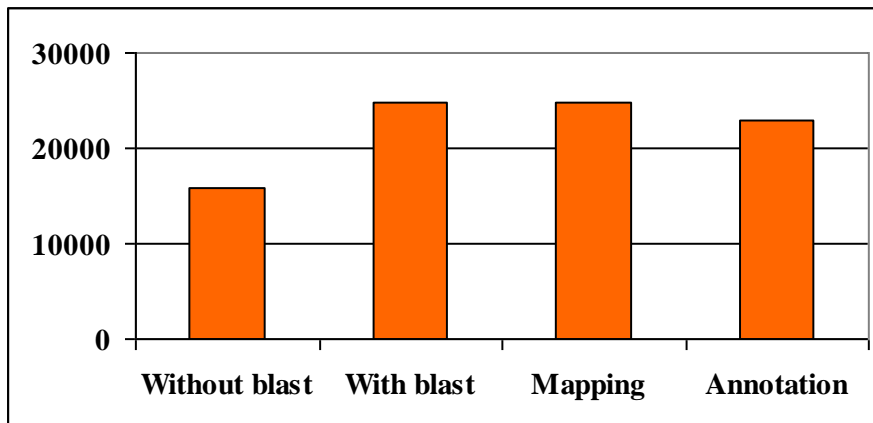| Treatment	n° reads	n° reads filtered	n° reads post-trimming
Bouche de Bétizac infested	59.773.686	49.052.146	49.025.350
Bouche de Bétizac not infested	110.150.508	91.004.052	90.928.827
Madonna infested	75.214.958	62.772.968	62.743.383
Madonna not infested	115.813.288	95.603.602	95.552.469

Figures

Fig. 1. Sequences with and without a match in database and accuracy of identification in subsequent steps (mapping, and annotation).
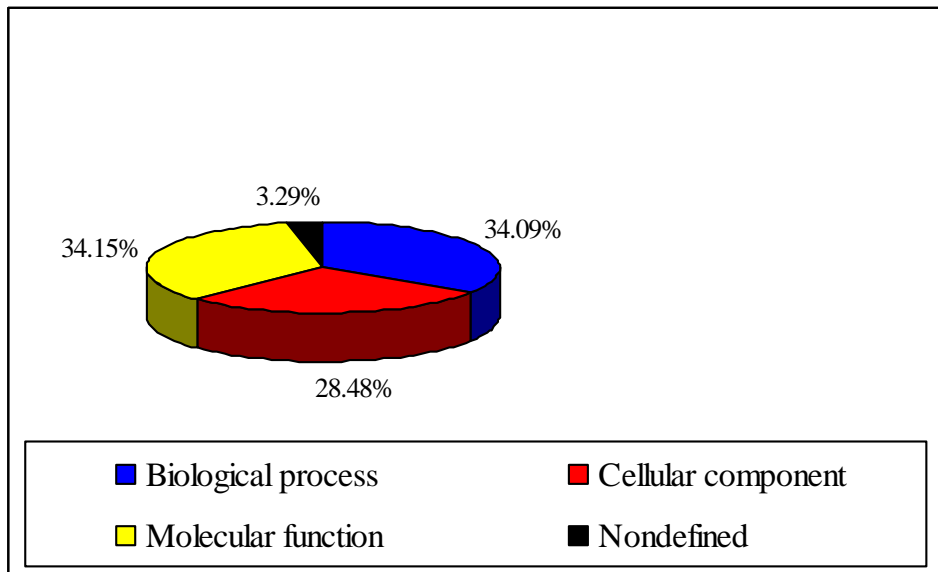
Fig.2. The three broad gene ontology (GO) classifications from the functional associations.
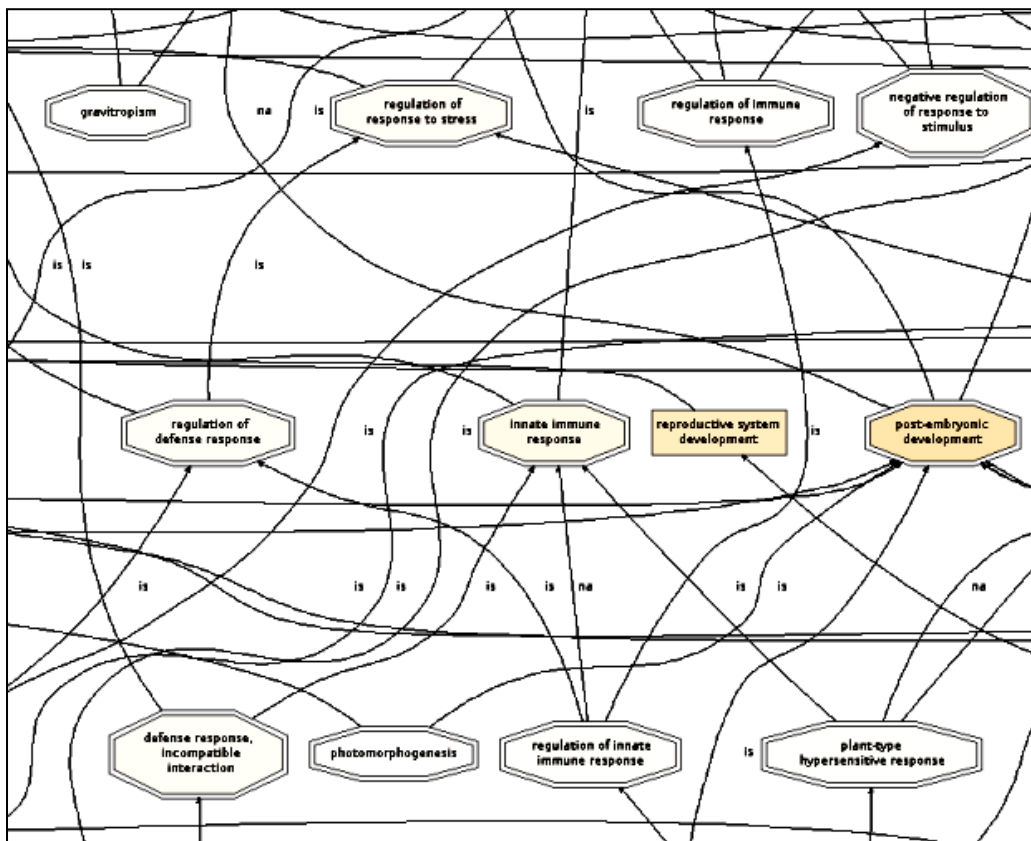

Fig. 3. Categories of sequences identified within the sequenced transcriptome of chestnut buds involved in the response to pathogens