7th International Horticulture Research Conference

Progress towards the complete haplotype-phased fig (*Ficus carica* L.) genome

Gabriele Usai (gabriele.usai@agr.unipi.it), Flavia Mascagni, Alberto Vangelisti, Tommaso Giordani, Andrea Cavallini, Lucia Natali

Department of Agriculture, Food and Environment, University of Pisa, Pisa, Italy.

Introduction

The availability of genome sequence is a key prerequisite to apply modern breeding procedures to crops. Although many plant genomes are currently available, it is increasingly important, especially for highly heterozygous cultivated species, to decipher the sequence of the two haplotypes that make up the genome. This is all the more true for fruit trees, whose condition of heterozygosity is maintained in the various cultivars thanks to the clonal propagation. In these crops, the heterozygosity can be the basis of favorable traits that must be maintained and/or improved. The fig tree (*Ficus carica* L.), one of the oldest known domesticated species (Figure 1), has a great potential for commercial expansion thanks to its valuable nutritional and nutraceutical characteristics (Veberic et al., 2008), combined with its ability to adapt well to marginal soils and difficult environmental conditions (Vangelisti et al., 2019). However, only in recent years high-quality genomic resources have been released (Mori et al., 2017; Usai et al., 2019). Here, we report our work-in-progress haplotype-phased genome assembly achieved combining the last published genome reference based on the single-molecule, real-time sequencing technology of Pacific Biosciences (PacBio) with the latest *in silico* methodologies relied on chromosome conformation capture. The released of a haplotype-phased genome reference is a pivotal resource to study allele-specific expression and epigenetic regulation, thus relaunching the genetic improvement in fig.

Figure 1. Vegetative and reproductive structures of fig.

Methods

The fig primary assembly and haplotigs produced by Usai et al. (2019) were curated by using Purge Haplotigs (Roach et al., 2018) in order to reassign mis-placed allelic contigs. Hi-C read pairs were obtained from young leaf tissue by using the ArimaGenomics kit and sequenced through Illumina platform. The dataset was filtered by using the Arima pipeline (ArimaGenomics). FALCON-Phase (Kronenberg et al., 2019) was used for the Hi-C-based phasing process to create phased, diploid contigs. The Hi-C read pairs were aligned to the produced output by using BWA-MEM (Li and Durbin, 2010) and SALSA (Ghurye et al., 2017) was run to perform the scaffolding process. Subsequently, FALCON-Phase was run for a second iteration. Finally, we used the Illumina SSR- and SNP-based scaffolds of the Horaishi assembly to order and orientate the produced sequences to the 13 chromosomes of fig (Mori et al., 2017).

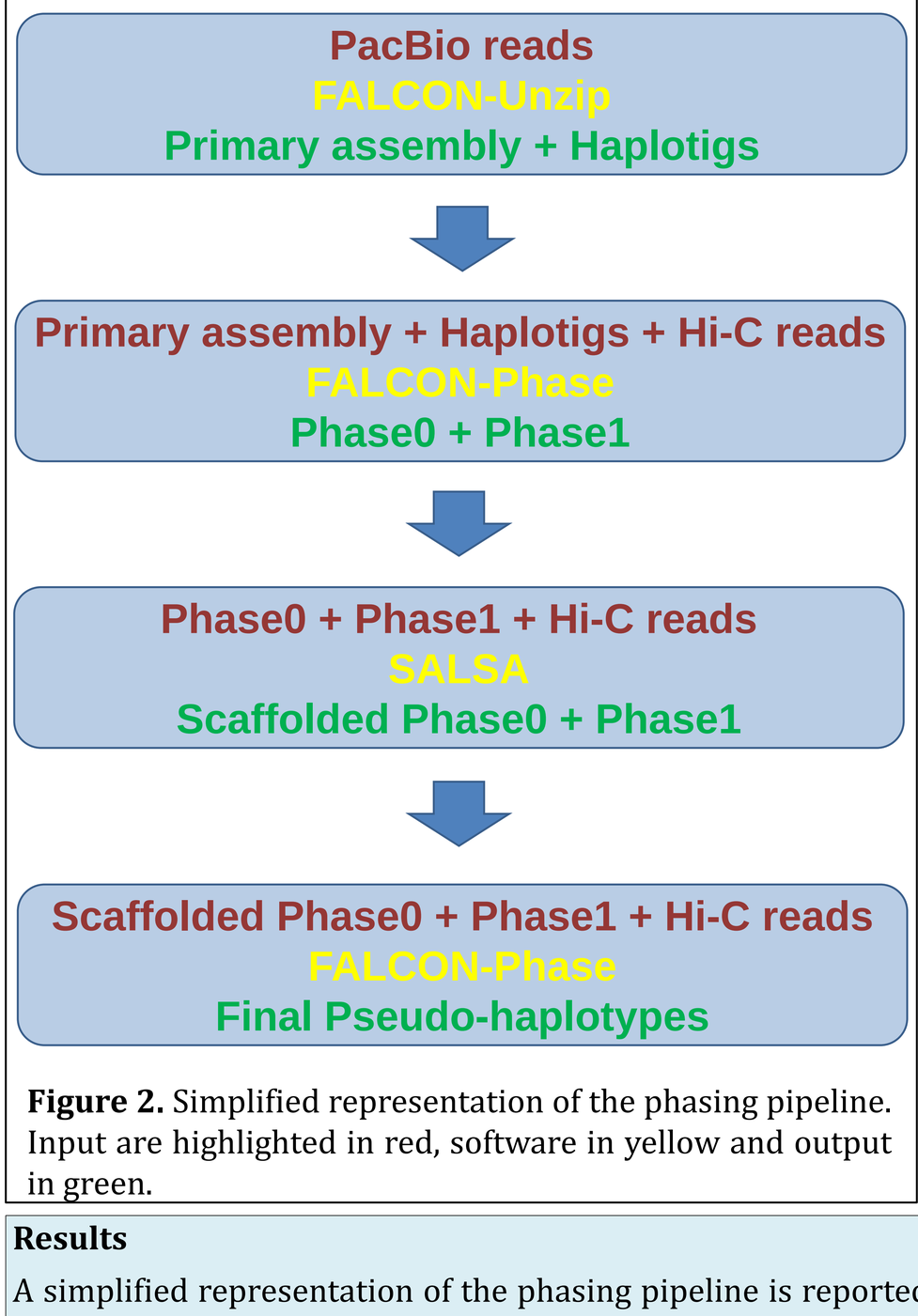

Figure 2. Simplified representation of the phasing pipeline. Input are highlighted in red, software in yellow and output in green.

(a)

FALCON-Unzip: Primary assembly	FALCON-Unzip: Haplotigs
Number of sequences (№): 905	Number of sequences (№): 6,933
Total size of sequences (bp): 333,400,567	Total size of sequences (bp): 408,159,944
Longest sequence (bp): 5,010,936	Longest sequence (bp): 1,220,129
Sequences > 100 kbp (№): 595	Sequences > 100 kbp (№): 953
Mean sequence size (bp): 368,398	Mean sequence size (bp): 58,872
N50 sequence length (bp): 823,517	N50 sequence length (bp): 89,539

(b)

Hi-C read pairs dataset
Total number of reads (№): 131,870,834
Read length (bp): 150
Total size of the reads (bp): 19,780,625,100
Coverage (x): ~55

(c)

FALCON-Phase: Pseudo-haplotype 0	FALCON-Phase: Pseudo-haplotype 1
Number of sequences (№): 538	Number of sequences (№): 538
Total size of sequences (bp): 351,483,585	Total size of sequences (bp): 349,982,723
Longest sequence (bp): 7,234,680	Longest sequence (bp): 8,569,698
Sequences > 100 kbp (№): 304	Sequences > 100 kbp (№): 302
Mean sequence size (bp): 653,315	Mean sequence size (bp): 650,526
N50 sequence length (bp): 1,991,136	N50 sequence length (bp): 1,927,249

Figure 3. (a) Statistics of fig primary assembly and haplotigs produced by FALCON-Unzip. (b) Hi-C read pairs dataset statistics. (c) Statistics of the final fig pseudo-haplotypes produced by FALCON-Phase. Intermediate results are not shown.

Results

A simplified representation of the phasing pipeline is reported in Figure 2. The most relevant results are shown in Figure 3. A total of ~55× Hi-C reads were obtained from the sequencing process. Those data were integrated with the previously produced fig assembly and haplotigs, resulting in two contiguous and proportionate pseudo-haplotypes of 538 sequences with mean size of 0.65 Mb and N50 of 1.99 Mb and 1.93 Mb, respectively. The pseudo-haplotypes represented ~98% of the estimated 356 Mb fig genome (Loureiro et al., 2007). Finally, 400 out of 538 sequences (~96% of both pseudo-haplotypes) were associated to the 13 corresponding chromosomes producing our version 2.0 of fig genome.

Conclusions and next research goals

In conclusion, the production of a high-quality, diploid-phased, reference genome will be a valuable genomic resource for the improve of fig breeding but also for the investigation of other tree species and in general for all species presenting a highly heterozygous genome. The next steps for the fig genome 2.0 will be the complete annotation by using *ad hoc* approaches including RNA-seq data, protein alignment and *de novo* prediction. The same level of accuracy will be achieved to analyze the repetitive DNA of this species. Furthermore, a genome-wide methylation analysis will be performed on both pseudo-haplotypes using the newest *in silico* methodologies. Further analysis will be conducted to investigate how heterozygosity and methylation affect the expression of different alleles.

References

[1] **Ghurye et al.** (2017) BMC Genomics; [2] **Kronenberg et al.** (2019) Biorxiv; [3] Li **and Durbin** (2010) Bioinformatics; [4] **Loureiro et al.** (2007) Ann. Bot.; [5] **Mori et al.** (2017) Sci. Rep.; [6] **Roach et al.** (2018) BMC Bioinformatics; [7] **Usai et al.** (2019) Plant J.; [8] **Vangelisti et al.** (2019) Sci. Rep.; [9] **Veberic et al.** (2008) Food Chemistry.