

# An insight into structure and composition of the fig genome

E. Barghini<sup>a</sup>, F. Mascagni<sup>a</sup>, T. Giordani, L.J. Solorzano Zambrano, L. Natali, A. Cavallini<sup>b</sup>

Dept. of Agriculture, Food, and Environment, University of Pisa (DAFE), Via del Borghetto 80, I-56124 Pisa, Italy

## Abstract

***Ficus carica* L. is a diploid species, with a genome size of 0.36 pg/2C, still poorly characterized at genetic and genomic level. With the aim of analysing the fig genome structure, we used Illumina technology to produce 25.64 genome equivalents of 35-511 nt long MiSeq sequences and 12.96 genome equivalents of 25-100 nt long HiSeq paired-end reads. The two libraries were subject to a first assembly run separately, then a hybrid assembly was performed; finally, contigs and supercontigs were scaffolded. This first rough assembly is composed of 264,088 scaffolds, up to 41,760 nt in length, covering 323,708,138 nt, that corresponds to 87.5% of the fig genome, with N50 = 2,523. Masking the scaffolds with a transcriptome of Rosaceae, from which sequences related to repetitive elements were removed, allowed us to establish that coding genes account for at least 6.8% of the fig genome. Gene prediction analysis produced 44,419 putative genes. A sample of around 5,000 predicted genes were annotated with regard to gene ontology and function. Concerning the repetitive component, the fig genome resulted composed for around 58% of repeated sequences, of which none was especially redundant. Among identified repeats, the most represented were LTR-retrotransposons, with *Gypsy* elements more frequent than *Copia*.**

**Keywords:** *Ficus carica*, Illumina sequencing, genome structure, Repetitive DNA, gene prediction

## INTRODUCTION

The cultivation of fig trees (*Ficus carica* L.) dates back to ancient time. Fig is now widely grown throughout the temperate world, both for its fruit and as an ornamental plant. In recent years, large interest arose on the nutraceutical properties of fig fruit, especially dried (Vinson et al., 2005). The figs point to a high anthocyanin content, mainly cyanidin-3-rutinoside, flavonols such as quercetin-rutinoside, phenolic acids such as chlorogenic acid and flavones like luteolin, 6C-hexose-8C-pentose and apigenin-rutinoside. Higher concentrations of total phenolics were found in skin than in flesh (Vallejo et al., 2012).

Extracts of darker varieties show higher contents of phytochemicals compared to lighter colored varieties. These varieties contain the highest levels of polyphenols, flavonoids, anthocyanins and have the highest antioxidant capacity (Solomon et al., 2006). The modern pharmaceutical industry is paying more attention to medicinal plants since scientists have rediscovered that plants are an almost infinite resource for medicine development. Thus, *F. carica* has been included in occidental Pharmacopoeias and in therapeutic guides of herbal medicines (Barolo et al., 2014).

Despite its economic, cultural and ecological importance in many areas of the world, fig is still a poorly characterized species at genetic and genomic level among other fruit tree crops. *Ficus* is one of the thirty-seven genera of the Moraceae family. The fig species of greatest commercial importance is *Ficus carica* L., which consists of numerous varieties with significant genetic diversity (Barolo et al., 2014). *Ficus carica* L. is a diploid species ( $2n = 2x = 26$ ), with a genome size of about 0.36 pg/2C (Falistocco, 2009).

<sup>a</sup> E. Barghini and F. Mascagni contributed equally to this work

<sup>b</sup> andrea.cavallini@unipi.it

The identification of important genes involved in agronomic and productive traits affecting fruit production and quality, biotic and abiotic stress resistance, synthesis and accumulation of metabolites, along with the characterization of their expression can offer a significant amount of tools and open new opportunities for improvement. Over the past years, the Next Generation Sequencing (NGS) technology has emerged as a cutting edge approach for high-throughput sequence determination and this has dramatically improved the efficiency and speed of both structural and functional genomics studies (Ansong, 2009). In this work, we used the Illumina NGS technology to generate several millions of DNA sequence reads that were *de novo* assembled and annotated to produce a first insight into the fig genome structure and composition.

## **MATERIALS AND METHODS**

### ***De novo* assembly**

Fig DNA was isolated from epidermal and sub-epidermal tissues of young fruits of the cv. Dottato, a common parthenocarpic variety, using the CTAB protocol described by Cavallini et al. (1996).

Nuclear DNA was used for the construction of paired-end (insert size of 500-600 bp) libraries according to the standard protocol (Illumina, San Diego, CA, USA). The DNA sequencing was carried out with a MiSeq and with a HiSeq2000 sequencer (Illumina) at Institute of Applied Genomics (Udine, Italy). HiSeq and MiSeq paired reads were first tested for quality using fastQC and trimmed using Trimmomatic (Bolger et al., 2014) to remove adapters and low quality regions. Duplicated reads were removed using CLC-BIO Genomic Workbench 8.0 (CLC-BIO, Aarhus, Denmark). The HiSeq reads that passed the quality check were analysed with KmerGenie (Chikhi and Medvedev, 2014) to 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).

In another experiment, raw reads output from Illumina MiSeq were used to reconstruct long reads by 3' overlapping with a minimum overlap of 30bp and maximum mismatch ratio of 0.4. Errors on ends were mutually corrected by best scoring bases. The MiSeq reads that passed the quality check 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 Mismatch cost = 2, Insertion cost = 3, Deletion cost = 3, Length fraction = 0.5, Similarity fraction = 0.8, Word size = 57).

Finally, a hybrid assembly was performed using all contigs previously assembled by CLC-BIO Genomic Workbench 8.0, using Minimus2 (-D REFCOUNT=158440 -D MINID=90), a tool from the AMOS toolbox (Sommer et al., 2007). Contigs and supercontigs with organellar read contamination were removed by masking against a Rosaceae organellar database using RepeatMasker (-s -no\_is -nolow -X -lib) (<http://www.repeatmasker.org>). After organellar removal, scaffolds were obtained from the pre-assembled sequences using SSPACE 2.0 software (-k 5 -a 0.70 -T 5 -n 15 -p 1) (Boetzer et al., 2011).

### **Analysis of the repetitive DNA**

HiSeq paired-end reads were trimmed to 100 nt to ensure length uniformity, and low-complexity sequences were identified and removed using PRINSEQ version 0.20.4 (Schmieder and Edwards, 2011). Random read sets corresponding to 0.1 X genome equivalents were then subjected to clustering using RepeatExplorer (Novak et al., 2013).

### **Gene prediction**

Gene prediction was performed on scaffolds longer than 1,000 nt using AUGUSTUS (Stanke and Waack, 2003) with Arabidopsis gene models and default parameters. Predicted CDS were subject to BLAST2GO (Conesa et al., 2005) in order to identify gene ontologies and functions, which were classified using AgriGO (Du et al., 2010).

## RESULTS AND DISCUSSION

After trimming sequences for low quality and adapters, we recovered a total of 38.60 genome equivalents of Illumina reads, composed of 25.64 genome equivalents of 35 to 511 nt long MiSeq sequences and 12.96 genome equivalents of 25 to 110 nt long HiSeq paired-end reads.

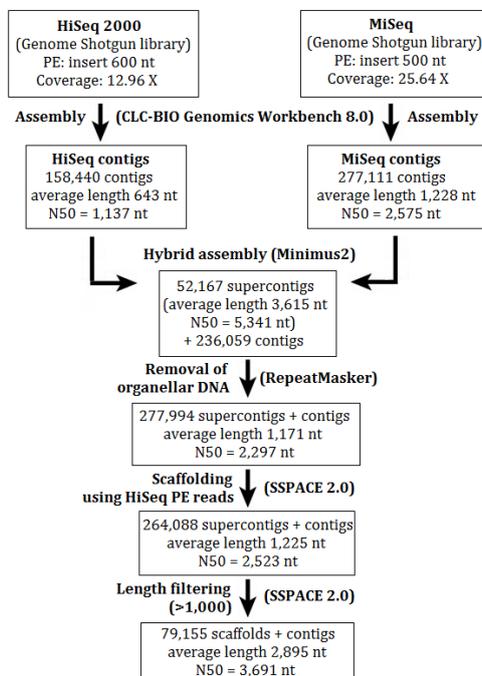


Figure 1. Strategy for sequence assembly. The tools used at each passage are indicated in parentheses.

The assembly procedure is schematized in Figure 1. The two libraries (HiSeq and MiSeq) were subject to a first assembly run, separately, by using the CLCBIO assembler. HiSeq reads produced 158,440 contigs, with N50 = 1,137 nt, MiSeq reads produced 277,111 contigs, with N50 = 2,575 nt. The overall 435,551 contigs assembled by CLC-BIO were further assembled using Minimus2, obtaining 52,167 supercontigs (mean length 3,615 nt, N50 = 5,341 nt) and 236,059 single contigs. After removal of sequences corresponding to organellar DNA, 274,994 supercontigs and contigs were collected (mean length 1,171 nt, N50 = 2,297 nt). Scaffolding of these sequences produced 264,088 scaffolds with average size = 1,225 nt (max size = 41,760), N50 = 2,523 nt and GC content = 33.6%. Overall, 323,708,138 nt of sequence were produced, corresponding to 87.5% of the fig genome size.

To gain an insight into genome structure, Illumina HiSeq reads were analysed using RepeatExplorer (Novak et al., 2013). The fig genome resulted composed for 41.7% of single or lowly repeated sequences and 58.3% of repeated sequences. This relatively low proportion of repeated sequences was expected, because of the small size of fig genome. None of repeat family was especially redundant. The composition of fig genome is reported in Figure 2. Unclassified repeats represents around 1/5 of the genome, probably because of the scarcity of fig sequences in the databases used for annotation. Among identified repeats, the most represented were LTR-retrotransposons (25.8% of the genome), with *Gypsy* elements much more frequent than *Copia*. In angiosperms, different ratios between the frequencies of *Gypsy* and *Copia* retroelements were reported, ranging from 5 : 1 in the genome of papaya to 1 : 2 in that of grapevine (Vitte et al., 2014). Ribosomal DNA constitutes 6.9% of the genome.

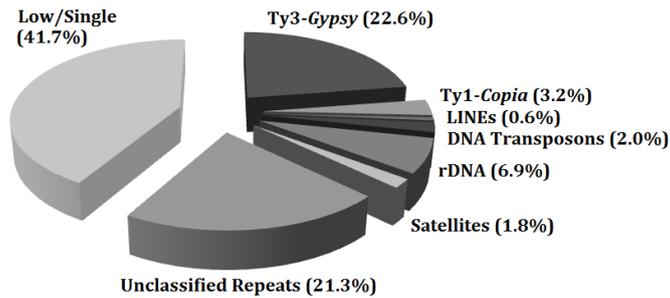


Figure 2. The fig genome composition according to sequence clustering by RepeatExplorer.

Considering the gene portion of the genome, masking the scaffolds with a transcriptome of Rosaceae, from which sequences related to transposable and other repetitive elements were removed, allowed us to establish that coding genes (without introns and promoters) account for at least 6.8% of the fig genome.

Gene prediction was performed on the scaffolds. A total of 44,419 predicted genes were found, with a coding sequence average length of 2,059 bp, average exon length of 330 bp and average intron length of 272 bp (Figure 3). Coding sequence and intron lengths are similar to those reported for another Moracea, *Morus notabilis*, whose genome was recently sequenced (He et al., 2013). On the contrary, exons resulted on average longer than those of a number of dicot species, including *M. notabilis* (He et al., 2013). For 9,660 putative genes no introns were predicted.

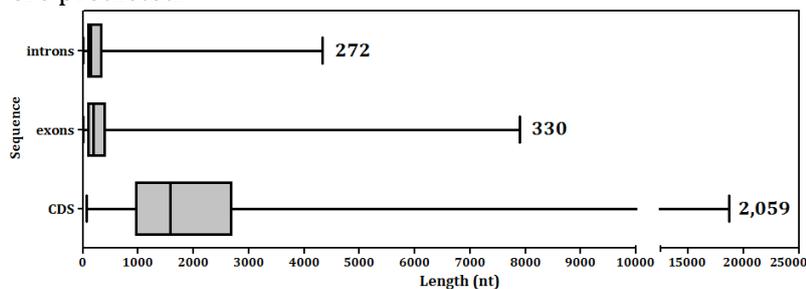


Figure 3. Box and whiskers plots of the lengths of predicted coding sequences (CDS), exons, and introns. The boxes represent the 20–80%, whiskers represent the whole range of values, and lines in the box represent the medians of the distributions. For each class, the mean is reported.

A first sample of 5,043 predicted proteins were annotated for their gene ontology and function. For 3,049 of these, at least one gene ontology was attributed. The distribution of gene ontologies in this sample of predicted proteins, subdivided into biological processes, cellular component, and molecular function is reported in Figure 4.

A large number of predicted genes (at least 313/3,049) encode transcription factors. Many of the other predicted genes in the sample belong to gene families as NBS-LRR domain containing proteins (130/3,049), *i.e.*, one of the largest plant gene families, involved in plant defense (McHale et al., 2006). Another gene family encodes pentatricopeptide-repeat containing proteins (139/3,049), a very heterogeneous class of proteins, often targeted to mitochondria or chloroplasts, and involved in RNA editing, with effects on many characters, for example plant development and environmental adaptation (Barkan and Small, 2014). Another large family is that encoding cytochrome P450 (39/3,049), mostly related to catalyze the oxidation of organic substances and widespread in both prokaryotes and eukaryotes. As observed in other tree species, as, for example, *Olea europaea* (Barghini et al., 2014), specific gene families are largely represented, for example those encoding WD40 repeat-containing proteins (84 genes), ankyrins (37), and ABC-transporters (36).

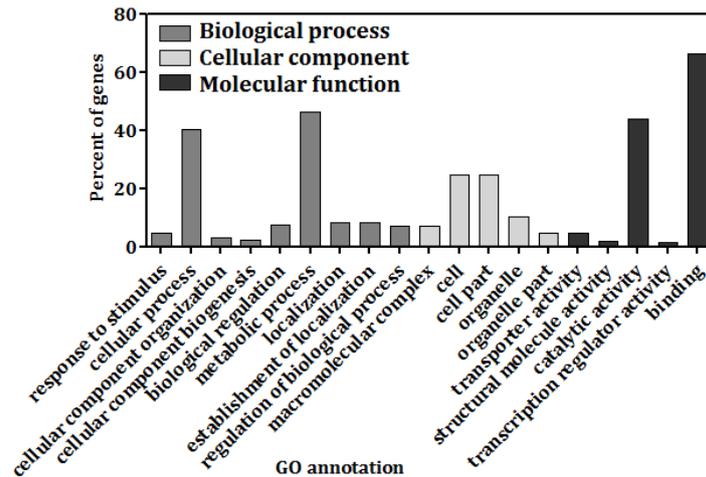


Figure 4. GO annotation, by comparison to *A. thaliana* coding sequences, of a sample of around 3,000 predicted genes of *Ficus carica*.

Among predicted proteins involved in pathogen response, beyond NBS-LRR domain containing proteins, we found proteins, as cysteine (11 genes), aspartic (9), and serine proteases (6), which are believed to be involved in the defence system against microbes or herbivores and have been found in the latex of plant species (An et al., 2002; Konno et al., 2003; He et al., 2013). Functional studies on these genes will allow us to expand our knowledge on fig defense mechanisms.

## CONCLUSIONS

Overall, the reported fig genome structure is similar to that of other small-genome-sized plant species, with repetitive DNA accounting for more than 50% of the genome and a relatively large number of LTR-retrotransposons.

Studies are in progress to improve assembly and characterization of fig genome and to identify important genes involved in agronomic and productive traits affecting fruit production and quality, biotic and abiotic stress resistance, and synthesis and accumulation of metabolites. These data will offer a significant amount of tools and open new opportunities for fig breeding.

## DATA ACCESSIBILITY

Raw reads of Illumina sequencing are available upon request. Assembled sequence set and annotations are available at the repository sequence page of the DAFE (<http://www.agr.unipi.it/ricerca/plant-genetics-and-genomics-lab/sequence-repository>).

## ACKNOWLEDGEMENTS

Research work funded by Department of Agriculture, Food, and Environment, project PLANTOMICS.

## Literature cited

- An, C.I., Fukusaki, E., and Kobayashi, A. (2002). Aspartic proteinases are expressed in pitchers of the carnivorous plant *Nepenthes alata* Blanco. *Planta* 214, 661–667 <http://dx.doi.org/10.1007/s004250100665>. PubMed
- Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology* 25, 195-203 <http://dx.doi.org/10.1016/j.nbt.2008.12.009>. PubMed
- Barghini, E., Natali, L., Cossu, R.M., Giordani, T., Pindo, M., Cattonaro, F., Scalabrin, S., Velasco, R., Morgante, M., and Cavallini, A. (2014). The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome, *Genome Biol. Evol.* 6, 776–791 <http://dx.doi.org/10.1093/gbe/evu058>. PubMed
- Barkan, A., and Small, I. (2014). Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant. Biol.* 65, 415-442 <http://dx.doi.org/10.1146/annurev-arplant-050213-040159>. PubMed

- Barolo, M.I., Ruiz Mostacero, N., and López, S.N. (2014). *Ficus carica* L. (Moraceae): an ancient source of food and health. *J. Food Chemistry* 164, 119-127 <http://dx.doi.org/10.1016/j.foodchem.2014.04.112>. PubMed
- Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578-579 <http://dx.doi.org/10.1093/bioinformatics/btq683>. PubMed
- Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* *btu170*, 1-7 <http://dx.doi.org/10.1093/bioinformatics/btu170>. PubMed
- Cavallini, A., Natali, L., Giordani, T., Durante, M., and Cionini, P.G. (1996). Nuclear DNA changes within *Helianthus annuus* L.: variations in the amount and methylation of repetitive DNA within homozygous progenies. *Theor. Appl. Genet.* 92, 285-291 <http://dx.doi.org/10.1007/BF00223670>. PubMed
- Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30, 31-37 <http://dx.doi.org/10.1093/bioinformatics/btt310>. PubMed
- Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674-3676 <http://dx.doi.org/10.1093/bioinformatics/bti610>. PubMed
- Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucl. Acids Res.* 38, W64-W70 <http://dx.doi.org/10.1093/nar/gkq310>. PubMed
- Falisticco, E. (2009). Presence of triploid cytotypes in the common fig (*Ficus carica* L.). *Genome* 52, 919-925 <http://dx.doi.org/10.1139/G09-068>. PubMed
- He, N., Zhang, C., Qi, X., et al. (2013). Draft genome sequence of the mulberry tree *Morus notabilis*. *Nature Comm.* 4, 2445 <http://dx.doi.org/10.1038/ncomms3445>. PubMed
- Konno, K., Hirayama, C., Nakamura, M., Tateishi, K., Tamura, Y., Hattori, M., and Kohno, K. (2003). Papain protects papaya trees from herbivorous insects: role of cysteine proteases in latex. *Plant J.* 37, 370-378 <http://dx.doi.org/10.1046/j.1365-313X.2003.01968.x>. PubMed
- McHale, L., Tan, X., Koehl, P., and Micheltore, R.W. (2006). Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* 7, 212 <http://dx.doi.org/10.1186/gb-2006-7-4-212>. PubMed
- Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics* 29, 792-793 <http://dx.doi.org/10.1093/bioinformatics/btt054>. PubMed
- Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863-864 <http://dx.doi.org/10.1093/bioinformatics/btr026>. PubMed
- Solomon, A., Golubowicz, S., Yablówic, Z., Grossman, S., Bergman, M., Gottlieb, H.E., Altman, A., and Flaishman, M.A. (2006). Antioxidant activities and anthocyanin content of fresh fruits of common fig (*Ficus carica* L.). *J. Agricultural Food Chem.* 54, 7717-7723 <http://dx.doi.org/10.1021/jf060497h>. PubMed
- Sommer, D.D., Delcher, A.L., Salzberg, S.L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 8, 64 <http://dx.doi.org/10.1186/1471-2105-8-64>. PubMed
- Stanke, M., and Waack, S. (2003). Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics* 19, ii215-ii225 <http://dx.doi.org/10.1093/bioinformatics/btg1080>. PubMed
- Vinson, J.A., Zubik, L., Bose, P., Samman, N., and Proch, J. (2005). Dried fruits: excellent *in vitro* and *in vivo* antioxidants. *J. Amer. College Nutrition* 24, 44-50 <http://dx.doi.org/10.1080/07315724.2005.10719442>. PubMed
- Vallejo, F., Marín, J.G., and Tomás-Barberán, F.A. (2012). Phenolic compound content of fresh and dried figs (*Ficus carica* L.). *Food Chemistry* 130, 485-492 <http://dx.doi.org/10.1016/j.foodchem.2011.07.032>. CrossRef
- Vitte, C., Fustier, M.A., Alix, K., and Tenailon, M.I. (2014). The bright side of transposons in crop evolution. *Brief. Funct. Genomics* 13, 276-295 <http://dx.doi.org/10.1093/bfpg/elu002>. PubMed