High-quality, haplotype-phased de novo assembly of the highly heterozygous fig genome, a major genetic resource for fig breeding

G. Usai[1], F. Mascagni[1], T. Giordani[1], A. Vangelisti[1], E. Bosi[2], A. Zuccolo[3], M. Ceccarelli[4], R. King[5], K. Hassani-Pak[5], L. Solorzano Zambrano[6], A. Cavallini[1,a] and L. Natali[1]

[1]Department of Agriculture, Food and Environment, University of Pisa, Pisa, Italy; [2]Department of Biomedical Experimental and Clinical Sciences, University of Florence, Florence, Italy; [3]Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy; [4]Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy; [5]Rothamsted Research, Harpenden, United Kingdom; [6]Facultad de Ciencias Zootécnicas, Universidad Técnica de Manabí, Portoviejo, Ecuador.

Abstract

The genome assembly of allogamous perennial species can be very challenging due to the high heterozygosity and repeat content they present. In fruit trees, many important phenotypic traits of a specific genotype lie in its heterozigosity, maintained by a widespread clonal propagation. The fig tree (*Ficus carica* L.) has a great potential for expansion thanks to valuable nutritional and nutraceutical characteristics, combined with the ability to adapt well to marginal soils and difficult environmental conditions. However, the fig is still poorly characterized at genomic level, and only a preliminary genome sequence (of the Japanese cultivar 'Horaishi') has been released. Here we report a de novo high-quality assembly of the typical Italian fig cultivar 'Dottato' obtained by single-molecule, real-time sequencing (SMRT). PacBio reads (with average length of 12,364 nt and corresponding to about 74 genome equivalents) allowed us to obtain sequence contiguity and resolve the repetitive component of the genome. The assembly, of approximately 333 Mb and N50 of 823 kb, was haplotype-phased using FALCON-Unzip and it is composed by 905 sequences of which 407 were arranged in 13 chromosome-related pseudomolecules. This new reference genome improved the assembly N50 of the previous short-read based fig assembly of about 5-fold. A curated genome annotation analysis resulted in the identification of 37,840 protein-coding genes and 1,685 non-coding genes. Furthermore, we found that the amount of repetitive sequences accounted for the 37.39% of the assembly. The production of a high-quality haplotype-phased reference genome sequence of fig offers interesting insights into the genomics structure of this species, opening great opportunities for speeding up the development of new cultivars and for the application to this species of genome editing, a technology which seems especially suitable to change the specific traits that are currently limiting the success of this ancient species.

Keywords: *Ficus carica*, genome annotation, genome assembly, single-molecule real-time sequencing

INTRODUCTION

Ficus is one of the 37 genera of the *Moraceae* family. *Ficus carica* L., or common fig, is the most commercially important species of the genera, being widely grown throughout the temperate world for its fruit. *F. carica* is a highly heterozygous diploid species (2*n*=2*x*=26), with a genome size of about 0.36 pg/2C (Mori et al., 2017).

In recent years, large interest arose on fig thanks to its valuable nutritional and nutraceutical characteristics (Solomon et al., 2006; Veberic et al., 2008) and the ability to adapt to marginal soils and difficult environmental conditions (Vangelisti et al., 2019). As a matter of fact, the pharmaceutical industry is paying more attention to species like fig,

[a]E-mail: andrea.cavallini@unipi.it

considered a very important resource for medicine development (Cavero et al., 2013). Thus, *F. carica* has been included in occidental Pharmacopoeias and in therapeutic guides of medicines (Barolo et al., 2014).

However, despite its economic, cultural and ecological importance, fig is still a poorly characterized species at both genetic and genomic level. A high-quality reference genome would provide an important resource for genetic improvement and breeding programmes, as a source of information regarding genes involved in agronomic and productive traits, for example the rapid perishability of fruits and biotic or abiotic stresses which is limiting fig fruits world distribution and commercial success. To date, only a preliminary genome sequence (of the Japanese cultivar 'Horaishi') has been released (Mori et al., 2017), produced with short-read sequencing and so affected by the typical deficiencies of this technology.

Over the past years, the emergence of third generation technologies (Ansorge, 2009), including single-molecule, real-time (SMRT) sequencing, has led to the production of high-quality genome assemblies (Vogel et al., 2018; Ye et al., 2019). Long-read sequencing has helped to overcome the problems of short-read sequencing, i.e., resolution of repetitive sequences along the assembly (Veeckman et al., 2016) or the heterozygosity ambiguity (Low et al., 2019). Actually, resolving heterozygosity ambiguity by generating a haplotype-phased reference genome is necessary for a proper comprehension of species biology but also to manage genetic diversity (Meuwissen et al., 2013). In addition, collapsing different haplotypes in genome assemblies can result in sequence errors due to differences between homologous chromosomes (Korlach et al., 2017) and lead to gene annotation issues, as well as breaks in the genome due to differences of the repeat content in intergenic regions between haplotypes. In this work, we present a haplotype-phased reference genome of the most important Italian fig cultivar, 'Dottato', providing a high-quality genomic resource for future genetic studies, potentially useful for the breeding of this promising commercial crop.

MATERIALS AND METHODS

DNA extraction and genome sequencing

Young buds (0.5-1 cm in diameter) were collected from a single female plant of the Italian *F. carica* cultivar 'Dottato'. Genomic DNA was isolated by using a standard protocol (Mascagni et al., 2018). SMRT sequencing was performed on a Pacific Biosciences (PacBio) Sequel system (V2 chemistry) at the Arizona Genomics Institute (Tucson, AZ, USA). A total of 6 SMRT cells were sequenced in accordance with the manufacturer protocols (Pacific Biosciences).

Genome assembly and curation

The assembly of the PacBio reads was performed using FALCON-Unzip v0.5 (Chin et al., 2016). We set up specific parameters by comparing different length_cutoff and seed_coverage values. The assembly was scaffolded using FinisherSC v2.0 (Lam et al., 2015) and polished using Arrow v2.2.2 (Chin et al., 2013) and Pilon v1.13 (Walker et al., 2014). Pilon was run using Illumina HiSeq 2000 paired-end reads (Solorzano Zambrano et al., 2017). After that, MEGAN v6.5.8 (Huson et al., 2007) was used to verify fungal and bacterial contamination. BUSCO v3.0.2 (Simão et al., 2015) was used to measure the gene completeness. To order and orientate the contigs into the 13 chromosomes of fig we used the Illumina scaffolds of the *F. carica* 'Horaishi' assembly, ordered by SSRs and SNPs map (Mori et al., 2017).

Gene prediction and annotation

Structural prediction and annotation of protein-coding genes was carried out using MAKER v2.31.10 (Holt and Yandell, 2011) and Blast2GO v5 (Conesa et al., 2005). First, Trinity v2.5.1 (Grabherr et al., 2011) was run to obtain a de novo transcriptome of fig by using already available RNA sequencing data (Vangelisti et al., 2019). AUGUSTUS v3.3.1 (Stanke et al., 2006) was trained by using the transcriptome as evidence. The repeat-masked assembly was used to train GeneMark-ES v4.32 (Ter-Hovhannisyan et al., 2008). The MAKER pipeline was run integrating the trained AUGUSTUS and GeneMark-ES programs by using the transcriptome as

evidence. Predicted genes were functionally annotated by using Blast2GO. The sequences were first searched for homologous sequences by a BLASTP analysis against the whole NCBI non-redundant (nr) database, then GO terms were obtained following the Blast2GO pipeline. The sequences were also submitted to InterPro within Blast2GO (Jones et al., 2014) to annotate with protein domain functional information, and classifying into families. Finally, tRNAscan-SE v2.0 (Lowe and Chan, 2016) and RfamScan v1.1.2 (Kalvari et al., 2018) were run to annotate the non-coding RNAs.

Repeats prediction and annotation

LTRharvest v1.5.10 (Ellinghaus et al., 2008) was run to identify LTR retroelements ranging from 1.5 to 25 kbp. The identified sequences were firstly annotated using LTRdigest v1.5.10 (Steinbiss et al., 2009). After that, LTR retroelements were submitted to the DANTE tool (http://repeatexplorer.org) for further classification via a phylogenetic approach. The Class I SINEs were identified using SINE_scan v1.1.1 (Mao and Wang, 2017). The Class II MITEs were collected using MITE-Hunter v11-2011 (Han and Wessler, 2010). The Class II Helitron elements were collected using HelitronScanner v1.0 (Xiong et al., 2014). All the libraries were used to mask the assembly using RepeatMasker v4.0.3 (Smit et al., 2015). Missed elements were identified by using RepeatModeler v1.0.11 (Smit and Hubley, 2008) on the masked assembly. The library obtained was used to mask the assembly again. Finally, Phobos v3.3.12 (Mayer, 2010) was used to identify tandem repeats (-u 1 -U 500 --minScore 12 --mismatchScore -5 --indelScore -5 -M imperfect -r 5).

Centromeric analysis

Putative centromeric regions were identified by searching for the characteristic repetitions tandemly arranged of these regions. Tandem repeats with a monomer unit length ranging between 80 and 350 bp were considered. The most abundant tandem repeats were used to mask the 13 pseudomolecules by using RepeatMasker. Only the tandem repeat occurring in each pseudomolecule was retained. The regions highly masked by that tandem repeat were considered as putative centromeric regions and were masked with a library of LTR *Chromovirus* elements previously identified to obtain the profile of their abundance in these regions.

RESULTS AND DISCUSSION

The availability of a high-quality reference genome is a prerequisite to accelerate innovation via breeding. With the introduction of high-throughput sequencing technologies, the number of genome sequencing projects has increased, leading to genome assemblies for organisms previously considered too low priority. For instance, high-quality genome assemblies were released for *Potentilla micrantha* (Buti et al., 2018), *Cuscuta campestris* (Vogel et al., 2018) and *Casuarina equisetifolia* (Ye et al., 2019).

The sequencing process generated 2,140,959 long-reads (minimum length = 1,000 bp; maximum length = 111,426 bp; average length = 12,364 bp) with an N50 value of 18,419 bp, corresponding to about 74-fold genome coverage of the 356-Mb fig haploid genome (Mori et al., 2017).

Despite fig has a predicted small genome size (Mori et al., 2017), its high heterozygosity represents a substantial challenge. The number of diploid (haplotype-phased) reference genomes have begun to increase due to leveraging the potential of third generation long-reads which can be used to generate contigs with enough sensitivity to define the alternative haplotypes, as shown for *Durio zibethinus* (Teh et al., 2017), *Scutellaria baicalensis* (Zhao et al., 2018) and *Prunus* × *yedoensis* (Shirasawa et al., 2019).

The FALCON-Unzip assembler produced a primary set of contigs and a set of linked haplotigs that represent alternative genome structures of the primary contigs (i.e., SNPs and structural variations). Only the primary assembly was considered in this work.

Overall, the assembly process produced 905 contigs with mean size of 368,398 bp (min size = 20,012 bp, max size = 5,010,936 bp) and N50 of 823,517 bp. We produced a total of 333,400,567 bp of the fig genome sequence, corresponding to about 95% of the estimated

size. A total of 407 contigs, corresponding to the 80% of the total assembly (266,522,563 bp), were associated to fig chromosomes, producing a set of 13 pseudomolecules.

After excluding fungal and bacterial contamination, BUSCO recovered 1,283 of the 1,375 (93.3%) highly conserved Embryophyta core genes, of which 1,177 (85.6%) were complete and single-copy and 106 (7.7%) were complete and duplicated. Thirty-five genes (2.5%) resulted fragmented and 57 (4.2%) missing.

In total, we identified 37,840 protein-coding genes, which represent 27.91% of the total genome assembly, and 1,685 non-protein-coding genes, representing the 0.15%, respectively (Figure 1).

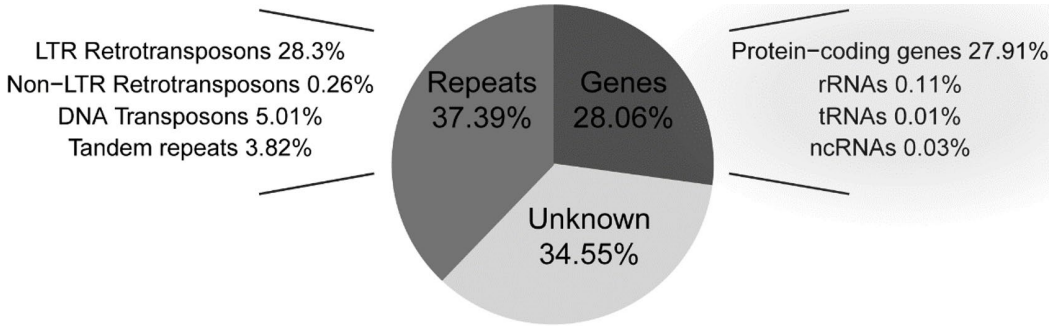

Figure 1. Pie chart showing the percentage of genes (i.e., protein-coding genes, rRNAs, tRNA and ncRNAs) (left), repeats (i.e., LTR retrotransposons, non-LTR retrotransposons, DNA transposons and tandem repeats) (right) and unknown sequences.

These data are similar to those reported by Usai et al. (2017). Gene average length was 2,460 bp and CDS average length was 956 bp. The average exon number per gene was 4.56, with an average length of 251 bp, while intron average length was 367 bp. Functional annotation showed that 28,737 of the predicted protein-coding genes (76% of the total) had a BLAST hit (e-value <0.001) in NCBI nr.

The number of protein-coding genes and their gene ontology (GO) annotation is shown in Figure 2, categorized by the three ontologies cellular component, molecular function and biological process.

Comparing our results to the previous fig genome draft (Mori et al., 2017), it is evident that the use of long-read sequencing and high depth coverage produced a genome of high sequence contiguity and accuracy (Table 1).

Furthermore, a full characterization of the repetitive component is a crucial step to decipher the genome structure and, by using PacBio long-reads we were able to significantly increase TE detection across the assembly. We identified a total of 123.8 Mb of repeat sequences, representing 37.39% of the genome assembly (Figure 1). Transposable elements (TEs) represent the most abundant repeats, covering the 33.57% of the assembly, while tandem repeats represent the 3.82%. The most abundant TEs are retrotransposons or Class I elements, representing 84.95% of the repetitive content and 28.3% of genome assembly. In particular, long-terminal repeat retrotransposons (LTR-REs) are the most represented, accounting for 99.06% of this class and 28.03% of the total genome assembly, whereas non-LTR retrotransposons (LINEs and SINEs) account for 0.94% of Class I. Among the LTR-REs, *Gypsy* and *Copia* are the most abundant superfamilies, representing 16.36 and 8.74% of the genome assembly, respectively. DNA transposons or Class II elements represent 15.05% of the repetitive content and 5.01% of the genome assembly.

Another goal in using long-read sequencing was to characterize the centromeric regions, which remain largely unknown in short-read assemblies, being difficult to assemble due to the highly repetitive composition. Exploiting the contiguity of the 13 pseudomolecules produced, we identified 13 putative centromeric regions and provided one of the first evaluations into the structure of these chromosomal regions.

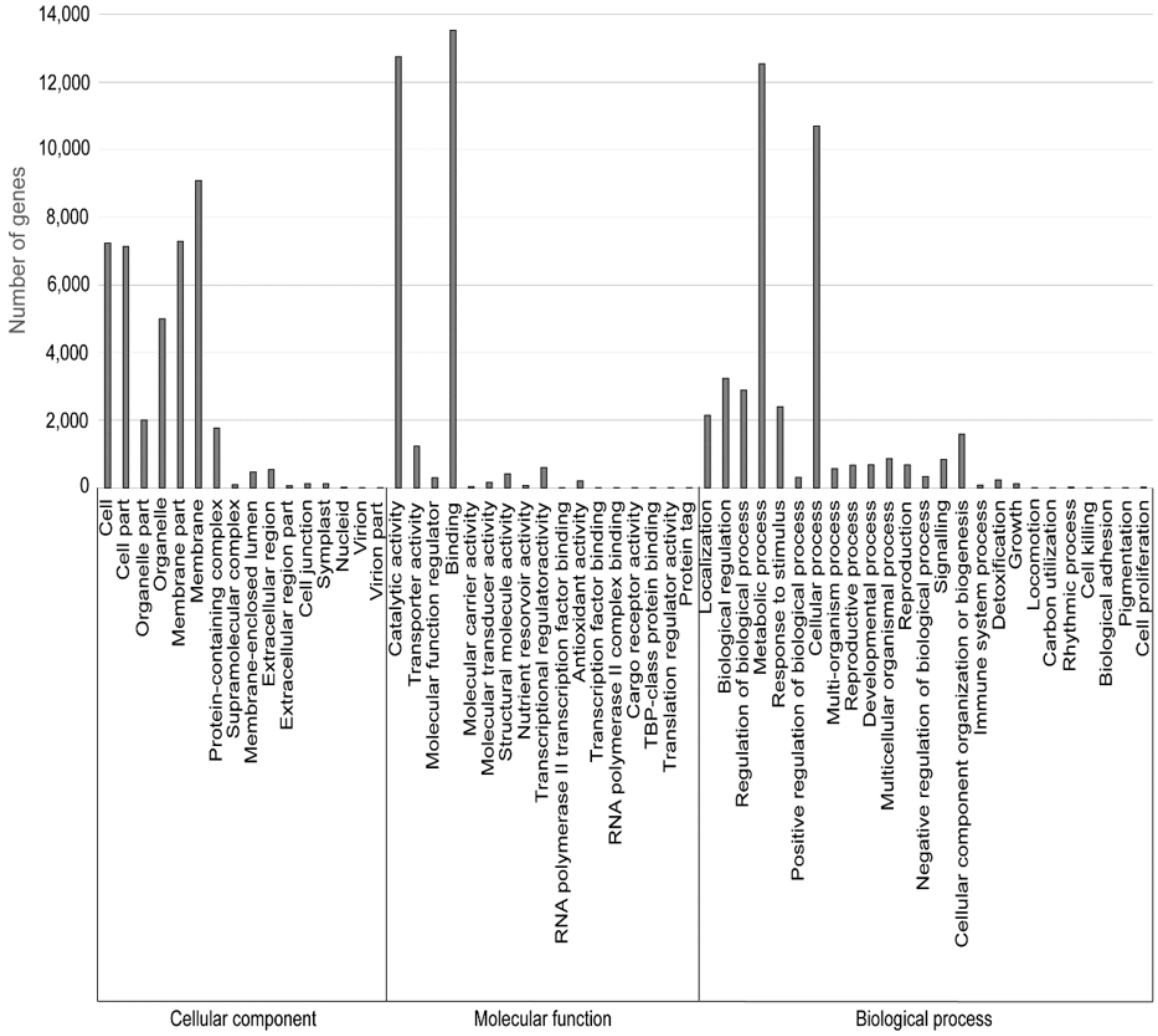

Figure 2. Distribution of GO terms including the number of genes attributed to the GO terms (i.e., cellular component, molecular function and biological process).

Table 1. Comparison between the 'Dottato' and 'Horaishi' assemblies.

	Dottato	Horaishi (Mori et al., 2017)
Genome representation (%)	95	70
Number of sequences (N°)	905	27,995
Total size of the assembly (bp)	333,400,567	247,090,738
Longest sequence (bp)	5,010,936	1,764,766
Shortest sequence (bp)	20,012	479
Number of sequences >10 kbp (N°)	905 (100%)	2,081 (7.4%)
Number of sequences >100 kbp (N°)	595 (65.7%)	671 (2.4%)
Number of sequences >1 Mb (N°)	81 (9.0%)	8 (0.0%)
Mean sequence size (bp)	368,398	8,826
Median sequence size (bp)	167,241	893
N50 sequence length (bp)	823,517	166,092
L50 sequence length (N°)	121	374
N sequence content (%)	0.00	14.72
BUSCO assessment (%)	93.3	90.5

In fact, a site containing highly-repetitive 103 bp-long tandem repeats was identified once for each pseudomolecule and considered as the putative centromeric repeat. This observation was further supported by the high abundance of *Chromovirus* elements, which are typically found in the centromeric structures (Neumann et al., 2011), in the above-mentioned sites. We isolated a total of 42 centromeric contigs, representing the 13 putative centromere regions of fig. Each region was represented by an average number of 3.23 contigs and had an average length size of 2.86 Mb.

CONCLUSIONS

The production of a high-quality reference genome sequence for *Ficus carica* L. presents interesting insights into the genomic structure of a highly heterozygous plant diploid species. Further analysis will be conducted in order to define the sequence of the alternative haplotype and to characterize its heterozygosity at sequence and methylation levels, investigating the effects of allelic changes on gene expression. In addition, both heterozygosity and homozygosity identification in offspring can shorten breeding programs for desirable and commercially known traits.

DATA ACCESSIBILITY

The raw reads, the fig genome and annotations are available from the corresponding author upon request.

ACKNOWLEDGEMENTS

Research work funded by the Department of Agriculture, Food and Environment (University of Pisa) project "Plantomics".

Literature cited

Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. N. Biotechnol. *25* (*4*), 195–203 https://doi.org/10.1016/j.nbt.2008.12.009. PubMed

Barolo, M.I., Ruiz Mostacero, N., and López, S.N. (2014). *Ficus carica* L. (Moraceae): an ancient source of food and health. Food Chem. *164*, 119–127 https://doi.org/10.1016/j.foodchem.2014.04.112. PubMed

Buti, M., Moretto, M., Barghini, E., Mascagni, F., Natali, L., Brilli, M., Lomsadze, A., Sonego, P., Giongo, L., Alonge, M., et al. (2018). The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry). Gigascience *7* (*4*), 1–14 https://doi.org/10.1093/gigascience/giy010. PubMed

Cavero, R.Y., Akerreta, S., and Calvo, M.I. (2013). Medicinal plants used for dermatological affections in Navarra and their pharmacological validation. J. Ethnopharmacol. *149* (*2*), 533–542 https://doi.org/10.1016/j.jep.2013.07.012. PubMed

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10* (*6*), 563–569 https://doi.org/10.1038/nmeth.2474. PubMed

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13* (*12*), 1050–1054 https://doi.org/10.1038/nmeth.4035. PubMed

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21* (*18*), 3674–3676 https://doi.org/10.1093/bioinformatics/bti610. PubMed

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics *9* (*1*), 18 https://doi.org/10.1186/1471-2105-9-18. PubMed

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29* (*7*), 644–652 https://doi.org/10.1038/nbt.1883. PubMed

Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. *38* (*22*), e199–e199 https://doi.org/10.1093/nar/gkq862. PubMed

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for

second-generation genome projects. BMC Bioinformatics *12* (*1*), 491 https://doi.org/10.1186/1471-2105-12-491. PubMed

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17* (*3*), 377–386 https://doi.org/10.1101/gr.5969107. PubMed

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30* (*9*), 1236–1240 https://doi.org/10.1093/bioinformatics/btu031. PubMed

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. *46* (*D1*), D335–D342 https://doi.org/10.1093/nar/gkx1038. PubMed

Korlach, J., Gedman, G., Kingan, S.B., Chin, C.S., Howard, J.T., Audet, J.N., Cantin, L., and Jarvis, E.D. (2017). *De novo* PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience *6* (*10*), 1–16 https://doi.org/10.1093/gigascience/gix085. PubMed

Lam, K.K., LaButti, K., Khalak, A., and Tse, D. (2015). FinisherSC: a repeat-aware tool for upgrading *de novo* assembly using long reads. Bioinformatics *31* (*19*), 3207–3209 https://doi.org/10.1093/bioinformatics/btv280. PubMed

Low, W.Y., Tearle, R., Bickhart, D.M., Rosen, B.D., Kingan, S.B., Swale, T., Thibaud-Nissen, F., Murphy, T.D., Young, R., Lefevre, L., et al. (2019). Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nat. Commun. *10* (*1*), 260 https://doi.org/10.1038/s41467-018-08260-0. PubMed

Lowe, T.M., and Chan, P.P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. *44* (*W1*), W54–W57 https://doi.org/10.1093/nar/gkw413. PubMed

Mao, H., and Wang, H. (2017). SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics *33* (*5*), 743–745 https://doi.org/10.1093/bioinformatics/btw718. PubMed

Mascagni, F., Vangelisti, A., Giordani, T., Cavallini, A., and Natali, L. (2018). Specific LTR-retrotransposons show copy number variations between wild and cultivated sunflowers. Genes (Basel) *9* (*9*), 433 https://doi.org/10.3390/genes9090433. PubMed

Mayer, C. (2010). https://www.rub.de/ecoevo/cm/cm_phobos.htm.

Meuwissen, T., Hayes, B., and Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. Annu. Rev. Anim. Biosci. *1* (*1*), 221–237 https://doi.org/10.1146/annurev-animal-031412-103705. PubMed

Mori, K., Shirasawa, K., Nogata, H., Hirata, C., Tashiro, K., Habu, T., Kim, S., Himeno, S., Kuhara, S., and Ikegami, H. (2017). Identification of RAN1 orthologue associated with sex determination through whole genome sequencing analysis in fig (*Ficus carica* L.). Sci. Rep. *7* (*1*), 41124 https://doi.org/10.1038/srep41124. PubMed

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J., and Macas, J. (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob. DNA *2* (*1*), 4 https://doi.org/10.1186/1759-8753-2-4. PubMed

Shirasawa, K., Esumi, T., Hirakawa, H., Tanaka, H., Itai, A., Ghelfi, A., Nagasaki, H., and Isobe, S. (2019). Phased genome sequence of an interspecific hybrid flowering cherry, 'Somei-Yoshino' (*Cerasus* × *yedoensis*). DNA Res. *26* (*5*), 379–389 https://doi.org/10.1093/dnares/dsz016. 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

Smit, A.F.A., and Hubley, R. (2008). http://www.repeatmasker.org/RepeatModeler.

Smit, A.F.A., Hubley, R., and Green, P. (2015). http://www.repeatmasker.org.

Solomon, A., Golubowicz, S., Yablowicz, Z., Grossman, S., Bergman, M., Gottlieb, H.E., Altman, A., Kerem, Z., and Flaishman, M.A. (2006). Antioxidant activities and anthocyanin content of fresh fruits of common fig (*Ficus carica* L.). J. Agric. Food Chem. *54* (*20*), 7717–7723 https://doi.org/10.1021/jf060497h. PubMed

Solorzano Zambrano, L., Usai, G., Vangelisti, A., Mascagni, F., Giordani, T., Bernardi, R., Cavallini, A., Gucci, R., Caruso, G., D'Onofrio, C., et al. (2017). Cultivar-specific transcriptome prediction and annotation in *Ficus carica* L. Genom. Data *13*, 64–66 https://doi.org/10.1016/j.gdata.2017.07.005. PubMed

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. *34* (*Web Server*), W435–W439 https://doi.org/10.1093/nar/gkl200. PubMed

Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of *de novo*

predicted LTR retrotransposons. Nucleic Acids Res. *37* (*21*), 7002–7013 https://doi.org/10.1093/nar/gkp759. PubMed

Teh, B.T., Lim, K., Yong, C.H., Ng, C.C.Y., Rao, S.R., Rajasegaran, V., Lim, W.K., Ong, C.K., Chan, K., Cheng, V.K.Y., et al. (2017). The draft genome of tropical fruit durian (*Durio zibethinus*). Nat. Genet. *49* (*11*), 1633–1641 https://doi.org/10.1038/ng.3972. PubMed

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res. *18* (*12*), 1979–1990 https://doi.org/10.1101/gr.081612.108. 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/10.12871/00021857201745.

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.

Veeckman, E., Ruttink, T., and Vandepoele, K. (2016). Are we there yet? Reliably estimating the completeness of plant genome sequences. Plant Cell *28* (*8*), 1759–1768 https://doi.org/10.1105/tpc.16.00349. PubMed

Vogel, A., Schwacke, R., Denton, A.K., Usadel, B., Hollmann, J., Fischer, K., Bolger, A., Schmidt, M.H.W., Bolger, M.E., Gundlach, H., et al. (2018). Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. Nat. Commun. *9* (*1*), 2515 https://doi.org/10.1038/s41467-018-04344-z. PubMed

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9* (*11*), e112963 https://doi.org/10.1371/journal.pone.0112963. PubMed

Xiong, W., He, L., Lai, J., Dooner, H.K., and Du, C. (2014). HelitronScanner uncovers a large overlooked cache of *Helitron* transposons in many plant genomes. Proc. Natl. Acad. Sci. USA *111* (*28*), 10263–10268 https://doi.org/10.1073/pnas.1410068111. PubMed

Ye, G., Zhang, H., Chen, B., Nie, S., Liu, H., Gao, W., Wang, H., Gao, Y., and Gu, L. (2019). *De novo* genome assembly of the stress tolerant forest species *Casuarina equisetifolia* provides insight into secondary growth. Plant J. *97* (*4*), 779–794 https://doi.org/10.1111/tpj.14159. PubMed

Zhao, Q., Yang, J., Liu, J., Cui, M., Fang, Y., Qiu, W., Shang, H., Xu, Z., Wei, Y., Yang, L., and Hu, Y. (2018). A draft reference genome sequence for *Scutellaria baicalensis* Georgi. BioRxiv *398032*, https://doi.org/10.1101/398032.