Haplotype-phased genome assembly for *Ficus carica* breeding

T. Giordani[a], G. Usai, M. Castellacci, A. Vangelisti, F. Mascagni, M. Ventimiglia, S. Simoni, L. Natali and A. Cavallini

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

Abstract

The availability of the genome sequence is a fundamental prerequisite to applying modern breeding procedures to crops. It is increasingly important to obtain information on the variations between the two haplotypes, which represent a fundamental resource for studying allele-specific expression. Fig (*Ficus carica* L.) has great potential for commercial expansion due to its valued nutritional and nutraceutical characteristics, along with its ability to adapt well to harsh environmental conditions. In this work, the fig genome was the starting point to identify intergenic and intragenic structural variations better to understand their impact from a functional point of view. A total of 2,700,243 single nucleotide polymorphisms (SNPs), 1,488,669 insertions/deletions (INDELs), and 8,360 structural variations (SVs) were identified. Overall, intragenomic diversity was estimated to be approximately 0.4%. 540 syntenic regions were identified, corresponding to approximately 95% of the fig genome. Among the syntenic regions, 22,120 gene pairs were considered reliable allelic genes. Of these, 15,927 gene pairs showed genetic mutations, including putative high-impact mutations identified in 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 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 at splice sites (1.13%). Considering mutations in coding regions, 18,047 missense mutations (54.48%), 14,875 synonymous mutations (44.9%), and 204 nonsense mutations (0.62%) were classified. These genomic resources will be the prerequisite for genome-wide association studies (GWAS) to identify genes linked to traits of agronomic interest and environmental adaptation for use in fig genetic improvement.

Keywords: fig tree, genome assembly, genetic variability

INTRODUCTION

Ficus carica L. (the common fig) represents one of the most important commercial species in 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, making them poorly resistant to handling and transport over long distances. Moreover, in recent years, the number of fig cultivars has increased 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. 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; consequently, this crop's sustainability 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

[a]E-mail: tommaso.giordani@unipi.it

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., 2017, 2020, 2021a, b), 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

Genomic variation analysis

The set of phased contigs from an updated version of the fig genome (Usai et al., in preparation) was used to identify and categorize global genomic variation. To do this, the programs LASTZ (Harris, 2007) and MUMmer (Kurtz et al., 2004) were implemented. LASTZ was used to align homologous contigs 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). Programs were concatenated into a single pipeline using internal scripts. The parameters of the programs were set to default.

Finally, we calculated intragenomic diversity, the estimated level of heterozygosity shown as the number of genomic variations per 1,000 bases.

Identification of syntenic regions and construction of the gene map

The set of phased contigs was analyzed for preliminary identification of syntenic regions. Syntenic regions are regions of chromosomes that share homologous genes derived from a common ancestor. First, all fig CDS sequences were subjected to alignment, all against all, using the BLAST program with default parameters (Altschul et al., 1990). Next, MCScanX (Wang et al., 2012) was used to identify syntenic regions with the following parameters: match_score 50; match_size 4; gap_penalty-1; overlap_window 5; evalue 1e-05; max_gaps 25.

In addition, a gene map was obtained by implementing the pipeline proposed by Zhou et al. (2020). The procedure involved reformatting and filtering the output produced by MCScanX, leading to the identification of homologous genes and, in particular, allelic gene pairs, representing the final gene map.

Genetic mutations annotation

Identified allelic gene pairs were aligned using MUMmer with default parameters. Genetic mutations identified between gene pairs were annotated using the SnpEff program (Cingolani et al., 2012) with default parameters. SnpEff annotates mutations based on their location within intronic regions, exonic regions, splice sites, UTRs (untranslated regions), and upstream and downstream regions of the genes. SnpEff distinguishes between two types of effects.

The first is the effect by impact, referring to SNPs, INDELs, and SVs and specific to genes. The effect by impact is divided into four categories: a) high impact, where the mutation is assumed to have a high impact on the protein, including loss of function; b) moderate impact, where the mutation is assumed to have a low impact on the protein, but may change the function; c) low impact, where the mutation is assumed to have a harmless impact on the protein; d) modifier, i.e., mutations in non-coding regions where predictions are difficult, or there is no evidence of impact.

The second is the effect by function, referring only to SNPs and specific to CDSs and

proteins. The effect by function is divided into three categories: a) nonsense mutation, assigned to point mutations that result in 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 (synonymous) mutation, assigned to point mutations that cause a change in the codon, but not an amino acid change or a new stop codon.

RESULTS AND DISCUSSION

Over the years, with the increasing efficiency of sequencing technologies, the quality of plant genomes has undoubtedly increased. However, only recently have these technologies begun to enable genome assemblies with separate haplotypes (Michael and Van Buren, 2020). To the best of our knowledge, plant genomes with published haplotypes available to date are diploid potato (Zhou et al., 2020), vanilla (Hasing et al., 2020), and hydrangea (Nashima et al., 2021).

The characterization work of Zhou et al. (2020) on diploid potato is certainly the closest to what is reported in this paper. The diploid potato genome was composed of approximately 1.6 billion bases, 800 million bases per haplotype, distributed over a chromosomal set of 12 pairs. The difference in size from the fig genome, about 356 million bases per haplotype on a chromosomal set of 13 pairs, cannot be overlooked but allowed us to make some considerations. The difference in size is most likely due to the larger genomic rearrangements related to the higher amount of repeat sequences present in the potato genome (Zhou et al., 2020).

Analysis of genomic variation 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. Differences from the diploid potato genome in SNPs and SVs were evident, with 12,299,445 SNPs and 38,999 SVs identified. On the other hand, it is interesting how the number of INDELs is higher in fig than in diploid potato, with an amount of 1,488,669 INDELs versus 1,393,680 INDELs, respectively (Zhou et al., 2020).

The estimated intragenomic diversity of fig was about 0.4%, similar to the estimated diversity 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), and lower than that estimated for pear (1.02%) (Wu et al., 2013), jojoba (1.90%) (Liu et al., 2014), and diploid potato (2.1%) (Zhou et al., 2020). This information confirmed fig as a moderately heterozygous species.

The small size of the fig genome allowed us to associate approximately 95% of the fig genome, distributed in 540 syntenic regions. In contrast, approximately 80% of the genome in diploid potato was allocated in syntenic regions. Based on the synteny results, 50,894 genes were identified as homologous among the phased set of contigs, and 44,240 genes, corresponding to 22,120 pairs, were considered reliable allelic genes. These gene pairs represented the gene map of fig. However, 17,092 genes could not be associated. These genes will be the subject of future analysis and manual curation to recover the remaining allelic gene pairs and evaluate possible hemizygosity situations. In diploid potato, 59,907 homologous genes were recovered, whereas the number of allelic gene pairs identified was 20,583 (Zhou et al., 2020). This similarity with our results could be due to the same pipeline implementation, where the same stringency and filtering parameters were used; however, it could also represent the sharing of homologous genes that are evolutionarily expected to be shared in plant genomes (Simão et al., 2015).

Genetic mutation annotation performed on the provisional 22,120 allelic gene pairs revealed the presence of mutations in 15,927 pairs. Of these pairs, 5,997 showed mutations presumed to have a potentially large effect on proteins, including loss of function. For now, 6,193 gene pairs were found to be free of mutations.

Quantitatively, 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%). Mutations were analyzed according to their abundance in five genomic regions: downstream of genes, in exonic regions, intronic regions, splice sites, and upstream of genes. Most of these mutations were identified within the intronic regions (42.84%). The remaining mutations were

identified with decreasing abundance downstream of genes (24.99%), upstream of genes (18.31%), in exonic regions (12.73%) and splice sites (1.13%), respectively (Figure 1).

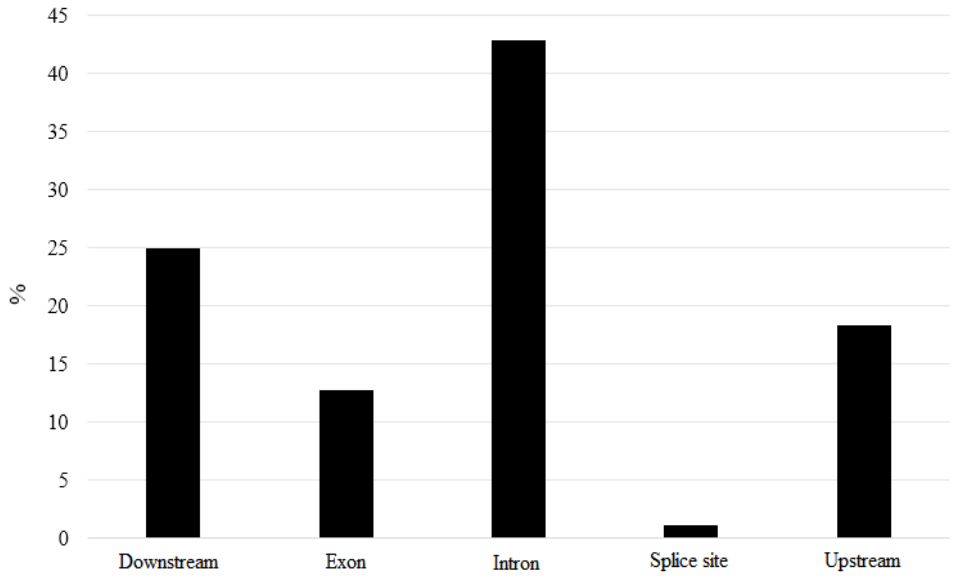

Figure 1. Distribution of genetic mutations between the provisionally identified allelic gene pairs of fig in downstream, exonic, intronic, splice site, and upstream regions.

Genetic mutations were then classified according to their putative impact effect. A total of 19,021 mutations were classified as having a high impact on proteins, 18,858 mutations were classified as having a moderate impact, 19,875 mutations were classified as having a low impact, and 354,698 mutations were classified as modifiers (no impact or impact not calculable) (Table 1). It is important to note that the counts reported here refer to all gene transcripts on 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. Also, a mutation can affect multiple genes; for example, the same mutation can affect both upstream of one gene and downstream of another nearby gene.

Table 1. The number of effects by type and impact.

Type	Impact	Count (nr)	Percent (%)
frameshift_variant	High	17,994	4.362
stop_gained	High	374	0.091
splice_donor_variant	High	286	0.069
splice_acceptor_variant	High	193	0.047
stop_lost	High	141	0.034
start_lost	High	33	0.008
missense_variant	Moderate	17,972	4.357
conservative_inframe_deletion	Moderate	448	0.109
disruptive_inframe_deletion	Moderate	438	0.106
synonymous_variant	Low	14,872	3.606
splice_region_variant	Low	4,999	1.212
stop_retained_variant	Low	3	0.001
initiator_codon_variant	Low	1	0.001
intron_variant	Modifier	178,601	43.3
downstream_gene_variant	Modifier	101,620	24.636
upstream_gene_variant	Modifier	74,477	18.056

Finally, genetic mutations were also classified according to their putative effect by function, in this case considering only SNPs located on the CDS sequences of allelic gene pairs, thus obtaining a more detailed description of the effects on protein structures. Overall, 18,047 mutations (54.48%) were classified as missense, 14,875 mutations (44.9%) as synonymous, and 204 mutations (0.62%) as nonsense. Again, these observations apply to all possible isoforms.

CONCLUSIONS

Once the identification process is complete, we will intersect the genomic variation data obtained in this work with the localization and annotation of allelic gene pairs to assess their distribution at the level of gene families. The process will highlight which functional families are the most conserved and those more prone to accumulate mutations (and what type of mutations), elucidating possible implications from a functional perspective.

Furthermore, the updated version of the fig genome, together with these data, will be the basis for assessing the genetic variability of available fig varieties belonging to Spanish, Tunisian and Turkish collections through a GBS-based GWAS analysis to discover genes or molecular markers related to traits of agronomic interest and environmental adaptation both in the perspective of climate change and for fig genetic improvement.

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.

Literature cited

Al-Dous, E.K., George, B., Al-Mahmoud, M.E., Al-Jaber, M.Y., Wang, H., Salameh, Y.M., Al-Azwani, E.K., Chaluvadi, S., Pontaroli, A.C., DeBarry, J., et al. (2011). *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat Biotechnol *29* (*6*), 521–527 https://doi.org/10.1038/nbt.1860. PubMed

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215* (*3*), 403–410 https://doi.org/10.1016/S0022-2836(05)80360-2. PubMed

Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6* (*2*), 80–92 https://doi.org/10.4161/fly.19695. PubMed

Harris, R.S. (2007). Improved Pairwise Alignment of Genomic DNA (The Pennsylvania State University).

Hasing, T., Tang, H., Brym, M., Khazi, F., Huang, T., and Chambers, A.H. (2020). A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. Nat. Food *1* (*12*), 811–819 https://doi.org/10.1038/s43016-020-00197-2.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol *5* (*2*), R12 https://doi.org/10.1186/gb-2004-5-2-r12. PubMed

Liu, M.J., Zhao, J., Cai, Q.L., Liu, G.C., Wang, J.R., Zhao, Z.H., Liu, P., Dai, L., Yan, G., Wang, W.J., et al. (2014). The complex jujube genome provides insights into fruit tree biology. Nat Commun *5* (*1*), 5315 https://doi.org/10.1038/ncomms6315. PubMed

Michael, T.P., and VanBuren, R. (2020). Building near-complete plant genomes. Curr Opin Plant Biol *54*, 26–33 https://doi.org/10.1016/j.pbi.2019.12.009. PubMed

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature *452* (*7190*), 991–996 https://doi.org/10.1038/nature06856. PubMed

Nashima, K., Shirasawa, K., Ghelfi, A., Hirakawa, H., Isobe, S., Suyama, T., Wada, T., Kurokura, T., Uemachi, T., Azuma, M., et al. (2021). Genome sequence of *Hydrangea macrophylla* and its application in analysis of the double flower phenotype. DNA Res *28* (*1*), dsaa026 https://doi.org/10.1093/dnares/dsaa026. PubMed

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31* (*19*), 3210–3212 https://doi.org/10.1093/bioinformatics/btv351. PubMed

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science *313* (*5793*), 1596–1604 https://doi.org/10.1126/science.1128691. PubMed

Usai, G., Vangelisti, A., Solorzano Zambrano, L., Mascagni, F., Giordani, T., Cavallini, A., and Natali, L. (2017). Transcriptome comparison between two fig (*Ficus carica* L.) cultivars. Agrochimica *61*, 340–354 https://doi.org/10.12871/00021857201735.

Usai, G., Mascagni, F., Giordani, T., Vangelisti, A., Bosi, E., Zuccolo, A., Ceccarelli, M., King, R., Hassani-Pak, K., Zambrano, L.S., et al. (2020). Epigenetic patterns within the haplotype phased fig (*Ficus carica* L.) genome. Plant J *102* (*3*), 600–614 https://doi.org/10.1111/tpj.14635. PubMed

Usai, G., Mascagni, F., Giordani, T., Vangelisti, A., Bosi, E., Zuccolo, A., Ceccarelli, M., King, R., Hassani-Pak, K., Solorzano Zambrano, L., et al. (2021a). High-quality, haplotype-phased de novo assembly of the highly heterozygous fig genome, a major genetic resource for fig breeding. Acta Hortic. *1310*, 21–28 https://doi.org/10.17660/ActaHortic.2021.1310.4.

Usai, G., Vangelisti, A., Simoni, S., Giordani, T., Natali, L., Cavallini, A., and Mascagni, F. (2021b). DNA modification patterns within the transposable elements of the fig (*Ficus carica* L.) genome. Plants (Basel) *10* (*3*), 451 https://doi.org/10.3390/plants10030451. PubMed

Vangelisti, A., Zambrano, L.S., Caruso, G., Macheda, D., Bernardi, R., Usai, G., Mascagni, F., Giordani, T., Gucci, R., Cavallini, A., and Natali, L. (2019). How an ancient, salt-tolerant fruit crop, *Ficus carica* L., copes with salinity: a transcriptome analysis. Sci Rep *9* (*1*), 2561 https://doi.org/10.1038/s41598-019-39114-4. PubMed

Veberic, R., Colaric, M., and Stampar, F. (2008). Phenolic acids and flavonoids of fig fruit (*Ficus carica* L.) in the northern Mediterranean region. Food Chem. *106* (*1*), 153–157 https://doi.org/10.1016/j.foodchem.2007.05.061.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res *40* (*7*), e49–e49 https://doi.org/10.1093/nar/gkr1293. PubMed

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., Khan, M.A., Tao, S., Korban, S.S., Wang, H., et al. (2013). The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res *23* (*2*), 396–408 https://doi.org/10.1101/gr.144311.112. PubMed

Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W., Tao, Y., Wang, J., Yuan, Z., Fan, G., et al. (2012). The genome of *Prunus mume.* Nat Commun *3* (*1*), 1318 https://doi.org/10.1038/ncomms2290. PubMed

Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J.P., Visser, R.G.F., Bachem, C.W.B., Robin Buell, C., Zhang, Z., et al. (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat Genet *52* (*10*), 1018–1023 https://doi.org/10.1038/s41588-020-0699-x. PubMed