Proceedings of the LXII SIGA Annual Congress
Verona, Italy – 25/28 September, 2018
ISBN **978-88-904570-8-1**

Poster Communication Abstract – 7.12

DECIPHERING THE GENOME OF A HIGHLY HETEROZYGOUS, NON-MODEL SPECIES, *FICUS CARICA* L., USING LONG-READ SEQUENCING

USAI G.*, MASCAGNI F.*, GIORDANI T.*, ZUCCOLO A.**, CECCARELLI M.***, KING R.****, HASSANI-PAK K.****, NATALI L.*, CAVALLINI A.*

*) Dept. of Agriculture, Food and Environment, University of Pisa, Via del Borghetto 80, 56124 Pisa (Italy)
**) Institute of Life Sciences, ScuolaSuperioreSant'Anna, Pisa (Italy)
***) Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia (Italy)
****) Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ

Ficus carica, genome sequence, PacBio sequencing, heterozygosity

Sequencing and assembly of non-model plant genomes can be very challenging depending on the high heterozygosity and repeat content. Especially in fruit trees, many important phenotypic traits of a specific genotype lie in its heterozygosity, which is maintained because of widespreadclonal propagation of these species.

The fig tree (*Ficus carica* L., *Moraceae*) is widely grown for its fruit throughout the temperate world. This crop has a great potential for expansion thanks to valuable nutritional and nutraceutical characteristics, combined with the ability to adapt well to marginal soils and severe environmental conditions. However, some features, such as the rapid perishability of its fresh fruits, prevent fig diffusion and commercial success. *F.carica* is a diploid species, with a relatively small genome size (0.36 Gbp) and is still poorly characterized compared to other fruit tree crops. In fact, only a very preliminary genome sequence (of the Japanese cv. Horaishi) has been recently released. The aim of this work was to generate a high-quality and phased reference genome of the typical Italian fig cultivar Dottato.

After DNA isolation, the Pacific Biosciences (PacBio) SMRT sequencing technology was carried out on a Sequel System platform. We sequenced 6 PacBio SMRT cells, obtaining a total of 2,140,959 raw reads (maximum length = 111,426 nt; average length = 12,364 nt) with an N50 value of 18,419 nt and corresponding to about 74 genome equivalents.*De novo* assembly was performed by the diploid-aware assembler FALCON. Then, the haplotypes were phased and resolved using the FALCON-Unzip module. FALCON-Unzip is able to phase structural variations and SNPs into distinct haplotype blocks, allowing the identification of allelic differences between parental chromosomes. The whole process origins a primary set of contigs (primary assembly) with the linked haplotigs (haplotype assembly) that represent the alternative genome structures of the primary contigs. Both set of contigs were polished using Arrow software. Then, the primary assembly was upgraded using FinisherSC tool. Finally, both set of contigs were polished again using Pilon and checked for contamination using MEGAN.We produced a primary set of 933 contigs (maximum length = 5,010,936 nt; average length = 359,668 nt) with an N50 value of 813,023 nt. This assembly corresponds to about 94.3% of the estimated genome size. The haplotype assembly consists of 7,010 haplotigs contigs (maximum length = 1,220,129 nt; average length = 58,637 nt).This assembly has been annotated using both the MAKER2 and the TEdenovo pipelines, for genes and repetitive elements, respectively.

Since the low level of genetic improvement of the present fig cultivars, the decryption of the fig genome will open great opportunities for speeding up the development of new cultivars and for the application to this species of genome editing, a new technology which seems especially suitable to change the specific traits currently limiting the success of this ancient species.