Agrochimica, Vol. 65 - Special Issue (2021)

Genomics and breeding of the fig tree, an ancient crop with promising perspectives

Gabriele Usai[1,*], Tommaso Giordani[1], Marco Castellacci[1], Alberto Vangelisti[1], Flavia Mascagni[1], Maria Ventimiglia[1], Samuel Simoni[1], Lucia Natali[1], Andrea Cavallini[1]

[1] Department of Agriculture, Food and Environment, University of Pisa, Pisa, Italy

Keywords: breeding, fig tree, fig genome, genome assembly, genetic variability

Abstract: The availability of the genome sequence is a key prerequisite to apply modern breeding procedures to crops, and it is increasingly important to obtain the genomic variation data between the two haplotypes, representing a pivotal resource to study allele-specific expression. The fig tree (*Ficus carica* L.) has a great potential for commercial expansion thanks to its esteemed nutritional and nutraceutical characteristics, combined with its ability to adapt well to difficult environmental conditions. In this work, the fig genome represented the starting point to identifying intergenic and intragenic structural variations to better understand their functional impact. 540 syntenic regions were detected, corresponding to 95% of the fig genome. 2,700,243 single nucleotide polymorphisms (SNPs), 1,488,669 insertions/deletions (INDELs) and 8,360 structural variations (SVs) were identified between the syntenic regions. Overall, the intragenomic diversity was estimated at around 0.4%. 22,120 gene pairs were considered reliable allelic genes. Of these, 15,927 gene pairs showed genetic mutations, including presumed high impact mutations that were identified on 5,997 gene pairs. Specifically, a total of 230,612 mutations were identified, divided into 121,028 SNPs (52.48%) and 109,584 INDELs (47.52%). Most of these mutations were identified within the intronic regions (42.84%), with the remaining ones located downstream of genes (24.99%), upstream of genes (18.31%), in exonic regions (12.73%), and in splice sites (1.13%). Considering mutations in coding regions, 18,047 mutations (54.48%) were classified as missense, 14,875 (44.9%) were classified as synonymous and 204 (0.62%) were classified as nonsense. These genomic resources will be the prerequisite for genome-wide association studies (GWAS), with the ultimate aim of uncovering genes linked to traits of agronomic interest and environmental adaptation to be used in the genetic improvement of the fig tree.

[*] Corresponding author: gabriele.usai@agr.unipi.it

DOI 10.12871/000218572022012

Introduction

Ficus carica L. (the common fig) represents one of the most important commercial species belonging to the Moraceae family. This species is characterized by high resilience to environmental changes (Vangelisti et al., 2019) and exceptional nutritional and pharmacological activities (Veberic et al., 2008). However, fig fruit production has drastically decreased in recent decades mainly due to the rapid ripening of the fruits which makes them poorly resistant to handling and transport over the long-term distance. Moreover, in recent years, the number of fig varieties have been increasingly reducing due to biotic and abiotic stresses, intensive urbanization, single-variety crops, and migration from rural to urban areas.

The fig tree has undergone a low level of genetic improvement and most of the production is still based on old accessions, grown locally, the result of the empirical selection made by farmers, showing phenotypic plasticity to the different environmental conditions. Over 80% of fig cultivation depends on the availability of rain and, consequently, the sustainability of this crop in the context of climate change could be jeopardized.

The ecological characteristics of the fig tree make it a promising species and at the same time highlight critical issues that require applied research and genetic improvement interventions. In this sense, research in the genomics field is central, starting with the sequencing and characterization of the genome, which will allow modern and effective approaches to overcome the various critical issues. In particular, it is increasingly important to obtain the genomic variation data between the haplotypes that make up the genome, a fundamental resource for studying allele-specific variability and its functional implications. This is particularly crucial for fruit trees, such as the fig, whose heterozygous condition is maintained through clonal propagation.

After sequencing and characterization of the fig genome (Usai et al., 2020), the work has focused on the identification of the genomic variations at intergenic and intragenic levels, thus obtaining a solid knowledge for subsequent genotyping by sequencing (GBS) approaches and genome-wide association studies (GWAS) to characterize the hypothetical impact of these variations from a functional point of view.

Materials and Methods

Identification of syntenic regions and construction of the gene map
The set of phased contigs of an updated version of the fig genome (Usai et al., in preparation) allowed us to perform a preliminary identification of syntenic regions. Syntenic regions are defined as regions of chromosomes that share homologous genes deriving from a common ancestor. First, all the CDS sequences were subjected to alignment, all vs. all, using the BLAST program with default parameters (Atschul et al., 1990). After that, MCScanX (Wang et al., 2012) was used to identify the syntenic regions with the following parameters: match_score 50; match_size 4; gap_penalty -1; overlap_window 5; evalue 1e-05; max_gaps 25.

Furthermore, a gene map was obtained implementing the pipeline proposed by Zhou et al. (2020). The procedure involved python reformatting and filtering steps of the output produced by MCScanX, resulting in the number of associated genes and the specific allelic gene pairs, representing the final gene map.

Genomic variation analysis between syntenic regions
The genomic variation analysis between the syntenic regions was carried out using LASTZ (Harris, 2007) and MUMmer (Kurtz et al., 2004) programs. LASTZ was used to align the syntenic regions and identify single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs; length between 1 and 50 bases), while MUMmer was used to identify large structural variations (SVs; length greater than 50 bases). The programs were joined by three main scripts (Zhou et al., 2020). The first script allowed to align the syntenic regions. The second script allowed to identify SNPs, INDELs and SVs resulting from the alignments. The third script allowed to calculate the genomic variations statistics. The programs' parameters were set to default. Finally, we calculated the intragenomic diversity, i.e., the estimated level of heterozygosity indicated as the number of genomic variations every 1,000 bases.

Genetic mutations annotation
The allelic gene pairs of the map were aligned using MUMmer with default parameters. The identified genetic mutations between the gene pairs were annotated using SnpEff (Cingolani et al., 2012) with default parameters. SnpEff annotates the mutations based on their position in intronic and exonic regions, splice sites, UTRs (untranslated regions), upstream and downstream regions. SnpEff distinguishes two types of effects. The first is the impact effect, referred to SNPs, INDELs, and

SVs and specific to the genes. The impact effect is divided into four categories: a) high impact, the mutation is assumed to have a high impact on the protein, including loss of function; b) moderate impact, the mutation is assumed to have a low impact on the protein, but it could modify the function; c) low impact, the mutation is assumed to have a harmless impact on the protein; d) modifier, mutations in non-coding regions where predictions are difficult or there is no evidence of impact. The second is the effect by function, referred only to SNPs and specific for CDS and proteins. The effect by function is divided into three categories: a) non-sense mutation, assigned to point mutations that determine the creation of a new stop codon; b) missense mutation, assigned to point mutations that cause an amino acid change, but not a new stop codon; c) silent mutation, assigned to point mutations that cause a change in the codon, but not a change in the amino acid or a new stop codon.

Results

Identification of syntenic regions and construction of the gene map

Five hundred and forty syntenic regions were identified between the phased fig contigs. In particular, the syntenic regions covered about 95% of the fig genome. Based on the synteny results, 50,894 genes were identified as having homologs between the phased set of contigs, and 44,240 genes, corresponding to 22,120 pairs, were considered reliable allelic genes, representing the gene map of fig. It was not possible to associate 17,092 genes, which will be the subject of future analyzes to evaluate any situations of hemizygosity.

Genomic variation analysis between syntenic regions

The genomic variation analysis on the identified syntenic regions allowed us to locate 2,700,243 SNPs, 1,488,669 INDELs, and 8,360 SVs, for a total of 4,197,272 genomic variations. Finally, the level of intragenomic diversity was estimated around 0.42%.

Genetic mutations annotation

The annotation process carried out on the 22,120 gene pairs revealed the presence of mutations on 15,927 gene pairs. Of the pairs, 5,997 showed mutations that are presumed to have a high effect on proteins, including loss of function. 6,193 gene pairs resulted free of mutations.

From a quantitative point of view, 230,612 total genetic mutations were identified, divided into 121,028 SNPs (52.48%), 54,806 insertions (23.77%) and 54,778 deletions (23.75%). The mutations were clustered

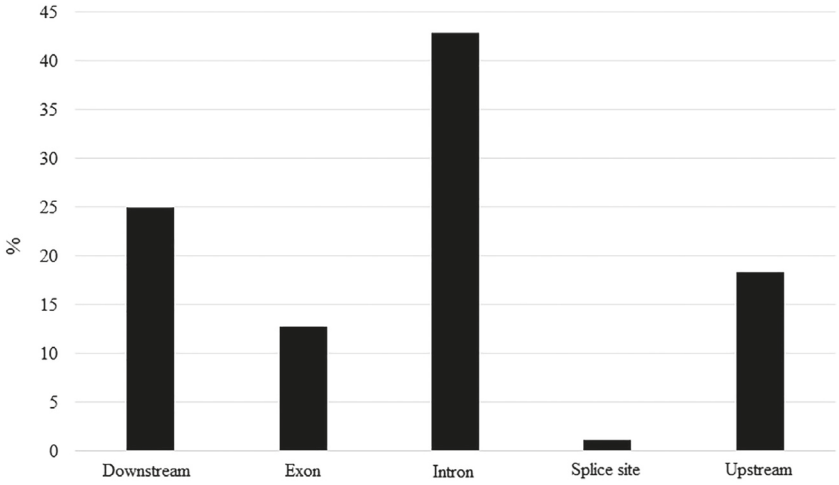

Fig. 1. Distribution of genetic mutations between the mutated allelic gene pairs of fig in downstream, exonic, intronic, splice site, and upstream regions.

into five genomic regions: downstream of genes, on exonic regions, on intronic regions, on splice sites and upstream of genes. Most of these mutations were identified within the intronic regions (42.84%), with the remaining ones located downstream of genes (24.99%), upstream of genes (18.31%), in exonic regions (12.73%), and splice sites (1.13%) (Fig. 1).

Genetic mutations were then classified based on their impact effect. A total of 18,750 mutations were classified as having a high impact on proteins, 18,846 mutations were classified as having a moderate impact, 18,743 mutations were classified as having a low impact, and 350,376 mutations were classified as modifiers (no impact or non-computable impact). It is important to underline that the reported counts refer to all the transcripts of each gene in the map. This means that if a gene encodes 3 transcripts that give rise to 3 isoforms, each sequence mutation in that gene will have 3 calculated consequences in total, one for each isoform. Additionally, a mutation can affect multiple genes, for example, a mutation can be both upstream of one gene and downstream of another gene.

Finally, genetic mutations were also classified according to their effect by function, considering only SNPs located on the CDS sequences of the allelic gene pairs, thus giving a more specific description of the

effects on the protein structures. Overall, 18,047 mutations (54.48%) were classified as missense, 14,875 mutations (44.9%) were classified as synonymous, and 204 mutations (0.62%) were classified as nonsense. Also in this case, these observations apply to all the possible isoforms.

Discussion

Over the years and with the increasing efficiency of sequencing technologies, the quality of plant genomes has undoubtedly increased, but only recently these technologies have begun to allow the production of genomic assemblies with separated haplotypes (Michael and VanBuren, 2020). As far as we know, the published haplotype-phased plant genomes available to date are diploid potato (Zhou et al., 2020), vanilla (Hasing et al., 2020), and hydrangea (Nashima et al., 2021).

The characterization work carried out by Zhou et al. (2020) on diploid potato is certainly the one that comes closest to what was done in our work. The genome of the diploid potato consisted of about 1.6 billion bases, 800 million bases per haplotype, distributed over a chromosomal set of twelve pairs. The difference in size concerning the fig genome, which is about 356 million bases per haplotype on a chromosomal set of thirteen pairs, cannot be neglected, but allowed us to make some considerations.

The reduced fig genome size allowed us to associate approximately 95% of the fig genome in syntenic regions, while in potato approximately 80% was attributed in syntenic regions. This difference is most likely due to the greater genomic rearrangements probably related to the greater quantity of repeated sequences present in potato (Zhou et al., 2020).

Differences are also evident in the number of genomic variations, with 2,700,243 SNPs and 8,360 SVs identified in fig versus 12,299,445 SNPs and 38,999 SVs identified in potato. On the other hand, it is interesting how the number of INDELs was higher in fig than in potato, with an amount of 1.488.669 INDELs against 1.393.680 INDELs, respectively (Zhou et al., 2020).

The estimated intragenomic diversity of fig of about 0.4% was similar to that of palm (0.46%) (Al-Dous et al., 2011), higher than that of poplar (0.26%) (Tuskan et al., 2006), papaya (0.06%) (Ming et al., 2008) and *Prunus mume* (0.03%) (Zhang et al., 2012), but lower than the intragenomic diversity of pear (1.02%) (Wu et al., 2013), jojoba (1.90%) (Liu et al., 2014) and the diploid potato (2.1%) (Zhou et al., 2020), thus confirming the fig as a moderately heterozygous species.

Based on synteny data, 50,894 genes were identified as having homologs between the two set of phased contigs, while in potato the number of homologous genes identified was 59,907 (Zhou et al., 2020). This similarity could be due to the same pipeline implementation, with the same levels of stringency and filtering, but it could also represent the sharing of homologous genes which evolutionarily are expected to be shared in plant genomes (Simão et al., 2015). Similar results were obtained also regarding the allelic gene pair identification, with 22,120 gene pairs identified in fig against the 20,583 gene pairs associated in potato, representing the respective gene maps. Finally, 15,927 fig gene pairs and 17,092 potato gene pairs, respectively, showed genetic mutations with different levels of impact (even high) on proteins.

In further analysis, the updated fig genome along with the genomic variation data produced in this work will be the basis for evaluating the genetic variability of available figs varieties belonging to Spanish, Tunisian and Turkish collections through a GBS-based GWAS analysis to discover genes or molecular markers linked to traits of agronomic interest and environmental adaptation both in the perspective of climate change and for the genetic improvement of the fig tree.

Acknowledgements

This research was supported by the FIGGEN project. FIGGEN is part of the PRIMA Programme supported under Horizon 2020, the European Union's Framework Programme for Research and Innovation.

REFERENCES

Al-Dous E.K., George B., Al-Mahmoud M.E., Al-Jaber M.Y., Wang H., Salameh Y.M. et al. (2011). *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat. Biotechnol. 29, 521.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Cingolani P., Platts A., Wang L.L., Coon M., Nguyen T., Wang L. et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6, 80-92.

Harris R.S. (2007). *Improved Pairwise Alignment of Genomic DNA*. Ph.D. Thesis, The Pennsylvania State University, USA.

Hasing T., Tang H., Brym M., Khazi F., Huang T., Chambers A.H. (2020). A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. Nat. Food 1, 811-819.

Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C. et al. (2004). Versatile and open software for comparing large genomes. Genome Biol. 5, 1-9.

Liu M.J., Zhao J., Cai Q.L., Liu G.C., Wang J.R., Zhao Z.H. et al. (2014). The complex jujube genome provides insights into fruit tree biology. Nat. Commun. 5, 1-12.

Michael T.P., VanBuren R. (2020). Building near-complete plant genomes. Curr. Opin. Plant Biol. 54, 26-33.

Ming R., Hou S., Feng Y., Yu Q., Dionne-Laporte A., Saw J.H. et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452, 991-996.

Nashima K., Shirasawa K., Ghelfi A., Hirakawa H., Isobe S., Suyama T. et al. (2021). Genome sequence of *Hydrangea macrophylla* and its application in analysis of the double flower phenotype. DNA Res. 28, 1-10.

Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210-3212.

Tuskan G.A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U. et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313, 1596-1604.

Usai G., Mascagni F., Giordani T., Vangelisti A., Bosi E., Zuccolo A. et al. (2020). Epigenetic patterns within the haplotype phased fig (*Ficus carica* L.) genome. Plant J. 102, 600-614.

Vangelisti A., Zambrano L.S., Caruso G., Macheda D., Bernardi R., Usai G. et al. How an ancient, salt-tolerant fruit crop, *Ficus carica* L., copes with salinity: a transcriptome analysis. Sci. Rep. 9, 1-13 (2019).

Veberic R., Colaric M., Stampar F. (2008). Phenolic acids and flavonoids of fig fruit (*Ficus carica* L.) in the northern Mediterranean region. Food Chem. 106, 153-157.

Wang Y. Tang H. DeBarry J.D., Tan X., Li J., Wang X. et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, 49.

Wu J., Wang Z., Shi Z., Zhang S., Ming R., Zhu S. et al. (2013). The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res. 23, 396-408.

Zhang Q., Chen W., Sun L., Zhao F., Huang B., Yang W. et al. (2012). The genome of *Prunus mume*. Nat. Commun. 3, 1-8.

Zhou Q., Tang D., Huang W., Yang Z., Zhang Y., Hamilton J.P. et al. (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat. Genet. 52, 1018-1023.