Genomics Data 13 (2017) 64–66

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in brief

Cultivar-specific transcriptome prediction and annotation in *Ficus carica* L.

Liceth Solorzano Zambrano, Gabriele Usai, Alberto Vangelisti, Flavia Mascagni, Tommaso Giordani, Rodolfo Bernardi, Andrea Cavallini, Riccardo Gucci, Giovanni Caruso, Claudio D'Onofrio, Mike Frank Quartacci, Piero Picciarelli, Barbara Conti, Andrea Lucchi, Lucia Natali*

Department of Agriculture, Food and Environment, University of Pisa, Via del Borghetto 80, I-56124 Pisa, Italy

A B S T R A C T

The availability of transcriptomic data sequence is a key step for functional genomics studies. Recently, a repertoire of predicted genes of a Japanese cultivar of fig (*Ficus carica* L.) was released. Because of the great phenotypic variability that can be found in this species, we decided to study another fig genotype, the Italian cv. Dottato, in order to perform comparative studies between the two cultivars and extend the pan genome of this species. We isolated, sequenced and assembled fig genomic DNA from young fruits of cv. Dottato. Then, putative gene sequences were predicted and annotated. Finally, a comparison was performed between cvs. Dottato and Horaishi predicted transcriptomes. Our data provide a resource (available at the Sequence Read Archive database under SRP109082) to be used for functional genomics of fig, in order to fill the gap of knowledge still existing in this species concerning plant development, defense and adaptation to the environment.

Specifications	
Organism/cell line/tissue	*Ficus carica*/Cv. Dottato/developing fruit (2 cm in diameter) epidermal and sub-epidermal tissue
Sex	F
Sequencer or array type	Illumina MiSeq and HiSeq2000
Data format	Raw data: FASTQ files, processed data: txt files
Experimental factors	Genomic DNA
Experimental features	gDNA-seq dataset for genome assembly and gene prediction
Consent	N/A
Sample source location	43°35′22.1″N, 10°38′27.9″E, Capannoli, Pisa, Italy

1. Direct link to deposited data

Deposited data can be found at: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP109082.

2. Experimental design, materials and methods

2.1. Sample collection, DNA isolation, generation and trimming of sequence data

Epidermal and sub-epidermal tissues of young fruits of the cv. Dottato, a common parthenocarpic variety, were isolated under a dissection microscope. Then, fig DNA was isolated using the CTAB protocol described by Mascagni et al. [1].

Nuclear DNA was used for the construction of paired-end libraries (insert size of 500–600 bp) using the TruSeq DNA sample kit (Illumina Inc., San Diego, CA, USA) according to the standard Illumina protocol. The DNA sequencing was carried out with two different sequencers: MiSeq and HiSeq2000 sequencer (Illumina). HiSeq and MiSeq paired reads were trimmed using Trimmomatic [2] to remove adapters and low quality regions, using the following parameters: ILLUMINACLIP:2:30:10; LEADING:20; TRAILING:20; SLIDINGWINDOW:4:20; and MINLEN:25. Duplicated reads were discarded using CLC-BIO Genomic Workbench 8.0 (CLC-BIO, Aarhus, Denmark).

2.2. Sequence assembly

The HiSeq reads that passed the quality check (12.96 genome equivalents, 25 to 110 nt long) were analysed with KmerGenie [3] to

* Corresponding author.
E-mail address: lucia.natali@unipi.it (L. Natali).

http://dx.doi.org/10.1016/j.gdata.2017.07.005
Received 19 June 2017; Received in revised form 30 June 2017; Accepted 4 July 2017
Available online 05 July 2017

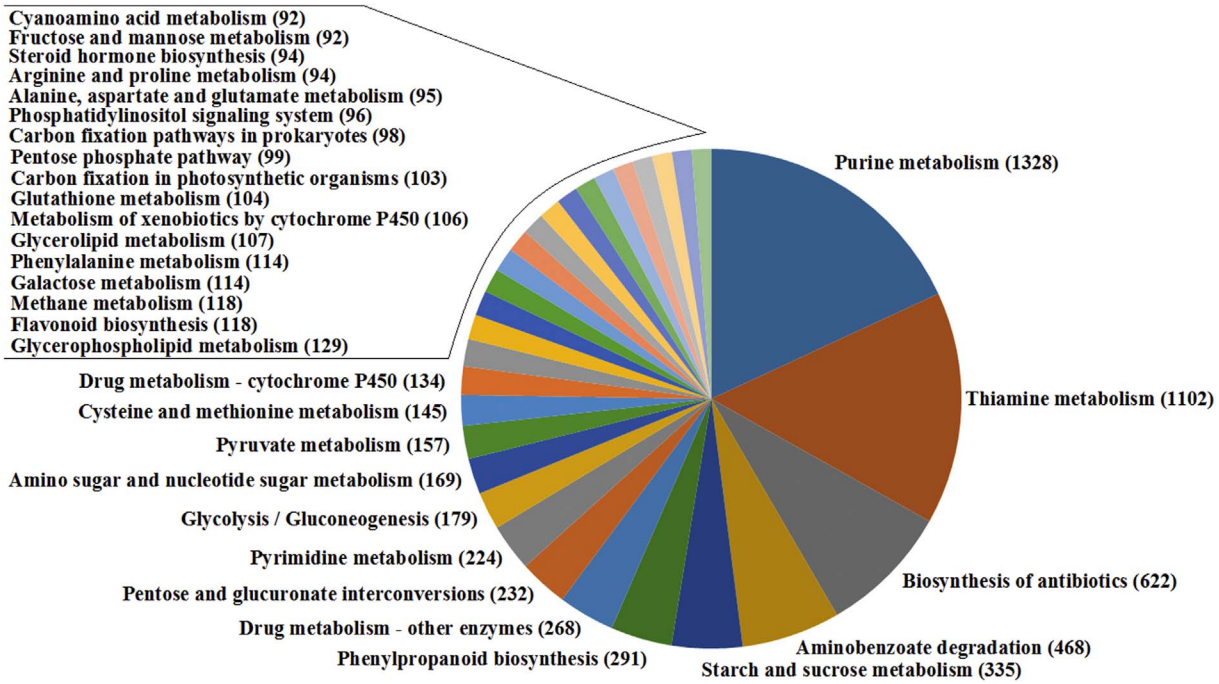

Fig. 1. Top 30 KEGG metabolic pathways in *F. carica* cv. Dottato predicted transcriptome.

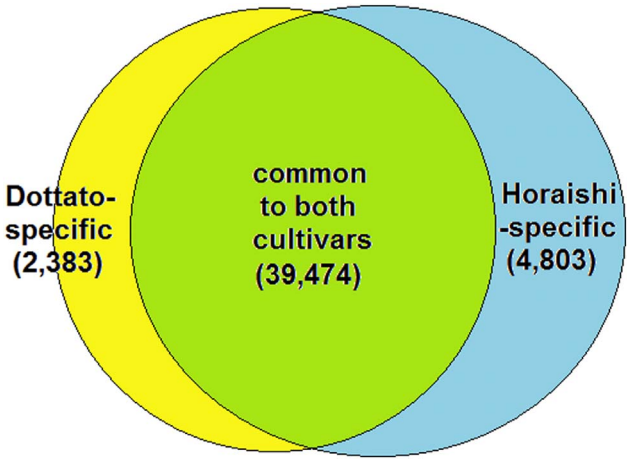

Fig. 2. Venn diagram showing a comparison between the predicted fig transcriptomes of cv. Horaishi and cv. Dottato.

detect the best k-mer for the assembly (best k = 25). De novo assembly of these reads was performed using CLC-BIO Genomic Workbench 8.0 (with mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.5, similarity fraction = 0.8, word size = 25). HiSeq reads produced 158,440 contigs, with N50 = 1137 nt.

The MiSeq reads were used to reconstruct long reads by 3′ overlapping with a minimum overlap of 30 bp and a maximum mismatch ratio of 0.4. Errors on ends were mutually corrected by best scoring bases. After quality check, reads (25.64 genome equivalents, 35 to 511 nt long) were analysed with KmerGenie to find the best k-mer for the assembly (best k = 57). Assembly was then performed using CLC-BIO Genomic Workbench 8.0 (with same parameters as above but Word size = 57). MiSeq reads produced 277,111 contigs, with N50 = 2575 nt.

A hybrid assembly was then performed using all contigs previously assembled by CLC-BIO Genomic Workbench 8.0, using Minimus2 (-D REFCOUNT = 158,440 -D MINID = 90), a tool from the AMOS toolbox [4], and obtaining 52,167 supercontigs (mean length 3615 nt, N50 = 5341 nt) and 236,059 single contigs. Contigs and supercontigs with organellar read contamination were removed by masking against a Rosaceae organellar database using RepeatMasker (-s -no_is -nolow -X -lib) [5]. After organellar removal, scaffolds were obtained from the pre-assembled sequences using the SSPACE 2.0 software (-k 5 -a 0.70 -T 5 -n 15 -p 1) [6]. This produced 264,088 scaffolds with average size = 1225 nt (max size = 41,760), N50 = 2523 nt and GC content = 33.6%. Overall, 323,708,138 nt of sequence were produced, corresponding to 87.5% of the fig genome size. Whole DNA-Seq data were submitted to the NCBI Sequence Read Archive (accession number SRP109082).

2.3. Gene prediction and annotation

Gene prediction was performed on scaffolds and supercontigs longer than 1000 nt using AUGUSTUS [7] with Arabidopsis gene models and default parameters. After retaining only the best score for each predicted gene, a total of 41,857 predicted genes were found, with a gene average length of 2135 bp and an average CDS length of 1230 bp. Total predicted gene length was 89,366,702 nt (corresponding to 24.2% of the genome). Total putative intron length was 33,896,665 nt, corresponding to 37.9% of the gene portion. A fasta file with the predicted genes of cv. Dottato is available at the Department of Agriculture, Food, and Environment of the University of Pisa repository website (http://www.agr.unipi.it/index.php/ricerca/plant-genetics-and-genomics-lab/sequence-repository).

Predicted CDSs were subject to BLAST2GO [8] for finding similarities with known protein sequences and collecting the corresponding gene ontologies. In order to identify the biological pathways active in *F. carica*, the predicted CDSs were also annotated with corresponding EC numbers against the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways database [9]. By mapping EC numbers to the reference canonical pathways, a total of 6608 contigs (15.8%) were assigned to 146 KEGG biochemical pathways.

Fig. 1 shows the 30 KEGG metabolic pathways mostly represented by unique sequences of *F. carica*. The most abundant pathways include purine metabolism, thiamine metabolism and biosynthesis of antibiotics.

2.4. Comparative analysis between predicted transcriptomes of two fig cultivars

A BLASTN analysis was performed to evaluate the differences between the predicted transcriptomes of the two fig cultivars - the Italian cv Dottato (this article) and the Japanese cv. Horaishi [10]. Whilst the vast majority of genes were found in both cultivars, a number of genes were recovered specifically in either the Japanese cultivar or the Italian cultivar (Fig. 2). Obviously, genes predicted only in the Japanese genotype could simply be missing in the Italian fig genome because of the lower sequence coverage used in our experiments. By contrast, predicted genes specific to cv. Dottato might represent genes that are not present (or are largely different) in the cv. Horaishi genome. Among KEGG pathways, significantly over-represented in the cv. Dottato predicted transcriptome, we found phosphoglycerolipid metabolism, involved in membrane composition and signal transduction, and cyano amino acid metabolism, involved in the chemical defense against herbivores and pathogens.

3. Discussion

The availability of a gDNA-based reference transcriptome is the best option for RNA-seq analyses of gene expression. Such a transcriptome was used, for example, in tree species under abiotic stress [11–13]. Such a reference transcriptome is available for fig [10]. However, it is known that differences in the genome (and even in the transcriptome) composition can occur among genotypes of the same species. For example, large variations in the coding portion of the genome were found between maize inbreds [14]. In this sense, the availability of the predicted transcriptome of a specific genotype allows a more precise and complete analysis of gene expression in that genotype. Moreover, extending the number of reference transcriptomes of a species allows the characterization of the pan-genome of that species.

Overall, 41,857 predicted genes of *F. carica* cv. Dottato were included in the fig reference transcriptome. Predicted genes were characterized by gene ontology and metabolic pathway. Among KEGG metabolic pathways, the most represented was purine metabolism (1328 members), a metabolic pathway of central significance in plant growth and development [15].

Differences were observed in the predicted gene repertoire of the two cultivars, with 4803 and 2383 genes specifically found in the Horaishi and in the Dottato predicted transcriptomes, respectively. Interestingly, many genes specific to the cv. Dottato predicted transcriptome are related to the chemical defense against herbivores and pathogens.

Our data serves as a resource for fig functional genomics and can be employed to address existing questions in this plant species relating to development, defense and adaptation to the environment.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgements

This research work was supported by the University of Pisa, Italy, project “Progetto di Ricerca di Ateneo 2017: Genomica, fisiologia e difesa del fico (*Ficus carica* L.), una specie antica con grandi prospettive”.

References

[1] F. Mascagni, E. Barghini, T. Giordani, L.H. Rieseberg, A. Cavallini, L. Natali, Repetitive DNA and plant domestication: variation in copy number and proximity to genes of LTR-retrotransposons among wild and cultivated sunflower (*Helianthus annuus*) genotypes, Genome Biol. Evol. 7 (2015) 3368–3382.
[2] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120.
[3] R. Chikhi, P. Medvedev, Informed and automated k-mer size selection for genome assembly, Bioinformatics 30 (2014) 31–37.
[4] D.D. Sommer, A.L. Delcher, S.L. Salzberg, M. Pop, Minimus: a fast, lightweight genome assembler, BMC Bioinf. 8 (2007) 64.
[5] A. Smit, R. Hubley, P. Green, RepeatMasker open 3. http://www.repeatmasker.org, (1996).
[6] M. Boetzer, C.V. Henkel, H.J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE, Bioinformatics 27 (2011) 578–579.
[7] M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel, Bioinformatics 19 (Suppl. 2) (2003) ii215–ii225.
[8] A. Conesa, S. Götz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization, and analysis in functional genomics research, Bioinformatics 10 (2005) 3674–3676.
[9] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome, Nucleic Acids Res. 32 (2004) D277–D280.
[10] K. Mori, K. Shirasawa, H. Nogata, C. Hirata, K. Tashiro, T. Habu, S. Kim, S. Himeno, S. Kuhara, H. Ikegami, Identification of RAN1 orthologue associated with sex determination through whole genome sequencing analysis in fig (*Ficus carica* L.), Sci. Rep. 7 (2017) 41124.
[11] R.M. Cossu, T. Giordani, A. Cavallini, L. Natali, High-throughput analysis of transcriptome variation during water deficit in a poplar hybrid: a general overview, Tree Genet. Genomes 10 (2014) 53.
[12] E. Barghini, R.M. Cossu, A. Cavallini, T. Giordani, Transcriptome analysis of response to drought in poplar interspecific hybrids, Genomics Data, 3 2015, pp. 143–145.
[13] T. Giordani, R.M. Cossu, F. Mascagni, F. Marroni, M. Morgante, A. Cavallini, L. Natali, Genome-wide analysis of LTR-retrotransposon expression in leaves of *Populus* × *canadensis* water-deprived plants, Tree Genet. Genomes 12 (2016) 75.
[14] S. Brunner, K. Fengler, M. Morgante, S. Tingey, A. Rafalski, Evolution of DNA sequence nonhomologies among maize inbreds, Plant Cell 17 (2005) 343–360.
[15] R. Zrenner, M. Stitt, U. Sonnewald, R. Boldt, Pyrimidine and purine biosynthesis and degradation in plants, Annu. Rev. Plant Biol. 57 (2006) 805–836.