Genome-wide analysis of gene heterozygosis and allele differential expression in the fig tree

geneticaagraria.it

USAI G., VANGELISTI A., CASTELLACCI M., SIMONI S., MASCAGNI F., NATALI L., CAVALLINI A. and GIORDANI T.

1) Department of Agriculture, Food and Environment, University of Pisa, Via del Borghetto 80, 56124 Pisa (Italy)

6.21

BACKGROUND

Heterozygosity plays a significant role in agriculture, especially in developing hybrid cultivars and heterozygous crops. However, studying the genome-wide effects of heterozygosity is challenging due to the complexities involved in analyzing diploid heterozygous individuals (Yuan et al., 2021). In this study, we investigated whole-genome heterozygosity and its impact on gene and allele expression in *Ficus carica* (fig tree), an important fruit tree known for its resilience to environmental changes (Vangelisti et al., 2019).

RESULTS

a) Genome sequencing, assembly and annotation

We sequenced and assembled the fig genome using long DNA read sequencing and chromosome conformation capture. Hi-C reads of ~55x coverage were integrated with the previously produced fig assembly (Usai et al., 2019), resulting into the two pseudo-haplotypes of the approximately 356 Mbp fig genome (Table 1). De novo prediction, RNA-seq analysis, and protein alignment resulted in a total of 33,954 and 33,379 protein-coding genes per pseudo-haplotype, of which approximately 82% were functionally annotated.

b) Pseudo-haplotypes diversity analysis and allelic genes characterization

We identified the genomic variations between the two pseudo-haplotypes of fig. After that, through synteny analysis, this data was integrated with the identified 20,441 allelic gene pairs. Considering the CDS, 13,331 allelic gene pairs were homozygous, while 7,110 exhibited heterozygosity. The heterozygous gene pairs were further categorized into three functional categories: 3,311 structural proteins, 2,664 enzymes, 477 transcription factors, and 658 remained uncharacterized.

c) Allelic differential expression analysis

Gene expression was then studied in the leaves of plants subjected or not to saline stress for 48 days. Out of the 7,110 heterozygous gene pairs, 5,067 were found to be expressed (Table 2). Among the expressed genes, 14.41% in the control group and 18.77% in the salt-treated group exhibited differential allelic expression (DAE), occurring in either the control, the salt treatment, or in both conditions. The percentage was higher for genes encoding structural proteins and lower for those encoding transcription factors. Generally, the most expressed allele was found to be the same in both control and treated groups. Only 0.14% of the differentially expressed genes (DEGs) showed DAE only in the control or only in the treated group, indicating that only one of the two alleles was regulated by saline stress. In addition, a reduced impact of sequence variations in the promoter regions related to allelic expression was observed (Figure 1). At genome-wide level, less variations in upstream gene promoters were apparently related to higher DAE levels than more variations in the same regions.

Table 1 - Statistics of the *F. carica* genome assembly.

Chromosome number (2n)	2n = 2x = 26		
Estimated genome size (1n)	~356 Mbp		
		Pseudo-haplotype 0	Pseudo-haplotype 1
Contig assembly	Genome representation (%)	~99	~97
	Total size of the assembly (bp)	355,244,677	346,221,631
	Number of sequences (No.)	538	538
	Mean sequence size (bp)	660,306	643,535
	N50 sequence length (bp)	1,989,800	1,927,249
Anchored assembly	Genome representation (%)	~97	~95
	Total size of the assembly (bp)	346,881,609	338,526,026
	Number of sequences (No.)	13	13
	Mean sequence size (bp)	26,683,201	26,040,464
	N50 sequence length (bp)	27,941,851	27,454,058
Annotation	BUSCO assessment (%)	93.6	92.7
	Protein-coding genes proportion (%)	30.02	30.01
	Predicted protein-coding genes (No.)	33,954	33,379
	Annotated protein-coding genes (No.)	27,916	27,558
	Rate of annotated protein-coding genes (%)	82.22	82.56
	Average exon per gene (No.)	4.57	4.59
	Average intron per gene (No.)	3.57	3.59
	LAI assessment (No.)	10.94	13.85
	Transposable elements proportion (%)	48.68	48.59

Table 2 - Percentage of heterozygous genes showing differential allelic expression (DAE) and differentially expressed genes (DEGs) between control (C48) and treated plants (S48). The percentages represent the various intersections of DAE, DEG, C48 and S48 along with their functional characterization.

	Structural protein	Enzyme	Transcription factor	NA	Tot
DAE C48 tot	7.32	5.59	0.77	0.73	14.41
DAE S48 tot	10.10	6.81	1.05	0.81	18.77
DEG tot	1.26	1.14	0.32	0.12	2.84
DAE C48, DAE S48 and DEG	0.14	0.02	0.00	0.00	0.16
DAE C48 and DAE S48	5.53	4.01	0.61	0.55	10.70
DAE C48 and DEG	0.02	0.12	0.00	0.00	0.14
DAE S48 and DEG	0.02	0.08	0.04	0.00	0.14
DAE C48 only	1.64	1.44	0.16	0.18	3.41
DAE S48 only	4.42	2.70	0.39	0.26	7.78
DEG only	1.09	0.93	0.28	0.12	2.41

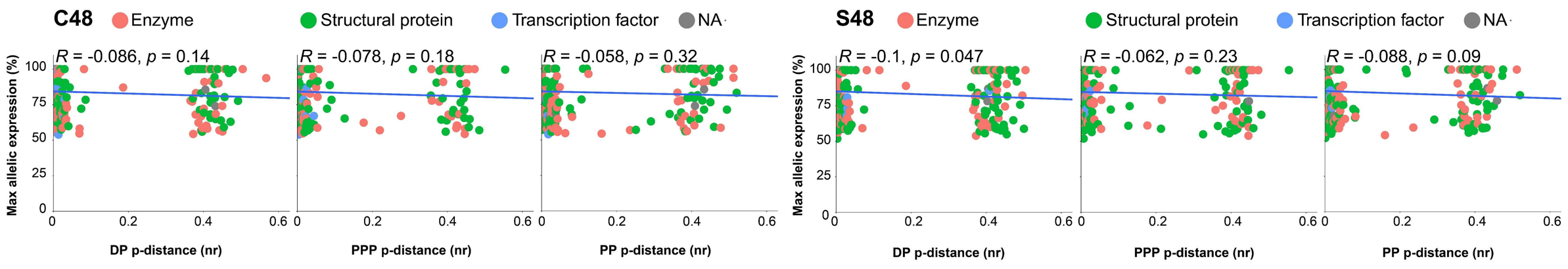

Figure 1 - Correlation between p-distance values of three 1,000 bp-defined promoter regions (PP = proximal, PPP = peri-proximal, DP = distal) and the maximum allelic expression value among allelic gene pairs in the fig genome. Data for control (C48) and treated plants (S48) are presented, with different colors indicating functional characterization.

CONCLUSIONS

This genome-wide analysis represents a preliminary step towards understanding the broader implications of genome heterozygosity. In conclusion, although limited to one treatment (48 days of salt stress), our findings evidence that, at genome-wide level, a significant fraction of heterozygous genes show DAE. In these conditions, the occurrence of sequence variations in the proximal promoter regions does not appear, on average, indispensable in determining DAE. However, a number of analyses, in different tissues and environmental conditions, are necessary to deduce a general rule. Moreover, it remains to be studied if such sequence variations in promoters does affect DAE in specific gene families. Finally, investigations are required to evaluate the contributions of distant enhancers and epigenetic changes to DAE.

MATERIALS & METHODS

1) Assembly process: Phalcon-phase (Kronenberg *et al*., 2021)
2) Scaffolding process: SALSA2 (Ghurye *et al*., 2017)
3) Gene prediction: EVidenceModeler (Haas *et al*., 2008)
4) Gene annotation: Blast2GO (Conesa *et al*., 2005)
5) Synteny analysis: Zhou *et al*. (2020) pipeline
6) Differential expression analysis: Kallisto (Bray *et al*., 2016)

REFERENCES

Bray *et al*. (2005) Nat. Biotechnol.; Conesa *et al*. (2005) Bioinformatics; Ghurye *et al*. (2017) BMC Genomics; Kronenberg *et al.* (2021) Nat. Commun.; Usai *et al*. (2019) Plant J.; Vangelisti *et al*. (2019) Sci. Rep.; Yuan *et al*. (2021) Plant Biotechnol. J.; Zhou *et al*. (2020) Nat. Genet.

The PRIMA programme is supported under Horizon 2020, the European Union's Framework Programme for Research and Innovation.

66th ANNUAL CONGRESS ITALIAN SOCIETY OF AGRICULTURAL GENETICS
Bari, 5-8 September 2023