64th ITALIAN SOCIETY OF AGRICULTURAL GENETICS ANNUAL CONGRESS

UNIVERSITÀ DI PISA

1.17 - Towards a new phased genome reference of fig (*Ficus carica* L.), a crucial resource for fig breeding

Usai G., Giordani T., Vangelisti A., Mascagni F., Ventimiglia M., Simoni S., Natali L. and Cavallini A.
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). This species has esteemed nutritional and nutraceutical characteristics (Veberic *et al*., 2008), combined with its ability of adaptation to marginal soils and difficult environmental conditions (Vangelisti *et al*., 2019). 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 ~55× Hi-C reads were obtained (Table 1). Those data were integrated with the previously produced fig assembly resulting in two pseudo-haplotypes of 538 sequences with mean size of 0.65 Mb and N50 of 1.99 Mb and 1.93 Mb, respectively (Table 2). 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.

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

Table 1. Hi-C read pairs dataset statistics.

Pseudo-haplotype 0	Pseudo-haplotype 1
Sequences (№): 538	Sequences (№): 538
Size of sequences (bp): 351,483,585	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

Table 2. Statistics of the fig pseudo-haplotypes.

Proper approaches based on *de novo* prediction, RNA-seq data and protein alignment allowed us to predict 34,029 and 33,957 protein-coding genes per pseudo-haplotype, of which 28,053 and 28,058 were functionally annotated, respectively.

To provide an evaluation of the divergence between the two fig haplotypes, we identified polymorphisms between the 13 homologous chromosome pairs. We detected 540 syntenic blocks carrying 2,700,243 SNPs, 1,488,669 indels (1-50 bp) and 8,360 structural variants (> 50 bp).

Based on synteny, 50,894 genes were identified as having homologs on the two haplotypes. 22,120 pairs were considered as reliable allelic genes, and 15,927 pairs showed mutations. Variant annotation on the 15,927 pairs revealed that 53.76% of variants were SNPs and 46.24% were indels. The majority of these occurred within intronic regions (Figure 1).

Of the coding variants, the majority were missense mutations (53.93%), followed by synonymous mutations (45.53%) with a little occurrence of nonsense mutations (0.52%).

Functional analysis both for allelic genes and transposable elements are ongoing.

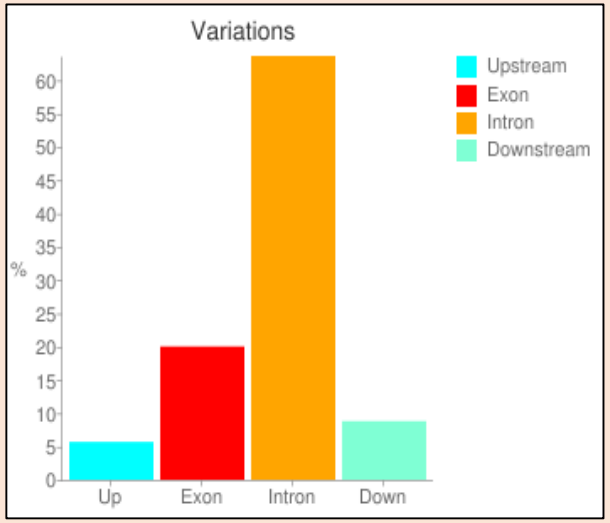

Figure 1. Variations distribution among paired upstream, exon, intron, and downstream regions of allelic genes.

MATERIALS AND METHODS

Hi-C reads were obtained from leaf tissue using the ArimaGenomics kit and sequenced through Illumina platform. FALCON-Phase (Kronenberg *et al*., 2021) was used for the phasing process starting from the fig assembly produced by Usai *et al*. (2020). SALSA (Ghurye *et al*., 2017) was used for the scaffolding process. FALCON-Phase was run for a second iteration. Illumina SSR- and SNP-based scaffolds of the Horaishi assembly were used to order and orientate the produced sequences to the 13 chromosomes of fig (Mori *et al*., 2017). Gene prediction and annotation was performed according to Usai *et al*. (2020) pipeline. Haplotypes comparison was performed according to Zhou *et al*. (2020) pipeline.

SUMMARY AND CONCLUSIONS

The high-quality phased genome reference will be the basis to assess the genetic variability of fig varieties on available Spanish, Tunisian and Turkish fig collections using a genotyping by sequencing approach in the frame of a PRIMA (Partnership for Research and Innovation in the Mediterranean Area) project, FIGGEN. This data will be the prerequisite for genome-wide association studies (GWAS) with the final purpose of unveil genes or molecular markers linked to traits related to fruit quality and to environmental adaptation to difficult conditions, consequence of climate change, leading to the genetic improvement of fig.

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.; Vangelisti *et al*. (2019) Sci. Rep.; Veberic *et al*. (2008) Food Chem.; Zhou *et al*. (2020) Nat. Genet.