4.09 – The haplotype-phased genome assembly for *Ficus carica* L.: an ancient crop with promising perspectives

Castellacci M., Usai G., Vangelisti A., Ventimiglia M., Simoni S., Natali L., Cavallini A., Mascagni F. and Giordani T.
Department of Agriculture, Food and Environment, University of Pisa, Pisa, Italy

BACKGROUND

Deciphering the sequence of the two haplotypes which constitute the genome is crucial to apply modern breeding procedures. This is even more true for fruit trees, whose condition of heterozygosity is maintained through clonal propagation. The fig tree (*Ficus carica* L.) has a great potential for commercial expansion, but high-quality genomic resources have been released only in recent years (Usai *et al*., 2020). Here we report our work-in-progress haplotype-phased assembly achieved combining the last published reference produced through single-molecule, real-time (SMRT) sequencing and Hi-C technique.

RESULTS

A total of ~55x Hi-C reads were obtained. These data were integrated with the previously produced fig assembly resulting in two pseudo-haplotypes (Table 1). The pseudo-haplotypes represented ~98% of the estimated 356 Mb fig genome (Loureiro *et al*., 2007). 400 out of 538 sequences (~96% of both pseudo-haplotypes) were associated to the 13 corresponding chromosomes of fig.

Table 1. Statistics of the fig pseudo-haplotypes and intra-genomic comparison.

Pseudo-haplotype 0	Pseudo-haplotype 1	Intra-genomic comparison
Size of sequences (bp): 355,244,677	Size of sequences (bp): 346,221,631	SNP: 832,619
Mean sequence size (bp): 660,306	Mean sequence size (bp): 643,535	INDEL: 996,026
N50 sequence length (bp): 1,989,800	N50 sequence length (bp): 1,927,249	Structural variants (SVs): 308

Proper approaches based on *de novo* prediction, RNA-seq data and protein alignment allowed us to predict 33,954 and 33,379 protein-coding genes per pseudo-haplotype, of which 27,916 and 27,558 were functionally annotated (about 82%), respectively. To provide an evaluation of the divergence between the two fig pseudo-haplotypes, we identified genomic variation between the 13 homologous chromosome pairs (Table 1).

Based on synteny, 53,024 genes were identified as having homologs on the two pseudo-haplotypes. 26,512 pairs were considered as reliable allelic genes, and 15,050 pairs showed mutations (Figure 1).

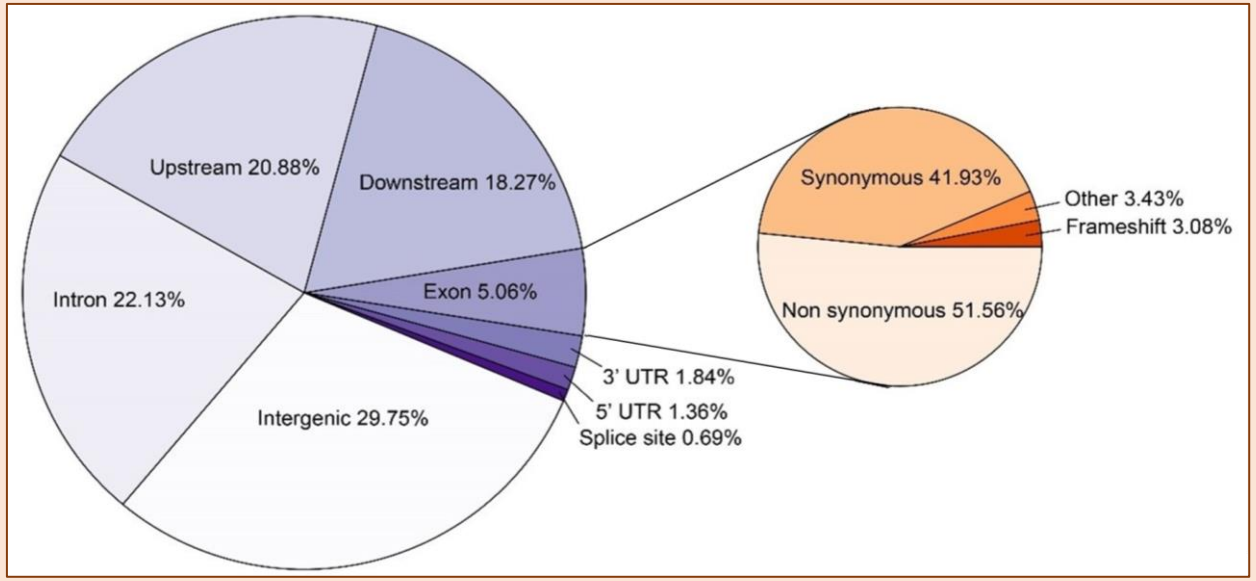

Figure 1. Genomic variation distribution and effect between the allelic paired genes of *F. carica* genome (reported as the percentage of total SNPs and INDELs). The left pie chart indicates the variations distribution. The right pie chart indicates the putative effect of SNPs on the exon regions.

MATERIALS AND METHODS

1) Assembly Process: Phalcon-phase
2) Scaffolding Process: SALSA
3) Annotation Process: Usai *et al*., (2020) pipeline
4) Intra-genomic comparison: Zhou *et al*., (2020) pipeline

SUMMARY AND CONCLUSIONS

In this work, we obtained a completely phased assembly genome of the *Ficus carica* L. cultivar Dottato. Comparison of the two pseudo-haplotype led to a more refined characterization of the genomic variants inside the allelic genes and the putative effects of their mutations. Interestingly, allelic genes showed more non-synonymous variants than synonymous ones. These data will be the basis for future fig breeding programs.

REFERENCES

Ghurye *et al*. (2017) BMC Genomics; Kronenberg *et al*. (2021) Nat. Commun.; Loureiro *et al*. (2007) Ann. Bot.; Mori *et al*. (2017) Sci. Rep.; Usai *et al*. (2020) Plant J.; Zhou *et al*. (2020) Nat. Genet.

Funding: FIGGEN/PRIMA project www.figgen.eu