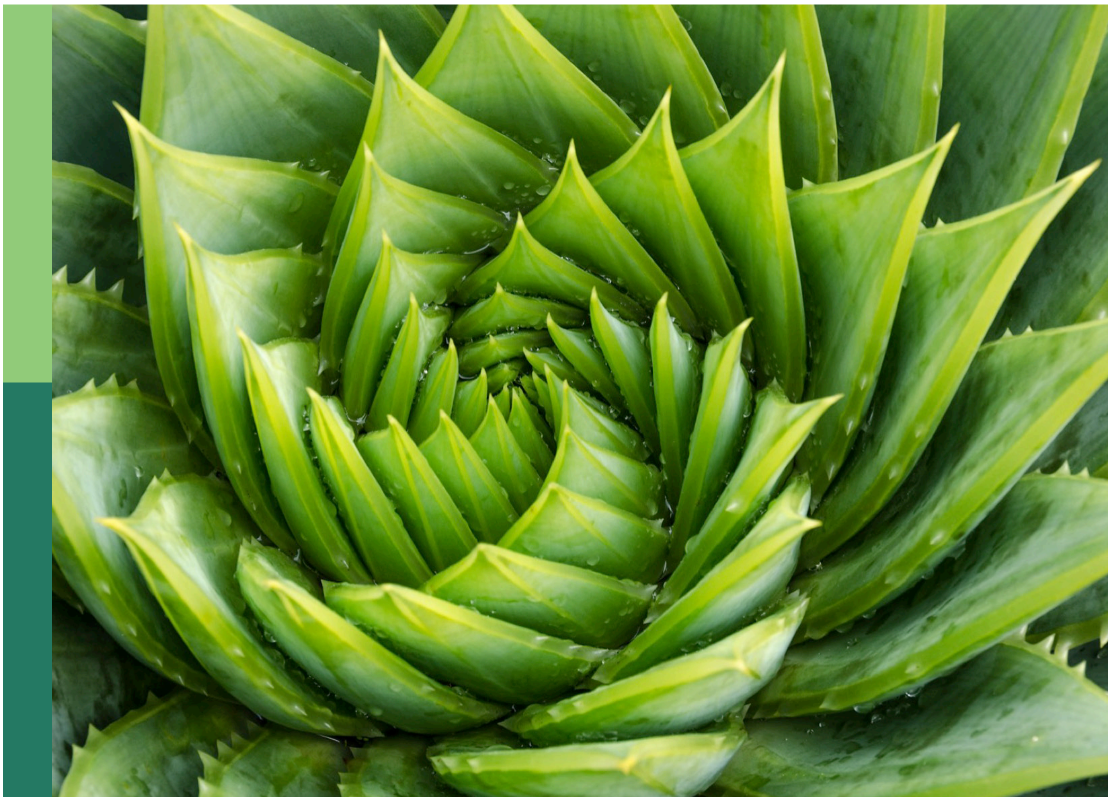
frontiers | Research Topics

Functional genomics in fruit trees: From 'omics to sustainable biotechnologies, volume II

Edited by
Concetta Licciardello, Giorgio Gambino, Manuel Talón,
Riccardo Velasco and Irene Perrone

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-2333-9
DOI 10.3389/978-2-8325-2333-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Functional genomics in fruit trees: From ‘omics to sustainable biotechnologies, volume II

Topic editors

Concetta Licciardello — CREA Research Centre for Olive, Fruit and Citrus Crops, Italy
Giorgio Gambino — Institute for Sustainable Plant Protection, National Research Council (CNR), Italy
Manuel Talón — Valencian Institute for Agricultural Research (IVIA), Spain
Riccardo Velasco — Research Centre of Viticulture and Oenology (CREA), Italy
Irene Perrone — Institute for Sustainable Plant Protection, National Research Council (CNR), Italy

Citation

Licciardello, C., Gambino, G., Talón, M., Velasco, R., Perrone, I., eds. (2023). *Functional genomics in fruit trees: From ‘omics to sustainable biotechnologies, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2333-9

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Table of contents

05 Editorial: Functional genomics in fruit trees: from ‘omics to sustainable biotechnologies, volume II
Concetta Licciardello, Irene Perrone, Giorgio Gambino, Riccardo Velasco and Manuel Talón

08 *VvEPFL9-1* Knock-Out via CRISPR/Cas9 Reduces Stomatal Density in Grapevine
Molly Clemens, Michele Faralli, Jorge Lagreze, Luana Bontempo, Stefano Piazza, Claudio Varotto, Mickael Malnoy, Walter Oechel, Annapaola Rizzoli and Lorenza Dalla Costa

25 Tissue-specific expression of *Ruby* in Mexican lime (*C. aurantifolia*) confers anthocyanin accumulation in fruit
Roger Thilmony, Kasturi Dasgupta, Min Shao, Daren Harris, Jake Hartman, Leslie A. Harden, Ron Chan and James G. Thomson

43 Transcriptome analysis of the pulp of citrus fruitlets suggests that domestication enhanced growth processes and reduced chemical defenses increasing palatability
Estela Perez-Roman, Carles Borredá, Francisco R. Tadeo and Manuel Talon

67 Transgene-free genome editing and RNAi ectopic application in fruit trees: Potential and limitations
Satyanarayana Gouthu, Christian Mandelli, Britt A. Eubanks and Laurent G. Deluc

80 Macadamia germplasm and genomic database (MacadamiaGGD): A comprehensive platform for germplasm innovation and functional genomics in *Macadamia*
Pan Wang, Yi Mo, Yi Wang, Yuchong Fei, Jianting Huang, Jun Ni and Zeng-Fu Xu

93 Building a foundation for gene family analysis in Rosaceae genomes with a novel workflow: A case study in *Pyrus* architecture genes
Huiting Zhang, Eric K. Wafula, Jon Eilers, Alex E. Harkess, Paula E. Ralph, Prakash Raj Timilsena, Claude W. dePamphilis, Jessica M. Waite and Loren A. Honaas

110 A dual sgRNA-directed CRISPR/Cas9 construct for editing the fruit-specific β-cyclase 2 gene in pigmented citrus fruits
Fabrizio Salonia, Angelo Ciacciulli, Helena Domenica Pappalardo, Lara Poles, Massimo Pindo, Simone Larger, Paola Caruso, Marco Caruso and Concetta Licciardello

124 PlantTribes2: Tools for comparative gene family analysis in plant genomics
Eric K. Wafula, Huiting Zhang, Gregory Von Kuster, James H. Leebens-Mack, Loren A. Honaas and Claude W. dePamphilis

141 Construction of a high density genetic linkage map to define the locus conferring seedlessness from Mukaku Kishu mandarin

Krishan Kumar, Qibin Yu, Dharminder Bhatia, Chitose Honsho and Frederick G. Gmitter

157 Role of long non-coding RNA in regulatory network response to *Candidatus* Liberibacter asiaticus in citrus

Xiaokang Zhuo, Qibin Yu, Riccardo Russo, Yi Zhang, Xu Wei, Yuanzhi Zimmy Wang, Paige Marie Holden and Fred G. Gmitter Jr.

TYPE Editorial
PUBLISHED 19 April 2023
DOI 10.3389/fpls.2023.1188832

OPEN ACCESS

EDITED AND REVIEWED BY
Adnane Boualem,
Université Paris-Saclay, CNRS, INRAE,
Université Evry, Institute of Plant Sciences
Paris-Saclay (IPS2) Gif sur Yvette, France

*CORRESPONDENCE
Concetta Licciardello
concetta.licciardello@crea.gov.it

SPECIALTY SECTION
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

RECEIVED 17 March 2023
ACCEPTED 03 April 2023
PUBLISHED 19 April 2023

CITATION
Licciardello C, Perrone I, Gambino G, Velasco R and Talón M (2023) Editorial: Functional genomics in fruit trees: from 'omics to sustainable biotechnologies, volume II. *Front. Plant Sci.* 14:1188832. doi: 10.3389/fpls.2023.1188832

Editorial: Functional genomics in fruit trees: from 'omics to sustainable biotechnologies, volume II

Concetta Licciardello[1]*, Irene Perrone[2], Giorgio Gambino[2], Riccardo Velasco[3] and Manuel Talón[4]

[1]Research Center for Olive Fruit and Citrus Crops, Council for Agricultural Research and Economics, Acireale, Italy, [2]Institute for Sustainable Plant Protection, National Research Council (CNR), Torino, Italy, [3]Research Center for Viticulture and Enology, Council for Agricultural Research and Economics, Conegliano, Italy, [4]Departamento de Genomica, Instituto Valenciano de Investigaciones Agrarias, Valencia, Spain

KEYWORDS

gene-editing, seedlessness, palatability, stomatal density, plant architecture, database, huanglongbing, fruit trees

Editorial on the Research Topic
Functional genomics in fruit trees: from 'omics to sustainable biotechnologies, volume II

The Research Topic collects 10 manuscripts focused on cutting-edge topics in functional genomics of fruit trees, that can be grouped into three main themes: advancements in genome editing technologies, availability of tools and approaches for understanding gene family structure and function, and study of gene regulation through long non-coding RNA (lncRNA), transgenesis, and transcriptomic analysis (Figure 1). The first set of manuscripts showcases perspectives and limitations of transgene-free genome editing in fruit trees, the use of CRISPR/Cas9 constructs for the reduction of stomatal density in grapevine, and the generation of edited citrus varieties enriched in antioxidant compounds. A second group describes the use of comparative gene family analysis tools, of novel workflows for the Rosaceae, the creation of a comprehensive platform for germplasm innovation and functional genomics in Macadamia, and the construction of a high-density genetic linkage map to identify genetic loci responsible for seedlessness in mandarin. The third set of articles includes studies on citrus focused on the role of lncRNA in response to Huanglongbing (HLB), the fruit-specific expression of Ruby to improve anthocyanin accumulation, and the exploitation of the transcriptome relating to growth and palatability.

The recent advancements of new gene-editing tools opens new opportunities for functional studies in fruit trees. Gouthu et al. report the use of a transgene-free gene editing *via* Ribonucleoprotein (RNP) delivery and the ectopic application of RNA-based products; these approaches are mainly addressed to a sustainable and an eco-friendlier environment for a crop production system that could potentially replace the use of chemicals. Both technologies are strictly dependent on the foundational knowledge of gene-to-trait relationships, and the potential and limitations are carefully reviewed.

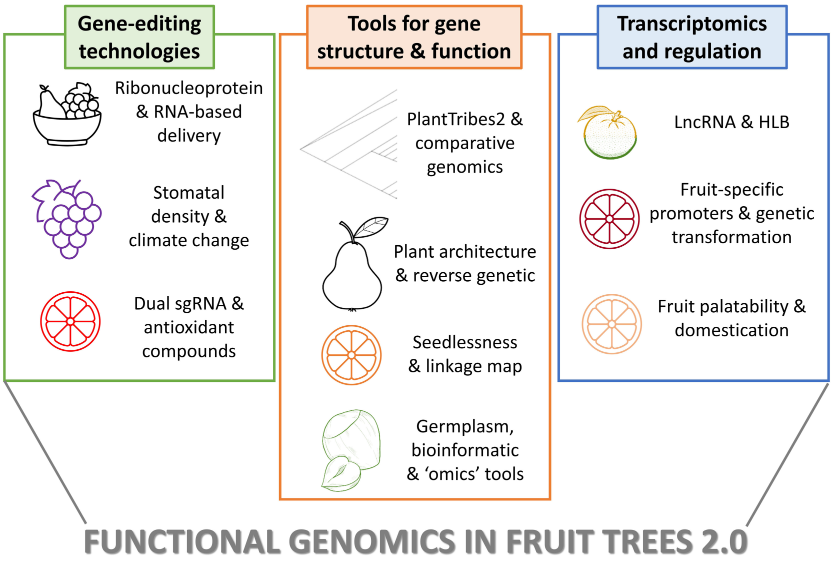

Gene-editing technologies
Ribonucleoprotein & RNA-based delivery
Stomatal density & climate change
Dual sgRNA & antioxidant compounds
Tools for gene structure & function
PlantTribes2 & comparative genomics
Plant architecture & reverse genetic
Seedlessness & linkage map
Germplasm, bioinformatic & 'omics' tools
Transcriptomics and regulation
LncRNA & HLB
Fruit-specific promoters & genetic transformation
Fruit palatability & domestication
FUNCTIONAL GENOMICS IN FRUIT TREES 2.0

FIGURE 1
Methods and applications showed in the Editorial Functional Genomics in fruits trees 2.0. The figure summarizes the main approaches reported in the 10 papers collected in Research Topic, organized in the three main groups as summerized in the Editorial. The plant species and the applications are briefly reported.

Through a genome editing approach based on CRISPR/Cas9 technology, Clemens et al. showed the potential of manipulating stomatal density for optimizing grapevine adaptation under changing climate conditions. By inactivation of the *VvEPFL9-1*, a positive regulator of stomata formation, different edited lines of the table grape variety 'Sugraone' with a significant reduction in stomatal density and a significant increase in pore length were produced. Interestingly, epfl9-1 mutants showed an improved intrinsic water-use efficiency, a desirable trait to improve plant water conservation and to delay early sugar accumulation.

Salonia et al. used a dual sgRNA approach to knockout the fruit-specific *β-LCY2* to introduce lycopene in five different Tarocco and Sanguigno sweet orange varietal groups. The approach revealed to be highly efficient in introducing point or short mutations, large deletions and the inversion of the region between the cutting site of both sgRNAs. No altered phenotype in vegetative tissues of edited plants has been observed. This work represents the first example of the use of a genome editing approach to potentially improve qualitative traits of citrus fruit.

In an effort to address accessibility and computational challenges in genome-scale research and to rely on comparative genomic approaches that integrate across plant community resources and data types, Wafula et al. provided a valuable tool for the research community working on plant genomics. PlantTribes2 is a scalable, easily accessible, highly customizable, and broadly applicable bioinformatic framework useful for comparative and evolutionary analyses of gene families from any type of organism, including fungi, microbes, animals, and plants. Examples of application are the evaluation of targeted gene family assembly and genome quality. Such as example, Zhang et al. showed an application of PlantTribes2 making simpler the acquisition and the analysis of genome-scale data, through an iterative processes of reverse genetics aimed to understand pear architecture genes. To individuate putative architecture genes in pear, it could be possible to start with genes of interest and the workflow proposed provides a comparative genome approach to efficiently identify, investigate, and then improve and/or validate genes of interest across genomes and genome resources.

Macadamia is an important nut crop, but it's becoming difficult for researchers to process and use the vast amount of genomic data available. As a central portal, Wang et al. have developed MacadamiaGGD, a database integrating data from germplasm, genomes, transcriptomes, genetic linkage maps, and SSR markers. The database is freely available online and includes bioinformatic tools to conveniently analyze data of interest. The database is expected to broaden the understanding of the germplasm, genetics, and genomics of macadamia species and facilitate molecular breeding efforts.

In citrus, Kumar et al. identified two closely associated SNPs, AX-160417325 and AX-160536283, in Fs-locus on LG5 of 'Mukaku Kishu' mandarin. These SNPs reduced the population size and positively predicted seedlessness in 25.0-91.9% of the progenies in studied populations. These markers should be strategic in reducing the effective population size at seedling stage in crosses involving 'MK' paternity. Further work will be done, but the availability of these SNPs opens the way in the production of seedless citrus fruits, highly appreciated by consumers.

LncRNAs serve as crucial regulators in plant response to various diseases. Zhuo et al. identified and characterized 8,742 lncRNAs among HLB-tolerant rough lemon and HLB-sensitive sweet orange.

LNC_28805 was identified as one of the most important candidate lncRNAs; on the other hands, WRKY33 and SYP121 are two candidate genes targeted by miRNA5021 developing a key role in the bacteria pathogen responses based on the prediction of protein-protein interaction network. This study will be useful in understanding the role of lncRNAs involved in citrus HLB regulation and opens the road for further investigation of their regulatory functions.

Tissue specific promoters are important tools for the precise genetic engineering of crops. Thilmony et al., in the framework of four fruit-preferential promoters, found that CitWax exhibited high fruit-preferential expression of Ruby in Mexican lime. In some of the transgenic trees with high levels of flower and fruit anthocyanin accumulation, leaves deeply coloured at juvenile phase, lost the coloration at maturity. CitWax promoter could control the expression of Ruby increasing the nutritional value and health benefits of citrus fruit.

Pérez-Roman et al. analyzed the transcriptomes of developing fruitlets of wild and domesticated citrus to identify key traits brought about by domestication. Domestication promoted growth processes at the expense of chemical defenses, also impacting in nitrogen and carbon allocation, presumably leading to major differences in organoleptic properties. The production of unpleasant secondary metabolites and acidity, for instance, decreased considerably improving palatability. The results also appear to suggest that domesticated mandarins evolved through progressive refining of other relevant palatability properties.

Author contributions

All authors drafted the manuscript, revised and approved the final version.

Acknowledgments

We greatly appreciate the active contribution of all authors and reviewers for their precious work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 17 May 2022
doi: 10.3389/fpls.2022.878001

VvEPFL9-1 Knock-Out via CRISPR/Cas9 Reduces Stomatal Density in Grapevine

Molly Clemens[1,2,3], Michele Faralli[1†], Jorge Lagreze[1], Luana Bontempo[1], Stefano Piazza[1], Claudio Varotto[1], Mickael Malnoy[1], Walter Oechel[2,4], Annapaola Rizzoli[1] and Lorenza Dalla Costa[1*]*

[1] Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy, [2] Global Change Research Group, San Diego State University, San Diego, CA, United States, [3] Department of Viticulture and Enology, University of California Davis, Davis, CA, United States, [4] Department of Geography, University of Exeter, Exeter, United Kingdom

OPEN ACCESS

Edited by:
*Giorgio Gambino,
Institute for Sustainable Plant Protection (CNR), Italy*

Reviewed by:
*Claudio Lovisolo,
University of Turin, Italy
Fatemeh Maghuly,
University of Natural Resources and Life Sciences Vienna, Austria*

***Correspondence:**
*Michele Faralli
michele.faralli@unitn.it
Lorenza Dalla Costa
lorenza.dallacosta@fmach.it*

†Present address:
*Michele Faralli,
Center Agriculture Food Environment (C3A), University of Trento, San Michele all'Adige, Italy*

Specialty section:
*This article was submitted to Plant Breeding,
a section of the journal
Frontiers in Plant Science*

Received: *17 February 2022*
Accepted: *11 April 2022*
Published: *17 May 2022*

Citation:
*Clemens M, Faralli M, Lagreze J, Bontempo L, Piazza S, Varotto C, Malnoy M, Oechel W, Rizzoli A and Dalla Costa L (2022) VvEPFL9-1 Knock-Out via CRISPR/Cas9 Reduces Stomatal Density in Grapevine.
Front. Plant Sci. 13:878001.
doi: 10.3389/fpls.2022.878001*

Epidermal Patterning Factor Like 9 (EPFL9), also known as STOMAGEN, is a cysteine-rich peptide that induces stomata formation in vascular plants, acting antagonistically to other epidermal patterning factors (EPF1, EPF2). In grapevine there are two *EPFL9* genes, *EPFL9-1* and *EPFL9-2* sharing 82% identity at protein level in the mature functional C-terminal domain. In this study, CRISPR/Cas9 system was applied to functionally characterize *VvEPFL9-1* in 'Sugraone', a highly transformable genotype. A set of plants, regenerated after gene transfer in embryogenic calli *via Agrobacterium tumefaciens*, were selected for evaluation. For many lines, the editing profile in the target site displayed a range of mutations mainly causing frameshift in the coding sequence or affecting the second cysteine residue. The analysis of stomata density revealed that in edited plants the number of stomata was significantly reduced compared to control, demonstrating for the first time the role of EPFL9 in a perennial fruit crop. Three edited lines were then assessed for growth, photosynthesis, stomatal conductance, and water use efficiency in experiments carried out at different environmental conditions. Intrinsic water-use efficiency was improved in edited lines compared to control, indicating possible advantages in reducing stomatal density under future environmental drier scenarios. Our results show the potential of manipulating stomatal density for optimizing grapevine adaptation under changing climate conditions.

Keywords: *Vitis vinifera*, stomata, genome editing, climate change, water-use efficiency

INTRODUCTION

Drought is a threat to the quality and yield of grapevine in the world's important wine grape growing regions (Mosedale et al., 2016; Van Leeuwen and Destrac-Irvine, 2017; Van Leeuwen et al., 2019). These regions are expected to have decreased precipitation with associated risks of developing soil water deficit in coming years (IPCC, 2014; Sherwood and Fu, 2014; Scholasch and Rienth, 2019). One adaptation strategy seen in plants to tolerate water limitation involves stomatal regulation of water loss (Hunt et al., 2010; Hughes et al., 2017; Bertolino et al., 2019; Caine et al., 2019; Dayer et al., 2020; Gambetta et al., 2020). Stomata are pores mainly located in the leaf epidermis. The opening of these pores controls leaf gas exchange (CO_2 uptake for photosynthesis and water loss *via* transpiration) and is regulated by changes in turgor pressure in the guard cells surrounding these pores. The two guard cells respond to a range of environmental signals, often in conflict with each other, and sometimes rapidly changing (e.g., humidity, CO_2 concentration, light). In drought-stressed grapevine, stomatal closure is triggered by hydraulic signals and maintained by abscisic acid following re-watering (Lovisolo et al., 2010; Tombesi et al., 2015). Genotypic variation for stomatal sensitivity to reduced water availability has been shown to exist in grapevine (Schultz, 2003; Soar et al., 2006; Bota et al., 2016; Villalobos-González et al., 2019; Faralli et al., 2021).

Stomatal density and distribution in the epidermal tissue also plays a critical role in determining transpiration rate per unit of leaf area (Hunt et al., 2010). Previous work focusing on natural variation for stomatal anatomical features provided evidence of a close negative relationship between plant water-use efficiency and stomatal density (Bertolino et al., 2019; Faralli et al., 2019). According to extensive studies carried out in Arabidopsis (Doheny-Adams et al., 2012; Franks et al., 2015; Hepworth et al., 2015; Lee et al., 2015), stomatal density and distribution are under the control of small cysteine-rich peptides (CRP) called epidermal patterning factors (EPFs) highly conserved in a wide range of higher plants (Lu et al., 2019). Three members of this family play a key role in the formation of stomata: EPF1, EPF2 and EPFL9. EPF2 and EPF1 are expressed in the epidermis, in the earlier and later stages of leaf development, respectively. EPF2 inhibits the formation of cells considered the precursors of stomata guard cells, while EPF1 inhibits the subsequent differentiation of these same precursors and induces asymmetric cell division (Hara et al., 2009). Epidermal Patterning Factor Like 9 (EPFL9), also known as STOMAGEN, plays an antagonist role with respect to EPF1 and EPF2 as it induces stomata formation (Kondo et al., 2010). EPF-peptides interact with two transmembrane receptors of epidermal cells, ERECTA and Too Many Mouths (TMM). While EPF1 and EPF2 activate the receptor complex which in turn induces a MAPKs (Mitogen-Activated Protein Kinases) cascade (Morales-Navarro et al., 2018; Zoulias et al., 2018) leading to the destabilization of important transcription factors involved in the formation of stomata (SPEECHLESS, MUTE, FAMA; Pillitteri et al., 2007; Chen et al., 2020), STOMAGEN inactivates it. STOMAGEN is the only known positive regulator of stomata produced in mesophyll, and was confirmed to act independently of EPF1 and EPF2 (Hunt et al., 2010; Kondo et al., 2010; Sugano et al., 2010; Ohki et al., 2011). Its activity is antagonized by that of EPF2, however, it is not well understood if the antagonistic action is due to the sharing of an identical binding site in the common receptor or to other mechanisms (Ohki et al., 2011). An evolutionary model suggests that EPFL9 may derive from the duplication of EPF1/2 with a subsequent alteration in the function (Shimada et al., 2011). This is confirmed by the fact that EPF1/2 are more widespread in higher plants compared to EPFL9 (Lu et al., 2019). Despite the different amino acid composition among the CRP different sub-classes and across species, the members of CRPs have in common a small size, a conserved N-terminal region that include an apoplast secretion signal and a functional C-terminal domain containing cysteine residues (Marshall et al., 2011).

Several functional genomics studies, based on the ectopic expression or silencing of EPF1, EPF2, or EPFL9, have recently demonstrated a highly conserved functional paradigm in Arabidopsis and cereals. In barley, Hughes et al. (2017) proved that *HvEPF1* overexpression limits stomatal development. In a hexaploid bread wheat, Dunn et al. (2019) decreased stomatal density (SD) *via* the overexpression of *TaEPF1* and *TaEPF2* orthologues and demonstrated improvements in water-use efficiency without affecting yield when SD reduction was moderate. Similarly, in rice Caine et al. (2019) and Mohammed et al. (2019) elucidated the function of *OsEPF1* adopting an over-expression approach. Adding to the studies on rice, Lu et al. (2019) confirmed the role of *OsEPF1*, *OsEPF2* and *OsEPF9* by a dual strategy, both over-expression and down-regulation *via* RNA interference. Yin et al. (2017) were the first to apply the genome editing technology in rice to disrupt *OsEPFL9*.

Gene editing *via* the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) (Jinek et al., 2012) is to date the most powerful tool for functional genomics studies in plants (Liu et al., 2016). CRISPR/Cas9 system can efficiently produce nucleotide mutations into precise positions in the genome through the combined action of a specific guide RNA and the Cas9 nuclease which cleaves the DNA eliciting the non-homologous end-joining (NHEJ) pathway for DNA repair (Podevin et al., 2013). NHEJ may produce knock-out (KO) mutants with random insertion or deletion (indels) of variable lengths at the Cas9 cleavage site causing frameshift mutations or loss of amino acids in protein-coding sequences. These KO mutants are perfect systems to prove the function of a candidate gene (Jain, 2015). This technology is steadily boosting (Hess et al., 2017; Anzalone et al., 2019) and, coupled with the advancements of *in-vitro* culture practices, represents a knowledge-based strategy for the genetic improvements of cultivated plants, with relevant advantages compared to traditional breeding (Chen et al., 2019).

In grapevine, CRISPR/Cas9 technology has been successfully applied to evaluate the function of genes involved in susceptibility or tolerance to diseases, mainly caused by fungal pathogens (Malnoy et al., 2016; Giacomelli et al., 2019; Li et al., 2020; Wan et al., 2020; Chen et al., 2021; Scintilla et al., 2021), or to enhance tolerance to cold stress (Wang et al., 2021).

In this study, we inactivated *VvEPFL9-1* in a grapevine table grape variety, 'Sugraone', adopting a genome editing approach based on CRISPR/Cas9 technology. Different edited lines with a significant reduction in stomatal density were produced and three of them were analyzed to investigate how reducing stomatal density affects grapevine physiological performance under different environmental conditions.

MATERIALS AND METHODS

Search for the Orthologous Gene of *AtEPFL9* in Grapevine Genomic Databases and Experimental Confirmation in a Set of Grapevine Genotypes

AtEPFL9 sequence (AT4G12970) was used as a query to interrogate the publicly available genomic databases of *Vitis* spp. (**Supplementary Table 1**). To experimentally confirm the presence of two *VvEPFL9* paralogs in a set of grapevine genotypes, DNA was extracted from leaf tissue of 'Chardonnay', 'Merlot', 'Syrah', 'Cabernet Sauvignon', 'Touriga National', 'Pinot Noir clone Entav 115', 'Pinot Noir PN40024', 'Sugraone' and 'Riparia Glorie de Montpellier' using Nucleospin Plant II kit (Macherey–Nagel, Düren, Germany) following the manufacturer's instruction. Genomic DNA was quantified using Nanodrop 8,800 (Thermo Fischer Scientific, Waltham, MA, United States) and diluted to a final concentration of 30 ng/μL. Two PCR reactions were performed in 25 μl final volume containing 1 × PCR BIO (Resnova, Rome, Italy), 30 ng of genomic DNA and 0.5 μM of primers in order to amplify *VvEPFL9-1* (primer VvEPFL9-1_fw and VvEPFL9-1_rv, see **Supplementary Table 2**) and *VvEPFL9-2* (primer VvEPFL9-2_fw and VvEPFL9-2_rv, see **Supplementary Table 2**). Amplification products were checked on agarose gel, purified using CleanNGS magnetic beads (CleanNA, Waddinxveen, Netherlands) and sequenced by Sanger sequencing (FEM Sequencing Platform Facility, San Michele all'Adige, Italy). Sequencing outputs were analyzed with Blast online tool[1] and for the alignment of the sequences the software MEGAX (Kumar et al., 2018) was used.

Plant Material (Gene Transfer Experiments, *in-vitro* and Greenhouse Growth)

The CRISPR/Cas9 binary vector with the customized sgRNA was purchased from DNA Cloning Service (Hamburg, Germany). The nucleotide sequence of *SpCAS9* and of *NPTII* genes were codon optimized for the plant expression system and their sequences are available on the company website.[2] The sequence of the guide RNA carried by the vector was designed with CRISPR-P 2.0 software[3] and recognizes a region of 20 bp in the third exon of *VvEPFL9-1* (GCACATACAATGAATGCAAA, on-score = 0.7058). *Agrobacterium tumefaciens* (A.t.)-mediated gene transfer was performed on embryogenic calli of 'Sugraone' according to Dalla Costa et al. (2022). *NPTII* was used as selectable marker to confer resistance to kanamycin. Regenerated plants were screened by PCR for the presence of *SpCAS9* (to select plants which integrated T-DNA) in 20 μl final volume containing 1 × PCR BIO (Resnova, Rome, Italy), 0.5 μM of each primer (SpCAS9_Fw and SpCAS9_Rv, see **Supplementary Table 2**) and 30 ng of genomic DNA. DNA was extracted from freshly frozen leaf tissue (approximately 100 mg) using Nucleospin Plant II kit (Macherey–Nagel, Düren, Germany) following the manufacturer's instruction, quantified using Nanodrop 8,800 (Termo Fischer Scientific, Waltham, MA, United States) and diluted to a final concentration of 30 ng/μL.

Edited lines and WT control were propagated *in-vitro* in sterilized jars containing WP medium (McCown and Lloyd, 1981) in a growth chamber at 100 photosynthetic photon flux density (PPFD) ± 20 (μmol m^{-2} s^{-1}), 24°C and a 16/8 light/dark photoperiod. Four biological replicates of healthy developed edited lines and of the WT control were acclimatized in the greenhouse using 0.25 l plastic pots with three holes in the bottom to allow for water drainage, filled with a similar amount of growing substrate (Extra quality - Semina, TerComposti, Calvisano, Italy) and covered by parafilm on the top. Plants were kept in a growth chamber (PPFD 100 +/− 20 μmol m^{-2} s^{-1}, 24°C, 16/8 light/dark photoperiod) and after 1 week, holes were gradually made in the top of the parafilm over the course of 2 weeks. After 17 days, plants were repotted into 0.75 l pots all containing growing substrate (Extra quality - Special Cactus, TerComposti, Calvisano, Italy). Pots were kept in the same growth chamber for a subsequent 10 days before moving to the greenhouse. In the greenhouse, plants were grown under natural light supplemented by high-pressure sodium lamps system (PPFD 200–250 μmol m^{-2} s^{-1}) with a 16-h/8-h light–dark photoperiod. Environmental conditions including temperature and humidity during the growth chamber and greenhouse cultivation are shown in **Supplementary Figure 1**.

Molecular Characterization of Edited Lines

Transgene Copy Number Quantification

The quantification of *SpCAS9* copy number (CN) in grapevine lines was carried out according to real-time PCR method developed by Dalla Costa et al. (2009). Reactions were performed in a 96-well plate on a C1000 thermal cycler (Bio-Rad, Hercules, United States) equipped with CFX96 real-time PCR detection system (Bio-Rad, Hercules, United States). The real-time PCR singleplex reaction was carried out in a 10 μl final volume containing 1 × SsoAdvanced Universal Probes Supermix (Bio-Rad, Hercules, United States), 40 ng of genomic DNA, 0.3 μM primers (Sigma, Haver hill, UK) and a 0.2 μM specifc Taqman probe (Sigma, Haverhill, UK). The thermal protocol was as follows: polymerase activation for 3 min at 95°C followed by 40 cycles of denaturation of 10 s at 95°C, annealing of 5 s at 58°C and 5 s at 60°C and an elongation of 30 s at 72°C. Primers and Taqman probes used to amplify grapevine endogenous *VvCHI* (VvChiRT_fw; VvChiRT_rv; VvChiRT_Probe) and *SpCAS9* (SpCas9RT_fw; SpCas9RT_rv; SpCas9RT_Probe) were reported in **Supplementary Table 2**. The standard curves (four points, starting from 10^6 plasmid molecules and adopting a serial dilution of 1:5) were built with a plasmid pGEM-T easy (Promega, Madison, Wisconsin, United States), in which we cloned a fragment of *VvCHI* and *SpCAS9*. For each sample, the *SpCAS9* CN was calculated using the following formula: (transgene total copies / endogenous gene total copies) × 2. The total copies of transgene and endogenous gene were calculated on the basis of the mean values of the quantification cycles (Cq) of two technical replicates.

On- and Off-Target Editing Evaluation

In the grapevine lines integrating T-DNA, a region of the gene *VvEPFL9-1* containing the site targeted by the sgRNA/Cas9 complex, was amplified with primers VvEPFL9-1_fw and VvEPFL9-1_rv (see **Supplementary Table 2**) both elongated with overhang Illumina adapters. PCR was carried out in 20 μl final volume containing 1 × PCR BIO (Resnova, Rome, Italy), 0.4 μM of each primer and 30 ng of genomic DNA. The Illumina

[1] blast.ncbi.nlm.nih.gov

[2] https://www.dna-cloning.com/

[3] http://crispr.hzau.edu.cn/cgi-bin/CRISPR2/CRISPR

library was sequenced on an Illumina MiSeq (PE300) platform at the Sequencing Platform Facility of Fondazione Edmund Mach (San Michele all'Adige, Italy). CRISPResso2 pipeline[4] (Clement et al., 2019) was used to process the raw paired end reads with default parameters and to visualize the mutations profiles in the target sequences. For the analysis of the off-target site in the gene *VvEPFL9-2*, a PCR was carried out in 25 µl final volume containing 1 × PCR BIO (Resnova, Rome, Italy), 0.5 µM of each primer (VvEPFL9-2_fw and VvEPFL9-2_rv, see **Supplementary Table 2**) and 30 ng of genomic DNA. Amplification products were checked on agarose gel, purified using CleanNGS magnetic beads (CleanNA, Waddinxveen, Netherlands) and sequenced by Sanger sequencing (FEM Sequencing Platform Facility). Sequencing outputs were analyzed with Blast online tool.[5]

T-DNA Integration Site Identification

T-DNA integration points (IP) were determined following the method described in Dalla Costa et al. (2020). The library was sequenced by Illumina MiSeq (PE300) platform at the Sequencing Platform Facility of Fondazione Edmund Mach (San Michele all'Adige, Italy). The putative genomic regions identified were validated by PCR amplification. PCR was performed in a 20 µl final volume containing 1 × PCR BIO (Resnova, Rome, Italy), 40 ng of genomic DNA and 0.5 µM of the primers reported in **Supplementary Table 2**. Amplification products were checked on agarose gel, purified using PureLink Quick Gel Extraction (Invitrogen, Carlsbad, CA, United States) and sequenced by Sanger sequencing (FEM Sequencing Platform Facility). Sequencing outputs were analyzed with the Blast sequence server (using the database PN40024.v4_REF_genome) available online at the European network INTEGRAPE website.[6]

Experimental Conditions and Physiological Analysis

Experiment 1: Well-Watered (WW) Conditions in Greenhouse

Biological replicates of edited lines S-*epfl9*KO1 (n = 4) and S-*epfl9*KO2 (n = 4), and of 'Sugraone' WT (n = 4) kept in a greenhouse for 2 months were used. Pots were covered in aluminum foil and wrapped in plastic to limit soil evaporation (**Supplementary Figure 2**). All plants were measured daily for 14 days at the same time each morning for mass of water loss.

Experiment 2: Water-Stress (WS) Conditions in Greenhouse

The same plants used in Experiment 1 were used in Experiment 2. Control pots (soil-filled pots without plants) were placed at the end of each row in randomized positions, weighed by balance and returned to the same positions every day to assess soil evaporation. Pots dried down naturally for a subsequent 15 days.

Experiment 3: Well-Watered (WW) Conditions in an Automated High-Throughput Phenotyping Platform

Biological replicates of the edited line S-*epfl9*KO6 (n = 6) and 'Sugraone' WT (n = 4), maintained in greenhouse for 12 months, with a height range of 60–70 cm and a weight brought to 3,000 g (in 5 l pots) were used. Plants were moved inside the phenotyping platform (WIWAM, Ghent, Belgium) at the Plant Phenotyping Facility of Fondazione Edmund Mach where temperature was set to 28/25°C, photoperiod to 16/8 h and average PPFD to 300 μmol m^{-2} s^{-1} at apical leaf level. Plants were automatically watered every day at 6:00 AM to target weight (3,000 g) and pot weight was evaluated before and after watering for 12 days.

Soil Water Content, Transpiration, and Leaf Area Determination

In Experiment 1 and 2, total transpirable soil water (TTSW) was calculated as the difference between pot mass at day 1, fully watered (100% capacity), and the pot mass at the end of the natural dry down when transpiration reached a minimum. Fully watered plants (100% relative soil water content) were weighted after watering to capacity and allowing pots to drain for 2 h. The fraction of transpiration soil water (FTSW) was calculated as a daily ratio between the amount of soil water remaining in the pot left for transpiration and the TTSW using the equation: FTSW = (PMn − PMfinal)/TTSW, where PMn is the pot mass for each day, and PMfinal is the pot mass at the end of the day 11. FTSW data were reported in **Supplementary Figure 3**. At day 12 (i.e., after Experiment 1), plants were unwrapped from the aluminum and plastic coverings, re-watered to 100% of their initial weight using syringes and weighed as a starting mass for the stress application. In both Experiment 1 and 2, transpiration (g/cm^2) was measured as the grams of water lost daily, normalized by the relative leaf area for each individual [T = (mass 0 - mass 1)/relative leaf area, where 0 and 1 represent the days in consecutive order]. Growth was measured as a relative leaf area every other day for a period of 28 days using RGB imaging. The software Easy Leaf Area (Easlon and Bloom, 2014) was used for analysis. Photos of the plants were taken at the same distance and tripod angle (45°) to provide uniform and consistent assessment of relative leaf area (example in **Supplementary Figure 4A**). A biomass-leaf area estimated curve was constructed using eight plants of varying sizes validating the non-destructive approach (**Supplementary Figure 5**). In Experiment 3, daily water-use was automatically calculated as daily pot weight loss (g). In addition, projected leaf area (pixels) was calculated at the beginning and at the end of the experiment (day 1 and day 12 respectively) as the average green pixels in four RGB images collected at different pot angles and analyzed with the WIWAM software (example in **Supplementary Figure 4B**).

Stomatal Characterization

Samples for stomatal characterization were taken under well-watered conditions as well as at the end of the drought treatment (i.e., Experiment 1 and 2). Leaves were chosen with the same

[4] https://crispresso.pinellolab.partners.org/submission

[5] blast.ncbi.nlm.nih.gov

[6] https://integrape.eu/resources/genes-genomes/genome-accessions/

size and position, typically leaf three, unless abnormal. Clear gel nail polish was applied to the abaxial and adaxial surfaces of the leaf to create an imprint of the leaf surface and allowed to dry. Clear tape was used to peel off the nail polish, and the tape was mounted on a microscope slide. Slides were imaged using a compound microscope (DM, Leica Microsystems, Wetzlar, Germany) at 40x and at five different technical positions of the same area (0.3 mm^2) on the four biological replicates for a total of twenty measurements of stomata density per individual. Stomatal size (SS) was characterized from three technical replicates from three biological replicates for a total of 9 replicates per individual. These 9 replicates were averaged to create an average radius (r) for reach individual, and the stomatal size was subsequently calculated as $SS = 0.5\pi r^2$; stomatal size is equal to 0.5 multiplied by the average length of stomata squared multiplied by π.

Gas-Exchange Analysis, SPAD and Leaf Temperature

For Experiment 1, 2 and 3, gas-exchange measurements were carried out using a portable infra-red gas analyzer and a 2 cm^2 leaf cuvette with an integral blue–red LED light source (LiCOR 6,400-40XT, Lincoln, NE, United States). Inside the cuvette, flow rate was set at 400 $\mu mol\, s^{-1}$, leaf temperature at 24°C, PPFD to 1,500 $\mu mol\, m^{-2} s^{-1}$ and C_a of 400 $\mu mol\, mol^{-1}$. In Experiment 1, measurements of the response of photosynthesis (A) to sub-stomatal CO_2 concentrations (C_i) curves (A/C_i) were performed between 9:00 and 12:00, on the most expanded leaf from each plant. For A/C_i, C_a was sequentially decreased to 300, 200, 150, 75 and 50 $\mu mol\, mol^{-1}$ before returning to the initial concentration of 400 $\mu mol\, mol^{-1}$. This was followed by a sequential increase to 500, 700, 900, 1,100, 1,300, and 1,500 $\mu mol\, mol^{-1}$. Readings were recorded when A reached steady state. The maximum velocity of Rubisco for carboxylation (V_{cmax}) and the maximum rate of electron transport demand for Ribulose 1,5-bisphosphate (RuBP) regeneration (J_{max}) were estimated as described by (Duursma, 2015; Easlon and Bloom, 2014). A_{sat} represents CO_2 assimilation rate at saturating PPFD while g_s represents stomatal conductance at ambient CO_2 (C_a). Intrinsic water-use efficiency ($_iWUE$) was calculated as $= A_{sat} / g_s$. During Experiment 2, measurements of A and g_s were taken every day on fully expanded leaves for the first 3 days to record a baseline gas-exchange before water stress was applied. Subsequently gas-exchange data were recorded every 2 days in fully expanded leaves. In Experiment 3, gas-exchange parameters (A and g_s), leaf temperature and leaf chlorophyll content were measured at day 5 on the same leaves, respectively with LiCOR 6,400-40XT (Lincoln, NE, United States), an infra-red thermometer (62 MAX+, FLUKE Corporation, Everett, Washington, United States) and a SPAD (Minolta SPAD 502).

Carbon Isotope Composition

Carbon isotope composition was estimated in leaves with the same leaf size and position, count as leaf three unless abnormal. Samples for stomatal characterization were taken first, and the remaining fresh leaf tissue was dried at 80°C for 2 days to be used for $\delta^{13}C$ determination. $\delta^{13}C$ was analyzed in 2 mg aliquots of leaf sample weighed in tin capsules. Samples were combusted in an elemental analyzer (Thermo Flash EA 1112 Series, Bremen, Germany), CO_2 was separated by chromatography and directly injected into a continuous-flow isotope ratio mass spectrometer (Thermo Finnigan Delta V, Bremen, Germany) through the interface ConFlo IV dilutor device (Thermo Finnigan, Bremen, Germany). Samples were measured in duplicate. The isotope ratios were expressed in δ‰ against Vienna-Pee Dee Belemnite for $\delta^{13}C$ according to the following equation: $\delta‰ = (R_{SA} - R_{REF}) / R_{REF}$ where R_{SA} is the isotope ratio measured for the sample and R_{REF} is the international standard isotope ratio. The isotopic values were calculated using a linear equation against working in-house standards, which were themselves calibrated against the international reference materials L-glutamic acid USGS 40 (US Geological Survey, Reston, VA, United States), fuel oil NBS-22 and IAEA-CH-6. The uncertainty of measurement (calculated as 2 standard deviations) was 0.1‰.

Statistics

Statistical analyses were performed using R software (R Core Team, 2020). A one-way ANOVA was used to compare differences in cumulative transpiration, conductance, photosynthesis, and water use efficiency between edited and WT lines for each day of measurement. *Post hoc* comparisons using Fisher's LSD test were carried out to assess group differences. p values lower than 0.05 were considered significant.

RESULTS

Identification of *AtEPFL9* Orthologous Genes in Grapevine

Two *VvEPFL9* gene variants (hereinafter *VvEPFL9-1* and *VvEPFL9-2*) were found in contigs of publicly available genomes of different *Vitis vinifera* varieties and of some other species within the same genus (*Vitis sylvestris*, *Vitis arizonica*, *Vitis riparia*; **Supplementary Table 1**). In the last annotation of the PN40024 grapevine reference genome (PN40024.v4.1,[7] genome assembly version 12X.v4) *VvEPFL9-1* (Vitvi05g01370) was localized on chromosome 5 (position 20,461,188–20,461,813) while VvEPFL9-2 (Vitvi07g04390) on chromosome 7 (position 17,537,397–17,536,742). Interestingly, before the new version of reference genome and related annotation was made publicly available (INTEGRAPE Workshop, 2021) in November 2021, only *VvEPFL9-1* was localized on the genome while the position of *VvEPFL9-2* was not assigned (VCost.v3 annotation). According to gene prediction, *VvEPFL9-1*/−*2* coding sequence have a length of about 330/315 bp and are composed of three exons encoding for: an N-terminal region with a secretion signal for the apoplast [i.e., first 27 amino acid according to SignalP-5.0 software (Almagro Armenteros et al., 2019)[8] a central region likely

[7] https://integrape.eu/resources/genes-genomes/genome-accessions/
[8] http://www.cbs.dtu.dk/services/SignalP/

involved in the processing of the mature peptide and a C-terminal domain of 45 amino acids containing 6 conserved cysteines, that is the functional peptide. A check on genomic DNA extracted from a panel of genotypes (i.e., 'Pinot Noir PN40024', 'Riparia Glorie de Montpellier', 'Pinot Noir clone Entav 115', 'Cabernet Sauvignon', 'Chardonnay', 'Merlot', 'Sugraone', 'Syrah' and 'Touriga National'), confirmed the presence of both gene variants in all the analysed samples with a very high conservation among genotypes (**Supplementary Table 3**). In all the genotypes no SNPs were detected between the two alleles of both isoforms in the region coding for the functional domain, except in Cabernet Sauvignon where an allelic polymorphism in position 25 was detected in *VvEPFL9-1*, which leads to two different amino acids after the first cysteine of the array (serine or threonine, both polar uncharged). Considering only the region encoding for the C-terminal domain (135 bp), the identity between the two variants was 74%, with a large part of polymorphism leading to synonymous codons (**Figure 1A**). At the protein level, the alignment of the C-terminal domains encoded by the two variants showed an identity of 82%, with 8 out of 45 different amino acids (**Figure 1B**). In five positions (14, 25, 28, 40, and 42) substitutions are conservative, i.e., the pair of amino acids belong to the same class, while in the remaining three positions (5, 18, and 34) the substitutions are non-conservative. A comparison with *AtEPFL9* mature peptide revealed that the identity between *VvEPF9-1* and *AtEPFL9* is 82% while the identity between *VvEPF9-2* and *AtEPFL9* is 95% (**Supplementary Figure 6**). Moreover, the relationship of *VvEPFL9-1/−2* with the orthologues of some di- and monocotyledonous plant species including some perennial fruit trees (retrieved from Ensembl Plants genomic database),[9] is shown in **Figure 1C**.

The Knock-Out of *VvEPF9-1* Reduces Stomatal Density in Grapevine

A highly transformable genotype of *Vitis vinifera*, 'Sugraone' was used for gene transfer of the CRISPR/Cas9 machinery in order to obtain edited plants knocked-out for the *VvEPF9-1* gene. The sgRNA was designed to target a region of 20 nucleotides in the third exon, spanning across "TGC" triplets coding for the first and the second cysteine of the functional C-terminal domain (**Figures 1A,B**; **Supplementary Table 4**). In particular, the cleavage operated by Cas9 was expected to affect the "TGC" triplet coding for the second cysteine, this being located 3 nucleotides upstream of the PAM site (i.e., GGG; **Figure 1A**). The corresponding region of *VvEPF9*-2 has 3 mismatches compared with the target site on *VvEPF9-1*, in positions 6, 18 and 20, the last two in the seed region close to the PAM site (**Figure 1A**). Several shoots were regenerated from somatic embryos after 7–10 months from *Agrobacterium tumefaciens* co-culture (**Figure 2**), and nine of them were selected for molecular characterization. The Cas9 integration copy number varied in the transgenic lines, ranging from 1 integration copy for line S-*epfl9*KO7 to 5 integration copies for line S-*epfl9*KO1, with the majority of lines showing values close to one or two copies (**Figure 3A**). A T-DNA integration site was identified for 5 lines: S-*epfl9*KO1 (chr18: position 2,096,753), S-*epfl9*KO2 (chr01: position 4,310,437), S-*epfl9*KO3 (Chr13: position 5,599,304), S-*epfl9*KO6 (chr04: position 6,948,780), S-*epfl9*KO7 (chr03: position 405,924). Concerning T-DNA rearrangements, all the lines showed a trimming of several bases at the LB border, ranging from 31 bp of S-*epfl9*KO7 to 110 bp of line S-*epfl9*KO3, and a T-DNA tandem repeat was detected in line S-*epfl9*KO1. The analysis of the genomic "on-target" site in *VvEPF9-1* proved that all lines were edited, some completely while others showed a degree of wild-type target sequence, indicated as WT (**Figure 3B**; **Supplementary Table 4**). In general, the editing profile was highly heterogeneous, with a composite mutation profile for many lines (e.g., S-*epfl9*KO2, S-*epfl9*KO5, S-*epfl9*KO6, S-*epfl9*KO7, S-*epfl9*KO9), including deletions of increasing size (from 1 bp to more than 7 bp), insertions of 1 or 2 bp, and single base substitutions. The most frequent kind of mutations were deletions of 4 or 5 bp (**Figure 3B**). The resulting mutations in the protein sequence were frameshift mutations (FS) with or without the formation of premature stop codons (SC), or non-frameshift mutations with loss of the second cysteine due to deletion of 3 or 6 bp or to a single base substitution (**Figure 3C**). The analysis of stomatal density in leaves of greenhouse-cultivated plants (2 months old) showed a significant reduction in stomata number in transgenic lines compared to WT (**Figure 3D**). This reduction was significant even for the lines maintaining a remarkable rate of non-mutated *VvEPF9-1* (i.e., S-*epfl9*KO1, S-*epfl9*KO5, S-*epfl9*KO7) and for lines that went through the loss of the second cysteine of the 6-Cys-array, highlighting the crucial role of such residue (i.e., S-*epfl9*KO5 and S-*epfl9*KO8). The editing in the potential "off-target" site in *VvEPFL9-2* was assessed and no mutations were found in all the transgenic lines (**Supplementary Figure 7**). This proved that 3 mismatches with respect to the sgRNA, 2 of which close to the PAM site, were enough to avoid Cas9 unspecific cleavage at this site. *In vitro* and greenhouse edited plants did not show phenotypic defects due to pleiotropic effects (e.g., rate of growth, total leaf area, chlorophyll content) compared to the control plants (data not shown).

Analysis of stomatal anatomical features confirmed the significant differences for stomatal density and pore length between the selected S-*epfl9*KO1 and S-*epfl9*KO2 knock-out mutants and WT (**Figure 4**). S-*epfl9*KO1 had an average SD of 65 stomata mm^{-2} while SD for S-*epfl9*KO2 was 95 stomata mm^{-2}, both significantly lower values than that of 'Sugraone' WT (160 stomata mm^{-2}) respectively by 60 and 40%. Conversely, pore length was significantly higher in S-*epfl9*KO1 and S-*epfl9*KO2 than 'Sugraone' WT, by up to 30%.

The Knock-Out of *VvEPF9-1* Enhances Plant Water Use Efficiency Under Optimal Growth Conditions

A/C_i response curves (net CO_2 assimilation rate, A, versus calculated substomatal CO_2 concentration, C_i) were carried out under optimal environmental conditions and saturating light

[9] http://plants.ensembl.org/index.html

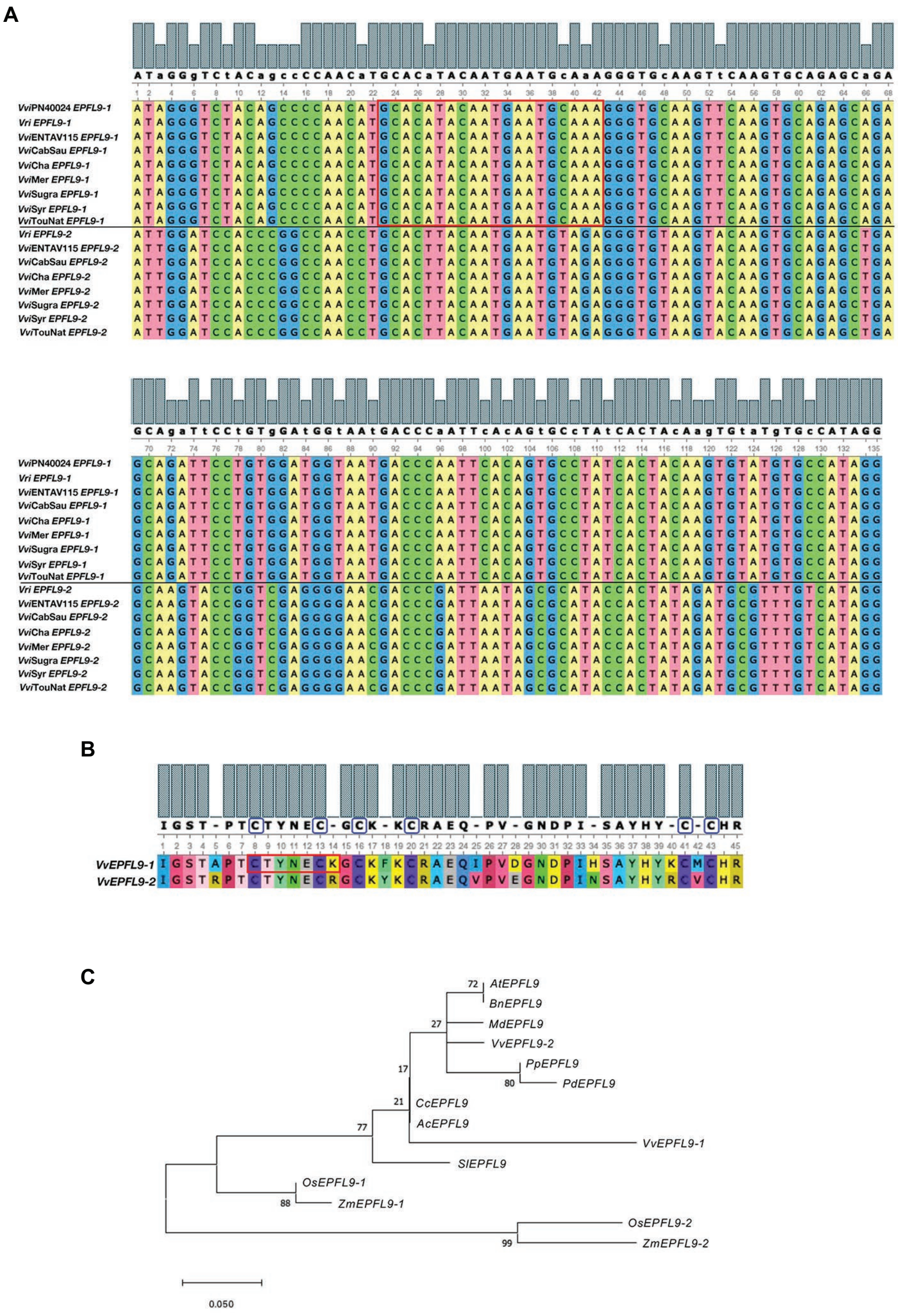


FIGURE 1 | Analysis of *VvEPFL9* paralogs. **(A)** Alignment of the nucleotide sequence encoding for the C-terminal domain (135 bp) obtained by Sanger sequencing of PCR fragments amplified on genomic DNA with primers VvEPFL9-1_fw; VvEPFL9-1_rv and VvEPFL9-2_fw; VvEPFL9-2_rv (see primer list in **Supplementary Table 2**). Genomic DNA was extracted from leaves of ‘Pinot Noir PN40024’, *Vitis riparia* ‘Riparia Glorie de Montpellier’, ‘Pinot Noir clone Entav 115’, ‘Cabernet Sauvignon’, ‘Chardonnay’, ‘Merlot’, ‘Sugraone’, ‘Syrah’, ‘Touriga National’. The red rectangle indicates the 20 bp-target site recognized by the sgRNA/Cas9 complex. **(B)** Alignment of the C-terminal protein domain of VvEPFL9-1 and VvEPFL9-2, translated from the 135 bp nucleotide sequences shown in **(A)**. Cysteine residues are circled in blue. The red rectangle indicates the peptide region corresponding to the target site. **(C)** Phylogenetic tree of the Arabidopsis *AtEPF9* mature peptide and its orthologs from some dicotyledonous (*Brassica napus*, *Malus × domestica*, *Vitis vinifera*, *Prunus persica*, *Prunus domestica*, *Prunus dulcis, Citrus clementina, Actinidia chinensis, Solanum lycopersicum*) and monocotyledonous (*Orytia sativa*, *Zea mays*) plant species. The alignments were generated with MUSCLE (MEGA X) and visualized with Unipro UGENE [http://ugene.net/faq.html (Okonechnikov et al., 2012)]. The phylogenetic tree was built with MEGA X using Maximum Likelihood (1,000 replicates bootstrap). Accession Numbers: *VvEPFL9-1* (*Vitis vinifera*; Vitvi05g01370); *VvEPFL9-2* (*Vitis vinifera*; contig VV78X057312.8. BioProject PRJEA18357); *AtEPFL9* (*Arabidopsis thaliana*; AT4G12970); *BnEPFL9* (*Brassica napus*; BnaA08g04900D-1); *OsEPFL9-1* (*Oryza sativa*; BGIOSGA005039-TA); *OsEPFL9-2* (*Oryza sativa*; BGIOSGA026626-TA); *ZmEPFL9-2* (*Zea mays*; Zm00001d049795_T001); *ZmEPFL9-1* (*Zea mays*; Zm00001d012079_T001); *SlEPFL9* (*Solanum lycopersicum*; Solyc08g066610.3.1); *MdEPFL9* (*Malus domestica*; mRNA:MD10G0128800); *CcEPFL9* (*Citrus clementina*; ESR50459); *PpEPFL9* (*Prunus persica*; ONH92727); *PdEPFL9* (*Prunus dulcis*; VVA33635); *AcEPFL9* (*Actinidia chinensis*; PSR86312).

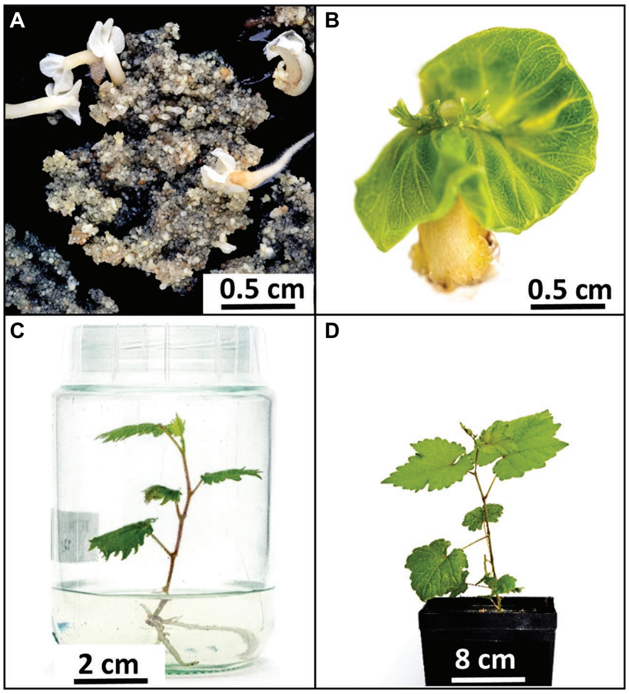


FIGURE 2 | Pipeline to obtain *epfl9*-1 mutants for physiological characterization. **(A)** Embryogenic callus of ‘Sugraone’ 7 months after co-cultivation with *Agrobacterium tumefaciens*. Some embryos are developing on a homogeneous callus mainly formed by small globular embryos. **(B)** Embryo producing shoot. **(C)** *In-vitro* plantlet cultivated in baby jar. **(D)** Greenhouse plant after 2 months from acclimatization of an *in-vitro* plantlet.

intensity assessed *via* light curves for selected edited lines S-*epfl9*KO1 and S-*epfl9*KO2 in Experiment 1 (**Supplementary Figure 8**). There were no significant differences for maximum rate of Rubisco-mediated carboxylation (V_{cmax}) between edited lines and WT control ($p > 0.05$, **Figure 5A**). Similarly, maximum electron transport rate for RuBP regeneration (J_{max}) did not vary between edited lines and WT control ($p > 0.05$, **Figure 5B**). On the contrary, significant reductions in CO_2 assimilation rate at saturating light (A_{sat}) were detected for S-*epfl9*KO1 and, in particular, S-*epfl9*KO2 when compared to WT and up to 50% ($p = 0.007$, **Figure 5C**). S-*epfl9*KO1 and S-*epfl9*KO2 had significantly lower conductance (g_s) than WT ($p < 0.001$) with S-*epfl9*KO2 showing the lowest values (0.030 mol m^{-2} s^{-1} on average, **Figure 5D**). This led to a significantly higher intrinsic water-use efficiency ($_{i}WUE$) for S-*epfl9*KO2 than ‘Sugraone’ WT ($p = 0.024$, **Figure 5E**). Accordingly, carbon isotope composition ($\delta^{13}C$) analysis detected for S-*epfl9*KO2 significant less negative $\delta^{13}C$ values compared to ‘Sugraone’ WT ($p = 0.046$), indicating a higher $_{i}WUE$ (**Figure 5F**). Gravimetric assessments of transpired water normalized for leaf area highlighted significant differences in cumulative transpiration between edited and WT lines. In general, both S-*epfl9*KO1 and S-*epfl9*KO2 used less water throughout a 14 day experimental period, by up to 21%, compared to ‘Sugraone’ WT (**Figure 5G**). Moreover, to expand our data

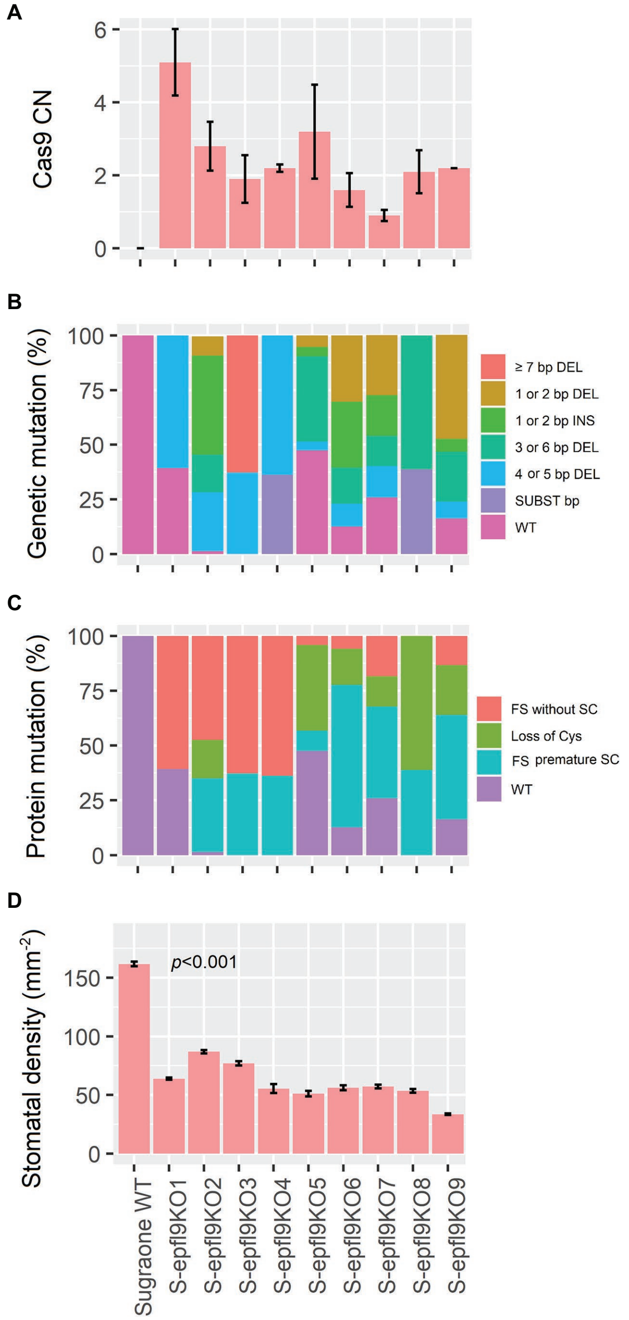


FIGURE 3 | Characterization of 9 ‘Sugraone’ transgenic lines. **(A)** Quantification of *SpCas9* copy numbers (CN) integrated in the plant genome. CN were calculated by Real-time PCR as the mean value of two measurements obtained for two *in-vitro* biological replicates (except for line S-*epfl9*KO9 for which only one value is available). **(B)** Bar plot indicating the mutation profile in the genomic target site on exon 3 of *VvEPF9-1* after CRISPR/Cas9 editing. The mutation pattern and rate (%) of a specific mutation (IN/DEL, insertion/deletion and SUB, substitution) were determined by the number of reads calculated by Illumina sequencing (see **Supplementary Table 4**). Different kinds of mutations are indicated with a different color. WT = the wild-type sequence. **(C)** Bar plot indicating the resulting mutation profile in the functional mature VvEPF9-1 peptide, predicted according to the nucleotide mutations in B (see **Supplementary Table 4**). The different outcomes at protein level are indicated with a different

(Continued)

FIGURE 3 | color. FS, frameshift mutations; SC, stop codons. **(D)** Measurements of stomatal density in the third leaf from the apex. For each plant, four leaves from four biological replicates were analyzed, each in five different technical positions of the same area for a total of twenty measurements.

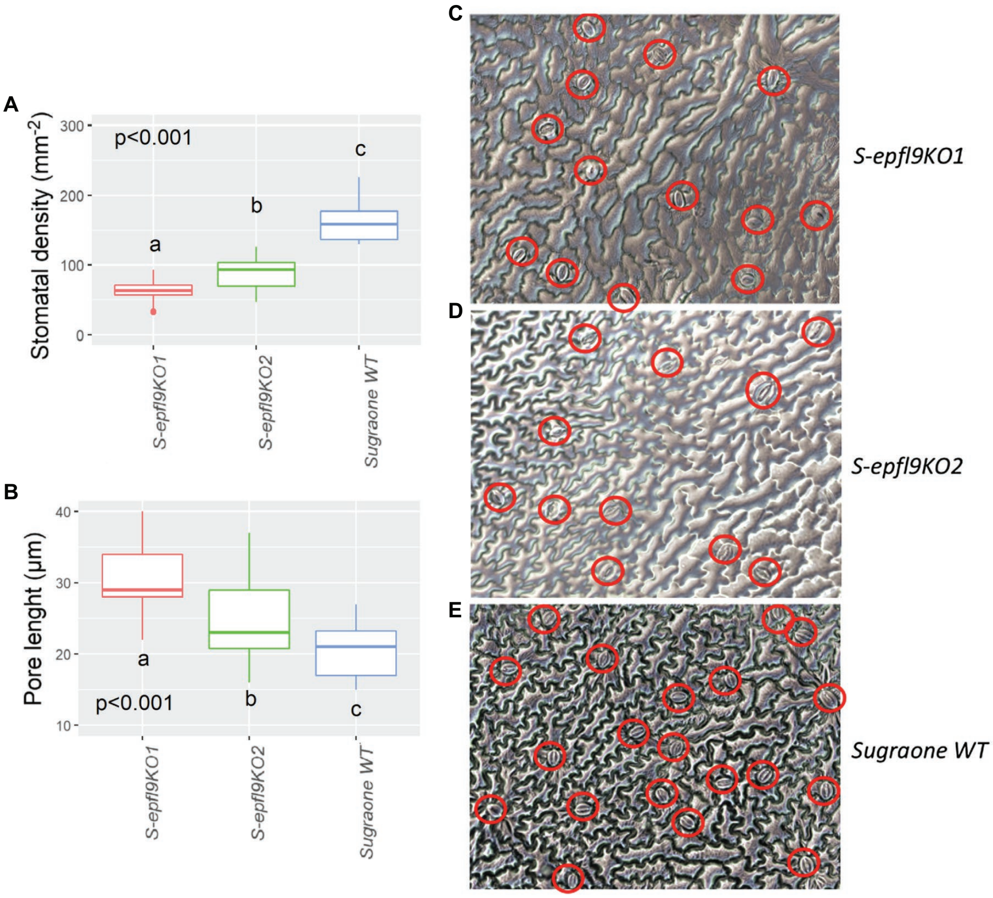


FIGURE 4 | Characterization of stomata in selected *epfl9-1* knock-out mutants. **(A)** Stomatal density for S-*epfl9*KO1, S-*epfl9*KO2 and Sugraone WT. **(B)** Pore length for S-*epfl9*KO1, S-*epfl9*KO2 and Sugraone WT. Whiskers indicate the ranges of the minimum and maximum values. Data were analysed with one-way ANOVA (n = 6–9). Different letters indicate significantly different values according to Fisher's test. **(C–E)** Images of nail polish printing of leaf tissue, respectively, from S-*epfl9*KO1, S-*epfl9*KO2 and Sugraone WT.

in well-watered conditions we evaluated the gas-exchange and transpiration performances of an additional line, S-epfl9KO6, maintained in greenhouse for 12 months (Experiment 3). S-*epfl9*KO6 showed similar SPAD values compared to 'Sugraone' WT (p = 0.607, **Figure 6A**) and trends were observed for leaf temperature with S-*epfl9*KO6 showing increased leaf temperature (p = 0.051, **Figure 6B**) compared to WT. This increase in leaf temperature was associated with a significant decrease in stomatal conductance (p = 0.042, **Figure 6D**) together with a non-significant difference for A_{sat} (p = 0.125, **Figure 6C**). This led to a significant increase in $_{i}WUE$ for S-*epfl9*KO6 compared to control (p = 0.034, **Figure 6E**). No significant differences were observed for projected leaf area (PLA; **Figures 6F,G**) and water use (WU; **Figure 6H**) between S-*epfl9*KO6 and 'Sugraone' WT although a trend was present for WU (p = 0.088).

The Knock-Out of *VvEPF9-1* May Reduce Impact of Water Stress in Grapevine

In vivo gas-exchange measurements at saturating light were carried out throughout the dry down Experiment 2 (**Figure 7**). ANOVA output for each DASA (Day After Stress Application) is shown in **Supplementary Table 5**. *In vivo* CO_2 assimilation rate (A) was significantly reduced by water stress (WS) in 'Sugraone' WT showing a steeper reduction than knock-out lines, although no significant differences were observed for each day and between lines (**Figure 7A**). S-*epfl9*KO1 and S-*epfl9*KO2 maintained a lower stomatal conductance (g_s) than 'Sugraone' WT (p = 0.0276, DASA 5, **Figure 7B**) but intrinsic water-use efficiency $_{i}WUE$ resulted not significantly different between the analysed plants (**Figure 7C**). Transpiration normalized on leaf area was significantly reduced during the WS and for all the lines (**Figure 7D**). The

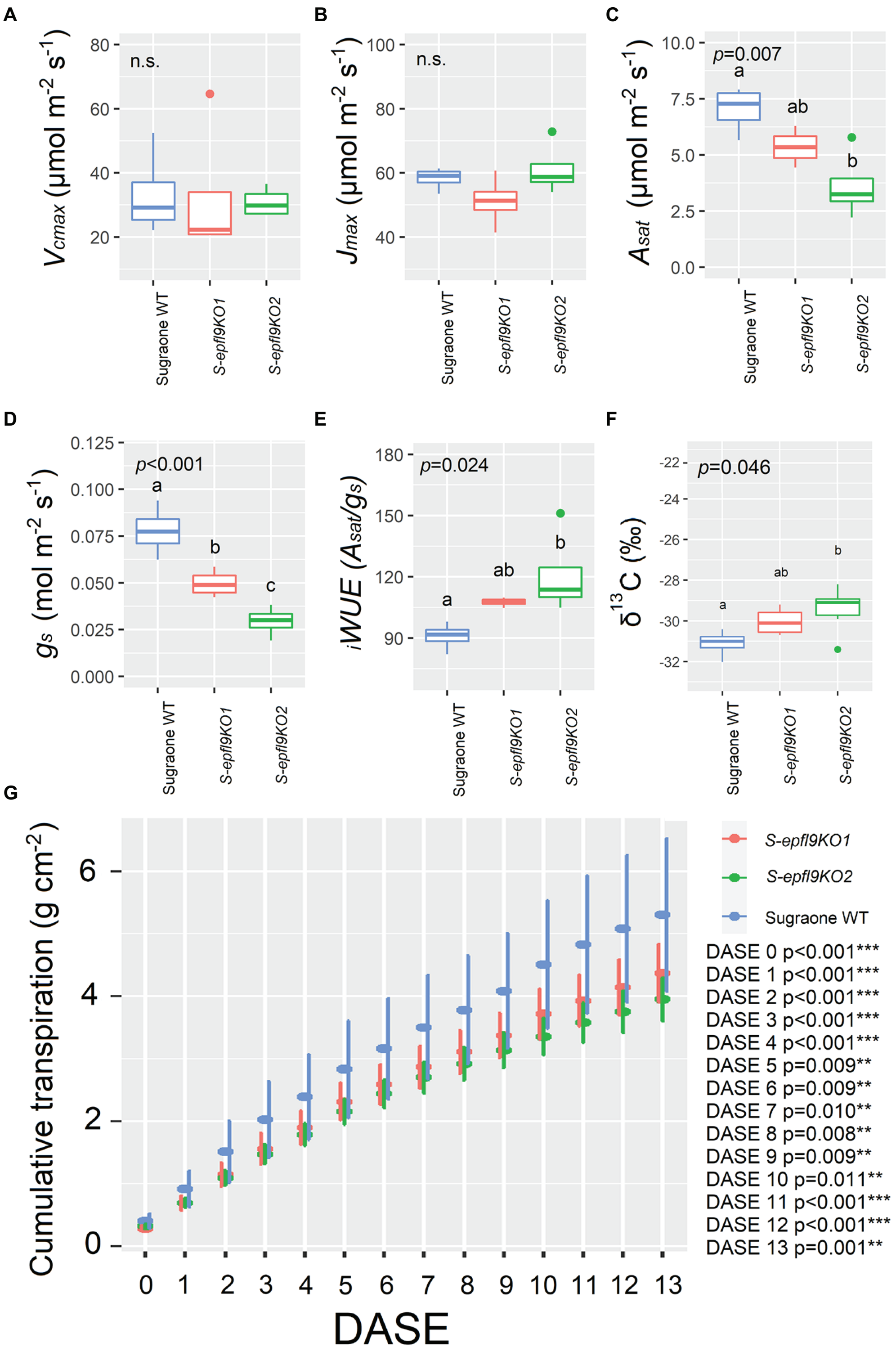


FIGURE 5 | Trait assessment under well-watered (WW) conditions (Experiment 1). **(A)** Maximum velocity of Rubisco carboxylation (V_{cmax}). **(B)** Maximum electron transport rate for RuBP regeneration (J_{max}) estimated with A/C_i curves and following curve fitting (Duursma, 2015; Easlon and Bloom, 2014). **(C)** CO_2 assimilation rate at saturating light (A_{sat}). **(D)** Stomatal conductance (g_s) extrapolated from A/C_i curves at 400 ppm CO_2 concentration and 1,500 μmol m^{-2} s^{-1}. **(E)** Intrinsic water-use efficiency ($_iWUE$) calculated as $_iWUE = A_{sat}/g_s$. **(F)** Carbon Isotope composition (δ^{13}C) analysis. Data were collected on fully expanded leaves of 20 cm tall plants on the twelfth day from the start of the experiment and were elaborated with one-way ANOVA ($n=4$ in **A–E**; $n=3$–6 in **F**). Whiskers indicate the ranges of the minimum and maximum values and different letters indicate significantly different values according to Fisher's test. **(G)** Cumulative water loss assessed gravimetrically and normalized for leaf area estimated *via* RGB imaging for a period of 14 days; DASE = Days After Start of the Experiment. Data were means ± standard error of the mean ($n=5$–6). Data were elaborated with one-way ANOVA for each day (***$p<0.001$, **$p<0.01$, *$p<0.05$). When present, different letters indicate significantly different values according to Fisher's test.

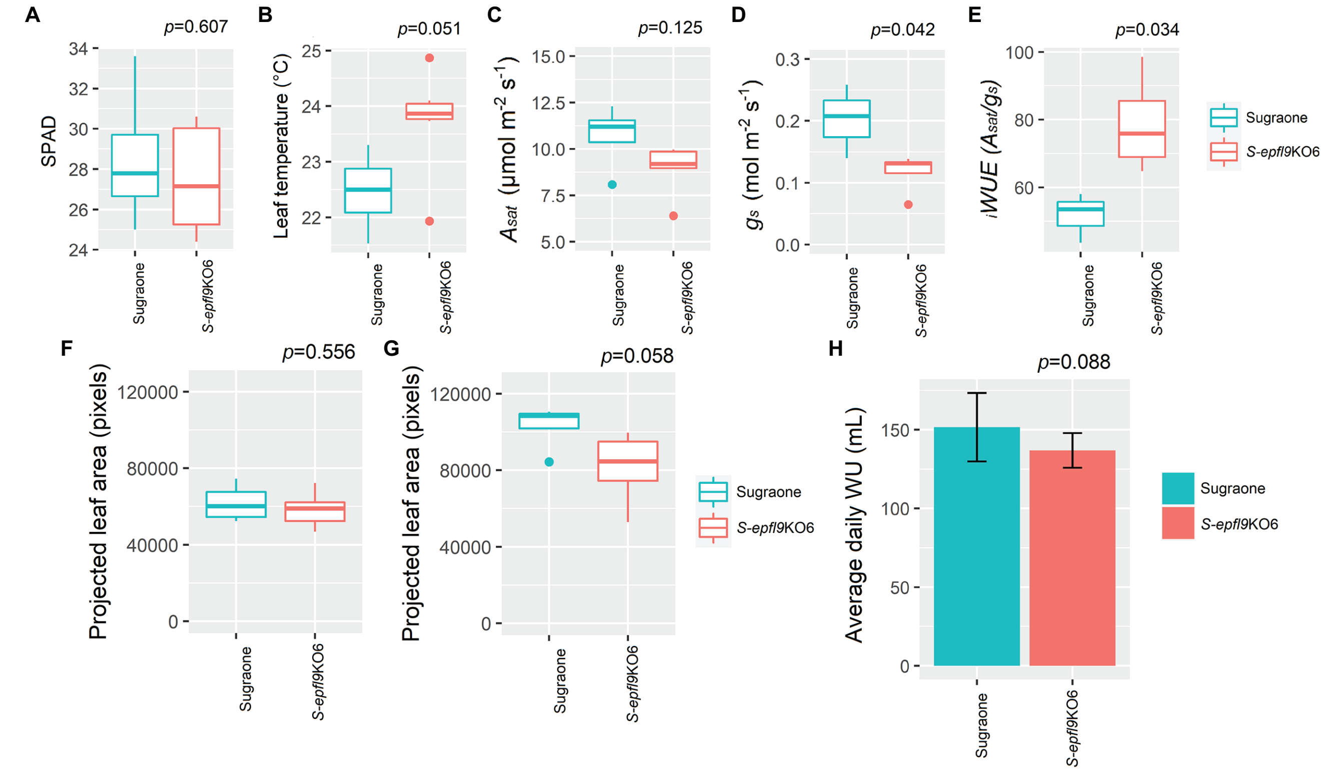


FIGURE 6 | Dynamics of gas exchange, projected leaf area and water-use under well-watered (WW) conditions (Experiment 3) for S-*epfl9*KO6 ($n=6$) and 'Sugraone' WT ($n=4$). **(A)** SPAD values. **(B)** leaf temperature. **(C)** CO_2 assimilation rate at saturating light (A_{sat}). **(D)** stomatal conductance (g_s). **(E)** Intrinsic water-use efficiency ($_iWUE$) calculated as $_iWUE = A_{sat}/g_s$. **(F,G)** projected leaf area (PLA, pixels) collected at day 1 and at day 12, respectively, and **(H)** average daily water-use. For gas exchange measurements, data were collected on fully expanded leaves and were analysed with one-way ANOVA. Whiskers indicate the ranges of the minimum and maximum values.

average fraction of transpirable soil water (FTSW) during the dry down is shown in **Supplementary Figure 3**. There were significant differences ($p<0.05$) between S-*epfl9*KO1 and 'Sugraone' WT, in particular in the first part of stress application (DASA 1 to 4). Trends ($p<0.1$) were observed under severe WS (DASA 10 to 12) with S-*epfl9*KO2 having higher transpiration than 'Sugraone' WT. Carbon isotope composition ($\delta^{13}C$) analysis showed that water stress led to less negative values for all the lines ($p<0.001$) although no significant differences were observed between edited lines and WT ($p=0.186$; **Supplementary Figure 9**).

DISCUSSION

Crops worldwide will experience warmer conditions in the next decades, followed by limited water availability and increasing atmospheric CO_2 concentration (McGranahan and Poling, 2018). Alteration of stomatal density and stomatal size through the genetic manipulation of epidermal patterning factors has been shown to be an effective approach to increase drought tolerance and reduce water loss in several species (Bertolino et al., 2019; Buckley et al., 2020). There is a lot of knowledge about *EPF* gene family in *Arabidopsis* and in domesticated grasses but in perennial crops, which present genetic and physiological differences compared to annual species due to ecological and agronomic peculiar features (Lundgren and Marais, 2020), no evidence has been collected on their role. The aim of our study was to shed light for the first time on the genetic basis of stomatal density traits in grapevine, a perennial woody fruit plant with a longer lifespan than previously studied crops (i.e., longer than 30 years).

Water conservation, higher $_i$*WUE* and enhanced tolerance to multiple stresses (e.g., drought stress combined with heat stress) were achieved in *Arabidopsis* and grasses overexpressing *EPF1/EPF2* or down-regulating *EPFL9*, due to a reduction in stomatal density (Franks et al., 2015; Hughes et al., 2017; Caine et al., 2019; Dunn et al., 2019; Lu et al., 2019). Between these two reverse genetics approaches, we have chosen the second, relying on the knock-out of *VvEPFL9* by the powerful CRISPR/Cas9 gene editing technology.

In the grapevine genus we found two *AtEPFL9* orthologs, we named *VvEPFL9-1* and *VvEPFL9-2*, identical at 82% in the protein region corresponding to the functional peptide and, respectively, sharing 82 and 95% identity with the same region of AtEPFL9 peptide. So far, two *EPFL9* paralogs have been found in maize and rice (Yin et al., 2017; Hepworth et al., 2018; Lu et al., 2019), showing, respectively, 84 and 73% (*ZmEPFL9-1* and *ZmEPFL9-2*) and 82 and 73% (*OsEPFL9-1* and *OsEPFL9-2*) identities to AtEPFL9 functional peptide. It has been suggested that *EPFL9* paralogs in cereals might

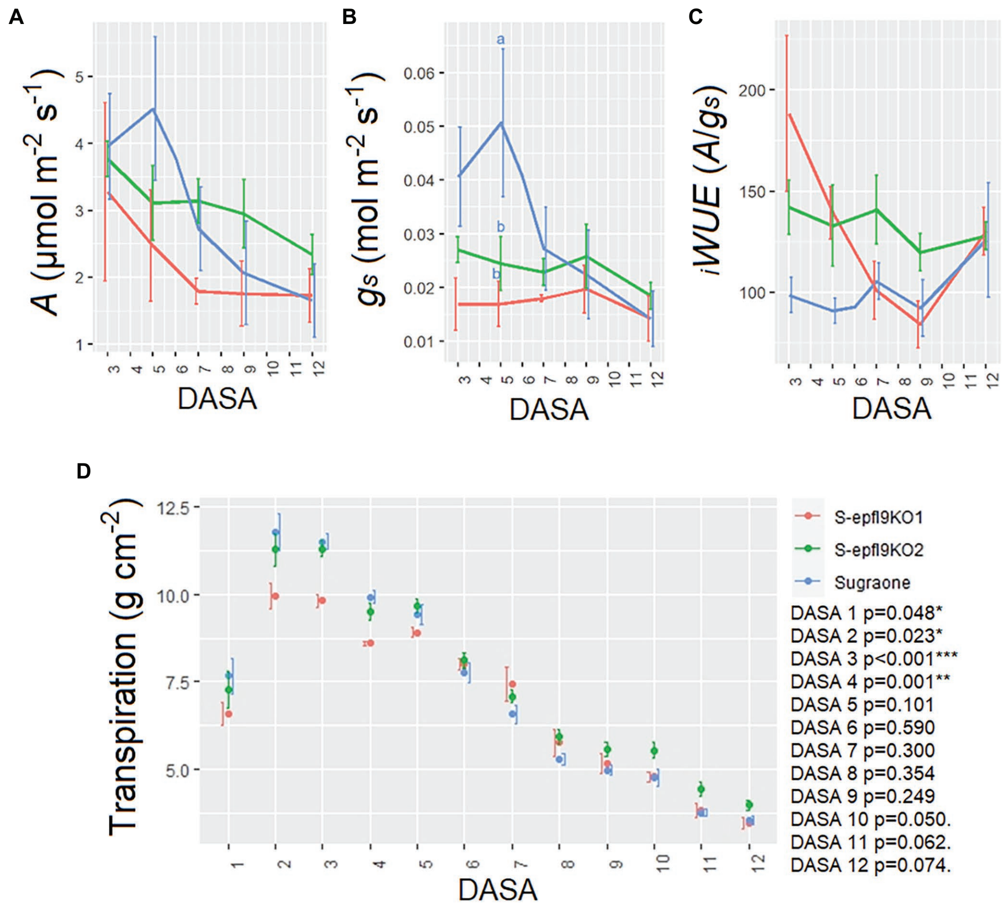


FIGURE 7 | Trait assessment under water stress (WS) conditions. **(A)** *In vivo* CO_2 assimilation rate at saturating light (A_{sat}). **(B)** Stomatal conductance (g_s). **(C)** Intrinsic water-use efficiency $_i$*WUE* calculated as $_iWUE = A_{sat}/g_s$. Data are the means ± standard error of the mean ($n = 4$–6). Data were analysed with one-way ANOVA (value of *p* in the **Supplementary Table 5**) while different letters indicate significant differences between lines according to Fisher's test. **(D)** Transpiration assessed gravimetrically and normalized for leaf area estimated *via* RGB imaging. Data are means ± standard error of the mean ($n = 5$–6). Data were analysed with one-way ANOVA (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $p < 0.1$) for each day. DASA, Days After Stress Application.

be functionally divergent (Lu et al., 2019) but definitive evidence indicating a different function has never been produced. In the study of Lu et al. (2019), the approach used to silence *OsEPF9-1* was RNA interference with a 450 bp-long hairpin RNA, which hardly discriminated between the two variants. In our study, we decided to focus on *VvEPFL9-1*, since at the time the experiment was designed, *VvEPFL9-2* was not anchored to any chromosome in the grapevine reference genome and this uncertainty oriented our choice on *VvEPFL9-1*. According to our data, the knock-out of *VvEPFL9-1* can reduce stomatal density by up to 60%, leading to the hypothesis that *VvEPFL9-1* and *VvEPFL9-2* could be both involved in stomatal induction with a redundant function. A similar approach based on CRISPR/Cas9 technology to knock-out *EPFL9* in rice achieved nearly 90% of stomatal density reduction compared to control by targeting a site on the first exon encoding for the signal peptide and thus not discriminating between *OsEPFL9* paralogs (Yin et al., 2017).

Our study also confirms the crucial role of cysteine residues in the C-terminal functional peptide. This is demonstrated by the lines S-*epfl9*KO5 and S-*epfl9*KO8 in which the loss of the second cysteine (due to a 3 bp-deletion or single base substitution) resulted in a stomatal density reduction similar to the one gained by a full frameshift of the coding sequence. This is consistent with the finding of Ohki et al. (2011) who observed that impairing the formation of a disulphide bond prevented the correct protein folding and function. The design of a sgRNA that directed Cas9 cleavage next to the nucleotide triplet coding for the second cysteine proved to be a good choice for effective 3- and 6- bp deletions. Moreover, our data showed that the retention of almost 50% functional *VvEPFL9-1* in some transgenic lines (S-*epfl9*KO1 and S-*epfl9*KO5) due to a partial editing of the target site, with a substantial maintenance of a WT peptide, still resulted in a significant decrease of SD, suggesting that a threshold amount of peptide may be required for EPFL9-1 to be functionally effective.

Reduction in stomatal density following *VvEPFL9-1* knock-out was significant, although partially compensated by an increase in stomatal size (SS, inferred by pore length measurements). The negative yet non-linear association between SD and SS has been frequently reported in many species (Franks and Beerling, 2009) and often linked to an improved economy of epidermal space allocation with the combination of low SD and high SS as a preferable strategy when low stomatal conductance is required (Doheny-Adams et al., 2012; Lawson and McElwain, 2016). In our work, however, the reduction in SD was accompanied by only a partial compensation for SS.

Stomata are the main drivers of transpiration but at the same time are pivotal for CO_2 uptake for mesophyll photosynthesis (Lawson and Blatt, 2014). For instance, in barley and wheat, a reduction in SD by 50% compared to WT led to a significant reduction in carbon assimilation (A_{sat}) and conductance (g_s) and to an enhanced water use efficiency ($_{\text{i}}WUE$) under optimal growth conditions (Hughes et al., 2017; Dunn et al., 2019). Similarly, in two-months-old 'Sugraone' at well-watered conditions (Experiment 1), we found that a 60% reduction in SD led to a reduced A_{sat} for the edited lines compared to the WT. Additionally, the reduction in g_s was even greater, leading to a higher value of $_{\text{i}}WUE$ (i.e., $A_{\text{sat}}/\ g_s$) in edited versus WT lines. Moreover, the reduction in A_{sat} was not concomitant to reductions in Rubisco velocity (V_{cmax}) or to impairment in electron transport chain (J_{max}) suggesting that the knock-out of *VvEPFL9-1* did not affect the photosynthetic machinery, at least at the conditions applied in this work. In an additional experiment (Experiment 3) carried out under well-watered conditions in a phenotyping platform on older plants than those used in experiment 1, *iWUE* confirmed to be significantly improved in the edited line (S-*epfl9*KO6) compared to 'Sugraone' WT while transpiration (WU) performances were not significantly different. The main sources of variation between the two experiments were plant age and environmental conditions. In Experiment 3 plants were older than in Experiment 1 (12- vs. 2-Months-old) and regarding light conditions, Experiment 1 was carried out under the natural fluctuating light of a greenhouse, while in Experiment 3 plants were subject to a steady-state light pattern. Our results suggest that canopy structure (over-saturation of apical leaves and basal leaves under the sub-saturating light intensities of the greenhouse) may play a role in defining the effectiveness of a reduced stomatal density phenotype. Furthermore, the conditions of dynamic light intensity such as those present in the greenhouse, may have contributed to accentuate the water saving behavior of the lines with lower stomatal density. Indeed, reducing stomatal density can limit stomatal clustering (Harrison et al., 2020) and therefore increase stomatal responsiveness to environmental cues (Faralli et al., 2021). Important differences for g_s were also observed between Experiments 1 and 3, suggesting that plant age and pot-effect significantly influences operating g_s, although the g_s values are inside the ranges shown by Lavoie-Lamoureux et al. (2017) for pot-grown grapevine. *Vitis vinifera* genotypes with reduced SD and, in turn, limited A_{sat} and greater $_{\text{i}}WUE$, may be desirable to improve plant water conservation and to delay sugar accumulation under current and future climatic scenarios (Kuhn et al., 2014; Arrizabalaga-Arriazu et al., 2021). Sugars and organic acids along with various secondary metabolites (e.g., tannins, flavonols, anthocyanins, aroma compounds) are determinants of grape berry quality and their accumulation during berry ripening is the result of the interaction between genotype and environment, a relationship made vulnerable by climate change (Bobeica et al., 2015; Rienth et al., 2021). It is known that grapevine physiology will be impacted by elevated carbon dioxide, increasing temperatures, and extreme heat events during the growing season (De Cortázar-Atauri et al., 2017; Delrot et al., 2020). In particular, high temperature and increasing CO_2 levels are already affecting viticulture (Cook and Wolkovich, 2016; Mosedale et al., 2016; Edwards et al., 2017; Droulia and Charalampopoulos, 2021) with an evident shift towards an earlier onset of phenological stages (Edwards et al., 2017; Alikadic et al., 2019) and accelerated berry ripening (Jones et al., 2005; Parker et al., 2020; Rienth et al., 2021). High temperatures and water stress slow down vine metabolism resulting in a lower accumulation of polyphenols and aromatic compounds in the berries (Tomasi et al., 2011; Jones, 2013; Pons et al., 2017; Venios et al., 2020). Thus, one of the consequences of a compressed phenology may be an earlier sugar accumulation in the berries that leads to anticipated harvest dates when the secondary metabolites content is sub-optimal (Palliotti et al., 2014; Edwards et al., 2017). Although currently several agronomic approaches of source-limitation (i.e., pre-flowering leaf removal, shading nets, anti-transpirant application, etc.) have been set up to delay sugar accumulation in ripening grapes in the field (Palliotti et al., 2014; Prats-Llinàs et al., 2020), stomatal manipulation may be a favorable genetic strategy for the future, that deserves to be further explored also under combined environmental stress and in field trials. In our study, we further applied a water stress experiment to test if and how a reduced stomatal density can affect plant behavior in drought conditions. During a progressive reduction in soil water availability, significant differences in transpiration rate were observed in edited lines compared to WT only under moderate water stress (i.e., DASA 3 and 4). Yet, under severe water stress (e.g., DASA 10–12), some trends ($p < 0.1$) were observed in edited lines showing higher transpiration rate followed by A_{sat} and g_s maintenance. Notably, the reduction in g_s and A_{sat} during the dry-down was evident for WT plants ($p < 0.001$) while this was not significant for edited lines. This conservative behavior induced by reduced SD has been previously associated with a longer period of transpiration maintenance during drought, leading to a prolonged carbon assimilation respect to WT (Caine et al., 2019). In rice, lines overexpressing the *OsEPF1* gene had higher yield than WT when water-stressed at flowering stage (Caine et al., 2019) confirming that water conservation during key-stages of yield formation may be desirable for yield maintenance (Faralli et al., 2019). In addition, limiting plant transpiration could be an advantage for irrigated vineyards in terms of a reduction in water input demand (Keller et al., 2016). In view of an increase in the number of grapevine growing regions where water resources will become limited (Schultz, 2000; Santillán et al., 2020), genotypes with reduced stomatal density will require less units of irrigation water for cultivation area, thus increasing crop water productivity for farmers (Scholasch and Rienth, 2019).

CONCLUSION

To our knowledge, this is the first study describing the function of *VvEPFL9-1* in a perennial fruit crop as well as the physiological advantages of *epfl9-1* knocked-out phenotype under different availability of soil water. In grapevine, reducing stomatal density *via VvEPFL9-1* loss of function can induce water conservation and increase $_{i}WUE$, although an impact of photosynthetic CO_2 absorbance (A_{sat}) was observed in some edited lines. While in several crops, reduced photosynthetic CO_2 uptake can decrease yield and biomass, we speculate that reduced A_{sat} and increased $_{i}WUE$ may be a favorable combination of physiological attributes in grapevine, especially under future climate change scenario. However, at this stage further trials in the field under standard management conditions are required as well as additional evaluations regarding the potential effects of reduced stomatal density under natural environmental fluctuations. To conclude, this work reinforces the concept that stomatal anatomical features constitute a promising target for designing climate change-resilient crops (Franks et al., 2015; Hughes et al., 2017; Bertolino et al., 2019; Caine et al., 2019; Dunn et al., 2019; Lu et al., 2019; Buckley et al., 2020) and provides evidence of this in grapevine, the most economically important fruit crop globally.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**. The original contributions presented in the study are publicly available. This data can be found here: NCBI Sequence Read Archive, BioProject accession number: PRJNA820619.

AUTHOR CONTRIBUTIONS

MC performed plant transformation experiments, plant molecular analysis, phenotyping, statistical analysis, and wrote the paper. MF contributed to the dry-down experiment design, carried out Experiment 3, supervised physiological analysis and statistical elaboration of the data, and wrote the paper. JL carried out alignments and phylogenetic tree and revised the manuscript. LB performed carbon isotope composition analysis and revised the manuscript. SP performed the analysis for the T-DNA integration point determination. CV, MM, WO, and AR supervised and revised the manuscript. LDC conceived the project, designed vectors for gene editing, performed paralogs analysis, transformation experiments, plant molecular characterization, took care of the plants in greenhouse, and wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Autonomous Province of Trento (Italy) in the framework of the Fondazione Edmund Mach (FEM) International PhD initiative and by the ERDF 2014–2020 Program of the Autonomous Province of Trento with EU co-financing (Fruitomics, CUP number: C49H18000000001).

ACKNOWLEDGMENTS

We thank Valentino Poletti for his technical help with culture media preparation and Lisa Giacomelli for sharing with us embryogenic calli of 'Sugraone'. We would like to thank Damiano Gianelle for lending the Li-Cor 6400. A special thanks to Claudio Moser for useful suggestions and discussion and to Nicola Busatto for taking photos for **Figure 2**.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.878001/full#supplementary-material

REFERENCES

Alikadic, A., Pertot, I., Eccel, E., Dolci, C., Zarbo, C., Caffarra, A., et al. (2019). The impact of climate change on grapevine phenology and the influence of altitude: A regional study. *Agric. For. Meteorol.* 271, 73–82. doi: 10.1016/j.agrformet.2019.02.030

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149–157. doi: 10.1038/s41586-019-1711-4

Arrizabalaga-Arriazu, M., Morales, F., Irigoyen, J. J., Hilbert, G., and Pascual, I. (2021). Growth and physiology of four *Vitis vinifera* L. cv. Tempranillo clones under future warming and water deficit regimes. *Aust. J. Grape Wine Res.* 27, 295–307. doi: 10.1111/ajgw.12494

Bertolino, L. T., Caine, R. S., and Gray, J. E. (2019). Impact of stomatal density and morphology on water-use efficiency in a changing world. *Front. Plant Sci.* 10:225. doi: 10.3389/fpls.2019.00225

Bobeica, N., Poni, S., Hilbert, G., Renaud, C., Gomès, E., Delrot, S., et al. (2015). Differential responses of sugar, organic acids and anthocyanins to source-sink modulation in Cabernet Sauvignon and Sangiovese grapevines. *Front. Plant Sci.* 6:382. doi: 10.3389/fpls.2015.00382

Bota, J., Tomás, M., Flexas, J., Medrano, H., and Escalona, J. M. (2016). Differences among grapevine cultivars in their stomatal behavior and water use efficiency under progressive water stress. *Agric. Water Manag.* 164, 91–99. doi: 10.1016/j.agwat.2015.07.016

Buckley, C. R., Caine, R. S., and Gray, J. E. (2020). Pores for thought: can genetic manipulation of stomatal density protect future rice yields? *Front. Plant Sci.* 10:1783. doi: 10.3389/fpls.2019.01783

Caine, R. S., Yin, X., Sloan, J., Harrison, E. L., Mohammed, U., Fulton, T., et al. (2019). Rice with reduced stomatal density conserves water and has improved drought tolerance under future climate conditions. *New Phytol.* 221, 371–384. doi: 10.1111/nph.15344

Chen, T., Peng, J., Yin, X., Li, M., Xiang, G., Wang, Y., et al. (2021). Importin-αs are required for the nuclear localization and function of the *Plasmopara viticola* effector PvAVH53. *Hortic. Res.* 8, 46–12. doi: 10.1038/s41438-021-00482-6

Chen, K., Wang, Y., Zhang, R., Zhang, H., and Gao, C. (2019). CRISPR/Cas genome editing and precision plant breeding in agriculture. *Annu.*

Rev. Plant Biol. 70, 667–697. doi: 10.1146/annurev-arplant-050718-100049

Chen, L., Wu, Z., and Hou, S. (2020). SPEECHLESS speaks loudly in stomatal development. *Front. Plant Sci.* 11:114. doi: 10.3389/fpls.2020.00114

Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., et al. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* 37, 224–226. doi: 10.1038/s41587-019-0032-3

Cook, B. I., and Wolkovich, E. M. (2016). Climate change decouples drought from early wine grape harvests in France. *Nat. Clim. Chang.* 6, 715–719. doi: 10.1038/nclimate2960

Dalla Costa, L., Piazza, S., Pompili, V., Salvagnin, U., Cestaro, A., Moffa, L., et al. (2020). Strategies to produce T-DNA free CRISPRed fruit trees via *Agrobacterium tumefaciens* stable gene transfer. *Sci. Rep.* 10:20155. doi: 10.1038/s41598-020-77110-1

Dalla Costa, L., Vaccari, I., Mandolini, M., and Martinelli, L. (2009). Elaboration of a reliable strategy based on real-time PCR to characterize genetically modified plantlets and to evaluate the efficiency of a marker gene removal in grape (*Vitis* spp.). *J. Agric. Food Chem.* 57, 2668–2677. doi: 10.1021/jf802740m

Dalla Costa, L., Vinciguerra, D., Giacomelli, L., Salvagnin, U., Piazza, S., Spinella, K., et al. (2022). Integrated approach for the molecular characterization of edited plants obtained via *Agrobacterium tumefaciens*-mediated gene transfer. *Eur. Food Res. Technol.* 248, 289–299. doi: 10.1007/s00217-021-03881-0

Dayer, S., Herrera, J. C., Dai, Z., Burlett, R., Lamarque, L. J., Delzon, S., et al. (2020). The sequence and thresholds of leaf hydraulic traits underlying grapevine varietal differences in drought tolerance. *J. Exp. Bot.* 71, 4333–4344. doi: 10.1093/jxb/eraa186

De Cortázar-Atauri, I. G., Duchêne, É., Destrac-Irvine, A., Barbeau, G., De Rességuier, L., Lacombe, T., et al. (2017). Grapevine phenology in France: from past observations to future evolutions in the context of climate change. *Oeno One* 51, 115–126. doi: 10.20870/oeno-one.2016.0.0.1622

Delrot, S., Grimplet, J., Carbonell-bejerano, P., Schwandner, A., Bert, P., Bavaresco, L., et al. (2020). "Genomic Designing of Climate-Smart Fruit Crops" in *Genomic Designing of Climate-Smart Fruit Crops*. ed. C. Kole (Switzerland: Springer International Publishing), 157–270.

Doheny-Adams, T., Hunt, L., Franks, P. J., Beerling, D. J., and Gray, J. E. (2012). Genetic manipulation of stomatal density influences stomatal size, plant growth and tolerance to restricted water supply across a growth carbon dioxide gradient. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 547–555. doi: 10.1098/rstb.2011.0272

Droulia, F., and Charalampopoulos, I. (2021). Future climate change impacts on european viticulture: a review on recent scientific advances. *Atmosphere (Basel).* 12:495. doi: 10.3390/atmos12040495

Dunn, J., Hunt, L., Afsharinafar, M., Meselmani, M.Al, Mitchell, A., Howells, R., et al. (2019). Reduced stomatal density in bread wheat leads to increased water-use efficiency. *J. Exp. Bot.* 70, 4737–4748. doi:10.1093/jxb/erz248.

Duursma, R. A. (2015). Plantecophys - An R package for analysing and modelling leaf gas exchange data. *PLoS One* 10, 1–13. doi: 10.1371/journal.pone.0143346

Easlon, H. M., and Bloom, A. J. (2014). Easy leaf area: Automated digital image analysis for rapid and accurate measurement of leaf area. *Appl. Plant Sci.* 2:1400033. doi: 10.3732/apps.1400033

Edwards, E. J., Unwin, D., Kilmister, R., Treeby, M., and Ollat, N. (2017). Multi-seasonal effects of warming and elevated CO2 on the physiology, growth and production of mature, field grown, shiraz grapevines. *J. Int. des Sci. Vigne Vin* 51, 127–132. doi: 10.20870/oeno-one.2016.0.0.1586

Faralli, M., Bontempo, L., Bianchedi, P. L., Moser, C., Bertamini, M., Lawson, T., et al. (2021). Natural variation in stomatal dynamics drives divergence in heat stress tolerance and contributes to seasonal intrinsic water-use efficiency in *Vitis vinifera* (subsp. *sativa* and *sylvestris*). *J. Exp. Bot.* doi: 10.1093/jxb/erab552 [Epub Ahead of Print]

Faralli, M., Matthews, J., and Lawson, T. (2019). Exploiting natural variation and genetic manipulation of stomatal conductance for crop improvement. *Curr. Opin. Plant Biol.* 49, 1–7. doi: 10.1016/j.pbi.2019.01.003

Franks, P. J., and Beerling, D. J. (2009). Maximum leaf conductance driven by CO2 effects on stomatal size and density over geologic time. *Proc. Natl. Acad. Sci. U. S. A.* 106, 10343–10347. doi: 10.1073/pnas.0904209106

Franks, P. J., Doheny-Adams, W., Britton-Harper, Z. J., and Gray, J. E. (2015). Increasing water-use efficiency directly through genetic manipulation of stomatal density. *New Phytol.* 207, 188–195. doi: 10.1111/nph.13347

Gambetta, G. A., Herrera, J. C., Dayer, S., Feng, Q., Hochberg, U., and Castellarin, S. D. (2020). The physiology of drought stress in grapevine: towards an integrative definition of drought tolerance. *J. Exp. Bot.* 71, 4658–4676. doi: 10.1093/jxb/eraa245

Giacomelli, L., Zeilmaker, T., Malnoy, M., van der Rouppe Voort, J., and Moser, C. (2019). Generation of mildew-resistant grapevine clones via genome editing. *Acta Hortic.* 1248, 195–200. doi: 10.17660/ActaHortic.2019.1248.28

Hara, K., Yokoo, T., Kajita, R., Onishi, T., Yahata, S., Peterson, K. M., et al. (2009). Epidermal cell density is autoregulated via a secretory peptide, EPIDERMAL PATTERNING FACTOR 2 in Arabidopsis leaves. *Plant Cell Physiol.* 50, 1019–1031. doi: 10.1093/pcp/pcp068

Harrison, E. L., Arce Cubas, L., Gray, J. E., and Hepworth, C. (2020). The influence of stomatal morphology and distribution on photosynthetic gas exchange. *Plant J.* 101, 768–779. doi: 10.1111/tpj.14560

Hepworth, C., Caine, R. S., Harrison, E. L., Sloan, J., and Gray, J. E. (2018). Stomatal development: focusing on the grasses. *Curr. Opin. Plant Biol.* 41, 1–7. doi: 10.1016/j.pbi.2017.07.009

Hepworth, C., Doheny-Adams, T., Hunt, L., Cameron, D. D., and Gray, J. E. (2015). Manipulating stomatal density enhances drought tolerance without deleterious effect on nutrient uptake. *New Phytol.* 208, 336–341. doi: 10.1111/nph.13598

Hess, G. T., Tycko, J., Yao, D., and Bassik, M. C. (2017). Methods and applications of CRISPR-mediated base editing in eukaryotic genomes. *Mol. Cell* 68, 26–43. doi: 10.1016/j.molcel.2017.09.029

Hughes, J., Hepworth, C., Dutton, C., Dunn, J. A., Hunt, L., Stephens, J., et al. (2017). Reducing stomatal density in barley improves drought tolerance without impacting on yield. *Plant Physiol.* 174, 776–787. doi: 10.1104/pp.16.01844

Hunt, L., Bailey, K. J., and Gray, J. E. (2010). The signalling peptide EPFL9 is a positive regulator of stomatal development. *New Phytol.* 186, 609–614. doi: 10.1111/j.1469-8137.2010.03200.x

INTEGRAPE Workshop (2021). *In XIth International Symposium on Grapevine Physiology and Biotechnology*. Stellenbosch, South Africa.

IPCC (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Switzerland: IPCC.

Jain, M. (2015). Function genomics of abiotic stress tolerance in plants: a CRISPR approach. *Front. Plant Sci.* 6:375. doi: 10.3389/fpls.2015.00375

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821. doi: 10.1126/science.1225829

Jones, G. (2013). "Winegrape phenology," in *Phenology: An Integrative Environmental Science*. ed. M. D. Schwartz (Dordrecht: Springer), 563–584.

Jones, G. V., White, M. A., Cooper, O. R., and Storchmann, K. (2005). Climate change and global wine quality. *Clim. Chang.* 73, 319–343. doi: 10.1007/s10584-005-4704-2

Keller, M., Romero, P., Gohil, H., Smithyman, R. P., Riley, W. R., Casassa, L. F., et al. (2016). Deficit irrigation alters grapevine growth, physiology, and fruit microclimate. *Am. J. Enol. Vitic.* 67, 426–435. doi: 10.5344/ajev.2016.16032

Kondo, T., Kajita, R., Miyazaki, A., Hokoyama, M., Nakamura-Miura, T., Mizuno, S., et al. (2010). Stomatal density is controlled by a mesophyll-derived signaling molecule. *Plant Cell Physiol.* 51, 1–8. doi: 10.1093/pcp/pcp180

Kuhn, N., Guan, L., Dai, Z. W., Wu, B. H., Lauvergeat, V., Gomès, E., et al. (2014). Berry ripening: recently heard through the grapevine. *J. Exp. Bot.* 65, 4543–4559. doi: 10.1093/jxb/ert395

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lavoie-Lamoureux, A., Sacco, D., Risse, P. A., and Lovisolo, C. (2017). Factors influencing stomatal conductance in response to water availability in grapevine: a meta-analysis. *Physiol. Plant.* 159, 468–482. doi: 10.1111/ppl.12530

Lawson, T., and Blatt, M. R. (2014). Stomatal size, speed, and responsiveness impact on photosynthesis and water use efficiency. *Plant Physiol.* 164, 1556–1570. doi: 10.1104/pp.114.237107

Lawson, T., and McElwain, J. C. (2016). Evolutionary trade-offs in stomatal spacing. *New Phytol.* 210, 1149–1151. doi: 10.1111/nph.13972

Lee, J. S., Hnilova, M., Maes, M., Lin, Y. C. L., Putarjunan, A., Han, S. K., et al. (2015). Competitive binding of antagonistic peptides fine-tunes stomatal patterning. *Nature* 522, 439–443. doi: 10.1038/nature14561

Li, M. Y., Jiao, Y. T., Wang, Y. T., Zhang, N., Wang, B. B., Liu, R. Q., et al. (2020). CRISPR/Cas9-mediated VvPR4b editing decreases downy mildew resistance in grapevine (*Vitis vinifera* L.). *Hortic. Res.* 7:149. doi: 10.1038/s41438-020-00371-4

Liu, D., Hu, R., Palla, K. J., Tuskan, G. A., and Yang, X. (2016). Advances and perspectives on the use of CRISPR/Cas9 systems in plant genomics research. *Curr. Opin. Plant Biol.* 30, 70–77. doi: 10.1016/j.pbi.2016.01.007

Lovisolo, C., Perrone, I., Carra, A., Ferrandino, A., Flexas, J., Medrano, H., et al. (2010). Drought-induced changes in development and function of grapevine (*Vitis* spp.) organs and in their hydraulic and non-hydraulic interactions at the whole-plant level: A physiological and molecular update. *Funct. Plant Biol.* 37, 98–116. doi: 10.1071/FP09191

Lu, J., He, J., Zhou, X., Zhong, J., Li, J., and Liang, Y. K. (2019). Homologous genes of epidermal patterning factor regulate stomatal development in rice. *J. Plant Physiol.* 235, 18–27. doi: 10.1016/j.jplph.2019.01.010

Lundgren, M. R., and Marais, D. L. (2020). Life history variation as a model for understanding trade-offs in plant – environment interactions. *Curr. Biol.* 30, R180–R189. doi: 10.1016/j.cub.2020.01.003

Malnoy, M., Viola, R., Jung, M.-H., Koo, O.-J., Kim, S., Kim, J.-S., et al. (2016). DNA-free genetically edited grapevine and apple protoplast using CRISPR/Cas9 ribonucleoproteins. *Front. Plant Sci.* 7:1904. doi: 10.3389/fpls.2016.01904

Marshall, E., Costa, L. M., and Gutierrez-Marcos, J. (2011). Cysteine-rich peptides (CRPs) mediate diverse aspects of cell-cell communication in plant reproduction and development. *J. Exp. Bot.* 62, 1677–1686. doi: 10.1093/jxb/err002

McCown, B. H., and Lloyd, G. (1981). Woody plant medium (WPM) - a mineral nutrient formulation for microculture of woody plant-species. *Hortic. Sci.* 16:453

McGranahan, D. A., and Poling, B. N. (2018). Trait-based responses of seven annual crops to elevated CO2 and water limitation. *Renew. Agric. Food Syst.* 33, 259–266. doi: 10.1017/S1742170517000692

Mohammed, U., Caine, R. S., Atkinson, J. A., Harrison, E. L., Wells, D., Chater, C. C., et al. (2019). Rice plants overexpressing OsEPF1 show reduced stomatal density and increased root cortical aerenchyma formation. *Sci. Rep.* 9, 5584–5513. doi: 10.1038/s41598-019-41922-7

Morales-Navarro, S., Pérez-Díaz, R., Ortega, A., de Marcos, A., Mena, M., Fenoll, C., et al. (2018). Overexpression of a SDD1-like gene from wild tomato decreases stomatal density and enhances dehydration avoidance in arabidopsis and cultivated tomato. *Front. Plant Sci.* 9:940. doi: 10.3389/fpls.2018.00940

Mosedale, J. R., Abernethy, K. E., Smart, R. E., Wilson, R. J., and Maclean, I. M. D. (2016). Climate change impacts and adaptive strategies: lessons from the grapevine. *Glob. Chang. Biol.* 22, 3814–3828. doi: 10.1111/gcb.13406

Ohki, S., Takeuchi, M., and Mori, M. (2011). The NMR structure of stomagen reveals the basis of stomatal density regulation by plant peptide hormones. *Nat. Commun.* 2:512. doi: 10.1038/ncomms1520

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., et al. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Palliotti, A., Tombesi, S., Silvestroni, O., Lanari, V., Gatti, M., and Poni, S. (2014). Changes in vineyard establishment and canopy management urged by earlier climate-related grape ripening: A review. *Sci. Hortic.* 178, 43–54. doi: 10.1016/j.scienta.2014.07.039

Parker, A. K., de Cortázar-Atauri, I. G., Trought, M. C. T., Destrac, A., Agnew, R., Sturman, A., et al. (2020). Adaptation to climate change by determining grapevine cultivar differences using temperature-based phenology models. *Oeno One* 54, 955–974. doi: 10.20870/OENO-ONE.2020.54.4.3861

Pillitteri, L. J., Sloan, D. B., Bogenschutz, N. L., and Torii, K. U. (2007). Termination of asymmetric cell division and differentiation of stomata. *Nature* 445, 501–505. doi: 10.1038/nature05467

Podevin, N., Davies, H. V., Hartung, F., Nogué, F., and Casacuberta, J. M. (2013). Site-directed nucleases: A paradigm shift in predictable, knowledge-based plant breeding. *Trends Biotechnol.* 31, 375–383. doi: 10.1016/j.tibtech.2013.03.004

Pons, A., Allamy, L., Schüttler, A., Rauhut, D., Thibon, C., and Darriet, P. (2017). What is the expected impact of climate change on wine aroma compounds and their precursors in grape? *Oeno One* 51, 141–146. doi: 10.20870/oeno-one.2016.0.0.1868

Prats-Llinàs, M. T., Nieto, H., DeJong, T. M., Girona, J., and Marsal, J. (2020). Using forced regrowth to manipulate chardonnay grapevine (*Vitis vinifera* L.) development to evaluate phenological stage responses to temperature. *Sci. Hortic.* 262:109065. doi: 10.1016/j.scienta.2019.109065

Rienth, M., Vigneron, N., Darriet, P., Sweetman, C., Burbidge, C., Bonghi, C., et al. (2021). Grape berry secondary metabolites and their modulation by abiotic factors in a climate change scenario–a review. *Front. Plant Sci.* 12:643258. doi: 10.3389/fpls.2021.643258

Santillán, D., Garrote, L., Iglesias, A., and Sotes, V. (2020). Climate change risks and adaptation: new indicators for Mediterranean viticulture. *Mitig. Adapt. Strateg. Glob. Chang.* 25, 881–899. doi: 10.1007/s11027-019-09899-w

Scholasch, T., and Rienth, M. (2019). Review of water deficit mediated changes in vine and berry physiology; consequences for the optimization of irrigation strategies. *Oeno One* 53, 423–444. doi: 10.20870/oeno-one.2019.53.3.2329

Schultz, H. R. (2000). Climate change and viticulture: A European perspective on climatology, carbon dioxide and UV-B effects. *Aust. J. Grape Wine Res.* 6, 2–12. doi: 10.1111/j.1755-0238.2000.tb00156.x

Schultz, H. R. (2003). Differences in hydraulic architecture account for near-isohydric and anisohydric behaviour of two field-grown *Vitis vinifera* L. cultivars during drought. *Plant Cell Environ.* 26, 1393–1405. doi: 10.1046/j.1365-3040.2003.01064.x

Scintilla, S., Salvagnin, U., Giacomelli, L., Zeilmaker, T., Malnoy, M. A., van der Voort, J. R., et al. (2021). Regeneration of plants from DNA-free edited grapevine protoplasts. *bioRxiv*. doi: 10.1101/2021.07.16.452503

Sherwood, S., and Fu, Q. (2014). A drier future? *Science* 343, 737–739. doi: 10.1126/science.1247620

Shimada, T., Sugano, S. S., and Hara-Nishimura, I. (2011). Positive and negative peptide signals control stomatal density. *Cell. Mol. Life Sci.* 68, 2081–2088. doi: 10.1007/s00018-011-0685-7

Soar, C. J., Dry, P. R., and Loveys, B. R. (2006). Scion photosynthesis and leaf gas exchange in *Vitis vinifera* L. cv. Shiraz: mediation of rootstock effects via xylem sap ABA. *Aust. J. Grape Wine Res.* 12, 82–96. doi: 10.1111/j.1755-0238.2006.tb00047.x

Sugano, S. S., Shimada, T., Imai, Y., Okawa, K., Tamai, A., Mori, M., et al. (2010). Stomagen positively regulates stomatal density in Arabidopsis. *Nature* 463, 241–244. doi: 10.1038/nature08682

Tomasi, D., Jones, G. V., Giust, M., Lovat, L., and Gaiotti, F. (2011). Grapevine phenology and climate change: relationships and trends in the Veneto region of Italy for 1964–2009. *Am. J. Enol. Vitic.* 62, 329–339. doi: 10.5344/ajev.2011.10108

Tombesi, S., Nardini, A., Frioni, T., Soccolini, M., Zadra, C., Farinelli, D., et al. (2015). Stomatal closure is induced by hydraulic signals and maintained by ABA in drought-stressed grapevine. *Sci. Rep.* 5, 1–12. doi: 10.1038/srep12449

Van Leeuwen, C., and Destrac-Irvine, A. (2017). Modified grape composition under climate change conditions requires adaptations in the vineyard. *Oeno One* 51, 147–154. doi: 10.20870/oeno-one.2016.0.0.1647

Van Leeuwen, C., Destrac-Irvine, A., Dubernet, M., Duchêne, E., Gowdy, M., Marguerit, E., et al. (2019). An update on the impact of climate change in viticulture and potential adaptations. *Agronomy* 9, 1–20. doi: 10.3390/agronomy9090514

Venios, X., Korkas, E., Nisiotou, A., and Banilas, G. (2020). Grapevine responses to heat stress and global warming. *Plan. Theory* 9, 1–15. doi: 10.3390/plants9121754

Villalobos-González, L., Muñoz-Araya, M., Franck, N., and Pastenes, C. (2019). Controversies in midday water potential regulation and stomatal behavior might result from the environment, genotype, and/or rootstock: evidence from Carménère and Syrah grapevine varieties. *Front. Plant Sci.* 10:1522. doi: 10.3389/fpls.2019.01522

Wan, D. Y., Guo, Y., Cheng, Y., Hu, Y., Xiao, S., Wang, Y., et al. (2020). CRISPR/Cas9-mediated mutagenesis of VvMLO3 results in enhanced resistance to powdery mildew in grapevine (*Vitis vinifera*). *Hortic. Res.* 7:116. doi: 10.1038/s41438-020-0339-8

Wang, Z., Wong, D. C. J., Wang, Y., Xu, G., Ren, C., Liu, Y., et al. (2021). GRAS-domain transcription factor PAT1 regulates jasmonic acid biosynthesis in grape cold stress response. *Plant Physiol.* 186, 1660–1678. doi: 10.1093/PLPHYS/KIAB142

Yin, X., Biswal, A. K., Dionora, J., Perdigon, K. M., Balahadia, C. P., Mazumdar, S., et al. (2017). CRISPR-Cas9 and CRISPR-Cpf1 mediated targeting of a stomatal developmental gene EPFL9 in rice. *Plant Cell Rep.* 36, 745–757. doi: 10.1007/s00299-017-2118-z

Zoulias, N., Harrison, E. L., Casson, S. A., and Gray, J. E. (2018). Molecular control of stomatal development. *Biochem. J.* 475, 441–454. doi: 10.1042/BCJ20170413

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 August 2022
DOI 10.3389/fpls.2022.945738

Tissue-specific expression of *Ruby* in Mexican lime (*C. aurantifolia*) confers anthocyanin accumulation in fruit

Roger Thilmony[1], Kasturi Dasgupta[1,2], Min Shao[1,2], Daren Harris[1], Jake Hartman[1], Leslie A. Harden[3], Ron Chan[1] and James G. Thomson[1]*

[1]Crop Improvement and Genetics, Western Regional Research Center, United States Department of Agriculture (USDA)-Agricultural Research Service (ARS), Albany, CA, United States, [2]Citrus Research Board, Visalia, CA, United States, [3]Produce Safety and Microbiology Research, Western Regional Research Center, United States Department of Agriculture (USDA)-Agricultural Research Service (ARS), Albany, CA, United States

OPEN ACCESS

EDITED BY
Concetta Licciardello,
CREA Research Centre for Olive, Fruit and Citrus Crops, Italy

REVIEWED BY
Mickael Malnoy,
Fondazione Edmund Mach, Italy
Laura Bassolino,
Council for Agricultural and Economics Research (CREA), Italy

*CORRESPONDENCE
James G. Thomson
James.Thomson@usda.gov

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 16 May 2022
ACCEPTED 19 July 2022
PUBLISHED 08 August 2022

CITATION
Thilmony R, Dasgupta K, Shao M, Harris D, Hartman J, Harden LA, Chan R and Thomson JG (2022) Tissue-specific expression of *Ruby* in Mexican lime (*C. aurantifolia*) confers anthocyanin accumulation in fruit.
Front. Plant Sci. 13:945738.
doi: 10.3389/fpls.2022.945738

Tissue specific promoters are important tools for the precise genetic engineering of crop plants. Four fruit-preferential promoters were examined for their ability to confer a novel fruit trait in transgenic Mexican lime (*Citrus aurantifolia*). The *Ruby* transcription factor activates fruit anthocyanin accumulation within Moro blood orange and has been shown to function in activating anthocyanin accumulation in heterologous plant species. Although the *CitVO1, CitUNK, SlE8,* and *PamMybA* promoters were previously shown to confer strong fruit-preferential expression in transgenic tomato, they exhibited no detectable expression in transgenic Mexican lime trees. In contrast, the *CitWax* promoter exhibited high fruit-preferential expression of *Ruby*, conferring strong anthocyanin accumulation within the fruit juice sac tissue and moderate activity in floral/reproductive tissues. In some of the transgenic trees with high levels of flower and fruit anthocyanin accumulation, juvenile leaves also exhibited purple coloration, but the color disappeared as the leaves matured. We show that the *CitWax* promoter enables the expression of *Ruby* to produce anthocyanin colored fruit desired by consumers. The production of this antioxidant metabolite increases the fruits nutritional value and may provide added health benefits.

KEYWORDS
anthocyanin, citrus (citrus sinensis), *CsMybA1*, *Ruby*, fruit biotechnology, promoter

Introduction

Biotechnology offers the potential to improve agricultural crop production and facilitate the development to design novel commodities for use as food, feed, or fuel. One avenue of biotechnological improvement is the use of gene expression control elements (promoters) to precisely control when, where and how the introduced genes/traits will be expressed. Although numerous promoters have been identified that confer constitutive, or inducible expression in transgenic plants, fewer organ- or tissue-specific promoters that confer expression within specific cell or tissue types have been identified, particularly in crop plants.

Fruit are important sources of nutrients, minerals, vitamins, and dietary fiber in the human diet, and as such, significant efforts have been made to breed for fruit with higher yield, better quality, and other desirable traits. Citrus is one of the most important fruit tree crops worldwide their fruit are considered healthy food because they are low in fat and rich in dietary fiber, vitamin C, vitamin B (thiamin, pyridoxine, niacin, riboflavin, pantothenic acid, and folate), vitamin A, carotenoids, flavonoids, and limonoids (Liu et al., 2012). Most citrus fruit have an orange or yellow color due to the presence of carotenoids, except blood orange fruit which exhibit a bright purple color due to the presence of anthocyanins. Anthocyanins are synthesized from flavonoid precursors through a complex expression anthocyanin biosynthetic enzymes and regulatory genes including WD-repeat proteins, and basic helix-loop-helix (bHLH) and MYB transcription factors (Holton and Cornish, 1995; Baudry et al., 2004) and provide pigmentation to many plant tissues. Anthocyanin-activating MYB transcription factors have been identified in many plant species including Arabidopsis, *Citrus sinensis*, *Prunus americana, Ipomoea batatas,* and *Vitis vinifera* (Dasgupta et al., 2017).

The *Citrus* genus has a limited number of species that express anthocyanins within their fruit. The trait can appear in the young shoots and floral tissues of many lemon cultivars (Fabroni et al., 2016), and in specific 'blood orange' cultivars like 'Tarocco, 'Moro,' and 'Sanguinello' fruit with purple flesh and rind produced, but anthocyanin accumulation is typically not observed within the young shoots or flowers. The presence of pigments like anthocyanins is believed to protect juvenile leaves from light stress and insect predation (Gould et al., 2000; Steyn et al., 2002; Karageorgou and Manetas, 2006). The Moro cultivar was previously shown to carry a novel copia-like retrotransposon sequence inserted within the promoter region of a MYB transcription factor *Ruby* and shown to be responsible for the activation of anthocyanin production in blood orange fruit (Butelli et al., 2012). Environmental factors also play a significant role in production of the blood orange trait and it appears to require hot, dry days followed by cool nights. Due to these specific conditions the blood orange is widely cultivated in southern Italy and Sicily. Mondello et al. (2000), Lo Piero (2015), but when grown in the subtropical areas, these varieties do not reliably accumulate sufficient anthocyanins to develop the purple color (Lee, 2002). The presence of anthocyanin pigments in these specific cultivars makes them attractive to consumers. In addition, anthocyanins have been shown to provide several health benefits if consumed in sufficient quantities (Bridle and Timberlake, 1997; Lo Piero, 2015). Previous studies have shown that anthocyanins are strong free radical scavengers (Sanchez-Moreno, 2002; De Beer et al., 2003) and have powerful antioxidant capacities (Deighton et al., 2000; Lila, 2004; Legua et al., 2022). Eating foods with anthocyanins has been linked to the prevention of a number of human health issues including obesity and diabetes (Tsuda et al., 2003). In addition to those health benefits, anthocyanins can also act as bacteriostatic agents (Naz et al., 2007) and are widely used as a natural source of food colorants (Manach et al., 2004). The accumulation of anthocyanin within plant tissues has also been correlated with enhanced drought, salt, and cold tolerance, and it has been shown to protect against insect herbivory and pathogen attack, UV-B, photo inhibition, and the accumulation of reactive oxygen species (Gould et al., 2000; Xu et al., 2017).

The overexpression of these MYB regulators has also been shown to induce anthocyanin accumulation in an array plant species (Borevitz et al., 2000; Gonzali et al., 2009; Meng et al., 2014; Xu et al., 2017). However, a suitable bHLH partner is often required to achieve full functionality in heterologous systems (Takos et al., 2006; Espley et al., 2007; Xiang et al., 2015). The beneficial properties of anthocyanins have motivated numerous researchers to enhance their production in various plants through metabolic engineering (Dixon et al., 2013). The development of genetically modified citrus with the ability to overexpress anthocyanins would potentially increase the health benefits of these fruits and allow the cultivation of these trees in more diverse environments. In addition, the modified citrus plants could also be used as ornamental plants or as alternative sources of easily extracted anthocyanins for use in food coloring or the nutritional enhancement of other foods (Scordino et al., 2015).

Our lab previously investigated the ability of a series of *MybA* transcription factors from Arabidopsis, citrus, grape and plum to confer anthocyanin accumulation in transgenic tobacco plants. The citrus MybA transcription factor *Ruby* from the Moro blood orange was shown to confer anthocyanin accumulation, seen as a bright fuchsia color, in transgenic tobacco and citrus plants (Dasgupta et al., 2017). The *Ruby* and *VvmybA1* transcription factor genes under the control of the constitutive CaMV35S promoter were also expressed in transgenic Mexican lime. These transgenic trees exhibited anthocyanin accumulation in multiple different tissues including leaves, flowers and fruit and some of the transgenic events with intense pigmentation had a stunted growth pattern and curled leaves (Dutt et al., 2016). Similar results have been observed in other species following constitutive accumulation of anthocyanins in transgenic plants (Goldsbrough et al., 1996; Bradley et al., 1998; Stover et al., 2013). The leaves of the *Ruby* transgenic trees exhibited a mottled purple color, pink flowers and fruit, but were typically less strongly pigmented than the *VvmybA1* transgenic events (Dutt et al., 2016; Hijaz et al., 2018).

To produce colored citrus fruit without need for specific environmental conditions or causing an unfavorable metabolic load and stunting plant growth our lab characterized several candidate promoters from citrus and plum that were found to confer fruit preferential expression in tomato (Dasgupta et al., 2020). The highest levels of transgene expression/activity were detected in fruit tissues including the pericarp, placenta, locule, and columella. Some of the promoters also exhibited weak activity in various reproductive or vegetative tissues (Dasgupta et al., 2020). The *CitWax* promoter exhibited expression that was similar to those for the tomato fruit ripening-specific promoters *SlE8* and *SlPG* (Xu et al., 1996; Lau et al., 2009). The *CitUNK, CitVO1,* and *PamMybA* promoters exhibited the strongest tomato fruit expression, but they also demonstrated expression in reproductive and/or vegetative tissues.

We investigate the use of four novel promoters (*CitWax, CitUNK, CitVO1,* and *PamMybA*) and compared to the published *SlE8* promoter for tissue specific expression of *Ruby* in Mexican lime. Transgenic plants were generated, molecularly and phenotypically characterized, and grown in the greenhouse for fruit production. The accumulation of anthocyanins within the fruit and other parts of the tree were evaluated. Results from this research will be informative for engineering citrus trees with robust and reliable fruit specific expression without compromising overall plant health. Results from this research and how these molecular tools may be useful in engineering citrus fruit with improved traits and nutritional quality will be discussed.

Results and discussion

Generation and molecular characterization of the transgenic Mexican lime trees

The *CitWax*, *CitVO1*, *CitUNK*, *PamMybA,* and control *SlE8* promoters were fused to the *Ruby* gene in a pCTAGII binary vector T-DNA (**Figure 1A**) and *Agrobacterium*-mediated transformation was used to generate transgenic Mexican lime trees. A total of 8 or more independent events were successfully transferred to soil and grown in a greenhouse for each promoter construct (**Table 1**). The transgenic events were each validated using PCR confirming the presence of the *codA* transgene and the junction between the promoter being tested and the *Ruby* sequence (**Figures 1B,C** and **Supplementary Figure 1**).

Multiple independent events for each construct were grown to reproductive maturity in the greenhouse and had their growth and development monitored. The transgenic trees exhibited no obvious differences in either vegetative or reproductive growth patterns compared to wildtype Mexican lime trees. The immature trees carrying the CitVO1, CitUNK, PamMybA, and SlE8 constructs did not exhibit visible anthocyanin accumulation in root, stem or leaf tissues, but some of the *CitWaxp-Ruby* events had newly developed leaves that displayed visible purple coloration. This coloration, although clearly noticeable in new growth, disappeared as the leaves matured (**Supplementary Figure 2**). At approximately three or more years of age, the trees began to flower and fruit, which is typical for the Mexican lime cultivar.

The *CitWaxp-Ruby* transgenic trees often produced flowers with pink or purple stigmas, styles, filaments and sometimes petals (**Figure 2A**), but none of the other transgenic trees carrying the other promoter constructs, nor did wildtype Mexican lime have this coloration. If cut open, the immature fruit the *CitWaxp-Ruby* transgenic events displayed a purple interior, while no coloration was visible in this tissue of wildtype or the other transgenic events (**Figure 2B**). The *CitWaxp-Ruby* transgenic events that exhibited a purple-colored stigma usually generated colored fruit (**Figure 2C**) and the strength of the pink or purple coloration of the flower and/or young leaves correlated well with the intensity of the coloration observed in the developing fruit. Although anthocyanins are known to transiently accumulate within the young shoots, leaves and floral tissues of some lemon cultivars [*Citrus limon* (L.) Burm. f.], citron (*Citrus medica* L.), and Rangpur lime (*C. limonia* Osbeck), this does not occur in Mexican lime (Rodrigo et al., 2013; Fabroni et al., 2016).

Within the *CitWaxp-Ruby* transgenic fruit, the pigmentation was only visible within the juice sac tissues, while coloration was not observed in the flavedo or albedo fruit peel tissues (**Figure 2C**). The seeds of these transgenic fruit also accumulated visible anthocyanins within the seed coat at the chalaza end (**Supplementary Figure 3A**). Seed germination was normal, and the transgene was inherited within the progeny as confirmed with genomic PCR (**Supplementary Figure 3B**). The transgenic events carrying the constructs with the *CitUNK*, *CitVO1*, *PamMybA,* or *SlE8* promoters fused to *Ruby* did not generate colored floral or fruit tissues, suggesting that these promoters failed to confer substantial expression in transgenic Mexican lime (**Figure 2C**, **Table 1**, and **Supplementary Table 1**).

Droplet digital PCR (ddPCR) was utilized to estimate the transgene copy number in the transgenic events (**Table 1**). The events carried between 1 and 9 copies of the *nptII* selection marker transgene with approximately 56% of the tested transgenic events carrying a single copy of the selection marker transgene. The overall growth and development of the transgenic trees did not appear to be impacted by the

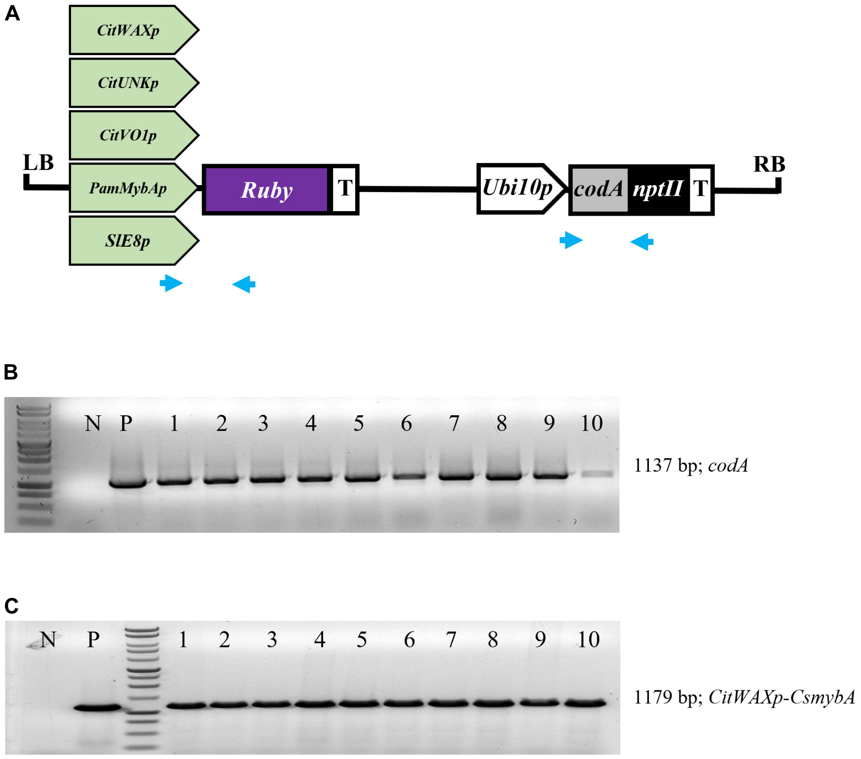


FIGURE 1
(A) Schematic representation of the promoter-*Ruby* construct T-DNAs. The *codA-npt*II selection marker gene is under the control of the Arabidopsis *Ubiquitin10* promoter (*Ubi10p*) and nopaline synthase transcriptional terminator (T). The test promoters from citrus (*CitWaxp*, *CitUNKp*, and *CitVO1p*) plum (*PamMybAp*) and tomato (*SlE8p*) are shown upstream of the *Ruby* (*CsmybA1*) gene. LB and RB designate the *Agrobacterium* left and right borders. Blue arrows indicate the approximate position of the *codA* and *Ruby* primers used for molecular validation. **(B)** PCR confirmation of the presence of *codA* transgene in 10 independent events. **(C)** PCR confirmation of the presence of the *CitWaxp-Ruby* sequence. P designates a sample with a plasmid DNA template. N indicates genomic DNA from wildtype Mexican lime. MW indicates molecular marker (Promega, 1 kb DNA ladder, G7541).

high transgene copy number. The growth and reproductive maturation for the high copy lines [e.g., CitVO1 1-1 (6.6 copies) and CitWax 9-37 (8.7 copies)] was indistinguishable from the single copy events (e.g., CitUNK 4-5, PamMybA 5-3A, and CitWax 7-2A) and wildtype Mexican lime trees (Figure 2D).

Physiological assessment of the transgenic trees

All transgenic trees grew normally in the greenhouse and appeared similar to wildtype Mexican lime trees. The only noticeable difference in vegetative growth was visible anthocyanin accumulation within the young shoots of some of the *CitWaxp-Ruby* transgenic events. Given that anthocyanin accumulation in transgenic citrus has previously associated with suboptimal growth (Dutt et al., 2016), a chlorophyll meter (SPAD-502, Spectrum Technologies, Inc. Aurora, IL, United States) and leaf porometer (SC-1, Meter Group Inc.) were employed to investigate the physiological status of the transgenic trees in further detail. SPAD values, which are proportional to leaf chlorophyll content, are calculated from the transmission of red light at 650 nm (which is absorbed by chlorophyll), and the transmission of infrared light at 940 nm (which is not absorbed by chlorophyll). The porometer estimates stomatal conductance by measuring water loss from the leaf, which is dependent on the rate of transpiration. Although some variability was noted in these measurements, they were all generally within a range similar to that of the wildtype Mexican lime trees (Table 1 and Supplementary Table 1). The variation observed did not correlate with the transgenic construct type, transgene copy number, or anthocyanin content. The introduction of a *Ruby*-containing T-DNA into the Mexican lime genome therefore did not appear to affect nitrogen management or transpiration rates compared to the wild-type control.

Characterization of the transgenic fruit

Brix is defined as the percent of sucrose and other total soluble solids by weight in a solution and is a measure of

TABLE 1 Summary of the tree and fruit characterization.

Genotype/ event	Copy #	Flower color	Fruit color	Anthocyanin	Brix	pH	SPAD value	Stomatal conductance
Mexican lime	0	–	–	0.00	7.31	2.05	40.4	223.5
CitVO1 1-1	6.6	–	–	0.00	6.90	2.22	28.8	152.7
CitVO1 1-5	1.1	–	–	0.00	7.54	2.02	35.3	161.1
CitVO1 1-6	1.1	–	–	0.00	8.19	2.04	52.2	275.7
CitVO1 1-11	1.3	–	–	0.00	7.90	2.06	44.5	181.2
CitVO1 1-16	1.1	–	–	0.00	8.48	2.04	51.8	133.7
CitVO1 1-22	1.1	–	–	0.00	7.09	2.13	47.6	154.8
CitVO1 1-24	1.1	–	–	0.00	7.45	2.07	56.5	201.0
CitVO1 1-25	1.1	–	–	0.00	7.45	2.14	66.5	405.8
CitVO1 1-26	1.0	–	–	0.00	8.56	2.09	49.3	191.4
CitUNK 4-3	2.9	–	–	0.00	8.35	2.13	39.3	288.5
CitUNK 4-5	1.0	–	–	0.00	8.57	2.05	47.8	147.9
CitUNK 4-37	4.1	–	–	0.00	6.84	2.22	43.1	424.0
CitUNK 4-38	1.1	–	–	0.00	8.74	2.14	42.9	165.2
CitUNK 4-40	1.1	–	–	0.00	7.95	2.02	41.6	413.0
CitUNK 4-42	1.1	–	–	0.00	7.29	2.16	40.2	405.0
CitUNK 4-44	2.4	–	–	0.00	7.45	2.08	55.3	326.5
CitUNK 4-47	1.1	–	–	0.00	8.59	2.17	57.9	534.6
CitUNK 4-56	1.4	–	–	0.00	8.26	2.04	53.3	463.2
CitUNK 4-57	1.0	–	–	0.00	8.00	2.09	40.9	179.7
CitUNK 4-70	0.9	–	–	0.00	8.84	2.07	32.8	190.0
CitWax 7-1	0.8	–	–	0.00	6.99	2.06	40.1	216.7
CitWax 7-2	2.1	+	+	0.05	7.06	2.14	43.0	167.7
CitWax 7-2A	1.0	+	+	0.08	8.04	2.47	54.7	230.0
CitWax 7-3	1.2	–	–	0.00	6.86	2.35	55.9	111.1
CitWax 7-5	1.1	+	+	0.00	7.67	2.13	49.2	176.4
CitWax 7-9	2.4	+	+	0.00	8.19	2.00	34.4	171.7
CitWax 7-10	2.2	+	+	0.19	7.83	2.04	48.0	152.9
CitWax 7-15	4.1	+	+	0.10	7.33	2.32	43.7	157.3
CitWax 7-16	4.0	+	+	0.24	7.24	2.12	47.8	363.0
CitWax 7-18	7.0	+	+	0.21	7.55	2.19	41.2	212.4
CitWax 7-19	4.0	+	+	0.30	6.56	2.23	48.1	236.2
CitWax 7-23	1.6	–	–	0.00	7.91	2.18	58.4	161.7
CitWax 7-24	6.3	+	+	0.66	7.93	2.13	27.7	261.7
CitWax 7-25	1.1	+	–	0.00	8.00	2.26	42.2	348.1
CitWax 9-1	1.1	+	+	0.21	6.90	2.32	54.7	185.2
CitWax 9-3	1.2	+	±	0.03	7.06	2.07	52.8	269.6
CitWax 9-6	1.1	+	+	0.04	7.95	2.08	50.1	157.2
CitWax 9-8	1.0	+	±	0.00	7.55	2.03	35.1	203.2
CitWax 9-8A	5.9	+	+	0.15	7.13	2.11	57.3	225.3
CitWax 9-9	6.5	+	+	0.41	5.66	2.15	57.1	310.0
CitWax 9-10	1.2	+	+	0.11	6.94	2.22	40.5	220.7
CitWax 9-12	6.6	+	+	0.30	7.14	2.24	55.6	302.0
CitWax 9-12A	2.3	+	+	0.10	7.62	2.15	46.5	160.6
CitWax 9-14	4.0	+	+	0.14	8.45	2.00	32.2	272.0
CitWax 9-15	6.1	+	+	0.42	6.89	1.97	45.3	253.3
CitWax 9-23	6.0	+	+	0.18	8.08	2.13	51.4	240.0
CitWax 9-24	7.5	+	+	0.30	7.79	2.01	48.7	452.8

(Continued)

TABLE 1 (Continued)

Genotype/ event	Copy #	Flower color	Fruit color	Anthocyanin	Brix	pH	SPAD value	Stomatal conductance
CitWax 9-27	6.9	+	+	0.31	7.78	2.07	57.6	155.7
CitWax 9-28	8.1	+	+	0.28	7.77	2.05	45.7	199.7
CitWax 9-31	2.3	+	±	0.00	7.75	2.14	56.6	188.5
CitWax 9-36	5.2	+	+	0.32	7.21	2.09	52.4	135.4
CitWax 9-37	8.7	+	+	0.45	7.44	2.01	43.2	195.2
CitWax 9-38	2.2	+	+	0.08	7.71	2.23	47.2	162.2
CitWax 9-40	2.2	+	+	0.04	7.44	2.28	54.2	253.2
CitWax 9-41	2.8	+	+	0.06	7.47	2.23	32.3	185.5
CitWax 9-43	2.2	+	+	0.00	8.28	1.94	39.8	194.2
CitWax 9-100	4.2	+	+	0.12	8.56	2.04	ND	ND
PamMybA 5-2	3.6	−	−	0.00	7.01	2.15	30.1	160.7
PamMybA 5-3A	1.1	−	−	0.00	7.49	2.19	39.2	166.8
PamMybA 5-4	1.1	−	−	0.00	7.46	2.12	38.8	326.9
PamMybA 5-5	1.8	−	−	0.00	7.10	2.18	36.9	212.6
PamMybA 5-9	1.2	−	−	0.00	8.07	1.97	28.6	234.8
PamMybA 5-11	1.1	−	−	0.00	6.95	2.14	48.9	205.5
PamMybA 5-14	1.2	−	−	0.00	8.07	2.09	42.0	160.2
PamMybA 5-18	1.2	−	−	0.00	8.45	2.20	43.5	163.8
SlE8 10-2	1.5	−	−	0.00	7.95	2.07	48.5	254.6
SlE8 10-3	0.6	−	−	0.00	8.13	2.09	43.7	157.3
SlE8 10-9	3.8	−	−	0.00	7.58	2.03	42.4	170.3
SlE8 10-10	ND	−	−	0.00	7.65	2.05	51.0	208.7
SlE8 10-11	1.8	−	−	0.00	7.95	2.07	54.6	474.6
SlE8 10-13	2.3	−	−	0.00	7.57	1.94	29.5	205.7
SlE8 10-14	4.5	−	−	0.00	7.28	2.09	29.8	324.8
SlE8 10-18	1.5	−	−	0.00	8.08	2.09	28.4	276.6
Blood Orange	0.0	−	+	0.82	11.74	3.35	66.5	207.9

The measured *nptII* transgene copy number is shown (Copy #). A plus sign (+) indicates visible anthocyanin accumulation within the stigma and/or other reproductive tissues (Flower color) and juice vesicles (Fruit color), while a minus sign (−) indicates no visible coloration. A ± sign indicates that either the tree appeared chimeric and produced some colored and colorless fruit, or had visible coloration in immature fruit, but lost that color as it matured. The measured levels of anthocyanins, acidity (pH) and total sugar content (Brix) of the juice from mature fruit is shown. The chlorophyll content (SPAD value) and porosity readings (Stomatal conductance) from leaves of each event is also shown. **Supplementary Table 1** includes graphs with the standard deviation values for analyzed samples.

general sweetness of fruits and their juices. This includes compounds such as organic acids, soluble amino acids, and other miscellaneous compounds, such as fat, minerals, alcohol, flavonoids (Vitamin C and Vitamin A), as well as the total sugar content. The Brix value is used as a measure of maturity, flavor, and level of sweetness in fruits that helps determine the time of harvest, sales, and processing (Ball, 2006; Kleinhenz and Bumgarner; Lannes et al., 2007). A standard Brix assay was used to evaluate the sweetness of the juice produced from the transgenic trees. Mature ripe fruit were harvested just as they began turning yellow and the degrees Brix of the juice was measured (**Table 1** and **Supplementary Table 1**). Wildtype Mexican lime fruit had juice with a 7.31 ± 0.20 degrees Brix and the juice from a blood orange was as expected sweeter with a 11.74 ± 0.20 degrees Brix. The transgenic events all produced juices with 6.6 to 8.8 degrees Brix, similar to wildtype Mexican lime and distinctly lower levels than blood orange (**Table 1** and **Supplementary Table 1**). The degrees Brix of the *CitWaxp-Ruby* events ranged from 6.6 to 8.6, very similar to the range exhibited by Mexican lime and the other transgenic events (**Table 1** and **Supplementary Table 1**). Taken together, the transgenic fruit all had similar Brix levels to that observed in Mexican lime, indicating that transgenesis and the expression of Ruby in the fruit did not significantly alter the overall sweetness of the limes.

Citrus flavor is influenced by the balance between sugars and acidity, mainly due to citric acid accumulation. As citrus fruit ripen, the acidity decreases while the sugars increase, so the amount of citric acid is among the major determinants of fruit maturity and the overall flavor (Strazzer et al., 2019). The pH of the transgenic lime fruit juices was measured to quantify their acidity. **Table 1** shows that the pH of the transgenic events ranged from 1.9 to 2.4 which was very similar to that of wildtype

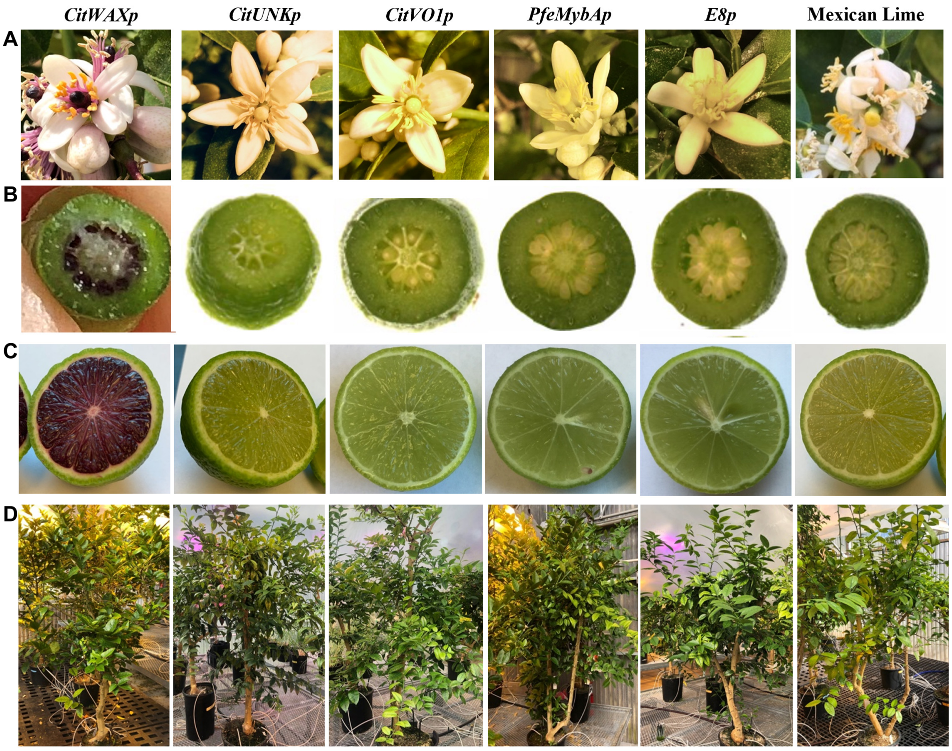


FIGURE 2
Visible phenotypes within the transgenic Mexican lime trees. **(A)** Coloration observed in wildtype and transgenic flowers. **(B)** Appearance of an immature fruit approximately 2 weeks after flower petal drop. **(C)** Mature fruit appearance approximately 3 months after flower petal drop. **(D)** Representative tree from each promoter-*Ruby* construct.

Mexican lime with a pH of 2.1 but is significantly lower than the juice from the blood orange with a pH of 3.4 (Table 1). These results are in general agreement with lime juice typically having a pH between 2.00 to 2.35 as listed by Clemson University which published the PDF entitled "pH Values of Common Foods and Ingredients." The *CitWaxp-Ruby* transgenic events had a pH similar to Mexican lime and the other transgenic events, indicating that the expression of Ruby within the fruit, and anthocyanin accumulation, did not noticeably affect the pH of the juice.

The anthocyanin content within the lime fruit juice from the transgenic trees was measured spectrophotometrically and compared to Moro blood orange and wildtype Mexican lime juices. The anthocyanin content was measured at wavelengths of 530 and 657 nm and the results are expressed as mg/g fresh weight as previously described (Neff and Chory, 1998; Chu et al., 2013). Wildtype Mexican lime has undetectable levels of anthocyanins in this assay, while Moro blood orange has 0.824 mg/g fresh weight. The *CitVO1p*-, *CitUNKp*-, *PamMybA1p*-, and the *SlE8p-Ruby* transgenic events did not have colored fruit (Figure 2C, Table 1, and Supplementary Table 1) and lacked detectable anthocyanin values in the assay, similar to wildtype Mexican lime. The *CitWaxp-Ruby* events displayed a variety of pink, fuchsia or purple hued fruits (Figures 2C, 3A,B, Table 1, and Supplementary Table 1) and a few events had undetectable anthocyanin coloration. The *CitWaxp-Ruby* events that had visible pigmentation of their fruit had anthocyanin values that ranged from 0.03 to 0.66 mg/g fresh weight, with the higher accumulating events having up to 75% of the level of anthocyanins found in blood oranges. The results also indicate that transgenic events with multiple copies of the introduced transgenes tended to exhibit more strongly pigmented fruit. Examples include CitWax event 7-24 (6.3 copies) with 0.633 mg/g fresh weight or 75% of the Moro blood orange value; event 9-37 (8.7 copies) and an absorbance of 0.455 mg/g fresh weight or 54%; and event 9-9 (6.6 copies) with absorbance of 0.406 mg/g fresh weight or 48% of the level found in blood orange. Conversely, those lines with one or two transgene copies, such as 7-5 and 9-40 provided a lightly colored fruit had anthocyanins levels that were below 5% of that found in blood orange. Seventeen of the top 20 anthocyanin accumulating events had two or more transgene copies.

The events that have low levels of anthocyanin in the juice, often had fruit that display a variable pattern of pigmentation (i.e., part of the fruit may be colored, while another part may not). In fact, for some events, there is a distinct lack of color surrounding the center of the fruit when sliced

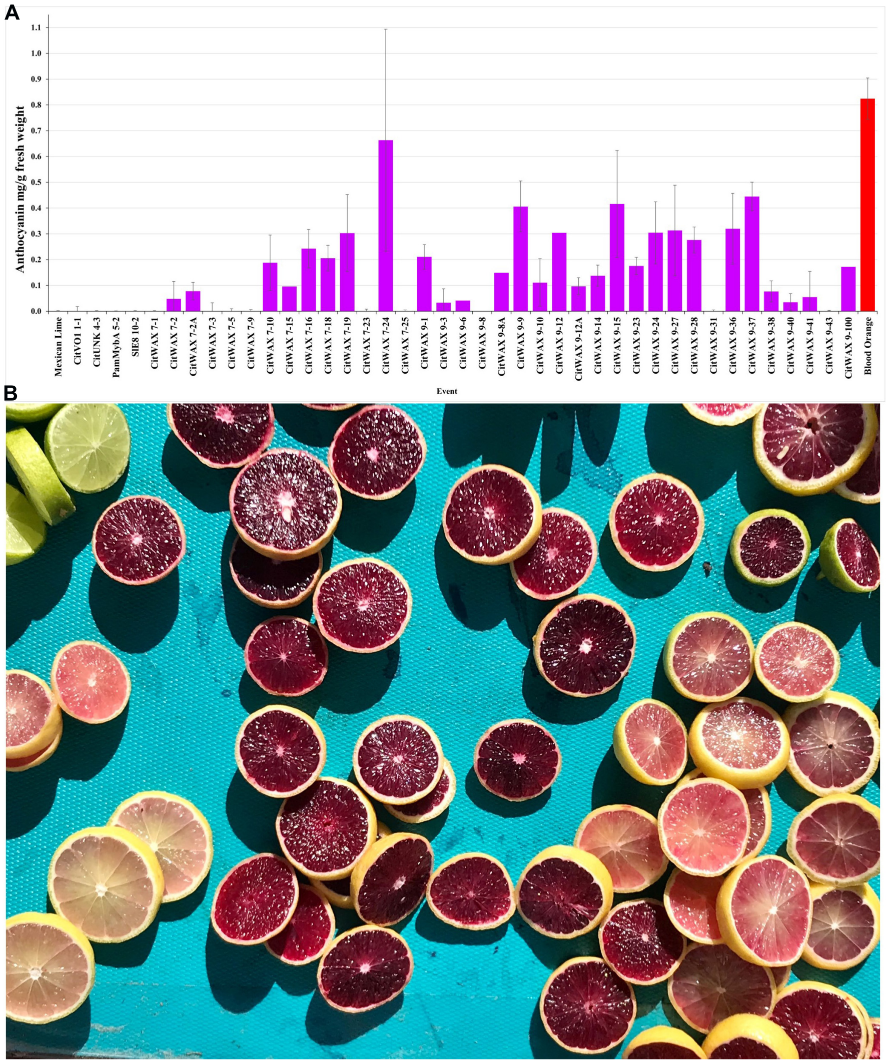


FIGURE 3
(A) Anthocyanin accumulation in Mexican lime fruit. Juice was collected from fruit and the anthocyanin content was determined spectrophotometrically. Wildtype Mexican lime (far left) and Moro blood orange (far right, red bar) are shown as comparators for the transgenic events (middle). A single representative event is shown for the CitVO1, CitUNK, PamMybA, SlE8 transgenic trees, while all of the characterized CitWax events are shown (fuchsia bars). **(B)** Mexican lime fruit derived from the *CitWax-Ruby* construct exhibited fruit of multiple purple hues. For comparison, wildtype Mexican lime fruit is shown in the upper left corner.

horizontally (**Supplementary Figure 4**), indicating a loss of anthocyanin in this area. Whether this is due to lack of expression in this region of the tissue or breakdown of the anthocyanins is unknown. *CitWaxp-Ruby* events 9-8 and 9-31 were unique in that they produced young fruit with easily visible anthocyanin pigmentation, but that pigmentation became weak or disappeared as the fruit matured (**Supplementary Figure 5**), and anthocyanin accumulation was not detected in the assay of their mature fruit (**Figure 3A**). The *CitWaxp-Ruby* events of potential commercial interest include those with low copy T-DNA inserts that exhibit substantial anthocyanin accumulation levels including CitWax 7-2A, 7-10, 9-1, and 9-10 (**Supplementary Figure 6**).

Flavonoids, including anthocyanidins, flavanols, flavanones and isoflavones are abundant compounds found in citrus fruit (Gattuso et al., 2007). Cyanidin 3-glucoside (C3G) and cyanidin 3-(6′′-malonyl)- β-glucoside (C3-6MG) are among the anthocyanins commonly found in citrus (Lee, 2002) and are abundant pigments in blood oranges (Dugo et al., 2003; Scordino et al., 2015). The anthocyanin profile in blood orange detected by our HPLC analytical chromatographic analysis shows that C3G and C3-6MG are easily recognizable peaks within the UV-vis spectra at 515 nm. The juice from selected transgenic fruit was examined using these same conditions and the results were compared to blood orange and wildtype Mexican lime fruit samples. As shown in **Figure 4A**, the C3G and C3-6MG peaks are prominent at 9.69 and 15.41 min, respectively (Figure 4A), consistent with previously published results (Lee, 2002; Scordino et al., 2015). Wildtype Mexican lime fruit juice exhibited no discernable peaks at these elution timepoints (Figure 4B), with similar seen in representative samples from the *CitVO1p*-, *CitUNKp*-, *PamMybA1p*-, and the *SlE8p-Ruby* transgenic events (Supplementary Figure 7), which is consistent with their fruit not exhibiting visible anthocyanin pigmentation.

In contrast, samples from the *CitWaxp-Ruby* events exhibit a series of novel peaks, and the two strongest elute at 9.70 and 15.47 min for the CitWax 7-19 event and 9.62 and 15.36 min for the CitWax 9-9 event (Figures 4C,D). These two strongest peaks correspond to C3G and C3-6MG based on the blood orange trace. Interestingly, in the *CitWaxp-Ruby* events, the strongest peak is C3G followed by the C3-6MG, while in the blood orange sample, this is reversed. The CitWax transgenic events also exhibit three other prominent peaks between 10 and 18 min that are not present in wildtype Mexican lime fruit juice. We hypothesize that this shift in peak strength and the presence of other peaks not seen in the blood orange profile is likely due to the unique metabolism of the lime fruit as compared to that of the sweet orange. We performed a similar analysis for 11 other *CitWaxp-Ruby* events and observed similar results (Supplementary Figure 7). These results demonstrate that the *CitWaxp-Ruby* events accumulate cyanidin glucosides (C3G and C3-6MG), while wildtype Mexican lime does not accumulate detectable levels of these compounds. Area under of the curve analysis was used to estimate the amount of cyanidin 3-glucoside (C3G) in the *CitWaxp-Ruby* events compared to the Moro blood orange. The events ranged from 7.5% (CitWax 7-5) to 55% (CitWax 9-24) of the amount of C3G in blood orange; these values correlated well with the anthocyanin levels measured using the spectrophotometric absorbance assay shown in Table 1 and Figure 3A. Results from the spectrophotometric absorbance assay for anthocyanin absorbance assay values (Table 1) were plotted against the under of the curve HPLC analysis (Figure 4). The trend of relative anthocyanin levels remains consistent where the low copy line 7-5 is observed as a weak expressor in the colorimetric assay and remains low in the HPLC relative value determination, while 9-24 a multicopy high expression line remains high for the HLPC relative value (Figure 5).

Multiple studies have previously examined blood orange juice anthocyanins using electrospray tandem mass spectrometry (MS-MS) and these efforts have provided the identification of the endogenous anthocyanin metabolites (Dugo et al., 2003; Scordino et al., 2015). Since anthocyanin aglycon compounds are typically unstable in the cellular environment, they are frequently observed in multiple glycosylated forms. MS-MS analyses enables the fragmentation of the anthocyanins and the identification of the specific glycosylated forms through the loss of the sugar moiety and the production of the positively charged aglycon. Representative CitWax events were analyzed along with Moro blood orange and wild-type Mexican lime samples using MS-MS in positive ion mode on a Thermo Orbitrap Elite instrument with a survey scan range of 200–2,000 m/z. The MS-MS spectra were reviewed for aglycone fragment ions and anthocyanin parent ions [e.g., m/z 287 ± 3 u Cyanidin aglycon m/z 449 ± 3 u Cyanidin-3-glucoside (C3G), at m/z 536 ± 3 u Cyanidin-3-malonyl glucoside (C3-6MG) and at m/z 596 ± 3 u Cyanidin-3-rutinoside (C3R)], (Figures 6A–C). Multiple anthocyanins were detected in this analysis including glycosylated forms of cyanidin, peonidin, delphinidin, and petunidin. The most abundant glycosylated forms detected in the *CitWaxp-Ruby* events include the glucoside, malonyl glucoside, and rutinoside forms of cyanidin (Table 2). Identification and peak assignment of anthocyanins were based on accurate mass measurements (within 10 ppm) of the calculated monoisotopic mass of both the MS1 anthocyanidin (Figure 6D) and its MS2 aglycone fragment (Figure 6E). The CitWax 7-19 sample is shown in Figure 6 and is representative of the *CitWaxp-Ruby* events (Supplementary Figure 8). The molecular weights of the three most abundant anthocyanidins and their sugar conjugates are shown in Table 2. Identification of selected metabolite peaks was performed by comparison of the retention time, the spectroscopic characteristics and observed mass spectra compared to that of blood orange juice (Supplementary Figure 8). The spectroscopic characteristics of the analytes detected in the blood orange juice samples are consistent with previously published results (Dugo et al., 2003; Scordino et al., 2015). Mass spectra results of the transgenic *CitWaxp-Ruby* events verify the presence of multiple anthocyanin metabolites in the isolated juice of the colored fruit. The cyanidin and peonidin glucosides and malonyl glucosides are seen at detectable levels in the *CitWaxp-Ruby* events, where none was observed in wildtype Mexican lime. Also, the peonidin and delphinidin rutinosides are present in the *CitWax* events, but not observed in the Moro blood orange sample (Table 2).

To further characterize the *CitWaxp-Ruby* transgenic fruit, the nutrient composition was analyzed for events 7-18 and 9-9 and compared to wildtype Mexican lime juice. The results shown in Table 3 indicate that other than a reduction in fructose and glucose content, the nutritional content of the transgenic fruit was overall very similar to wildtype fruit. The observed reduced levels of fructose and glucose may be at least partially due to differences in the ripeness of the samples and can be partially offset by the observed modest increases in sucrose. This result is generally consistent with Brix data for these events being similar to wildtype (Table 1 and Supplementary Table 1). The differences in the sugar levels may also be due to environmental differences between the greenhouse grown transgenic samples compared to the field grown wildtype samples used in the analysis, which is supported by previously published results showing that environmental conditions can

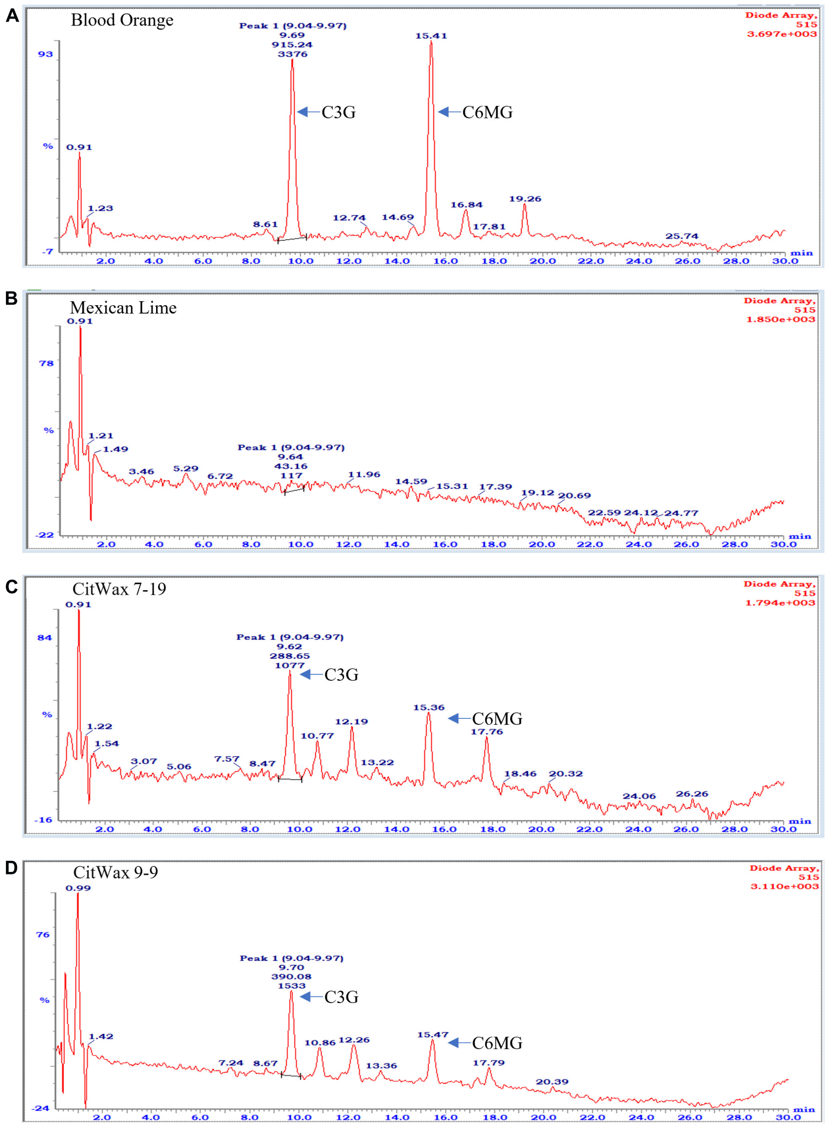


FIGURE 4

Chromatographic profile of anthocyanins found in citrus fruits. **(A)** Commercially bought Moro blood orange. **(B)** Wildtype Mexican lime. **(C)** Transgenic CitWax 7-19. **(D)** Transgenic CitWax 9-9. Cyanidin 3-glucoside (C3G) and cyanidin 3-(6′′-malonyl)- β-glucoside (C3-6MG) provide the two strongest and peaks at approximately 9.6 and 15.4 min with UV-Vis absorption at 515 nm.

affect sugar production within the citrus fruit (Zhang and Ritenhou, 2016; Sadka et al., 2019).

Materials and methods

Plant material and promoter isolation

Binary plasmid construction was as previously described (Dasgupta et al., 2020). The *GUS* gene was replaced by the *Ruby* (*CsMybA1*) coding sequence. The *Ruby* (*CsMybA*/*MoroMybA*) genomic sequence including native introns was synthesized (GenScript USA Inc.) to remove native restriction sites and to aid in cloning (Dasgupta et al., 2017). The complete *Ruby* sequence is available in **Supplementary Text 1** of a previous publication, (Dasgupta et al., 2017). The nucleotide sequences for the tested promoters are available from the GenBank under the following accession numbers *CitWaxp* (GenBank MK012380), *CitUNKp* (GenBank MK012381), *CitVO1p* (GenBank MK012383), (Belknap et al., 2015), *SlE8p* (GenBank KJ561284), and *PamMybAp* (GenBank MK012385). Each construct is derived from the pCTAGII-GUSPlus binary vector (GenBank MG818373) contains a unique promoter of interest that is used to express *Ruby*, along with the *codA-nptII* selection marker for transgenic tissue selection (**Figure 1A**).

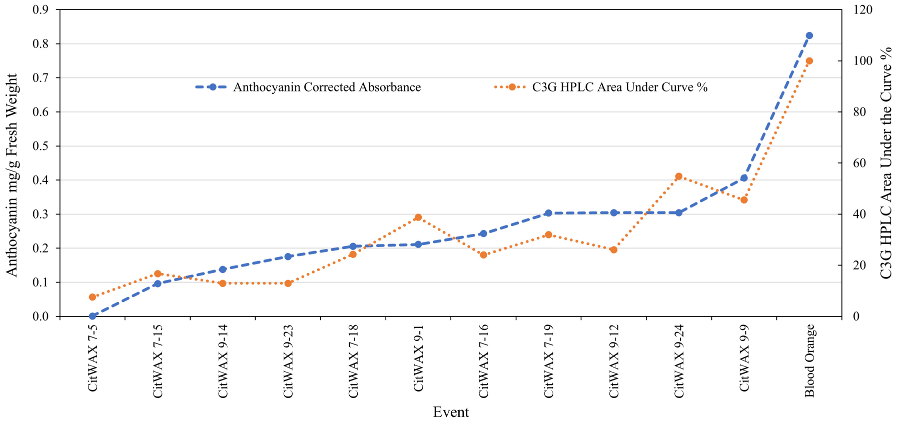


FIGURE 5

Comparison of anthocyanin absorbance content (**Table 1**) vs. chromatographic profile area under the curve values for cyanidin 3-glucoside (**Figure 4**) of select *CitWaxp-Ruby* fruits and blood orange control.

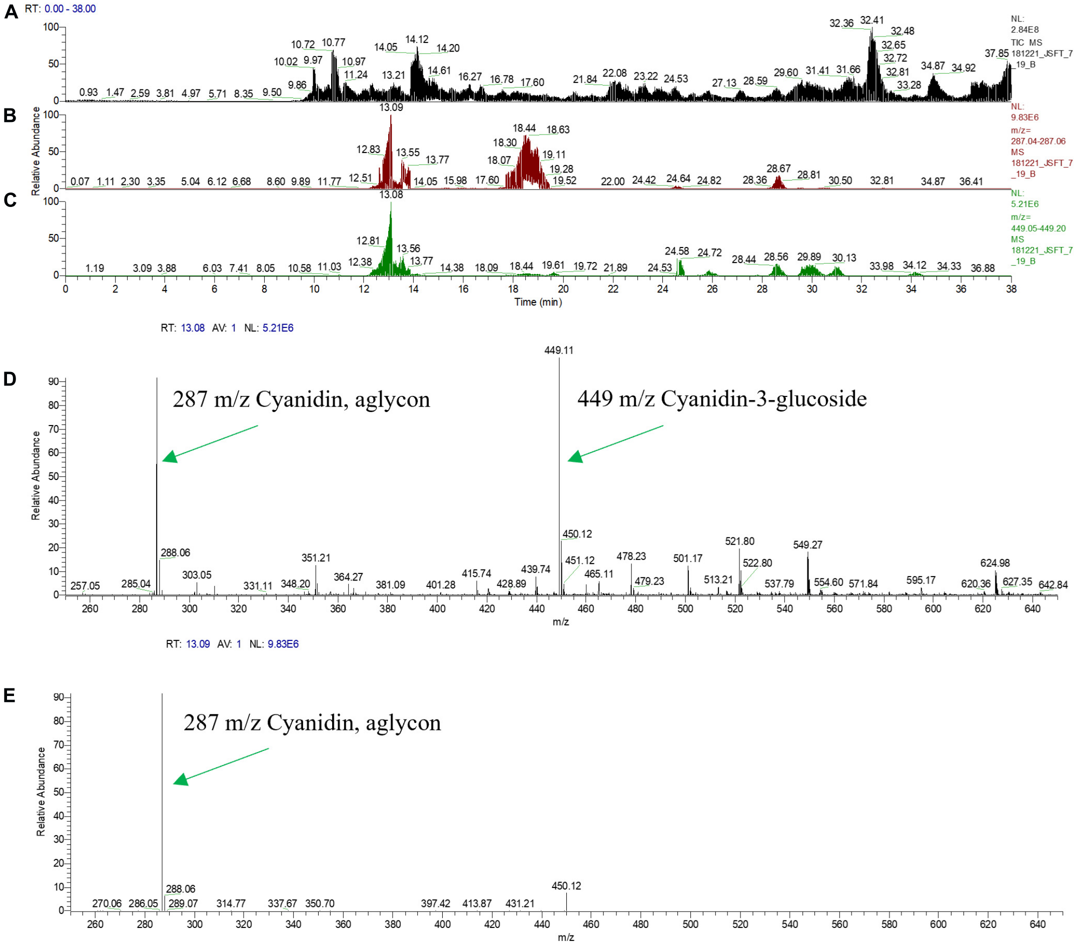


FIGURE 6

Mass Spectral Analysis (HPLC-MS-MS) profiles of fresh squeezed juice of line CitWax 7-19. **(A)** Top panel shows the ion capture of all charge in compounds from MS survey scan range of 200–2,000 m/z. **(B)** Second panel shows specific ion capture from 287.04 to 287.06 m/z. **(C)** Third panel shown specific ion capture from 449.05 to 449.20 m/z. **(D)** Forth panel shows the strongest specific peaks from the full MS1 survey scan range of 200–2,000 m/z at the retention time of 13.08 min. **(E)** Fifth panel shows MS2 of 449.11 m/z at retention time 13.09.

TABLE 2 MS-MS identified anthocyanins.

Anthocyanins and glycoforms	Calculated m/z	Blood orange	Mexican lime	CitWax 9-9	CitWax 7-19
Cyanidin aglycon	287.056				
Glucoside (C3G)	449.108	4.4^5	ND	9.6^5	5.2^6
Malonyl glucoside (C3-6MG)	535.109	4.5^5	ND	5.1^5	6.6^6
Rutinoside (C3R)	595.166	ND	1.0^6	1.2^6	2.6^6
Peonidin aglycon	301.071				
Glucoside	463.12	2.4^4	ND	6.8^5	3.6^6
Malonyl glucoside	549.124	4.4^4	ND	6.6^5	3.9^6
Rutinoside	609.182	ND	1.9^7	5.2^6	2.1^7
Delphinidin aglycon	303.051				
Glucoside	465.103	4.6^4	ND	ND	4.0^5
Malonyl glucoside	551.104	1.0^5	ND	ND	7.0^5
Rutinoside	611.161	ND	2.4^6	5.1^6	6.5^6
Petunidin	317.070				
Glucoside	479.123	ND	ND	ND	ND
Malonyl glucoside	565.123	ND	ND	ND	ND
Malonyl rutinoside	625.181	ND	2.0^6	1.5^6	7.8^6

Values for each sample are absolute values of peak height from reconstructed ion current profiles for the m/z values of the glycosylated forms are indicated. ND, not detected. Superscript numbers indicate scientific notation.

Production of transgenic Mexican lime

Mexican lime (*C. aurantifolia*) juvenile tissue was used for hypocotyl transformation as previously described (De Oliveira et al., 2009). Certified Mexican lime seed was obtained from Lyn Seed[1]. The seed was removed of its outer seed coat and surface sterilize the by soaking in 20% bleach for 20 min. Seeds were then rinsed three times in sterile water and 3–5 seeds were placed on citrus seed germination medium (4.4 g/L MS salts with vitamins, 25 g/L sucrose, adjusted to a pH of 5.8 with 1 N NaOH, 8.0 g/L Bacto Agar) and incubated in the dark at 28°C for 4–5 weeks. Etiolated seedlings that were 12–15 cm long were used for transformation. A 25 mL *Agrobacterium* culture was prepared in YEP medium with 50 mg/l kanamycin and 200 µM acetosyringone and incubated at 28°C shaker at 225 rpm overnight. The *Agrobacterium* culture was then centrifuged at 3,600 × *g* for 10 min at 22°C, the supernatant poured off and the pellet resuspended in liquid inoculation medium (4.4 g/L MS salts with vitamins, 30 g/L sucrose, adjusted to a pH of 5.8 with 1 N NaOH. After autoclaving add, 200 µM acetosyringone, 400 µl myo-inositol 100 mg/L, 1 mg/ml thiamine-HCL, 1 mg/mL pyridoxine, 1 mg/mL nicotinic acid and 1 mg/mL BAP) and adjusted to an OD_{600} of 0.2–0.4.

Young, healthy stems were cut into pieces approximately 10 mm in length with the ends cut with at least 45° angle so that each angle will be in the opposite direction of the other. The segments were incubated in the *Agrobacterium* suspension for 10 min and then blotted and allowed to dry on sterile filter paper for 5 min. They were then placed on co-cultivation medium [1.21 g/L Broadleaf Tree Basal Medium (B1396 – Phytotechnology Labs), 30 g/L sucrose, adjusted the pH 5.8 with 1 N NaOH, and then 8.0 g/L Bacto Agar. After autoclaving, 200 µM acetosyringone, 400 µl/L, myo-inositol 100 mg/L thiamine HCL, 1 mg/mL pyridoxine, 1 mg/ml, nicotinic acid, 1 mg/mL glycine 1 mg/L BAP, 1 mg/ml 2.4-D, and 1 mg/mL, NAA – Sigma-Aldrich] in sterile petri dishes and kept in the growth chamber at 28°C for 2 days in the dark. The infected explants were then transferred to selection regeneration medium (SRM; same as the co-cultivation medium with 70 mg/L of kanamycin added). They were incubated for 2 weeks in dark and then transferred to an incubator at 28°C/16 h light and 24°C/8 h dark. The explants were transferred to fresh SRM every 2–3 weeks.

Primary Mexican lime transformants regenerated on SRM with 50 mg/L kanamycin. Individual green shoots from explants were excised when approximately 10–12 mm in length and grafted onto Carrizo seeding rootstocks. Once the grafts were established in tissue culture, the plantlets were transferred to soil in a growth chamber at 16/8 light dark cycle and 28°C under a humidity dome and allowed to grow for another 4 weeks before being transferred to the greenhouse where they were maintained with a mist spray for 2 weeks and then allowed to harden to greenhouse conditions. Greenhouse conditions were 16/8 light dark cycle and 26°C.

Polymerase chain reaction

Genomic DNA was extracted by grinding a 1 cm^2 piece of citrus leaf in 400 µL of buffer (200 mm Tris–HCl pH

[1] https://lyncitrusseed.com

TABLE 3 Nutrient compositional analysis of CitWax lines 7-18 and 9-9 compared to store bought wildtype Mexican lime juice.

Analyte	Wildtype	CitWax 7-18	CitWax 9-9	Units
Carbohydrate	7.80	8.20	7.38	g/100 g
Fat	0.88	0.90	0.91	g/100 g
Moisture	90.8	90.4	91.2	g/100 g
Protein	0.33	0.31	0.33	g/100 g
Fructose	0.32	0.28	<0.01	g/100 g
Glucose	0.22	0.09	<0.01	g/100 g
Sucrose	0.00	0.01	0.05	g/100 g
Vitamin A	N.D.	N.D.	N.D.	mcg/100 g
Vitamin C	32.5	19.4	20.8	mg/100 g
Calcium (CA)	0.59	0.29	0.33	mg/100 g
Copper (Cu)	0.02	0.01	0.01	mg/100 g
Iron (Fe)	0.02	0.07	0.05	mg/100 g
Magnesium (Mg)	5.50	5.57	5.40	mg/100 g
Phosphorus (P)	1.78	2.19	2.54	mg/100 g
Potassium (K)	16.7	19.3	17.4	mg/100 g
Selenium (Se)	<0.5	<0.5	<0.5	mg/100 g
Sodium (Na)	0.55	0.36	0.37	mg/100 g
Zinc (Zn)	0.10	0.11	0.12	mg/100 g

Analysis was conducted by the USDA-AMS National Science Laboratory. <0.01 and <0.05 values mean below the limit of accurate detection. Mexican lime control fruit was field grown, store bought. CitWax lines 7-18 and 9-9 were greenhouse grown.

7.8, 250 mm NaCl, 25 mm EDTA, and 0.5% SDS). After centrifugation and isopropanol precipitation, the pellet was washed with 70% ethanol and resuspended in 50 μL of water with 1 mM RNase A. PCR amplification was performed using 0.5 μL (∼50 ng) of genomic DNA in reactions with a total volume of 20 μL. Platinum Superfi (Invitrogen) reagents and conditions were used as directed by manufacturer. Sequences of the PCR primers that were used for amplification are shown in **Table 4**. Droplet digital PCR (ddPCR) was performed following the methods described in Collier et al. (2017). Sequences of the ddPCR primers and probes used are shown in **Table 4**.

Imaging of plant tissues

Photographs of the plants and their tissues were recorded using a Nikon D7000 digital camera with an AF Micro Nikkor 60 mm 1:2.8 D lens or AF-S Nikkor 18–70 mm DX lens (Nikon Inc., Melville, NY, United States) under tungsten lamps (Philips, 120 V, 300 W). The camera was set manually for all parameters including ISO sensitivity, focus, f-stop and time. A photography gray card was used as a reference to get the correct exposure. The callus images in petri plates were observed and photographed in a Leica MZ16-F (Leica Microsystems, Inc., Buffalo Grove, IL, United States) stereo zoom light microscope equipped with a QImaging Retiga 2000 R fast cooled, digital color camera.

Juice extraction

Ripe lime fruit (just as they were beginning to turn yellow) were harvested from the trees as they became available. The fruit was rolled on a flat surface and then cut horizontally in half. The juice was squeezed into a beaker using a handheld 8-inch aluminum lime juicer. The juice was then strained through a fine mesh metal strainer and poured into 1.5 mL plastic tubes. These tubes are then spun in a centrifuge at 13,500 RPM for 5 min. The juice was the decanted into a clean beaker and syringe filtered (0.2 μm nylon filter) into 1.5 ml tubes. Tubes were labeled and stored in a −35°F freezer in the dark. The Moro blood orange control juice was obtained from field grown, store bought source that had relatively uniform pigmentation. A set of 10 oranges were bought together, juiced together, filtered, aliquoted, and frozen −35°C. This homogenized juice was used as the blood orange control for all subsequent experiments.

Spectrophotometric (colorimetric) anthocyanin assay

Total anthocyanin levels in the juice were determined using methanolic HCl and measured spectrophotometrically at wavelengths of 520 and 700 nm as described by Chu et al. (2013). All samples were measured in duplicates with six biological replicates for each independent line when available. The anthocyanin content is expressed as absorbance (Abs) in mg/g fresh weight as previously described (Neff and Chory, 1998; Chu et al., 2013). Two 500 μL samples of 0.2 μm filtered juice was prepared per citrus event sample examined. The first sample had HCl added to achieve a pH = 1.0. The second sample had NaOH added to achieve a pH of 4.5. The sample was vortexed to mix it and then let stand for 10–15 min covered or in the dark. A volume of 250 μL sample was loaded per well into a 96 well plate and a SpectraMax Plus 384 spectrophotometer (Molecular Devices Corporation, CA, United States) was used to read the absorbance. Samples were assayed in triplicate (i.e., 3 × pH 1.0 and 3 × pH 4.5). A 50% methanol 50% water solution was used as a blank. Sample absorbance from 400 to 710 nm in 10 nm steps was recorded and analyzed using SoftMax Pro software (Molecular Devices, CA, United States). The anthocyanins are calculated as cyanidin-3-glucoside equivalents, mg/L, using the equation:

where A = (A 520 nm − A 700 nm) pH 1.0 – (A 520 nm − A 700 nm) pH 4.5.

$A \times MW \times DF \times 103/\varepsilon \times l$

MW (molecular weight) = 449.2 g/mol for cyanidin-3-glucoside (C3G);

DF = dilution factor (1:10);

l = path length in cm;

ε = 26,900 molar extinction coefficient, in $L \times mol^{-1} \times cm^{-1}$, for C3G;

1,000 = factor for conversion from g to mg.

TABLE 4 Primers and probes used in PCR and ddPCR.

Primer	Sequence	Amplicon
CitUNK4p 375 F60	GGACTCAGCAACCCTACCCAAGTG	1,058 bp
CitVO1p 400 F60	CACATGCACTAACTTAACCATATAGAGCTGTTGACC	1,080 bp
CitWaxp 490 F61	GGACGATTGTGTTACAGAGAGCATTTAATAAAGCACC	1,179 bp
SlE8p 800 F60	GGTTTAGTCCACAAGTTTTAGTGAGAAGTTTTGC	1,574 bp
PamMybAp 700 F60	CAGCGGAGTCTAACATCCTACGAATAAACCG	1,395 bp
Ruby 670 R60	GGGTAGTTTATGTGTATGCTATATGTTGCTCAACC	
codA ORF 70 F60	CATCTGCAGGACGGAAAAAT	1,137 bp
codA ORF1137 R60	GATAATCAGGTTGGCGCTGT	
CsDehydrin F2	GCCACCGAGTTTGAGAAAG	134 bp
CsDehydrin R2	GAGCTAGAGCTGCTGGTG	
Cs1.1g026736m.g (FAM)	ATGTCTCTGAGCCTCAGCCA	
1116_nptII-F3	ACGTTGTCACTGAAGCG	100 bp
1117_nptII-R3	ATGGATACTTTCTCGGCAG	
nptII_Probe3 (HEX)	TCTCCTGTCATCTCACCTTGCTC	

Chromographic analysis of anthocyanins using HPLC at 515 nm

Analysis of anthocyanins was performed using an Alliance Waters 2695 Separation Module (Milford, MA, United States) coupled to a 996 PDA detector. The system is equipped with a Luna 5u C18 (Number 28; 150 mm × 3.0 mm) column with a guard column of the same material. Temperature of the column and pre-column is maintained at 40°C with an isocratic mobile phase consisting of the monobasic solution. The flow is 0.100 mL/min and injection volume of standards, control and samples. The total run time is 20 min for a blank, and 30 min for the samples. The analysis was conducted at 40°C with a 3.0 mm × 150 mm, 5 μM, Luna C-18 (2) column (Phenomenex, Torrance, CA, United States) employing a binary gradient of 2% formic acid: acetone (time 0–15 min, 2–12% acetone; time 15–20, 12–22% acetone; time 20–25, 22% acetone; time 25–30, 2% acetone). The flow rate was 1 mL/min. Masslynx (version 4.1) is used to control the HPLC system and for data analysis. Standard curve and control are injected at the beginning and at the end of the set of samples. Integration of individual peak areas is done using Quanlynx, and quantification is based on external calibration curves of anthocyanins at the wavelength of 515 nm.

Anthocyanin analysis using HPLC-mass spec-mass spec

Reverse phase HPLC was accomplished with an Eksigent Ekspert nanoLC 425 fitted with a PicoSlide column switching device and three Reprosil-PUR C18-AQ, 3 μM 120A; 105 mm PicoChip columns (New Objective, Woburn, MA, United States). Samples were loaded directly onto the column with 400 nL/min 2% for 60 min. Column elution was at 400 nL/min with a program of 3% for 6 min, followed by a 36-min linear ramp to 25%, followed by a 16 min linear ramp to 50%, followed by a 2 min linear ramp to 3%. Sample elution was followed by a sawtooth column wash step with four cycles to 90%, linear ramp from 5 to 90% over 3 min, hold at 90% for 3 min, then ramp to 5% over 3 min, followed by 24 min at 2%.

Mass Spectral analyses were performed in positive ion mode on a Thermo Orbitrap Elite (Thermo Fisher, Waltham, MA, United States), with mass resolution set to 30,000 for survey scans and, and fragment ion scans. Samples were loaded with an Eksigent Ekspert nanoLC 425 (SCIEX) directly onto a PicoChip C18 nanoflow column mounted on a PicoSlide ion source (New Objective, Woburn, MA, United States). The HPLC was programed with 400 nL/min 2% for 60 min and eluted accordingly. Mass Spectral Analysis was performed in positive ion mode on a Thermo Orbitrap Elite with a survey scan range of 200–2,000 m/z with MS2 spectra being collected for the most intense ions in a given survey scan. Electrospray source voltage was set to 3.5 kV and capillary inlet temperature set to 275°C. Fragmentation was performed in CID mode with 30 V normalized collision-induced dissociation (CID) of the top three most intense ions from the survey scan. Data were processed with Thermo Xcalibur software. The presence of aglycones was observed using reconstructed ion current profiles, and glycosylation was determined by m/z measurement of intact compounds, confirmed by MS-MS fragmentation of those compounds. Compounds detected were primarily mono- and di-saccharides of the respective anthocyanins. Anthocyanins detected were cyanidin, peonidin, delphinidin, petunidin, and another anthocyanin differing in molecular weight from delphinidin by 0.04 Da. The mass of 303.09 m/z is consistent with the mass of 5-Methyl-6-hydroxyluteolinidin 6,7,3′,4′-tetrahydroxy-5-methoxy-flavylium, however, this will require further investigation to confirm. The data was processed with Thermo Xcalibur software.

Plant physiological measurements

Leaves were measured for chlorophyll content using a SPAD-502Plus chlorophyll meter (Spectrum Technologies, Plainfield, IL, United States). SPAD values are calculated by the meter from the transmission of red light at 650 nm (which is absorbed by chlorophyll), and the transmission of infrared light at 940 nm (which is not absorbed by chlorophyll). All measurements were made at a central point on the leaf between the midrib and leaf margin for three randomly selected leaves per tree and averaged to generate a single value. A porometer (SC-1, Meter Group Inc.) was used to assess stomatal conductance in leaves. All measurements were made at a central point on the leaf between the midrib and leaf margin for three randomly selected leaves per tree and averaged to generate a single value. All trees including the control Moro blood orange and Mexican lime were greenhouse grown for analysis.

Juice analysis

The nutrient analysis of juice samples (175 mL) was conducted by the USDA-AMS National Science Laboratory (Gastonia, NC) for a comprehensive 17 component analysis. Samples were analyzed for Carbohydrate, Fat, Moisture, Protein, Fructose, Glucose, Sucrose, Vitamin C, Calcium (CA), Copper (Cu), Iron (Fe), Magnesium (Mg), Phosphorus (P), Potassium (K), Selenium (Se), Sodium (Na) and Zinc (Zn). Report ID #AU27038, AU27039, and AU27040. The National Science laboratory utilizes AOAC (Association of Official Analytical Chemists) methodologies for testing or a modified approach of an AOAC method. The laboratory is ISO/IEC 17025:2017 accredited and utilize methods of testing that are fit for purpose that have been validated http://www.eoma.aoac.org/#:~:text=AOAC%20is%20a%20leader%20in,and%20scientific%20information%20and%20opportunities.

The pH of the juices was determined using a Thermo Scientific Orion 3 Star pH Benchtop double junction micro pH meter. Due to the limited amount of juice available, pH readings were performed in 1.5 mL plastic tubes. Readings were taken in triplicate and the pH meter probe was thoroughly rinsed with purified water between readings.

A J47 refractometer (Rudolph Research Analytical, Hackettstown, NJ, United States) was used for juice Brix analysis. A 200 μL aliquot of purified water was used as a blank. A 200 μL sample of lime juice was used for analysis. Values were measured in triplicate and recorded. The refractometer prism was rinsed with water and then 50% ethanol between samples. The Mexican lime control fruit was field grown, store bought, while the CitWax lines 7-18 and 9-9 were greenhouse grown.

Conclusion

The research presented demonstrates that the Moro blood orange *Ruby* transcription factor, when expressed using the *CitWaxp* fruit-preferential promoter, confers substantial anthocyanin accumulation within Mexican lime fruit. Somewhat surprisingly, the other tested fruit-preferential novel promoters, *CitVO1p, CitUNKp, PamMybAp,* and control *SlE8p*, did not express sufficient levels of Ruby to confer anthocyanin accumulation in fruit or other Mexican lime tissues (**Figure 2**). The molecular and physiological characterization of the transgenic trees showed that they all grew similarly to wildtype Mexican lime trees in the greenhouse. The *CitWaxp-Ruby* transgenic Mexican lime trees, in addition to accumulating anthocyanins within the fruit juice sac tissues, also frequently produced colored flower tissues (primarily the pistil and filaments) and occasionally within young emerging shoots and leaves. The anthocyanin accumulation in these events ranged from undetectable to as high as 75% of levels detected in the Moro blood orange via colorimetric spectrophotometry (**Figure 3** and **Table 1**). It was noticed that the anthocyanin accumulation levels in the *CitWaxp-Ruby* transgenic events correlated to the transgene copy number of the event [higher copy events tended to have higher anthocyanin levels (**Figure 3** and **Table 1**)].

To further investigate the transgenic fruit, the level of anthocyanins within lime fruit juice was examined using HPLC and MS-MS analyses (**Figures 4**, **5**). The results establish that the *CitWaxp-Ruby* transgenic Mexican lime fruit contain cyanidin/peonidin glucosides and malonyl glucosides where not previously observed in wildtype Mexican lime. The observed levels of peonidin/delphinidin rutinosides in these events were also at higher levels than those found in the Moro blood orange. Nutritional analysis of the juice showed that 17 major components were within normal range and similar to the levels observed in wildtype Mexican limes (**Table 4**). A slight reduction in the sugar content was detected in this analysis, but the difference is likely to the greenhouse growing conditions in which the transgenic trees were grown. The fruit from the various transgenic lines were also examined for pH and Brix illustrating that they also were within the normal range of values for Mexican lime fruit (**Table 1**). Although the transgenic trees characterized in this study exhibited normal growth and fruit production, in our previous study, we had observed detrimental effects on two of our regenerated transgenic events that were darkly pigmented within vegetative tissues (Dasgupta et al., 2017). These plants displayed curled leaves, a lack of vigor, and stunted growth, in the greenhouse that appeared to correlate with apparent vegetative expression of a *mybA1* and substantial anthocyanin accumulation in leaves and other vegetative tissues (Dasgupta et al., 2017). In agreement with these observations, previous reports have shown that constitutive production of high levels of anthocyanins are toxic to the health of the transgenic plants and inhibit plant growth (Bradley et al., 1998; Stover et al., 2013; Dutt et al., 2016). Although the mechanism behind the reduction in plant growth is not fully understood, it is hypothesized that anthocyanin production may interrupt auxin transport, negatively impacting growth (Belknap et al., 2015). The research presented here demonstrated that utilizing a fruit preferential promoter to control *Ruby* expression, limits anthocyanin production to specific tissue types and generates trees that grew, matured and produced fruit similar to the wild-type control.

Previous studies have shown that blood orange cultivars require strong day-night temperature differentials for intense color formation in the fruit and therefore are dependent on the prevailing climate during fruit ripening for activation of the anthocyanin production (Dugo et al., 2003). However, the *CitWaxp-Ruby* transgene is not under environmental constraints and its tissue specific expression confers colored lime fruit by accumulating multiple different anthocyanin glycosides within the juice sac tissues. As anthocyanins exhibit antioxidant activity against harmful free radicals (Li et al., 2007; Reddy et al., 2007), act as bacteriostatic agents (Naz et al., 2007) and have been associated with the prevention of obesity and diabetes (Tsuda et al., 2003), this novel citrus fruit could be a beneficial addition to the human diet.

The *CitWaxp-Ruby* transgenic trees also exhibited novel coloration in the flowers and young leaves via anthocyanin overproduction. These lines could potentially be used as novel ornamentals, or the tissues could be used as a source of extracted anthocyanins to enhance the color and nutritional value of other foods. In addition, this research opens up the possibility for the development of other novel citrus cultivars that produce anthocyanin rich fruit under variety of environmental conditions. The *CitWax* promoter will be a potentially useful tool for engineering novel fruit traits within citrus and potentially other crops.

Significance statement

Multiple tissue-specific promoters were tested for their ability to confer anthocyanin accumulation with *C. aurantifolia* fruit. The results presented demonstrate the ability to precisely genetic engineer anthocyanin accumulation in Mexican lime fruit.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Author contributions

JT conceived the project. JT, RT, and KD designed the experiments, analyzed the data, and wrote the manuscript. JT, RT, KD, JH, DH, LH, and RC performed the experiments. All authors approved the final manuscript.

Funding

This research was financially supported by California Citrus Research Board (5200–141 and 5200-155). This research was also supported by USDA-Agricultural Research Service CRADA No. 58-2030-8-017 and Project No. 2030-21000-020-00D. Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the United States Department of Agriculture. USDA is an equal opportunity provider and employer.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.945738/full#supplementary-material

References

Ball, D. W. (2006). Concentration scales for sugar solutions. *J. Chem. Ed.* 83, 1489–1491.

Baudry, A., Heim, M. A., Dubreucq, B., Caboche, M., Weisshaar, B., and Lepiniec, L. (2004). TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in Arabidopsis thaliana. *Plant J.* 39, 366–380. doi: 10.1111/j.1365-313X.2004.02138.x

Belknap, W. R., McCue, K. F., Harden, L. A., Vensel, W. H., Bausher, M. G., and Stover, E. (2015). A family of small cyclic amphipathic peptides (SCAmpPs) genes in citrus. *BMC Genom.* 16:303. doi: 10.1186/s12864-015-1486-4

Borevitz, J. O., Xia, Y., Blount, J., Dixon, R. A., and Lamb, C. (2000). Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* 12, 2383–2394. doi: 10.1105/tpc.12.12.2383

Bradley, J. M., Davies, K. M., Deroles, S. C., Bloor, S. J., and Lewis, D. H. (1998). The maize Lc regulatory gene up-regulates the flavonoid biosynthetic pathway of *Petunia*. *Plant J.* 13, 381–392.

Bridle, P., and Timberlake, C. F. (1997). Anthocyanins as natural food colours - Selected aspects. *Food Chem.* 58, 103–109.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., et al. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* 3, 1242–1255. doi: 10.1105/tpc.111.095232

Chu, H., Jeong, J. C., Kim, W. J., Chung, D. M., Jeon, H. K., Ahn, Y. O., et al. (2013). Expression of the sweet potato R2R3-type IbMYB1a gene induces anthocyanin accumulation in Arabidopsis. *Physiol. Plant* 148, 189–199. doi: 10.1111/j.1399-3054.2012.01706.x

Collier, R., Dasgupta, K., Xing, Y. P., Hernandez, B. T., Shao, M., Rohozinski, D., et al. (2017). Accurate measurement of transgene copy number in crop plants using droplet digital PCR. *Plant J.* 90, 1014–1025. doi: 10.1111/tpj.13517

Dasgupta, K., Hotton, S., Belknap, W., Syed, Y., Dardick, C., Thilmony, R., et al. (2020). Isolation of novel citrus and plum fruit promoters and their functional characterization for fruit biotechnology. *BMC Biotechnol.* 20:43. doi: 10.1186/s12896-020-00635-w

Dasgupta, K., Thilmony, R., Stover, E., de Oliveira, M. L., and Thomson, J. (2017). Novel R2R3-myb transcription factors from Prunus Americana regulate differential patterns of anthocyanin accumulation in tobacco and citrus. *GM Crops Food* 8, 81–21. doi: 10.1080/21645698.2016.1267897

De Beer, D., Joubert, E., Gelderblom, W. C., and Manley, M. (2003). Antioxidant activity of South African red and white cultivar wines: Free radical scavenging. *J. Agric. Food Chem.* 51, 902–909. doi: 10.1021/jf026011o

De Oliveira, M. L. P., Febres, V. J., Costa, M. G. C., Moore, G. A., and Otoni, W. C. (2009). High-efficiency *Agrobacterium*-mediated transformation of citrus via sonication and vacuum infiltration. *Plant Cell Rep.* 28, 387–395. doi: 10.1007/s00299-008-0646-2

Deighton, N., Brennan, R., Finn, C., and Davies, H. V. (2000). Antioxidant properties of domesticated and wild *Rubus* species. *J. Sci. Food Agric.* 80, 1307–1313.

Dixon, R. A., Liu, C., and Jun, J. H. (2013). Metabolic engineering of anthocyanins and condensed tannins in plants. *Curr. Opin. Biotechnol.* 24, 329–335. doi: 10.1016/j.copbio.2012.07.004

Dugo, P., Mondello, L., Morabito, D., and Dugo, G. (2003). Characterization of the anthocyanin fraction of sicilian blood orange juice by micro-HPLC-ESI/MS. *J. Agric. Food Chem.* 51, 1173–1176. doi: 10.1021/jf026078b

Dutt, M., Stanton, D., and Grosser, J. W. (2016). Ornacitrus: Development of genetically modified anthocyanin-expressing citrus with both ornamental and fresh fruit potential. *J. Am. Soc. Horticult. Sci.* 141, 54–61.

Espley, R. V., Hellens, R. P., Putterill, J., Stevenson, D. E., Kutty-Amma, S., and Allan, A. C. (2007). Red colouration in apple fruit is due to the activity of the MYB transcription factor. MdMYB10. *Plant J.* 49, 414–427. doi: 10.1111/j.1365-313X.2006.02964.x

Fabroni, S., Ballistreri, G., Amenta, M., and Rapisarda, P. (2016). Anthocyanins in different Citrus species: An UHPLC-PDA-ESI/MS^n -assisted qualitative and quantitative investigation. *J. Sci. Food Agric.* 96, 4797–4808. doi: 10.1002/jsfa.7916

Gattuso, G., Barreca, D., Gargiulli, C., Leuzzi, U., and Caristi, C. (2007). Flavonoid composition of Citrus juices. *Molecules* 12, 1641–1673. doi: 10.3390/12081641

Goldsbrough, A. P., Tong, Y., and Yoder, J. I. (1996). Lc as a non-destructive visual reporter and transposition excision marker gone for tomato. *Plant J.* 9, 927–933.

Gonzali, S., Mazzucato, A., and Perata, P. (2009). Purple as a tomato: Towards high anthocyanin tomatoes. *Trends Plant Sci.* 14, 237–241. doi: 10.1016/j.tplants.2009.02.001

Gould, K. S., Markham, K. R., Smith, R. H., and Goris, J. J. (2000). Functional role of anthocyanins in the leaves of Quintinia serrata A. Cunn. *J. Exp. Bot.* 51, 1107–1115. doi: 10.1093/jexbot/51.347.1107

Hijaz, F., Nehela, Y., Jones, S. E., Dutt, M., Grosser, J. W., Manthey, J. A., et al. (2018). Metabolically engineered anthocyanin-producing lime provides additional nutritional value and antioxidant potential to juice. *Plant Biotechnol. Rep.* 12, 329–346. doi: 10.1007/s11816-018-0497-4

Holton, T. A., and Cornish, E. C. (1995). Genetics and Biochemistry of Anthocyanin Biosynthesis. *Plant Cell* 7, 1071–1083. doi: 10.1105/tpc.7.7.1071

Karageorgou, P., and Manetas, Y. (2006). The importance of being red when young: Anthocyanins and the protection of young leaves of Quercus coccifera from insect herbivory and excess light. *Tree Physiol.* 26, 613–621. doi: 10.1093/treephys/26.5.613

Kleinhenz, M. D., and Bumgarner, N. R. *Using brix as an indicator of vegetable quality: An overview of the practice*. Available online at: https://ohioline.osu.edu/factsheet/HYG-1650 (accessed July 27, 2022).

Lannes, S. D., Fernando, F. L., Adilson, S. R., and Casali, V. W. D. (2007). Growth and quality of Brazilian accessions of Capsicum chinense fruits. *Sci. Horticul.* 112, 266–270. doi: 10.1016/j.scienta.2006.12.029

Lau, J. M., Cooper, N. G., Robinson, D. L., and Korban, S. S. (2009). Sequence and In Silico Characterization of the Tomato Polygalacturonase (PG) Promoter and Terminator Regions. *Plant Mol. Biol. Rep.* 27, 250–256.

Lee, H. S. (2002). Characterization of major anthocyanins and the color of red-fleshed Budd Blood orange (*Citrus sinensis*). *J. Agric. Food Chem.* 50, 1243–1246. doi: 10.1021/jf011205+

Legua, P., Modica, G., Porras, I., Conesa, A., and Continella, A. (2022). Bioactive compounds, antioxidant activity and fruit quality evaluation of eleven blood orange cultivars. *J. Sci. Food Agric.* 102, 2960–2971. doi: 10.1002/jsfa.11636

Li, H., Flachowsky, H., Fischer, T. C., Hanke, M. V., Forkmann, G., Treutter, D., et al. (2007). Maize Lc transcription factor enhances biosynthesis of anthocyanins, distinct proanthocyanidins and phenylpropanoids in apple (*Malus domestica* Borkh.). *Planta* 226, 1243–1254. doi: 10.1007/s00425-007-0573-4

Lila, M. A. (2004). Anthocyanins and human health: An in vitro investigative approach. *J. Biomed. Biotechnol.* 2004, 306–313. doi: 10.1155/S111072430440401X

Liu, Y., Heying, E., and Tanumihardjo, S. A. (2012). History, global distribution, and nutritional importance of citrus fruits. *Compr. Rev. Food Sci. Food Saf.* 11, 530–545.

Lo Piero, A. R. (2015). The State of the art in biosynthesis of anthocyanins and its regulation in pigmented sweet oranges [(Citrus sinensis) L. Osbeck]. *J. Agric. Food Chem.* 63, 4031–4041. doi: 10.1021/acs.jafc.5b01123

Manach, C., Scalbert, A., Morand, C., Rémésy, C., and Jiménez, L. (2004). Polyphenols: Food sources and bioavailability. *Am. J. Clin. Nutr.* 79, 727–747. doi: 10.1093/ajcn/79.5.727

Meng, X., Yin, B., Feng, H. L., Zhang, S., Liang, X. Q., and Meng, Q. W. (2014). Over-expression of R2R3-MYB gene leads to accumulation of anthocyanin and enhanced resistance to chilling and oxidative stress. *Biol. Plant* 58, 121–130. doi: 10.1007/s10535-013-0376-3

Mondello, L., Cotroneo, A., Errante, G., Dugo, G., and Dugo, P. (2000). Determination of anthocyanins in blood orange juices by HPLC analysis. *J. Pharm. Biomed. Anal.* 23, 191–195. doi: 10.1016/s0731-7085(00)00269-7

Naz, S., Siddiqi, R., Ahmad, S., Rasool, S. A., and Sayeed, S. A. (2007). Antibacterial activity directed isolation of compounds from Punica granatum. *J. Food Sci.* 72, M341–M345. doi: 10.1111/j.1750-3841.2007.00533.x

Neff, M. M., and Chory, J. (1998). Genetic interactions between phytochrome A, phytochrome B, and cryptochrome 1 during Arabidopsis development. *Plant Physiol.* 118, 27–35. doi: 10.1104/pp.118.1.27

Reddy, A. M., Reddy, V. S., Scheffler, B. E., Wienand, U., and Reddy, A. R. (2007). Novel transgenic rice overexpressing anthocyanidin synthase accumulates a mixture of flavonoids leading to an increased antioxidant potential. *Metab. Eng.* 9, 95–111. doi: 10.1016/j.ymben.2006.09.003

Rodrigo, M. J., Alquezar, B., Alos, E., Lado, J., and Zacarias, L. (2013). Biochemical bases and molecular regulation of pigmentation in the peel of Citrus fruit. *Sci. Hortic.* 163, 46–62. doi: 10.1016/J.SCIENTA.2013.08.014

Sadka, A., Shlizerman, L., Kamara, I., and Blumwald, E. (2019). Primary metabolism in citrus fruit as affected by its unique structure. *Front. Plant Sci*. 10:1167. doi: 10.3389/fpls.2019.01167

Sanchez-Moreno, C. (2002). Review: Methods used to evaluate the free radical scavenging activity in foods and biological systems. *Food Sci. Tech. Int*. 8, 121–137. doi: 10.1106/108201302026770

Scordino, M., Sabatino, L., Lazzaro, F., Borzì, M. A., Gargano, M., Traulo, P., et al. (2015). Blood orange anthocyanins in fruit beverages: How the commercial shelf life reflects the quality parameter. *Beverages* 1, 82–94. doi: 10.3390/beverages1020082

Steyn, W. J., Wand, S. J. E., Holcroft, D. M., and Jacobs, G. (2002). Anthocyanins in vegetative tissues: A proposed unified function in photoprotection. *New Phytol*. 155, 349–361. doi: 10.1046/j.1469-8137.2002.00482.x

Stover, E., Avila, Y., Li, Z. T., and Gray, D. (2013). Transgenic expression in citrus of *Vitis* MybA1 from a bidirectional promoter resulted in variable anthocyanin expression and was not suitable as a screenable marker without antibiotic selection Proc. *Florida State Hort. Soc.* 126, 84–88.

Strazzer, P., Spelt, C. E., Li, S., Bliek, M., Federici, C. T., Roose, M. L., et al. (2019). Hyperacidification of Citrus fruits by a vacuolar proton-pumping P-ATPase complex. *Nat. Commun*. 10:744. doi: 10.1038/s41467-019-08516-3

Takos, A. M., Jaffé, F. W., Jacob, S. R., Bogs, J., Robinson, S. P., and Walker, A. R. (2006). Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples. *Plant Physiol.* 142, 1216–1232. doi: 10.1104/pp.106.088104

Tsuda, T., Horio, F., Uchida, K., Aoki, H., and Osawa, T. (2003). Dietary cyanidin 3-O-beta-D-glucoside-rich purple corn color prevents obesity and ameliorates hyperglycemia in mice. *J. Nutr.* 133, 2125–2130. doi: 10.1093/jn/133.7.2125

Xiang, L. L., Liu, X. F., Li, X., Yin, X. R., Grierson, D., Li, F., et al. (2015). A novel bHLH transcription factor involved in regulating anthocyanin biosynthesis in chrysanthemums (Chrysanthemum morifolium Ramat.). *PLoS One* 10:e0143892. doi: 10.1371/journal.pone.0143892

Xu, R., Goldman, S., Coupe, S., and Deikman, J. (1996). Ethylene control of E4 transcription during tomato fruit ripening involves two cooperative cis elements. *Plant Mol. Biol*. 31, 1117–1127. doi: 10.1007/BF00040829

Xu, Z. S., Feng, K., Que, F., Wang, F., and Xiong, A. S. (2017). A MYB transcription factor, DcMYB6, is involved in regulating anthocyanin biosynthesis in purple carrot taproots. *Sci. Rep*. 7:45324. doi: 10.1038/srep45324

Zhang, J., and Ritenhou, M. A. (2016). Sugar composition analysis of commercial citrus juice products. *Proc. Fla. State Hort. Soc*. 129, 178–180.

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fpls.2022.982683

Transcriptome analysis of the pulp of citrus fruitlets suggests that domestication enhanced growth processes and reduced chemical defenses increasing palatability

Estela Perez-Roman, Carles Borredá, Francisco R. Tadeo and Manuel Talon*

Centro de Genómica, Instituto Valenciano de Investigaciones Agrarias, Moncada, Spain

Check for updates

OPEN ACCESS

EDITED BY
Angela Roberta Lo Piero,
University of Catania, Italy

REVIEWED BY
Angelo Sicilia,
University of Catania, Italy
Marco Caruso,
Research Center for Olive, Citrus and Tree Fruit, Italy

*CORRESPONDENCE
Manuel Talon
talon_man@gva.es

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 30 June 2022
ACCEPTED 03 August 2022
PUBLISHED 02 September 2022

CITATION
Perez-Roman E, Borredá C, Tadeo FR and Talon M (2022) Transcriptome analysis of the pulp of citrus fruitlets suggests that domestication enhanced growth processes and reduced chemical defenses increasing palatability.
Front. Plant Sci. 13:982683.
doi: 10.3389/fpls.2022.982683

To identify key traits brought about by citrus domestication, we have analyzed the transcriptomes of the pulp of developing fruitlets of inedible wild Ichang papeda (*Citrus ichangensis*), acidic Sun Chu Sha Kat mandarin (*C. reticulata*) and three palatable segregants of a cross between commercial Clementine (*C. x clementina*) and W. Murcott (*C. x reticulata*) mandarins, two pummelo/mandarin admixtures of worldwide distribution. RNA-seq comparison between the wild citrus and the ancestral sour mandarin identified 7267 differentially expressed genes, out of which 2342 were mapped to 117 KEGG pathways. From the remaining genes, a set of 2832 genes was functionally annotated and grouped into 45 user-defined categories. The data suggest that domestication promoted fundamental growth processes to the detriment of the production of chemical defenses, namely, alkaloids, terpenoids, phenylpropanoids, flavonoids, glucosinolates and cyanogenic glucosides. In the papeda, the generation of energy to support a more active secondary metabolism appears to be dependent upon upregulation of glycolysis, fatty acid degradation, Calvin cycle, oxidative phosphorylation, and ATP-citrate lyase and GABA pathways. In the acidic mandarin, downregulation of cytosolic citrate degradation was concomitant with vacuolar citrate accumulation. These changes affected nitrogen and carbon allocation in both species leading to major differences in organoleptic properties since the reduction of unpleasant secondary metabolites increases palatability while acidity reduces acceptability. The comparison between the segregants and the acidic mandarin identified 357 transcripts characterized by the occurrence in the three segregants of additional downregulation of secondary metabolites and basic structural cell wall components. The segregants also showed upregulation of genes involved in the synthesis of methyl anthranilate and furaneol, key substances of pleasant fruity aroma and flavor, and of sugar transporters relevant for sugar accumulation. Transcriptome and qPCR analysis in developing and ripe fruit of a set of genes previously associated with citric acid accumulation, demonstrated that lower acidity is linked to downregulation of these regulatory genes in the segregants. The results suggest that the transition of inedible papeda to sour mandarin implicated drastic gene expression reprograming of pivotal pathways of the primary and secondary metabolism, while palatable mandarins evolved through progressive refining of palatability properties, especially acidity.

KEYWORDS

edible mandarins, fruit acidity, genetic admixture, secondary metabolism, wild citrus

Introduction

Our current knowledge on citrus domestication is still very unprecise (Deng et al., 2020; Kalita et al., 2021; Rao et al., 2021). During the last years, most attention has been paid to the identification of genes or gene families impacting palatability traits, fruit characteristics and reproductive behavior. Among the palatability traits, wide evidence indicates that domestication has modulated pivotal genes regulating major components of citrus flavor, such as acidity (Butelli et al., 2019; Strazzer et al., 2019; Feng et al., 2021; Borredá et al., 2022), bitterness (Chen et al., 2021) or sweetness (Li et al., 2019; Feng et al., 2021). Domestication also appears to have reduced certain chemical defenses in citrus (Gonzalez-Ibeas et al., 2021a,b; Rao et al., 2021), as reported in many crops (Köllner et al., 2008; Yactayo-Chang et al., 2020). It has also been proposed that the increased fruit size of cultivated citrus was acquired during the citrus domestication stage (Wu et al., 2018; Gonzalez-Ibeas et al., 2021b). Several reproductive characteristics closer linked to yield, have also been suggested to be key domestication targets, such as flowering (FT, TFL1 LEAFY or AP1; Rao et al., 2021), self-incompatibility (S-locus; Liang et al., 2020), and apomixis (RWP, Nakano et al., 2012; Wang et al., 2017).

In previous work, we proposed that citrus domestication was led by apomixis and hybridization phenomena (Wu et al., 2018, 2021), a combination that drove reticulate evolution (Hamston et al., 2018) and the formation of a syngameon (Buck and Flores-Rentería, 2022), in the genus *Citrus*. Apomixis, that gives rise to nucellar embryony (polyembryony), allows asexual reproduction of the maternal phenotype. The rise of apomixis in the ancestral mandarin lineage, provided the framework to select through clonal propagation, plants with highly appreciated organoleptic and agronomical characteristics. On the other hand, all edible citrus bear unmistakable hybridization signatures of ancestral mandarins and pummelos (Wu et al., 2014, 2018), indicating that relevant introgressed traits were selected and fixed during domestication. During the last centuries, a myriad of crosses between ancestral hybrids and admixtures, gave rise to the current basic types of edible citrus that were progressively improved through somatic mutations and recurrent selection (Talon et al., 2020). We have also used genomic and transcriptomic analyses on wild and domesticated citrus to discriminate major determinants of evolution and domestication (Gonzalez-Ibeas et al., 2021a,b) and to identify genes involved in relevant physiological processes for domestication (Borredá et al., 2022).

In the current work, we follow this approach comparing gene expression in the fruitlet pulp of wild inedible Ichang papeda (ICH; *C. ichangensis* Swingle), acidic Sun Chu Sha Kat mandarin (SCM; *C. reticulata*, Blanco; *C. erythrosa* Tanaka) and three palatable genetic admixtures derived from a cross between two modern commercial mandarins, namely, Clementine (CLM; *C. x clementina* Hort. ex Tanaka) and W. Murcott (WMU, *C. x reticulata* Blanco). Recent developments suggest that ICH, that is considered one of the most primitive wild forms of citrus (Swingle and Reece, 1967; Yang et al., 2017), split from the main citrus clade around 7 million years (Wu et al., 2018). The ancestor of SCM probably appeared during the last 1.4 million years, after the divergence of the two main subspecies of mainland Asian mandarins (Wu et al., 2021). The main objective of this study was to provide insights on the genetic regulation of major biological processes affected by domestication in citrus.

Materials and methods

Plant material and sample processing

The plant materials used in this work were wild inedible Ichang papeda (ICH; *C. ichangensis* Swingle), acidic Sun Chu Sha Kat mandarin (SCM; *C. reticulata*, Blanco; *C. erythrosa* Tanaka) and three palatable genetic admixtures (S1, S2, and S3) derived from a cross between two modern commercial mandarins, namely, Clementine (CLM; *C. x clementina* Hort. ex Tanaka) and W. Murcott (WMU, *C. x reticulata* Blanco). Developing fruitlets were harvested from adult trees grown under normal culture practices at the IVIA germplasm bank and experimental fields, following the protocol described in Cercós et al. (2006). In essence, homogeneous fruits were selected by uniformity of size, appearance and absence of abiotic and biotic stress symptoms. For the transcriptomic analysis, the pulp of developing fruitlets of those two pure species, ICH and SCM, and the three pummelo/mandarin admixtures, S1, S2, and S3, was collected. Fruitlets were peeled, and flavedo (exocarp) and albedo (mesocarp) discarded. The remaining tissue, the fruit flesh consisting of juice vesicles (endocarp) including the segments with their membranes and vascular bundles, was frozen under liquid nitrogen and stored at −80°C until analyses. Three biological replicates of each sample were taken on July, 3rd, 2020 (62 days after anthesis), at the end the cell division stage (phase I) of the development of citrus fruits (Cercós et al., 2006), a critical period for the establishment of pivotal characteristics of the citrus ripe fruits (Terol et al., 2019). A second set of pulp samples to be used in qPCR analysis determinations, consisting exclusively of juice vesicles, was collected from ripening fruits on November 21st (190 days after anthesis).

Phenotypical data

In order to study correlation between gene expression and acidity and because acid levels do not still show accumulation at early July, fruits were also collected once a month, during the ripening period (from October to February), when the maximum acid accumulation that generally occurs in September, has already taken place (Cercós et al., 2006). Biochemical parameters (acidity, °Brix, and maturity index) were registered in these samples. Citric acid equivalents (g/l) were determined by titration with 0.1 M sodium hydroxide and a phenolphthalein indicator. Soluble sugar content was measured with a refractometer ATAGO PR-1.

RNA extraction, library preparation and sequencing

Total RNA from pulp samples was extracted with acid phenol and precipitated with lithium chloride. Three biological replicates were used for each sample. Library preparation and sequencing were carried out by a commercial service following standard protocols. Essentially, samples enriched in mRNA were randomly fragmented and cDNA synthesized. After adapter ligation, size selection and PCR enrichment, samples were sequenced in an Illumina NovaSeq 6000 platform yielding 150 bp pair-end reads. On average, each biological replicate produced 7.04 Gb of sequence data in 23497582 raw reads.

RNA-seq read mapping and differential expressed genes analysis

The *C. clementina* reference genome (Wu et al., 2014) and its annotation data, as reported at the NCBI (National Center for Biotechnology Information [NCBI], 1988), were used for RNA analysis. First, raw reads were mapped using STARv2.7.6 (Dobin et al., 2013). Read counts were computed by featureCounts function in the Rsubread package (Liao et al., 2014). DESeq2 1.26 (Love et al., 2014) was used for expression analysis following author's recommendation. Differential expressed genes (DEGs) were detected performing pair-wise comparisons of ICH and each one of segregants, against SCM. In these comparisons, the three biological replicates of each sample were treated as a group. Two combinations of Log_2 Fold Change and alpha thresholds, either $Log_2FC = 0.58$ and $\alpha = 0.05$ or $Log_2FC = 1$ and $\alpha = 0.01$, were set for expression analysis.

Pathway analysis

The set of DEGs between ICH-SCM resulting of the comparison using the softer thresholds, $Log_2FC = 0.58$ and $\alpha = 0.05$, was annotated using *C. clementina* KEGG data (Kanehisa et al., 2016). Enzymatic information and KEGG identifiers of those DEGs that mapped to the set of KEGG metabolic pathways were retrieve and Pathview 4.1 (Luo and Brouwer, 2013) was used to represent the differentially expressed. For easy interpretation, differential expression was represented by three uniform colors (up, red; blue, down; yellow, undetermined) without indication of the Log_2FC. We reserved the term "undetermined" expression for genes that share the same functional annotation (i.e., same KEGG identifier), but showed opposite expression trends, as long as one of the biological replicates reached at least 100 reads. In a few cases where both the number of genes and reads with the same expression trend were unequivocally higher, the expression displayed by these genes was considered the dominant expression.

Functional annotation and category assignment

In order to functionally characterize and categorize relevant DEGs found in the previous ICH-SCM comparison, a second analysis using more restrictive thresholds, $Log_2FC = 1$ and $\alpha = 0.01$, was performed. From the set of DEGs obtained in this second comparison, genes that were included in previous KEGG analyses and genes that did not reach 100 reads in at least one replicate, were removed. In a first step, the remaining DEGs were functionally annotated according to the Uniprot database (UniProt Consortium, 2021) and current literature. The DEGs were grouped into 45 user-defined categories, according to their involvement in major physiological, biochemical and genetic processes, and these categories were in turn split twice as much in subsequent subcategories. Genes that could not been assigned to a specific process were grouped in clusters defined by their molecular function. A group of uncharacterized DEGs, with general and ambiguous annotations or with undescribed assignments in plants was also created.

In another experiment, DEGs between the three palatable segregants and SCM were identified. Pairwise comparisons of each segregant, S1, S2, and S3, against SCM ($Log_2FC = 0.58$ and $\alpha = 0.05$) were parsed as above and genes that did not reach at least 100 reads in one of the biological replicates were similarly filtered out. DEGs were grouped into the aforementioned categories and common DEGs in the three segregants were included in a single list and manually characterized.

RT-qPCR expression analysis

RT-qPCR was used to validate RNA-seq analysis in the pulp of developing fruitlets (**Supplementary Figure 1**) and to test expression of target genes in juice vesicles of ripening fruits (November). Two replicates of each sample/gene combination were performed in one-step reaction in a LightCycler Instrument. Each sample was incubated with the reverse transcriptase MultiScribe (Invitrogen) at 48°C during 30 min and with an RNAse Inhibitor (Applied Biosystems). Reaction mastermix also included LightCycler FastStart DNA Master Plus SYBR Green I kit for amplification step.

Relative gene expression was calculated using ΔΔCt method. CitUBC1 (Merelo et al., 2017) and CitACTIN11 (Strazzer et al., 2019) were used as housekeeping genes for data normalization. All primer sequences used are available in **Supplementary Table 1**.

DNA extraction, sequencing and mapping

For each segregant, a sample of fresh leaves was collected and DNA purified using CTAB extraction method. Library preparation and whole genome sequencing (WGS) were carried out by a commercial service following a standard protocol. In short, genomic DNA was randomly sheared into short fragments that, subsequently, were end-repaired, A-tailed and further ligated with Illumina adapters. Fragments with adapters were PCR amplified, size selected, and purified. Sequencing was run in an Illumina NovaSeq 6000 platform yielding 150 bp pair-ended reads. On average, each sample produced 104484177 raw reads, generating 31.35 Gb of sequence data. The sequences were mapped to the *C. clementina* haploid reference (Wu et al., 2014) using BWA-MEM tool (Li, 2013). Map files were sorted and indexed using Samtools (Li et al., 2009). The mean read depth of each segregant was 68.8x, 75.9x, and 70.7x.

WGS data for *C. maxima* (CHP, pummelo) and *C. reticulata* (SCM, mandarin), the two pure species that make up the genome of the three segregants, were retrieved from the Sequence Read Archive database (National Center for Biotechnology Information [NCBI], 1988) and processed as above.

Variant calling

Variant calling was performed using the GATK-4.0.0.0 software (Van der Auwera et al., 2013). We used the HaplotypeCaller tool to generate single-sample variant call format files that were combined using the CombineGVCF tool to get matrices including the samples of the study. Each site showing a quality value greater than 10 was genotyped by GenotypeGVCF and only calls tagged as SNP were filtered according to a set of standard filters specified in the variant caller practice guide. To include RNA-seq data, input files were reformatted to adapt the alignments that span introns using SplitNCigarReads tool. Raw matrices were filtered to get species informative markers (SIM). To this end, we retrieved sites holding fixed differences between *C. maxima* (CHP) and *C. reticulata* (SCM). The set of diagnostic markers only included sites with different alleles in homozygosis, supported by at least 20 reads.

Admixture pattern

The above set of SNPs was used to define local ancestry segments along the genome of each segregant. The haplotype (pummelo or mandarin) of the admixture stretches was determined, as in Wu et al. (2018), using windows of 1000 markers. Essentially, for each SIM, the copy number of both ancestral alleles (i.e., 2, 1, or 0) was recorded and the ancestry of the window was inferred from the most frequent one. Stable segments were considered when a minimum of five windows in a row exhibited the same inheritance. Otherwise, the ancestry of the nearest stable block was applied. These data were used to compute the distribution of expressed genes in each segregant. Only genes that were covered with a minimum of 10 reads in any of the biological replicates were considered. Haplotype combinations were named, MA/MA (mandarin/mandarin), PU/MA (pummelo/mandarin), and PU/PU (pummelo/pummelo).

Allele differential expression

Species informative markers were additionally used to study differentially expressed alleles in PU/MA regions. Only genes with heterozygous sites covered by a minimum average depth of 10x, showing identical genotype in the three biological replicates, were initially considered for this analysis. The final set of cDNAs used in the analysis of allele differential expression included genes, in which approximately the 80% of the SNPs spanning their sequences met these conditions. From this set of target genes, the occurrence of homozygous SIMs was assessed as exclusive expression of a certain allele.

Admixture validation

We used PCR and Sanger sequencing to validate both a sequence change causing an admixture pattern shift such as crossover, and an event of allele differential expression (**Supplementary Figures 2**, **3**).

Results

To study differential gene expression as related to domestication in the pulp of developing fruitlets, two independent comparisons were performed. In one of them, the transcriptome of inedible Ichang papeda (ICH) and acidic Sun Chu Sha Kat mandarin (SCM) were contrasted. In a second evaluation, this mandarin was compared with a group formed by three palatable genetic pummelo/mandarin admixtures derived from a cross between the commercial mandarins Clementine and W. Murcott.

Gene expression in inedible wild Ichang papeda versus acidic Sun Chu Sha Kat mandarin

The transcriptome comparison between the developing fruitlet pulp of ICH and SCM (Log$_2$ Fold Change threshold of 0.58 and alpha of 0.05) identified 7267 DEGs (**Supplementary Table 2**). Out of this set of genes, 2342 were mapped to 117 pathways, mostly related to metabolism, genetic information and cellular processes (**Supplementary Figures 4.001–4.117**), as defined in the KEGG database resource (Kanehisa et al., 2016). Genes that did not map to the KEGG collection, were compared using more stringent conditions (Log$_2$ Fold Change threshold of 1 and alpha of 0.01). This analysis rendered 2832 DEGs (**Supplementary Table 3**) that were, first, functionally annotated according to the Uniprot database (UniProt Consortium, 2021), and then grouped into 45 user-defined categories. **Figure 1** presents the 32 categories encompassing more than 10 members, excluding the group of uncharacterized processes. This classification overall indicates that only categories included in Secondary Metabolism and categories related to protein metabolism (Translation, Ubiquitination, Trafficking) were upregulated in the papeda while most of those grouped in Development, Regulation and Transport, Signaling, Gene Expression, Cellular Growth and Stress responses were downregulated. Thus, more than 60% of those 2832 DEGs were downregulated in the fruit flesh of ICH, a percentage rather identical to that observed in the group of uncharacterized genes. The results described below are restricted to categories potentially involved in domestication while the description and comments affecting to the rest of results are available in **Supplementary Material**.

Alkaloids

We grouped 23 DEGs into the Alkaloids category (**Supplementary Table 3**) while the KEGG pathway mapping added several additional genes to this cluster. According to the expression levels of transcripts associated with the synthesis of several principal alkaloids, the data suggest that these compounds are generally upregulated in ICH. In this species, for instance, caffeine synthase is upregulated (LOC112100740). The transcripts of codeine 3-*O*-demethylase, the last step in the biosynthesis of morphine (LOC18036840) in the Isoquinoline Alkaloid Pathway, also was upregulated. Similarly, synthesis of major indol alkaloids, i.e., ajmaline, vinblastine and vincristine and also that of iridoid compounds appears to be favored as suggested by the upregulation of relevant biosynthetic steps controlling the conversions of monoterpenoid precursors of indole alkaloid biosynthesis. These upregulated steps are geraniol 8-hydroxylase (5 genes out of 6), 8-hydroxygeraniol dehydrogenase (LOC18037214), the dehydrogenase involved in the biosynthesis of oxogeranial from hydroxygeraniol and 7-deoxyloganetin glucosyltransferase (LOC18042369), an iridoid glucosyltransferase involved in the synthesis of secologanin, one of the major intermediates in the indole alkaloid biosynthesis. It was also observed that a vinorine synthase (LOC18037906), a gene coding for the acetyltransferase catalyzing the formation of vinorine, a precursor of the monoterpenoid indole alkaloid ajmaline, was also upregulated. The synthesis of nicotine, a major alkaloid derived from nicotinic acid (**Supplementary Figure 4.067**), however, does not appear to be activated. In the acridone alkaloid biosynthesis, two methanol *O*-anthraniloyltransferases genes (LOC18050765 and LOC18050764), in principle implicated in the synthesis of methyl anthranilate, showed opposite expression tendencies. Likewise, evidence for the upregulation of the tropane biosynthesis, leading to alkaloids such as atropine, hyoscyamine and scopolamine was not obtained. The analysis of gene expression certainly showed upregulation of two tropinone reductase-like genes (LOC18039952 and LOC18039951) but the corresponding proteins do not appear to exhibit tropinone reductase activity (Jirschitzka et al., 2012). Interestingly, synthesis of both dopamine and serotonin (**Supplementary Figures 4.033, 4.089**), two amides that are considered bioactive alkaloid neurotransmitters, in principle, appear to be downregulated in ICH, since tyrosine/DOPA decarboxylase 5 (aromatic L-amino acid decarboxylase; LOC18043348, **Supplementary Table 3**, and LOC18046227, EC:4.1.1.28, **Supplementary Figures 4.033, 4.087** and

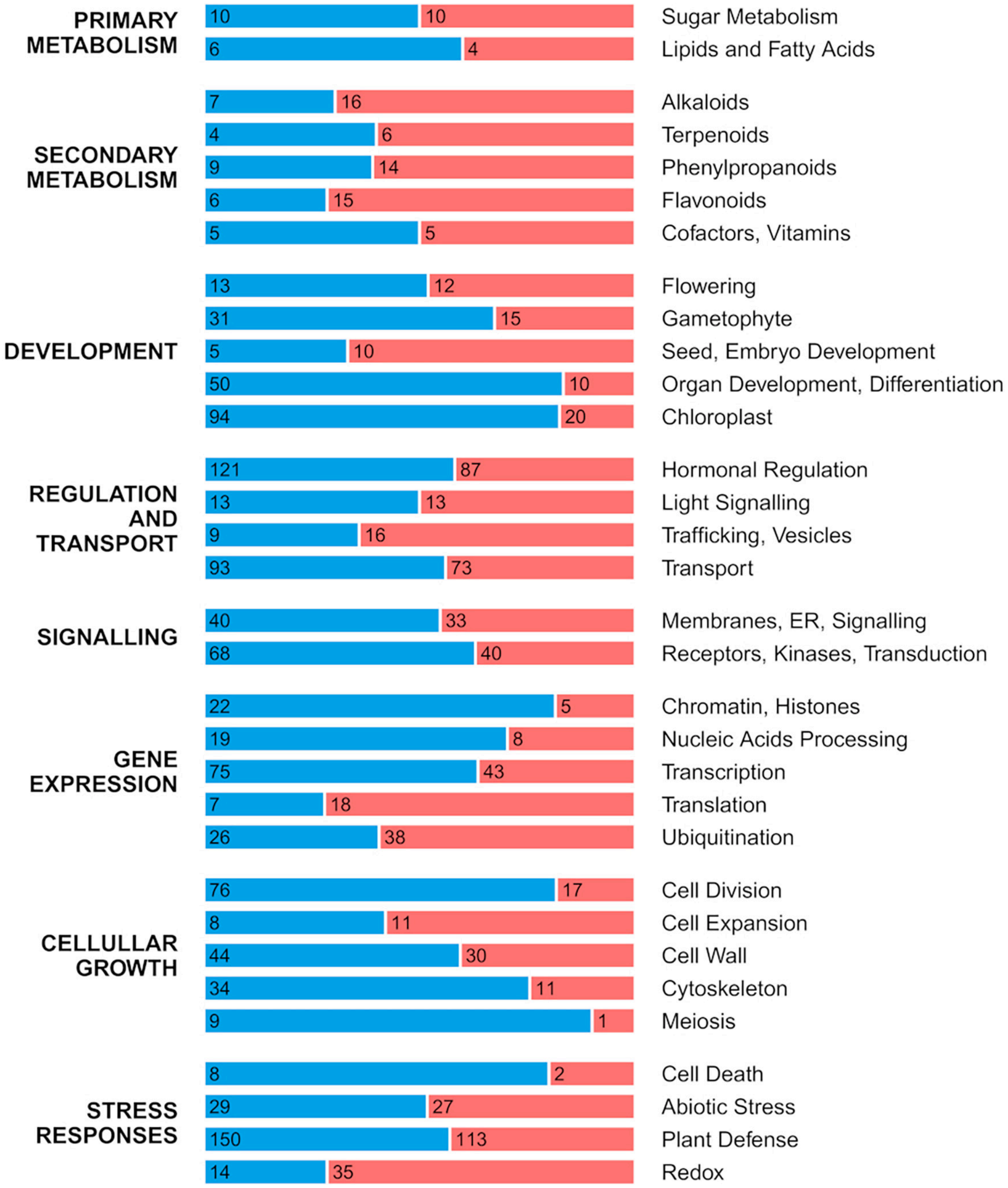


FIGURE 1
Distribution of up- and downregulated DEGs in the pulp of developing fruitlets of ICH as related to SCM. DEGs were manually classified into eight major groups that were subsequently split in 32 gene categories corresponding to the most populated categories presented in Supplementary Table 3, excluding the group of uncharacterized genes. Length of red and blue stretches represents the frequency of up- and downregulated genes, respectively, while numbers refer to the number of genes included in each category.

Supplementary Table 2) is repressed in this species. This enzyme (EC:4.1.1.28) may render tyramine, tryptamine, dopamine and serotonin. Likewise, a CYP71P1 gene (LOC18041681, EC:1.14.-.-, Supplementary Figure 4.033 and Supplementary Table 2) encoding tryptamine 5-hydroxylase that also catalyzes the conversion of tryptamine to serotonin, is repressed too.

Phenylpropanoids

The last steps in the formation of the derived aldehyde and alcohol of major phenylpropanoids (Supplementary Figure 4.081 and Supplementary Table 2), such as cinnamic, coumaric, caffeic, ferulic, hydroxyferulic and sinapic acids, are predominantly upregulated in ICH, as observed for instance, cinnamoyl-CoA reductase (LOC18035881, EC:1.2.1.44).

However, trans-cinnamate 4-monooxygenase, CYP73A (LOC18055509, EC:1.14.14.91) and 4-coumarate–CoA ligase (LOC18034975, LOC18050151, and LOC18036308, EC:6.2.1.12) the main steps in the synthesis of *p*-Coumaroyl-Co A, the precursor of flavonoids and the non-flavonoid polyphenols, stilbenoid, diarylheptanoic and gingerol biosynthesis, are downregulated. In this pathway, *trans*-resveratrol di-*O*-methyltransferase (LOC18051743 and LOC18031586, **Supplementary Tables 2, 3**), the last step in the biosynthesis of the antifungal phytoalexin pterostilbene, is expressed a higher level in the papeda. In **Supplementary Table 3**, there are listed other DEGs that participate in the regulation of the phenylpropanoid biosynthesis such several members of the Cytochrome P450 71A1 family, that appear to act as trans-cinnamic acid 4-hydrolases, or the MYB family transcription factor PHL11. The enzymes listed in this table, caffeic acid 3-*O*-methyltransferases, caffeoylshikimate esterases-like, cinnamoyl-CoA reductases, and shikimate *O*-hydroxycinnamoyltransferases, that are not mapped at the KEGG pathways, appear in principle to be associated with the phenylpropanoid pathway and at least some of them, with lignin biosynthesis.

Flavonoids

In the flavonoid pathway, the synthesis of the pivotal intermediate flavanone naringenin in ICH appears to be downregulated (**Supplementary Figure 4.082** and **Supplementary Table 2**), since chalcone isomerase (LOC18044429, EC:5.5.1.6), was repressed, although there were at least four different chalcone synthetases (LOC18042808, LOC18042812, LOC18033130, and LOC18051925, E.C:2.3.1.74), with opposite genetic expression levels. The synthesis of isoflavonoids (**Supplementary Table 3**) was mostly characterized by the upregulation of 2-hydroxyisoflavanone dehydratase-like (LOC112096719 and LOC112098436), that catalyze the final step in the formation of the isoflavonoid skeleton rendering daidzein, of isoflavone 4′-*O*-methyltransferase (LOC18053393) involved in the biosynthesis of formononetinin and of isoflavone 2′-hydroxylases (LOC18043085), that mediates the hydroxylation of daidzein and formononetin, to yield 2′-hydroxyisoflavones. In spite of downregulation of naringenin, the formation of the polyphenolic flavonols, kaempferol, quercetin, and myricetin appears to be upregulated (**Supplementary Figure 5**). **Supplementary Table 3**, for instance, reports several flavonoid 3′-monooxygenases (LOC18053376) and flavonoid 3′,5′-hydroxylases (LOC18048580) controlling the conversion of naringenin to eriodictyol, dihydrokaempferol, dihydroquercetin and dihydromyricetin, and **Supplementary Figure 4.082** shows upregulation of naringenin 3-dioxygenase (LOC18036490, EC:1.14.11.9) and flavonol synthase (LOC18037475, EC:1.14.20.6, **Supplementary Table 2**), two regulating enzymes of the synthesis of kaempferol, quercetin and myricetin. The synthesis of the anthocyanidins, pelargonidin, cyanidin and delphinidin, and their corresponding anthocyanins (anthocyanidin glycosides; Khoo et al., 2017) was similarly upregulated in the papeda. **Supplementary Table 2**, for example, shows that two limiting steps in the synthesis of anthocyanidins and anthocyanins, anthocyanidin synthase (LOC18047155, EC:1.14.20.4, **Supplementary Figure 4.082**) and anthocyanidin 3-*O*-glucosyltransferase (LOC18047244, EC:2.4.1.115, **Supplementary Figure 4.083**), respectively, were clearly upregulated. In addition, **Supplementary Table 3** enumerates a number of upregulated members of several gene families, coumaroyl-CoA:anthocyanidin 3-*O*-glucoside-6′′-*O*-coumaroyltransferase 1 (LOC18050842 and LOC18032737) malonyl-CoA:anthocyanidin 5-*O*-glucoside-6′′-*O*-malonyltransferase (LOC18055666, LOC18038126, LOC18044783, and LOC18046030) and putative anthocyanidin reductase (LOC18047966), suggesting that the metabolism of anthocyanins is very active in this species. It should be mentioned, that mandarins do not contain anthocyanins likely because the Ruby gene (synonymous AN2) is not functional in these varieties (Butelli et al., 2019; Wu et al., 2021). **Supplementary Table 3** also lists 2 flavonol-specific transcription activators, MYB11 and MYB111 (LOC18049115 and LOC18031574) involved in the regulation of several genes of the flavonoid biosynthesis. The synthesis of the phytoalexin glyceollin, however, does not appear to be promoted since 3,9-dihydroxypterocarpan 6A-monooxygenase (LOC18031687), a previous biosynthetic step, was downregulated.

Terpenoids

In the terpenoid pathway, there were clear-cut differences between both species. The papeda showed upregulation of pivotal genes of the mevalonate pathway, in detriment of the MEP pathway, that takes place in plastids (**Supplementary Figure 4.072**). Thus, as related to the biosynthetic regulation of the isoprenoid precursors, isopentenyl pyrophosphate was promoted against dimethylallyl pyrophosphate. In addition, the pivotal gene, geranylgeranyl diphosphate synthase, (dimethylallyltranstransferase, LOC18039078, EC:2.5.1.1, and LOC18039079, EC:2.5.1.29, **Supplementary Table 2**), giving rise to main precursors of the different terpenoid types, were also expressed at higher levels. The synthesis of monoterpenoids was not basically modified except for the upregulation of 2 out of 3 (R)-limonene synthases 1 (LOC112098486 and LOC112098571, **Supplementary Table 3**), catalyzing the conversion of geranyl diphosphate to (+)-(4R)-limonene. In the diterpenoid biosynthesis, trimethyltridecatetraene/dimethylnonatriene synthase, CYP82G1 (LOC18049179 and LOC18049178, EC:1.14.14.58, **Supplementary Figure 4.074** and **Supplementary Table 2**), catalyzing the production of the volatile homoterpenes DMNT and TMTT, was also upregulated. In the sesquiterpenoid pathway, the synthesis of farnesene appears to be downregulated since transcripts of α-farnesene synthase (LOC18053589, EC:4.2.3.46, Supplementary Figure 4.078 and Supplementary Table 2) are present at lower levels, while there were two (3S,6E)-nerolidol synthase 1 genes (LOC18033168 and LOC18051782, EC:4.2.3.48) expressed in opposite directions. Expression of premnaspirodiene oxygenase (LOC18052043), involved in the biosynthesis of the sesquiterpenoid, solavetivone, a potent antifungal phytoalexin, was upregulated, while that of α-copaene synthase-like (LOC112095677) that converts farnesyl diphosphate to the bicyclic olefins α-copaene and (*E*)-β-caryophyllene and participates in the synthesis of the macrocyclic sesquiterpene germacrene D, was downregulated (Supplementary Table 3). One important gene implicated in the synthesis of steroids and triterpenoids was squalene monooxygenase, SQLE (LOC18033947 and LOC18033838, EC:1.14.14.17, Supplementary Figures 4.012, 4.078 and Supplementary Table 2), that was downregulated in the papeda and therefore, probably limiting the flux toward these compounds. Consistently, the number of DEGs regulating triterpenoid metabolism were also scarce, since only the synthesis of β-amyrin appears to be upregulated (LOC18034001, EC:5.4.99.39, LOC18045727 and LOC18053646, Supplementary Tables 2, 3). Regarding carotenoids, only transcripts coding for enzymes mediating phytoene synthesis (LOC18051922, EC:2.5.1.32, Supplementary Figure 4.076 and Supplementary Table 2) were upregulated in the papeda, while the metabolism of α- and β-carotenes, including the synthesis of cryptoxanthin, lutein, astaxanthin and the xanthophyll cycle, was strongly repressed. Supplementary Table 3 also lists two upregulated β-D-glucosyl crocetin β-1,6-glucosyltransferases (LOC18045744 and LOC18044391), catalyzing the β 1-6 glucosylation of crocetin, a natural apocarotenoid. In addition, this table reports on other two glycosyltransferases (LOC18054015 and LOC18033356) conjugating diterpenes that are downregulated, and an upregulated gene of the diterpenoid metabolism, cytochrome P450 76M5 (LOC18031421), involved in the biosynthesis of oryzalexin, a class of phytoalexins. The terpenoid pathways implicating plant hormones such as GAs, ABA, cytokinins and brassinosteroids are described in the Hormonal category. Upregulation of the committed steps of the synthesis of tocopherol and tocotrienol (vitamin E), homogentisate phytyltransferase/geranylgeranyltransferase (LOC18055996, EC:2.5.1.115 and EC:2.5.1.116, Supplementary Figure 4.013 and Supplementary Table 2) was also found in the papeda.

Hormonal regulation

The comparison between the transcriptome of ICH and SCM also rendered a relatively high number of DEGs involved in hormone biosynthesis and action. According to the KEGG mapping of biosynthetic DEGs (Supplementary Table 2), the synthesis of active cytokinins, brassinosteroids and ethylene was downregulated in the papeda. Regarding cytokinins biosynthesis, transcripts for cytokinin dehydrogenase, CKX (LOC18042746 and LOC18033392, EC:1.5.99.12), an enzyme that inactivates isopentenyl adenine, were upregulated, in contrast to those of two glucosyltransferases conjugating zeatin, (LOC18031300, EC:2.4.1.215, and LOC18038024 and LOC18037288, EC:2.4.1.-), that were repressed (Supplementary Figure 4.077). The synthesis of brassinosteroids depending upon steroid precursors, also appears to be strongly downregulated (Supplementary Figure 6), since most steps, including last steps in the synthesis of brassinolide, such as brassinosteroid-6-oxidase 1, CYP85A1 (LOC18038016 and LOC18044268, EC:1.14.-.-) and PHYB activation tagged suppressor 1, CYP734A1 (LOC18054829, EC:1.14.-.-), were downregulated (Supplementary Figure 4.075). The generation of ethylene does not appear to be promoted either in the papeda, because the last step in its synthesis, 1-aminocyclopropane-1-carboxylate oxidase (LOC18050524, EC:1.14.17.4) was repressed, although the previous conversion catalyzed by 1-aminocyclopropane-1-carboxylate synthase (LOC18048242 and LOC18046436, EC:4.4.1.14), was upregulated (Supplementary Figure 4.024). Although there was upregulation of early steps in the gibberellin biosynthesis, including the conversions between ent-kaurene to inactive GA_{12} (LOC18046916, EC:1.14.14.107, Supplementary Figure 4.074) no differences in the expression of biosynthetic genes controlling the formation of active GAs between both species were found. A similar situation was observed in the synthesis of jasmonate and methyl-jasmonate, characterized by the upregulation of no less than 6 early biosynthetic steps (MFP2, LOC18032845, EC:4.2.1.17 and ACX, LOC18047109, EC:1.3.3.6, Supplementary Figure 4.053). The synthesis of xanthoxin, a precursor of ABA, and ABA degradation, was upregulated in the papeda, since genes coding for 9-*cis*-epoxycarotenoid dioxygenase (LOC18043465 and LOC18050641, EC:1.13.11.51) and abscisic acid 8′-hydroxylase (LOC18039758, EC:1.14.14.137, Supplementary Figure 4.076), the proteins controlling these conversions were expressed at higher levels. The data also suggest that auxin synthesis was also promoted because two amidases (LOC18033993 and LOC18034584, EC:3.5.1.4), an aldehyde dehydrogenase (LOC18036436, EC:1.2.1.3) and at least an indole-3-pyruvate monooxygenase (LOC18032700, EC:1.14.13.168) participating in the auxin synthesis, were expressed at higher levels (Supplementary Figure 4.033). In contrast, the amino acid derived polyamines, spermidine (LOC18039571, EC:1.5.3.17), spermine (LOC18043803, LOC18051913 and LOC18054425, EC:1.5.3.16) and putrescine (LOC18049691, EC:4.1.1.17) were apparently downregulated (Supplementary Figure 4.029 and Supplementary Table 4). Moreover, the expression levels of pivotal receptors, transporters and regulators implicated in the hormone signal transduction, indicate that the receptors of cytokinins, CRE1 (LOC18052080, EC:2.7.13.3), brassinosteroids, BRI1

(LOC18035850, EC:2.7.10.1, and EC:2.7.11.1) and ethylene, ETR (LOC18031847, EC:2.7.13.-) are repressed in the papeda, like most components of the auxin transduction pathway, including the auxin flux carrier AUX1 (LOC18034947, LOC18038157, and LOC18045480) and the regulators TR1 (LOC18052162) or GH3 (LOC18035901, LOC18053209, LOC18054772, LOC18033692, and LOC18041056). In contrast, JAR1 (LOC18049830, EC:6.3.2.52, ST 7267) and PP2C (LOC18043434, EC:3.1.3.16), major regulator of jasmonate and ABA responses, respectively and GID1 (LOC18049839 and LOC18043172), the receptor of GAs, were both upregulated (Supplementary Figure 4.109). Regarding MAPK Signaling Pathway, upregulation of MKK3 (LOC18051058, EC:2.7.12.2) and MPK6 (LOC18047683, EC:2.7.11.24, Supplementary Figure 4.107 and Supplementary Table 2), was the most significant observation as related to hormonal regulation.

In addition, Hormonal Regulation category included 208 DEGs, out of which more than 58% were downregulated in ICH (Supplementary Table 3). Although all phytohormones were represented in this set of genes, not all of them exhibited the same down/up regulation ratio. In particular, transcripts related to auxins (49/19, transport, homeostasis, ARFs, AUX/IAA, SAURs, response, biosynthesis, signaling), cytokinins (6/2, transport, receptors, transcription factors), gibberellins (13/4, biosynthesis, response, signaling, transcription factors) and jasmonic acid (13/8, transcripts related to biosynthesis, transport, response, receptor, induced response, signaling, transcription factors) had higher number of downregulated genes. Downregulation frequency of these categories ranked from 0.62 to 0.76. Genes linked to ethylene (6/18, transcription factors, induced responses, ERFs) displayed the opposite tendency, while transcripts associated with ABA (21/20, biosynthesis, response, receptor, induced response, signaling, transcription factors), polyamines (1/1, transporters), salicylic acid (8/10, transcription factors, induced responses, signaling, biosynthesis) and brassinosteroids (4/4, transcription factors, response, signaling, homeostasis) exhibited a down/up ratio that hardly departs from 0.5 (Supplementary Table 3).

Other categories related to growth, such as Gametophyte, Organ Development, Differentiation, Chloroplast, Cell division, Meiosis, Cytoskeleton, Cell Wall, Receptors and Protein Kinases, Chromatin, Histones, Transcription, and Nucleic Acids Processing were also downregulated in the papeda (Supplementary Material).

Primary metabolism

Regarding carbon metabolism, striking differences were found between both species, since practically all genes coding enzymes of central regulatory pathways such as glycolysis (Supplementary Figure 4.001), including the generation of pyruvate, acetyl CoA and acetaldehyde (Supplementary Figure 4.057), were clearly up-regulated in ICH. In the pentose phosphate pathway (Supplementary Figure 4.003), the synthesis of the pivotal intermediate, glyceraldehyde 3P, was similarly upregulated. Expression of genes involved in sucrose synthesis and degradation do not appear to be clearly modified, while the formation of ADP-glucose and amylose (Supplementary Figure 4.042), but not that of starch, also were upregulated in the papeda, as that of α amylase (LOC18043125 and LOC18045113, EC:3.2.1.1, Supplementary Figure 4.042 and Supplementary Table 2). In this species, it was also repressed a regulatory subunit of the probable trimeric SNF1-related protein kinase, (SnRK; LOC18052199, Supplementary Table 3) complex, that appears to play a role in the transduction cascade regulating gene expression and carbohydrate metabolism. In the papeda, other important differences were found in the metabolism of organic acids, especially the tricarboxylic cycle (TCA), that showed up-regulation of most genes coding for their regulatory enzymes (Supplementary Figure 4.002 and Supplementary Table 2), including those of the pyruvate dehydrogenase complex (LOC18045003 and LOC18037469, EC:2.3.1.12). A noticeable exception to this observation, however, was the conversion of oxoglutarate to succinyl-CoA that was down-regulated (LOC18044474, EC:1.2.4.2). Succinyl-CoA synthetase alpha subunit (LOC18045656, EC:6.2.1.4 and EC:6.2.1.5), on the other hand, was upregulated. Expression of genes regulating cytoplasmatic organic acid metabolism was similarly altered, since several subunits of ATP-citrate lyase, ACLY (LOC18032750 and LOC18043354, Supplementary Table 2 and LOC18039980, Supplementary Table 3), that converts citrate into oxaloacetate and cytosolic acetyl-CoA, were upregulated. Likewise, genes coding for a series of enzymes acting sequentially, such as aconitate hydratase, ACO3 (LOC18055416, EC:4.2.1.3, Supplementary Figure 4.058), one isocitrate dehydrogenase [NADP], IDH1, (LOC18031748, EC:1.1.4.2) and aspartate transaminase, GOT1 (LOC18054901, EC:2.6.1.1, Supplementary Figure 4.017), rendering the amino acid glutamate from 2-oxogluarate, were also upregulated. In addition, three additional enzymes, glutamate decarboxylase, GAD5 (LOC18046053 and LOC18052541, EC:4.1.1.15), butyrate pyruvate transaminase, POP2 (LOC18039191, EC:2.6.1.96) and succinate-semialdehyde dehydrogenase, SSADH (LOC18031917, EC:1.2.1.24), that together make up the GABA shunt, showed the same tendency (Supplementary Figure 4.021). The conversion of glutamate to glutamine (LOC18044424, EC:6.3.1.2) and pyrroline-carboxylate (LOC18045924, EC:1.2.1.88) also was favored in ICH (Supplementary Figure 4.021). The oxidative phosphorylation likewise seems to be more active in ICH, since all DEGs implicated in this process were upregulated (Supplementary Figure 4.014 and Supplementary Table 2). These genes code for several components of complex I, NADH hydrogenase, including NADH-quinone oxidoreductase subunit A (LOC18034396, EC:7.1.1.2); complex III, cytochrome c reductase, including ubiquinol-cytochrome c reductase cytochrome b/c1 subunit (LOC18051936, EC:7.1.1.8), complex IV, cytochrome oxidase, including cytochrome c oxidase cbb3-type subunit I, COX6A and COX6B (LOC18037730 and LOC18055702, EC:7.1.1.9) and complex V, ATP synthase, including H + -transporting ATPase (LOC18055993, LOC18039766, LOC18053876, and LOC18035736, EC:7.1.2.1). Major changes in the lipids and fatty acids pathways are discussed in **Supplementary Material**.

Amino acid metabolism

Amino acid metabolism and the synthesis of several derived compounds differ in both species. **Supplementary Table 4** speculates on the regulation of the synthesis of these compounds in each species, based on gene expression levels of the last regulatory steps (**Supplementary Figures 4.017, 4.021–4.036, 4.039, 4.042**). According to this information, the synthesis of amino acids tends to be upregulated in ICH, except for the production of valine, leucine, and isoleucine that was clearly repressed, and their degradation upregulated. Data related to amino acid-derived hormones and alkaloids are specified elsewhere in this section. In the cyanoamino acid metabolism, it is worth to mention that in ICH, linamarin synthase (LOC18036876, **Supplementary Table 3**), an UDP glycosyltransferase producing cyanogenic glucosides was upregulated. Consistently, β-glucosidase 13 (orthologous of β-glucosidase 12 of cassava, linamerase, LOC18044658, EC:3.2.1.21, **Supplementary Table 2**), that converts cyanogenic glucosides into acetone cyanohydrins such as mandelonitrile, and mandelonitrile lyase (LOC18037363, EC:4.1.2.10), that releases HCN, hydrogen cyanide, from the acetone were both downregulated (**Supplementary Figure 4.039**). The synthesis of glucosinolates may be promoted in the papeda, since a flavin-containing monooxygenase FMO GS-OX-like 4 (LOC18049889, **Supplementary Table 3**) and two mRNAs coding for cytochrome P450 83B1 (LOC18041363 and LOC18046366, ST 2832) that catalyze the oxime metabolizing step in indole glucosinolate biosynthesis were upregulated.

Transport

In the papeda, the Transport category was also enriched in downregulated genes (**Figure 1**), although specific differences were found among the wide range of transport systems included in this group (**Supplementary Table 3**). For instance, ABC transporters for glutathione S-conjugates, vacuolar ATPases, and copper, magnesium, sulfate, and zinc transporters were mostly upregulated, while transporters of amino acids, ascorbate, cation channels, proton antiporters, components of the mitochondrial electron transport chain, mechanosensitive channels, metal-nicotianamine transporters, nitrate, sodium, cadmium, nuclear import, oligopeptides and xenobiotics were mostly downregulated. Other transporters, such as aquaporins, purines, and transporters of boron, calcium, and potassium showed similar number of up- and downregulated genes. The number of sugar transporters was relatively high (19) and some genes were highly expressed in SCM mandarin in contrast to ICH, such as the vacuolar hexose transporter SWEET17 (LOC18032835) and other hexoses carriers (LOC18031330 and LOC18048094) or monosaccharide transporters (LOC18031593). ALTM4 (LOC18043583), an aluminum-activated malate transporter was also downregulated in papeda.

Abiotic stress

As related to stress responses, upregulation enrichment was observed in the papeda fruitlet flesh predominantly in three cases (**Supplementary Table 3**): in heat shock proteins (4/11), in cold responses (3/5), and in oxidative stress (6/35). Out of the five upregulated genes involved in cold responses, CORA, a cold and drought-regulated protein, showed the highest expression observed in this analysis (LOC18054952). Among the three downregulated transcripts, there were two transcription factors, the activator ICE1 (LOC18051975) and the repressor MYBS3 (LOC18055469), that regulate the cold-induced transcription of DREB1/CBF genes. Additional genes involved in cold responses mediated by ABA were also differentially expressed among both species. For instance, negative regulators of ABA such as MSI4 (LOC18036552) and ERD15 (LOC18045483) were expressed at lower levels. JUB1 (LOC18044819), another gene participating in the response to freezing was upregulated. As mentioned before, a further category enriched in the papeda with upregulated DEGs was that of Redox (**Figure 1** and **Supplementary Table 3**). It includes mostly genes coding for oxidoreductase enzymes, that play major roles in the antioxidant defense system, such as oxidases, reductases, peroxidases, cytochromes P450, mono- and dioxygenases and glutathione S-transferases. In the glutathione-ascorbate cycle, an efficient metabolic pathway to detoxify H_2O_2, monodehydroascorbate reductase (LOC18039033, E.C:1.6.5.4) was upregulated, while ascorbate peroxidase (LOC18039197, LOC18040244, LOC18037637, LOC18042802, and LOC18035392, E.C. 1.11.1.11) appears to be predominantly repressed (**Supplementary Figures 4.007, 4.041** and **Supplementary Table 2**). Two glutathione peroxidases (LOC18047364, EC:1.11.1.9, **Supplementary Figure 4.041**, **Supplementary Table 2** and LOC18047405, **Supplementary Table 3**), that were also upregulated, might further contribute to H_2O_2 removal. In addition, the data indicate that several sequential enzymes, such as nicotinamidase (LOC18034819, **Supplementary Table 3**), nicotinate phosphoribosyltransferase 1 (LOC18036859, EC:6.3.4.21), pyrimidine and pyridine-specific 5′-nucleotidase (LOC18050280, EC:3.1.3.-), nicotinamide/nicotinic acid mononucleotide adenylyltransferase (LOC18055145, EC:2.7.7.1), and NAD + kinase (LOC18042070, EC:2.7.1.23), implicated through the salvage pathway in the synthesis of the central coenzyme NAD + /NADH, were upregulated in the papeda (**Supplementary Figure 4.067** and **Supplementary Table 2**).

Gene expression in palatable pummelo/mandarin genetic admixtures versus acidic Sun Chu Sha Kat mandarin

To select mRNAs that were differentially expressed in the pulp of developing fruitlets of acidic and palatable mandarins, the transcriptomes of SCM and three palatable pummelo/mandarin genetic admixtures, coded S1, S2, and S3, were compared through RNA-seq analysis. Pairwise comparisons of each segregant against SCM ($Log_2FC = 0.58$ and $\alpha = 0.05$) were generated and the 357 DEGs showing similar expression patterns in the three palatable mandarins and opposite in SCM were selected following the criteria utilized in previous comparisons (**Supplementary Table 5**).

Alkaloids

Out of the eight DEGs clustered in the Alkaloids category, seven of them belonged to three groups that were previously identified in the comparison between ICH and SCM. Five of these genes (3/2) belonged to the group of tropinone reductase homologs, that as commented above do not appear to possess tropinone reductase activity. LOC18039754 and LOC112096160, the two genes with highest expression in this group showed opposite tendencies. The other two genes, geraniol 8-hydroxylase (LOC18040911), involved in the biosynthesis of terpenoid indole alkaloids and in the biosynthesis of flavonoids, and methanol *O*-anthraniloyltransferase (LOC18050771), that generates methyl anthranilate, were both upregulated in the segregants. The remaining gene was downregulated and is annotated as hyoscyamine 6-dioxygenase-like (LOC112100197), the limiting step in the synthesis of scopolamine in the tropane alkaloid biosynthesis.

Terpenoids

The Terpenoids category (4/2) was characterized by the upregulation of zeta-carotene desaturase (LOC112098137), chloroplastic/chromoplastic-like, that plays a crucial role catalyzing the conversion of zeta-carotene to lycopene in the biosynthesis of carotenoids. The three palatable mandarins also showed upregulation of (–)-germacrene D synthase-like (LOC112100727), suggesting an activation of the synthesis of the germacrin-type sesquiterpenoid, germacrene *D*, a class of volatile organic hydrocarbon with antimicrobial and insecticidal properties. The conversion of these sesquiterpenes to the derived lactones was probably downregulated, since α-copaene synthase-like (LOC112095677), that catalyzes several of these conversions, was repressed. In the acyclic sesquiterpenoid pathway, two additional downregulated genes were (*E*)-β-farnesene synthase-like (LOC112101583), a cyclase catalyzing the production of β-farnesene, and dimethylnonatriene synthase (LOC18049179), a cytochrome P450 82G1 involved in the biosynthesis of homoterpenes such as TMTT. (R)-limonene synthase 1 (LOC112098571), chloroplastic-like, that synthetizes the monoterpene limonene also was downregulated.

Sugar metabolism

In the three palatable mandarins, sugar metabolism (6/1) was characterized by the repression of sucrose synthase 2 (LOC18032959), probable galactinol-sucrose galactosyltransferase 2 (LOC18031328) and stachyose synthase (LOC18050124). The first gene is implicated in sucrose cleavage and the other two in the synthesis of raffinose, stachyose, and verbascose. Other repressed genes were probable trehalose-phosphate phosphatase F (LOC18049897), that produces free trehalose and phosphoenolpyruvate carboxylase kinase 1, that through decarboxylation renders oxalacetate to fuel the citric acid cycle. The only gene upregulated in this category was enolase (LOC18031514), that is responsible of the conversion of 2-phosphoglycerate (2-PG) to phosphoenolpyruvate (PEP).

Transport

This category grouped a set of genes (10/9) showing different patterns of expression. Relevant upregulated genes, for instance, were an aquaporin (LOC18045515), a calcium-transporting ATPase (LOC18042874), a potassium transporter (LOC18043735), and two different sugar transporters ERD6-like (LOC18040371 and LOC18046355), which are vacuolar H + /glucose symporters involved in the export of glucose to cytosol (Klemens et al., 2014). Downregulated genes were a boron transporter (LOC18036378), a nitrate reductase (LOC18041814), a zinc transporter (LOC18050857), and overall, an ATPase 10, plasma membrane-type (LOC18035736), that has been previously associated with citric acid accumulation in lemon juice (Aprile et al., 2011). There were also two members of the sodium/hydrogen exchanger gene family (LOC18040436 and LOC18039966), involved in acidity regulation in oranges (Wang et al., 2021), expressed in opposite directions.

Other categories are discussed in **Supplementary Material**, although two transcripts should be mentioned for their potential relevance. One of these is UDP-glycosyltransferase 74B1 (LOC18044914), that is involved in the biosynthesis of benzyl-glucosinolate and appears to be downregulated in the three segregants. The other gene is 2-methylene-furan-3-one reductase (LOC18050984), that codes for the enone oxidoreductase rendering furaneol, the key flavor compound in strawberries (Raab et al., 2006), and is expressed at higher levels in the palatable mandarins.

Gene expression as related to acidity in palatable pummelo/mandarin genetic admixtures versus Sun Chu Sha Kat mandarin

The finding that ATPase 10, plasma membrane-type (Aprile et al., 2011), a pivotal component of the vacuolar proton-pumping P-ATPase complex that regulates acidity in citrus (Strazzer et al., 2019), was downregulated in the three segregants, prompted us to focus our attention on the expression of the rest of components of this complex and of other related genes previously associated with this process (Huang et al., 2021). This set of genes, listed in **Supplementary Table 6**, included CitPH1 (LOC18037376, magnesium-transporting ATPase, P-type 1), two existing versions of CitPH5; namely, CitPH5.2 (LOC18035739, ATPase 10, plasma membrane-type) and CitPH5.1 (LOC18035736, ATPase 10, plasma membrane-type), CitAN1 (LOC18047507, basic helix-loop-helix protein A, synonymous Noemi), CitPH3 (LOC18038669, WRKY transcription factor 44), CitERF13 (LOC18047942, ethylene-responsive transcription factor 13), CitAN11 (LOC18032473, protein TRANSPARENT TESTA GLABRA 1), CitSO (LOC18039929, protein PIN-LIKES 6), CitVHA-c4 (LOC18041768, V-type proton ATPase 16 kDa proteolipid subunit), CitMAC9F1 (LOC18037289, uncharacterized LOC180372899), and CitPH4 (LOC18053295, transcription factor MYB34) (Shi et al., 2015; Li et al., 2016; Butelli et al., 2019; Shi et al., 2019; Strazzer et al., 2019; Huang et al., 2021, Wang et al., 2021). Although most of these genes showed relatively low TPM (Transcripts per Million) values (**Figure 2**), generally combined with high variability between replicates, raw data show that 8 out of these 11 transcripts, namely, CitPH1, CitPH5.2, CitMAC9F1, CitAN1, CitPH3, CitPH4, CitPH5.1, and CitERF13, were expressed at lower levels in the palatable mandarins. Further qPCR analyses confirmed these tendencies in the samples tested (**Supplementary Table 7**).

Correlation between expression of these genes and acidity, was studied in mature fruit, since the period of acid accumulation in mandarins usually starts about early July and reaches maximum acid levels around the end of September (Cercós et al., 2006). During ripening, fruits of SCM and the three palatable mandarins contained approximately the same sugar quantities (°Br), although SCM fruits were much more acidic, which resulted in lower, unacceptable maturity indices (**Supplementary Figure 7**). Total acidity in the segregants reached palatability levels similar to those of commercial Clementine and W. Murcott, while SCM still contained higher and unpleasant amounts of acids at the end of the ripening period. qPCR data from juice vesicles of samples collected in November showed that transcript levels of CitVHA-c4, CitPH1, CitPH5.2, CitAN1, CitPH3, CitSO, and CitERF13 were downregulated in the three segregants in comparison with SCM, as observed in Clementine and W. Murcott fruits (**Figure 3**). Taken together, these observations suggest that CitPH5.1, CitPH4, CitMAC9F1, CitAN11 appear to play a minor role controlling acidity during ripening.

Differential allele expression in pummelo/mandarin genetic admixtures

We additionally studied the influence of pummelo alleles on the differential gene expression of the three segregants as related to SCM. The data show that the percentage of genes expressed in each haplotype sequence, i.e., MA/MA, PU/MA, and PU/PU, in the three segregants was, on average, 73, 25, and 1%, respectively. However, the frequency of DEGs calculated for each haplotype was higher in PU/PU (0.33), intermediate in PU/MA (0.16) and lower in MA/MA (0.09) sequences (**Supplementary Figure 8**). As related to the differential expression of the pummelo and the mandarin alleles, in non-DEGs, the percentage of expressed alleles was slightly higher for MA (52–58%) than for PU (42–48%), while in the set of DEGs the PU allele was predominantly expressed (68-75%), in detriment of the MA allele (25–32%). There also was a clear-cut difference between both alleles when the expression trend is considered since most MA (8 out of 12) alleles were downregulated, whereas virtually all PU alleles (27 out of 28) were upregulated. In the three segregants, 4 genes located in pummelo introgressed areas (LOC18035377, tropinone reductase homolog At2g29170; LOC18042174, pectinesterase/pectinesterase inhibitor, PPE8B; LOC18045400, auxin response factor 4; LOC112098377, disease resistance RPP8-like protein 3), only exhibited expression (down or up) of the MA alleles (**Figure 4**). In contrast, the MA alleles of other 9 genes (LOC18040371, sugar transporter ERD6-like 18; LOC18042131, UDP-glycosyltransferase 83A1; LOC112096160, tropinone reductase homolog At2g29170-like; LOC18043735, potassium transporter 5; LOC112095422, probable pectinesterase/pectinesterase inhibitor 21; LOC18041614, protein DDB_G0271606; LOC18041176, 3-oxo-Delta(4,5)-steroid 5-beta-reductase; LOC1804023, probable linoleate 9S-lipoxygenase 5; LOC18040524, probable pectinesterase/pectinesterase inhibitor 25) were not expressed at all. These set of genes expressing only PU alleles were all upregulated.

Discussion

The main goal of this work was to identify major domestication traits in citrus, based on the comparisons of gene expression patterns in the pulp of developing fruitlets of inedible and edible citrus types. The citrus examined were wild inedible

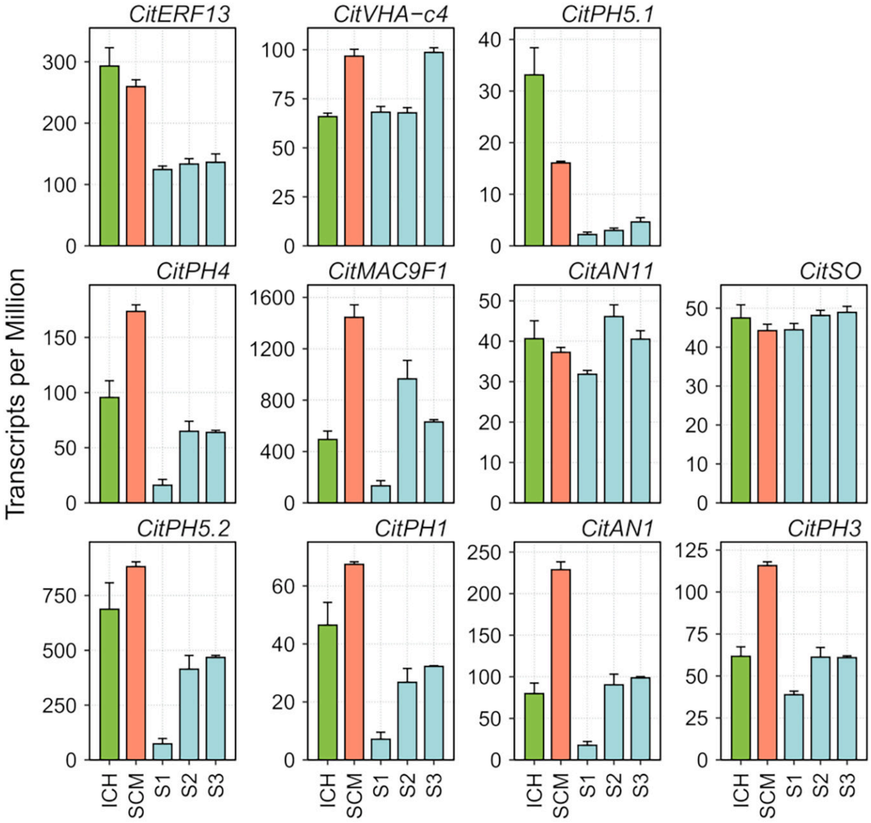


FIGURE 2
RNA abundance of relevant genes reported to be involved in acid regulation of citrus fruits (Aprile et al., 2011; Butelli et al., 2019; Strazzer et al., 2019; Huang et al., 2021), obtained in RNA-seq analyses of the pulp of developing fruitlets of ICH, SCM, and S1, S2, and S3 segregants. Transcripts per Million were computed using DESeq2 read count normalization. Vertical bars represent standard error from three biological replicates.

Ichang papeda (ICH; *C. ichangensis* Swingle), acidic Sun Chu Sha Kat mandarin (SCM; *C. reticulata*, Blanco; *C. erythrose* Tanaka) and three selected palatable genetic admixtures, S1, S2, and S3, derived from a cross between Clementine (CLM; *C. x clementina* Hort. ex Tanaka) and W. Murcott (WMU, *C. x reticulata* Blanco). According to Wang et al. (2017), there were two main mandarin domestication events that generated two mandarin subpopulations differentiated by the degree of acidy. The parentals CLM and WMU were selected for this study because they are representative commercial mandarins of the two adjacent clades of the low acidity subpopulation of mandarins. We carried out the comparison with the three segregants rather than with the parental varieties, to reduce the number of false positives that could be generated comparing two genetically related varieties. The three segregants were selected for the study because their fruits exhibited morphological parameters (**Figure 5**) and organoleptic traits (**Supplementary Figure 7**) in the range shown by the parent varieties. ICH, that grows in a truly wild state, is an endemic citrus thought to be originated in glacial refugia in Wuling Mountains and Ta-pa Mountains in southwestern and middle-west China (Yang et al., 2017). It is currently found in natural populations in these areas, is the most cold-resistant citrus and is also tolerant to both damp and drought conditions (Swingle and Reece, 1967; Yang et al., 2017). ICH is considered one of the most primitive wild forms of citrus, produces inedible fruits with very little flesh and juice, if any, and contains acrid and sour oils that release aroma reminiscent of lemons. Recent developments suggest that ICH split from the main citrus clade around 7 million years ago (Wu et al., 2018). According to Tanaka (1954), SCM is an antique mandarin that was very common in temperate China, occurred in Assam and was also cultivated in Japan. The small SCM fruits, as those of several other traditional mandarins (Wu et al., 2018), are acidic or acidic-sweet, moderately sharp to the taste and very spicy. The ancestor of SCM probably appeared during the last 1.4 million years, after the divergence of the two main subspecies of mainland Asian mandarins (Wu et al., 2021). Under a genomic point of view, ICH and SCM contain pure genomes, i.e., do not show foreign genome introgressions (Wu et al., 2018), while S1, S2, and S3, are genetic admixtures carrying

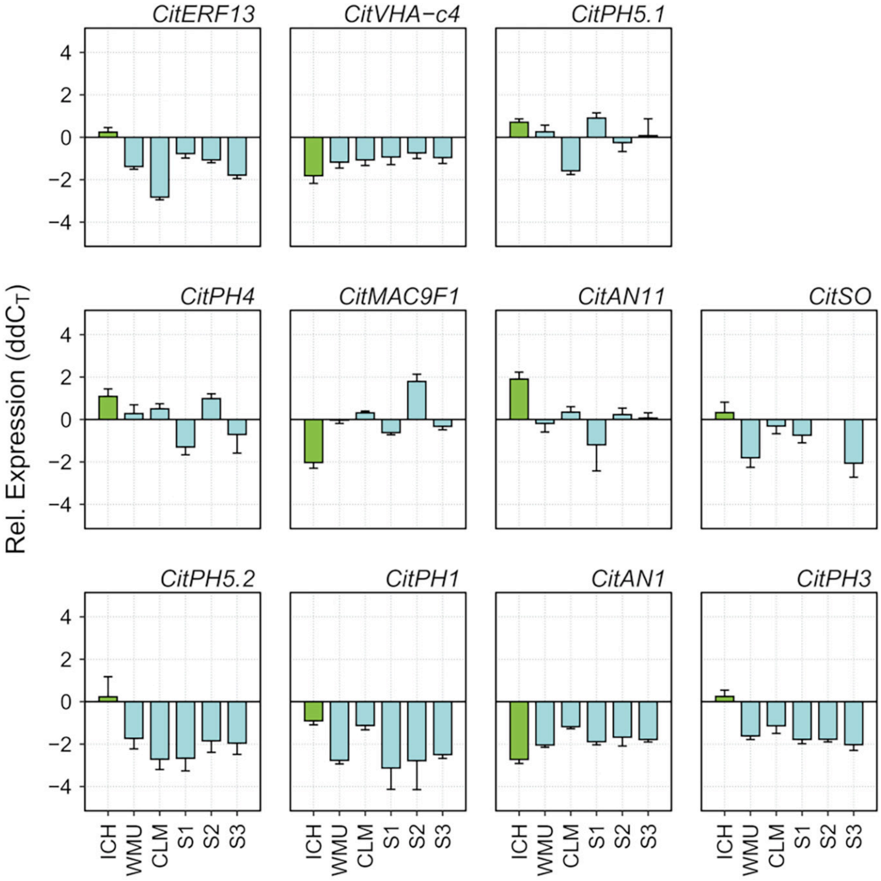


FIGURE 3
Relative expression of genes reported to be involved in acid regulation of citrus fruits (Aprile et al., 2011; Butelli et al., 2019; Strazzer et al., 2019; Huang et al., 2021), determined by RT-qPCR in juice vesicles of ripening fruits (November) of ICH, SCM, WMU, CLM and S1, S2, and S3 segregants. Vertical bars represent standard error from two technical replicates.

pummelo introgressions, in a mandarin genome background (Wu et al., 2014).

Samples used for RNA-seq analyses were collected during the transition between the phases of cell division and cell elongation of citrus fruits (Cercós et al., 2006; Tadeo et al., 2020), a period that appears to be critical for the establishment of major ripening characteristics (Terol et al., 2019). A first comparison between ICH and SCM rendered 7267 DEGs (**Supplementary Table 2**), that were mapped (**Supplementary Figures 4.001–4.117**) against the pathway collection of the KEGG database (Kanehisa et al., 2016). The remaining genes were filtered using more stringent criteria and a final set of 2832 genes were ranked according to the Uniprot database annotation (UniProt Consortium, 2021) in 45 categories (**Supplementary Table 3**). In a second analysis, the expression of the three palatable segregants, S1, S2, and S3, was studied as related to SCM. This analysis identified 357 DEGs that showed similar expression in the three palatable mandarins but opposite expression in SCM (**Supplementary Table 5**). The discussion that follows below highlights the most relevant results derived from those transcriptomic comparisons, while the rest of results are discussed in **Supplementary Material**.

Differential gene expression in Ichang papeda versus Sun Chu Sha Kat mandarin: Upregulation of secondary metabolism

In ICH, expression of pivotal genes controlling secondary metabolites, as illustrated in the categories Alkaloids, Terpenoids, Phenylpropanoids, Flavonoids, Glucosinolates and Cyanogenic glucosides (**Figure 1**), was typically upregulated. Regarding Alkaloids, the data indicate that there are several pathways that were promoted in the papeda. Thus, the synthesis of caffeine, a methylxanthin which acts as a natural defense compound (Nathanson, 1984), appears to be promoted, as is that of morphine (**Supplementary Table 3**) in the isoquinoline

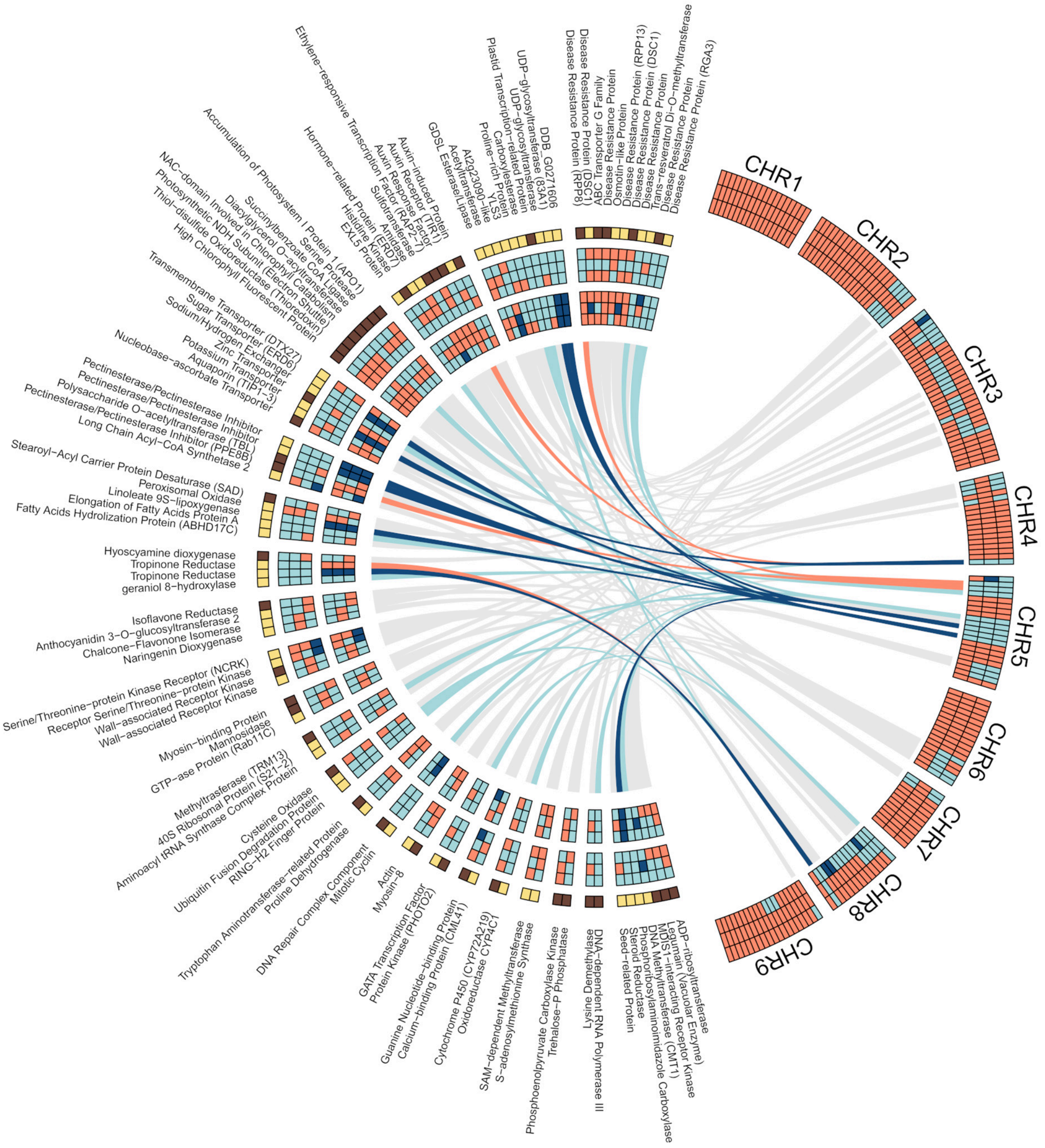


FIGURE 4

Shared DEGs in the pulp of developing fruitlets of segregants, S1, S2, and S3, as related to SCM. Right Circos. Chromosome admixture structure of the three segregants. From top to bottom, S3 (first row), S2 (second row), and S1 (third row); orange: mandarin/mandarin; light blue: pummelo/mandarin; dark blue: pummelo/pummelo. Left Circos. First row (from top to bottom): gene expression; yellow: up; dark brown: down. Second, third, and fourth rows: gene genotype; orange: mandarin/mandarin; light blue: pummelo/mandarin; dark blue: pummelo/pummelo. Fifth, six, and seventh rows: allele expression; orange: mandarin/mandarin; light blue: pummelo/mandarin; dark blue: pummelo/pummelo. Ribbons connect each gene which its physical position in the reference genome. DEGs with the same genotype in the three segregants are connected by colored ribbons; orange: mandarin/mandarin; light blue: pummelo/mandarin; dark blue: pummelo/pummelo. All other DEGs are connected to their respective positions by gray ribbons. DEGs at the outside of the Circos are clustered by categories (**Supplementary Table 5**) and ordered by the number of members of the category, from top to bottom: Plant defense (11); Uncharacterized process (10); Hormonal regulation (9); Chloroplast (8); Transport (7); Cell wall (5); Lipid and fatty acids (5); Alkaloids (4); Flavonoids (4); Receptors, kinases, transduction (4); Trafficking, vesicles (3); Translation (3); Ubiquitination (3); Aminoacid metabolism (2); Cell division (2); Cytoeskeleton (2); Light signaling (2); Membranes, RE, signaling (2); Redox (2); *S*-adenosylmethionine metabolism (2); Sugar metabolism (2); and Transcription (2). A final cluster included those categories with a single member: Seed, embryo development (1); Organ development, differentiation (1); Nucleobases (1); Nucleic acids processing (1); Gametophyte and fertilization (1); Cell death (1); and Abiotic stress (1). Results were plotted in R language and environment (R Core Team, 2021), using packages included in Tidyverse collection (Wickham et al., 2019) and circlize (Gu et al., 2014).

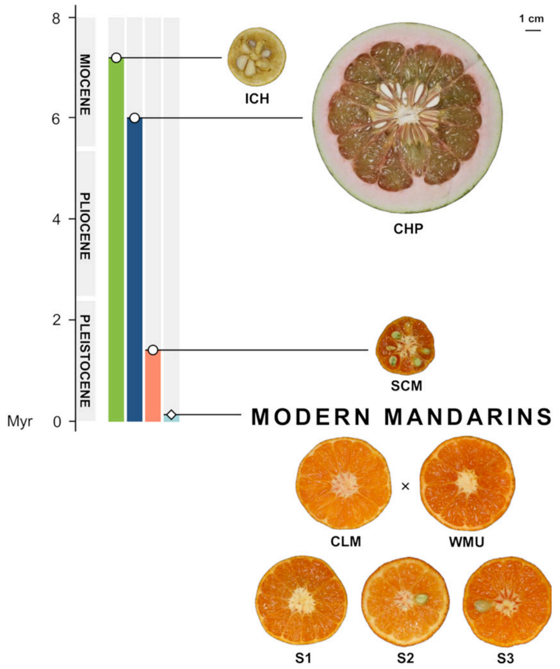


FIGURE 5
Timeline showing the emergence of the citrus mentioned in this study, according to Wu et al. (2018, 2021). ICH, Ichang papeda (*C. ichangensis*); CHP, Chandler pummelo (*C. maxima*); SCM, Sun Chu Sha Kat mandarin (*C. reticulata*); CLM, Clementine mandarin (*C. clementina*); WMU, W. Murcott mandarin (*C. reticulata*), and segregants S1, S2, and S3. The hybrid origin of these three segregants is marked with a cross.

alkaloid pathway. In this route, however, the data also suggest that the syntheses of both dopamine and tyramine and also of serotonin, an indoleamine, are downregulated. Interestingly, it has been recently shown that suppression of serotonin biosynthesis increases resistance to insect pests (Lu et al., 2018). Several biosynthetic genes controlling the conversions that render the iridoid glycoside secologanin, the building unit in the biosynthesis of indole and isoquinoline alkaloids, were upregulated in the monoterpenoid pathway (**Supplementary Table 3**). Iridoid glycosides show a broad defensive spectrum due to their deterrent character on herbivorous and the post-ingestive toxic effects on fungal pathogens (Biere et al., 2004). The upregulation of the synthesis of secologanin and that of other precursors of the indol alkaloids, such as ajmaline, vinblastine and vincristine, may also indicate that the indole alkaloid pathway, was similarly upregulated in the papeda.

This species also showed upregulation of practically all biosynthetic genes of the mevalonate pathway, leading to the isoprenoid intermediates, isopentenyl-PP and dimethylallyl pyrophosphate-PP. The data suggest that two genes encoding geranylgeranyl-PP, the enzyme responsible for the synthesis of the precursors of the main terpenoid groups, were also upregulated (**Supplementary Figure 4.072**). The synthesis of mono-, sesqui-, and homoterpenoids, the crucial groups of terpenoid volatiles operating attracting parasitoids or repelling herbivores, were characterized in the papeda by a relative high activity. Thus, most of the genes involved in the biosynthesis of the monoterpene limonene, a natural insecticide, antifeedant, antifungal and attractant for pollinators (Erasto and Viljoen, 2008), were upregulated, as it was the synthesis of the phytoalexin, solavetivone, a potent antifungal sesquiterpenoid. In this group, transcripts of α-farnesene synthase (**Supplementary Figure 4.078**), generating farnesene, that acts as pheromone, a natural insect repellent, and α-copaene synthase-like, that renders several bicyclic olefins and sesquiterpene hydrocarbons, were in contrast expressed at lower levels. There were also two (3S,6E)-nerolidol synthase 1 genes expressed in opposite directions (**Supplementary Figure 4.078** and **Supplementary Table 2**). This enzyme participates in the synthesis of the homoterpene 4,8-dimethyl-1,3,7-nonatriene, DMNT (Degenhardt and Gershenzon, 2000), an observation linked to the upregulation of trimethyltridecatetraene/dimethylnonatriene synthase, encoding the enzyme (EC: 1.14.14.58), generating 4,8,12-trimethyl-1,3(E),7(E),11-tridecatetraene, TMTT, in the Diterpenoid Pathway (**Supplementary Figure 4.074**). DMNT and TMTT are two irregular acyclic monoterpenes exhibiting pivotal roles attracting parasitoids and predators of herbivores (Tholl et al., 2011). The synthesis of oryzalexin, a diterpenoid phytoalexin, also appears to be upregulated in the papeda (**Supplementary Table 3**, Schmelz et al., 2014). In the triterpenoid group, the synthesis of β-amyrin (**Supplementary Figure 4.078** and **Supplementary Table 3**), a common plant saponin with important antimicrobial, antifungal, and anti-feedant properties (Faizal and Geelen, 2013), was similarly upregulated. Regarding carotenoids, only transcripts coding for enzymes mediating phytoene synthesis were upregulated in the papeda. Mandarins exhibit a wide range of carotenoids (**Figure 5**, Gross, 1987), and accordingly showed upregulation of this pathway (**Supplementary Figure 4.076**). In the phenylpropanoid pathway, main steps in the synthesis of the flavonoid precursors are downregulated, although the synthesis of flavonols, anthocyanidins and anthocyanins were clearly upregulated in the papeda (**Supplementary Figures 4**, **4.082, 4.083**). Citrus fruits contain a wide range of flavonoids (Nogata et al., 2006), although mandarins and most cultivated citrus species do not, because carry defectives alleles of Ruby gene encoding a MYB transcription factor controlling anthocyanin biosynthesis (Butelli et al., 2017; Wu et al., 2021). From data in **Supplementary Figure 4.082**, it is suggested that Ruby (synonymous AN2) may participate in the regulation of the expression regulation of naringenin 3-dioxygenase, flavonol synthase and anthocyanidin synthase, key players in the synthesis of anthocyanidins. Flavonoids show antipathogenic activity and participate in the defense against biotic stresses caused by herbivory and pathogenicity. For instance, many flavonoids including the flavonols, kaempferol, quercetin and myricetin may act as deterrents against insects. Flavonoids also reduce the effects of abiotic stresses, such as UV radiation and heat, and show relevant antioxidant properties (Mierziak et al., 2014). In the pathway of the non-flavonoid polyphenol, trans-resveratrol di-*O*-methyltransferase, the last step in the biosynthesis of the antifungal phytoalexin pterostilbene, is expressed at higher levels in the papeda (**Supplementary Table 3**). The data also suggest that the synthesis of glucosinolates and cyanogenic glucosides, two kinds of phytoanticipins, may be promoted in the papeda. These are constitutive chemicals, whose non-toxic forms and the catalyzing enzymes that release the toxic compounds are stored in different cells (glucosinolates) or subcellular compartments (cyanogenic glucosides) (Yactayo-Chang et al., 2020).

The depletion of defensive chemicals is a process generally linked to domestication (Moreira et al., 2018), and in SCM appears to have played a critical role in the production of tastier and more flavorful citrus, since these compounds are essentially of bitter taste and toxic to arthropods and vertebrates (Matsuura and Fett-Neto, 2015). It is interesting also to mention that while chemical defenses, secondary metabolites that represent a major barrier to herbivorous insects, are restricted in SCM, in the Plant defense category (**Supplementary Material**), populated by all kind of protein-based defenses against microbial pathogens, there are more genes up than downregulated.

Differential gene expression in Ichang papeda versus Sun Chu Sha Kat mandarin: Downregulation of growth

Another important difference in gene expression between both species is the prevalence in the papeda of downregulation of many genes (>60%), involved in practically all processes of growth and development (**Supplementary Table 3**). Categories enriched with downregulated genes included genes with roles in Development, Chloroplast, Hormonal Regulation, Signaling, Gene Expression, and Cellular Growth (**Figure 1**). While it may be reasonable to find expression of genes involved in chloroplast metabolism or even photosynthesis in the developing fruit pulp, where the transition chloroplast/chromoplast maybe still ongoing, the finding that a high number of genes unquestionably involved in processes such as flowering, fertilization, or organ development (leaves, gametophytes, roots, etc.) were also expressed in this fruit tissue, might be unexpected. There are no convincing explanations for this observation, except perhaps that certain transcripts have not been yet degraded, or that in addition to the reported functions, some genes could also be implicated in fruit growth in hitherto unknown roles. Repressed genes controlling papeda development were also associated with the biosynthesis of cytokinins, brassionosteroids, ethylene (**Supplementary Figures 4.077, 4.075, 4.024**), or polyamines (**Supplementary Table 4** and **Supplementary Figure 4.029**) and with the transduction of plant hormones in general (**Supplementary Figure 4.109**). In addition, **Supplementary Table 3** also reports that a majority of DEGs included in the Hormonal Regulation category, were downregulated in ICH, particularly those linked to auxins and gibberellins. The repression is also evident in groups of genes related to cellular growth (**Figure 1**), including the categories of Cell Division (**Supplementary Figures 4.099, 4.102–4.105**), Meiosis, Cytoskeleton and Cell Wall (**Supplementary Table 3**). Consistently, the data also revealed lower levels of gene expression in basic genetic processes regulating growth, such as nucleocytoplasmic transport (**Supplementary Figure 4.094**), mRNA surveillance pathway (**Supplementary Figure 4.095**), the processing of the nucleic acids or the chromatin condensation and transcription (**Supplementary Table 3**). Similarly, low expression was associated with signaling transduction pathways involving genes in Receptors, Kinases, Transduction category or GTPases and second messengers as reported in **Supplementary Table 3**. Biosynthesis of steroids, one of the most important components of the cellular membranes was also strongly repressed in the papeda (**Supplementary Figure 4.012**). Other categories enriched with downregulated genes were Transport and Plant Defense (**Supplementary Table 3**).

Differential gene expression in Ichang papeda versus Sun Chu Sha Kat mandarin: Activation of secondary metabolism versus growth stimulation

The RNA-seq analysis, overall, reveals that upregulation in the papeda was mostly associated with the increase of chemical defenses (**Table 1**), a situation that may imply a penalty in terms of energy and development, as suggested by the downregulation of relevant DEGs involved in a wide variety of growth processes. These conspicuous differences appear to be related to the control of the carbon flux through central pathways of the primary metabolism. Thus, the KEGG data show that gene expression of practically all genes encoding enzymatic activities involved in glycolysis (**Supplementary Figure 4.001**), cytoplasmatic citric acid degradation (**Supplementary Figures 4.017, 4.021, 4.058**), GABA shunt (**Supplementary Table 2**), fatty acid degradation (**Supplementary Figure 4.010**), TCA cycle (**Supplementary Figure 4.002**), and several subunits of the major regulatory complex of the oxidative phosphorylation process (**Supplementary Figure 4.014**) were upregulated in the papeda. As in SCM, in the highly acidic species lemon and citron, several genes involved in the TCA cycle and GABA shunt also displayed reduced expression during ripening

TABLE 1 Expression of genes involved in the biosynthesis of chemical defenses and associated compounds, in the pulp of developing fruitlets of ICH as related to SCM.

Chemical defense	Gene Id	Gene name	Compound	Expression
Alkaloids	LOC18037906	Vinorine synthase	Ajmaline	Up
Alkaloids	LOC112100740	Probable caffeine synthase 4	Caffeine	Up
Alkaloids	LOC18036840	Codeine *O*-demethylase	Morphine	Up
Alkaloids	LOC18037214	8-Hydroxygeraniol dehydrogenase	Vinblastine	Up
Alkaloids	LOC18042369	7-Deoxyloganetin glucosyltransferase	Secologanin	Up
Alkaloids	LOC18043348; LOC18046227	tyrosine/DOPA decarboxylase 5	Serotonin	Down
Alkaloids	LOC18041681	Tryptamine 5-hydroxylase (CYP71P1)	Serotonin	Down
Alkaloids	LOC18043348; LOC18046227	Tyrosine/DOPA decarboxylase 5	Dopamine	Down
Terpenoids	LOC112098571; LOC112098486	(R)-limonene synthase 1, chloroplastic-like	Limonene	Up
Terpenoids	LOC18045727	Beta-amyrin synthase	ß-amyrin	Up
Terpenoids	LOC18049179; LOC18049178	cytochrome P450 82G1	TMTT	Up
Terpenoids	LOC18051782	(3S,6E)-nerolidol synthase 1	DMNT	Up
Terpenoids	LOC18033168	(3S,6E)-nerolidol synthase 1	DMNT	Down
Terpenoids	LOC112095677	α-copaene synthase-like	α-copaene, (*E*)-β-caryophyllene and germacrene D.	Down
Terpenoids	LOC18053589	α-farnesene synthase	Farnesene	Down
Phytoalexins	LOC18031421	Cytochrome P450 76M5	Oryzalexin	Up
Phytoalexins	LOC18052043	Premnaspirodiene oxygenase	Solavetivone	Up
Phytoalexins	LOC18031586; LOC18051743	*Trans*-resveratrol di-*O*-methyltransferase	Pterostilbene	Up
Flavonoids	LOC18037475	Flavonol synthase/flavanone 3-hydroxylase	Kaempferol	Up
Flavonoids	LOC18037475	Flavonol synthase/flavanone 3-hydroxylase	Quercetin	Up
Flavonoids	LOC18037476	Flavonol synthase/flavanone 3-hydroxylase	Myricetin	Up
Cyanogenic glucosides	LOC18036876	Linamarin synthase 1	Linamarin	Up
Glucosinolate metabolism	LOC18049889	Flavin-containing monooxygenase FMO GS-OX-like 4	Methylsulfinylalkyl glucosinolates	Up

(Borredá et al., 2022). Based on these observations, we propose that in the papeda, the generation of energy in the ATP form, is stimulated through the increase of the carbon flux *via* both glycolysis and fatty acid degradation, generating pyruvate and acetyl CoA, respectively, to fuel the Krebs cycle (**Figure 6**). The activation of the TCA cycle increases the production of both succinate, a substrate of complex II of the mitochondrial electron transport chain, and citric acid, that after transport to the cytosol may increase cytoplasmatic acidity to a detrimental level for normal cellular functions. Citric acid may be, then, stored in the vacuole, catalyzed to Acetyl CoA or further metabolized into glutamate entering the GABA shunt, that finally restores the carbon pool that the TAC cycle requires. We proposed in a previous work (Cercós et al., 2006), that in cultivated Clementine the GABA shunt, a powerful proton consuming reaction, is a very efficiently way to reduce both citric acid and cytoplasmatic acidity in ripe fruit flesh, while this current work suggests that this mechanism is active in developing fruitlets of wild citrus. Current consensus agrees that the regulation of acid metabolism in citrus, is basically focused on the generation of citric acid in the TCA, and its storage in the vacuole and later reduction in the cytosol through the GABA and ATP citrate lyase pathways (Cercós et al., 2006; Li et al., 2019; Sadka et al., 2019, Feng et al., 2021). However, our suggestion expands this concept identifying the regulatory TCA as a hub linking catabolism of fatty acids, production of organic acids and activation of oxidative phosphorylation, for the generation of energy as requested by growth and/or environmental demands. In this view, citric acid appears to be the major player in a system, that is balanced modulating its concentration and compartmentalization through processes

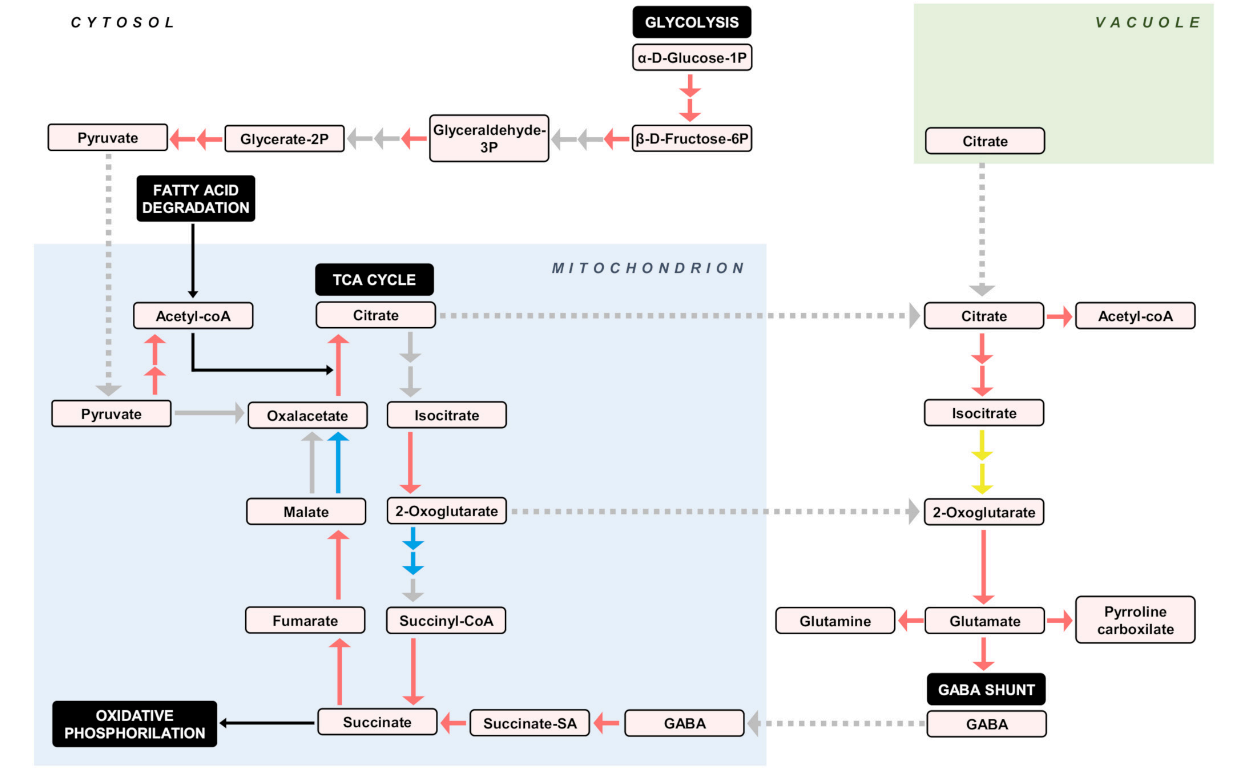


FIGURE 6
Proposed activity of central carbon metabolic pathways, deduced from DEGs in the pulp of developing fruitlets of ICH as related to SCM. Bigger solid arrows indicate gene regulation: red = upregulation, blue = downregulation, yellow = undetermined, and gray = no differential expression. Small solid black arrows represent substrate inputs, while gray dotted arrows indicate directional transport. Reaction products and pivotal pathways are embedded in colored and black boxes, respectively.

that ultimately determine fruit acidity, a pivotal citrus organoleptic trait.

These changes differentially affected nitrogen and carbon allocation in both species. The most important difference in amino acid production was probably related to the degradation of leucine, isolecucine and valine, that was promoted in the papeda (Supplementary Table 4). These amino acids are degraded to Acetyl CoA and succinyl CoA that may thus fuel the TCA (Supplementary Figures 4.002, 4.025). In addition, upregulation in the papeda, of practically all genes controlling fatty acid degradation (Supplementary Figure 4.010), also appears to contribute to provide higher amounts of Acetyl-CoA to fuel the citrate cycle. In addition, the degradation of leucine produces hydroxymethylglutaryl, an intermediate of the mevalonate pathway in the terpenoid pathway, that may then be reinforced (Supplementary Figure 4.072). This pathway also provides precursors for monoterpenoid and isoquinoline alkaloid (Supplementary Figure 4.087), whose synthesis appears to be upregulated in the papeda as suggested above. Synthesis of major intermediates participating in carbon primary metabolism, for instance, was apparently more active in ICH (Supplementary Figure 4.045). In contrast, the number of upregulated DEGs involved in transport of hexoses and monosaccharides, including the vacuolar hexose transporter SWEET17 and NDR1/HIN1-like protein 26, required for correct sugar partitioning between source leaves and sink organs, was higher in SCM mandarin than in ICH (Supplementary Table 3).

The above results, overall, indicate that growth and development is rather restricted in inedible wild ICH, while secondary metabolism and the production of chemical defenses in particular (Table 1), are clearly upregulated. The dichotomy between growth stimulation versus activation of secondary metabolism is a common situation in the plant kingdom, which poses to plants, as sessile organisms, a dilemma that is resolved balancing the cost of investment in chemical defense and the availability of resources for its development. The payment of these costs, which takes place in the form of energy and carbon and nitrogen supplies, implies a proportional reduction in the growth and development of the plant (Mithöfer and Boland, 2012). It has been indicated, for instance, that pathogen and insect tolerance and resistance of domesticated citrus has generally declined compared with wild relatives (Bernet et al., 2005). We have previously shown through genomic analysis that citrus domestication tended to reduce chemical defenses involving cyanogenesis and alkaloids (Gonzalez-Ibeas et al., 2021b), while in the current work we expand this concept and show evidence that practically all major groups of chemical defenses, including alkaloids, terpenoids, glucosinolates, and cyanogenic glycosides are repressed in SCM. Therefore, the results support the suggestion that in our system, the papeda restricts its growth to allocate resources and energy to the production of defensive chemicals to escape herbivory.

Differential gene expression in Ichang papeda versus Sun Chu Sha Kat mandarin: Upregulation of cold tolerance

The RNA-seq analysis, on the other hand, did not provide strong indications or evidence that both species behave differently facing abiotic stresses (**Figure 1**), except for several genes involved in cold and oxidative stresses (**Supplementary Table 3**). The set of DEGs related to cold was mostly characterized by the upregulation of genes implicated in cellular responses, including CORA, a cold and drought-regulated protein that showed the highest expression observed in this analysis (Jha et al., 2021). Key transcription factors governing cold response genes were also differentially expressed between both species. The most striking difference was related to the absence in the papeda of MYBS3 mRNA, a central transcription repressor that suppresses the DREB1/CBF-dependent signaling pathway regulating cold stress responses (Su et al., 2010). Also noticeable was the downregulation of negative regulators of abscisic acid such as MSI4 (Banerjee et al., 2017) and ERD15 (Kariola et al., 2006), components of stress responses, including freezing resistance. Furthermore, JUB1, a gene that modulates cellular H_2O_2 levels (Wu et al., 2012), enhancing tolerance to various abiotic stresses including cold (Fang et al., 2021), was upregulated (**Supplementary Table 3**). It is also worth to highlight that DEGs related to sphingolipid metabolism were repressed, except the conversion of ceramide to phytoceramide-1-phosphate (**Supplementary Figure 4.054**), that appears to be important for the resistance to cold (Dutilleul et al., 2015). Taken together, these observations might be related to the fact that ICH is the hardiest species in the genus Citrus (Swingle and Reece, 1967), tolerating both frost temperatures, even at −20°C, and damp conditions (Yang et al., 2017). It should be notice that these observations were made in samples not subjected to cold conditions, while the natural habitat of the papeda in montane regions of China (Yang et al., 2017) is rather chiller.

Since temperature stress increases the generation of ROS (Hasanuzzaman et al., 2013), this circumstance might be also connected with the enrichment of upregulated DEGs with several roles on antioxidant defense, that were detected in the Redox category (**Supplementary Table 3**), in the salvage pathway of the central coenzyme NAD + /NADH (**Supplementary Figure 4.067**) and in the synthesis of other coenzymes and vitamins, such as pantothenate (vitamin B5) and CoA (**Supplementary Figure 4.068**), riboflavin and flavin mononucleotide (**Supplementary Figure 4.065**), tocopherol and tocotrienol (vitamin E) (**Supplementary Figure 4.013**), and the compounds integrating vitamin B6 (**Supplementary Figure 4.066**).

Differential gene expression in mandarin admixtures versus Sun Chu Sha Kat mandarin: Palatability increment

The comparative RNA-seq analyses between the four mandarin transcriptomes was characterized by the predominant downregulation in the three segregants of genes involved in both Abiotic Stress and Plant Defense categories, mostly participating in central roles and general responses, such as SRG1, regulating plant immunity. Similarly, pivotal genes implicated in the synthesis of relevant terpenoids, alkaloids and glucosinolates, and hence, in chemical defense, i.e., the pheromone β-farnesene, the homoterpene TMTT, the antifeedant limonene, the alkaloid scopolamine, or benzyl-glucosinolate, were downregulated. It is worth to note that in the wild papeda the expression of cytochrome P450 82G1 and (R)-limonene synthase 1, chloroplastic-like, the regulatory genes controlling the synthesis of TMTT and limonene, was relatively high (**Supplementary Tables 2, 3**), while in SCM, these genes were expressed at lower levels and in the three palatable mandarins, their expression was hardly detected. However, there were other defense genes (**Supplementary Material**) upregulated in the three segregants, an effect perhaps related to specific responses to local pathogen attacks and/or to the contribution of pummelo.

The three palatable mandarins showed downregulation of important genes controlling structural components of cell wall such as cellulose, expansins, pectins, and lignans and lignin. Expression of cytochrome P450 84A1 (ferulate 5 hydroxylase) involved in lignin biosynthesis, for example, was high in ICH, lower in SCM (**Supplementary Table 2**) and even lower in the three segregants (**Supplementary Table 5**). These observations might suggest that cell wall stiffening of the juice sacs in the fruit pulp or the peel of the segments is reduced, a characteristic that may be associated to a higher degree of palatability.

Other DGEs that may also affect palatability and taste were methanol *O*-anthraniloyltransferase, and 2-methylene-furan-3-one reductase, that were both upregulated and glutathione S-transferase L3, that was repressed. The first gene is an acyltransferase that catalyzes the formation of methyl anthranilate in the acridone alkaloid pathway, a substance of pleasant aroma, involved in the fragrance of Concord grapes (Wang and De Luca, 2005), that has been used in flavoring foods as mandarin candies or soft drinks (cited in Lee et al., 2019;

Luo et al., 2019). Interestingly, expression of this gene is barely detectable in the inedible papeda, but its expression is relatively high in SCM and even higher in the three segregants. The enone oxidoreductase 2-methylene-furan-3-one reductase, on the other hand, renders furaneol, a key flavor compound in strawberries (Raab et al., 2006), that appear to be present only in fruits. The third gene, glutathione *S*-transferase L3 catalyzes the reduction of *S*-glutathionylquercetin to quercetin, a polyphenol that has a bitter flavor, in agreement with the suggestion that in comparison with wild citrus, cultivars show decreased secondary metabolite levels, such as bitterness compounds (Rao et al., 2021).

In citrus, taste is mainly dependent of the sugar and acid content of the juice. Regarding soluble sugars, the three palatable mandarins showed repression of sucrose synthase 2, probable galactinol-sucrose galactosyltransferase 2 and stachyose synthase, suggesting that sucrose conversion and catabolism in the three segregants is limited during this immature stage favoring sucrose accumulation. Citrus fruitlets appear to operate as main utilization sinks and sucrose synthase expression and activity at these immature fruit stages generally are relatively low (Iglesias et al., 2007; Li et al., 2019). During fruit development, upregulating of sucrose synthases enhances sink strength promoting sucrose and starch accumulation and increasing fruit size (Sadka et al., 2019; Feng et al., 2021). SCM also showed downregulation of two ERD6L sugar transporters that might be relevant for sugar accumulation in citrus fruits, especially sugar transporter ERD6-like 7, as reported in kumquats (Wei et al., 2021).

There were two sodium/hydrogen exchanger genes operating in low affinity electroneutral exchange of protons for cations, expressed in opposite directions in the three segregants. One of these, sodium/hydrogen exchanger 6 is the orthologous of CsNHX, that has been reported to be involved in the regulation of acidity levels in low-acid orange mutants (Wang et al., 2021). In addition, ATPase 10, plasma membrane-type, whose expression has been associated with citric acid accumulation in lemon juice sac cells (Aprile et al., 2011) was upregulated in the acidic mandarin.

Differential gene expression in palatable mandarin admixtures versus Sun Chu Sha Kat mandarin: Downregulation of acidity

Since acidity is one of the fundamental traits determining palatability of citrus fruits and therefore a critical trait for citrus domestication (Wang et al., 2018; Butelli et al., 2019; Rao et al., 2021) we examined expression of pivotal genes previously suggested or proposed to regulate acidity in citrus (**Supplementary Table 6**). The analysis of gene expression in fruit pulp revealed that total acidity in palatable mature fruits was linked to downregulation of 5 genes, CitPH1 (magnesium-transporting ATPase, P-type 1), CitPH5.2 (ATPase 10, plasma membrane-type), CitAN1 (basic helix-loop-helix protein A), CitPH3 (WRKY transcription factor 44) and CitERF13 (ethylene-responsive transcription factor 13), in both developmental stages analyzed, developing and ripening fruits (**Figure 3**) in all three segregants. Of the genes studied, CitAN11 (protein TRANSPARENT TESTA GLABRA 1) was not repressed in any of the two stages, CitSO (protein PIN-LIKES 6) and CitVHA-c4 (V-type proton ATPase 16 kDa proteolipid subunit) were only downregulated in ripening fruits, whereas CitMAC9F1 (uncharacterized LOC180372899, CitPH4 (transcription factor MYB34) and CitPH5.1, in contrast, were downregulated in developing fruitlets.

The current study, while providing data on mandarins, complements the proposal of Strazzer et al. (2019), that shows that CitPH1 and CitPH5, two major downstream genes involved in vacuolar acidification, are highly expressed in ripe fruits of acidic varieties of lemons, oranges, and pummelos. Expression of both CitPH5 (Shi et al., 2015, 2019; Feng et al., 2021) and AN1 (Butelli et al., 2019; Strazzer et al., 2019; Wang et al., 2021), has been associated with citric acid accumulation in a number of studies, a subject revised in Huang et al. (2021).

CitPH1 and CitPH5 appear to act together since in petunia, PH1 may bind to PH5 to promote PH5 proton-pumping activity (Faraco et al., 2014). CitPH1 and CitPH5 expression, in contrast, is strongly reduced in acidless varieties of citrus and this downregulation is associated with mutations that disrupt expression of CitPH4 (MYB), CitAN1 (HLH) and/or CitPH3 (WRKY) transcription factors (Strazzer et al., 2019). These authors also report that CitMAC9F1, a gene of unknown function, is activated by the same transcription factors as CitPH1 and CitPH5 and that CitSO does not contribute to the differences in acidity. Our analyses are in line with these results since in both developing and ripening mandarin fruits CitPH1, CitPH5.2, CitAN1, and CitPH3 were downregulated. In addition, developing fruits showed repression of CitMAC9F1, CitPH4, and CitPH5.1, while ripe fruits exhibited further downregulation of CiSO (**Figures 2**, **3**).

It has also been proposed that CitERF13 regulates citrate accumulation by directly activating the vacuolar proton pump gene CitVHA-c4102 (Li et al., 2016). In mandarins, CitERF13 was effectively downregulated in young and mature fruits while CitVHA-c4, that appears to be normally expressed during ripening (Feng et al., 2021) was repressed only in young fruit. The above observations suggest that although this set of genes, except CitAN11, is very likely involved in the control of acidity in the fruit pulp of mandarins, there may be precise patterns of regulation of gene expression, specifically controlling the accumulation and/or degradation of organic acids at each developmental stage.

In the work by Strazzer et al. (2019), the role CitPH1 and CitPH5, was mostly deduced comparing large differences in acidity between acidic and acidless varieties. They reported that those differences are produced by mutations disrupting the expression of those transcription factors that regulate the two ATPases. Consequently, the authors wonder if small acidity differences between varieties of the same group, may also be due to small differences in the expression of CitPH1/CitPH5. The evidence presented in this current work answers this question showing evidence that smaller differences in non-palatable acidic SCM and edible mandarins are equally correlated with the expression of CitPH1 and CitPH5 and that the range of this change is enough to determine its acceptance, suggesting that this circumstance was a pivotal domestication trait in citrus.

Differential allele expression in palatable mandarin admixtures

The study on differential allelic expression on the pummelo introgressed areas of the three segregants presented in Figure 4 indicates that the contribution of pummelo to acidity did not play a critical role. The analysis detected, however, genetic targets that in principle appear to be contributors to other relevance traits, such as sugar transporters, cell wall modifying pectinesterases or auxin responses. For instance, expression of a major regulator of auxin action, the auxin receptor TIR1 (LOC18052162), in the pulp of developing fruitlets is low in the papeda, relatively high in SCM and higher in the three segregants. Interestingly, differential allelic expression was mostly characterized by the downregulation of MA alleles and the upregulation of the PU ones (**Supplementary Figure 8**). This fact suggests that mandarin domestication in part was certainly based on the selection and substitution of mandarin alleles by pummelo genes.

In conclusion, we propose that during the transition of inedible papedas to sour mandarins, domestication involved a first phase of major changes in the gene regulation of central pathways of the primary and secondary metabolism, characterized by both growth stimulation and reduction of distasteful chemical defenses. It is intriguing, that this reduction appears to affect to all main alkaloids, except dopamine and serotonin, two brain neurotransmitters regulating mood and emotion in humans. We also suggest that in a second phase, several edible attributes of mandarins, especially acidity, were progressively improved through specific changes. Lastly, several observations might indicate that the strong resilience of ICH to frost could be related to the regulation of relevant genes governing cold response.

Data availability statement

The data presented in this study are deposited in the Sequence Read Archive (SRA database) repository, accession number PRJNA853264 (https://www.ncbi.nlm.nih.gov/sra/PRJNA853264).

Author contributions

EP-R collected the samples and phenotype data, performed the data analysis, plot the results, designed the figures and tables, and reviewed and edited the original draft. CB advised the data curation. FT assisted reviewing and editing the original draft. MT conceived and conceptualized the study, conducted and supervised the research, interpreted the results, and wrote and edited the original draft. All authors contributed to the article and approved the submitted version.

Funding

This research was co-funded by the Ministerio de Ciencia, Innovación y Universidades (Spain) through grant #RTI2018-097790-R-100 and by the European Union through the European Regional Development Fund (ERDF) of the Generalitat Valenciana 2014–2020, through grants #PROMETEO/2020/027 and #IVIA/52201A.

Acknowledgments

We thank Matilde Sancho and Antonio Prieto for laboratory tasks and sampling.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.982683/full#supplementary-material

References

Aprile, A., Federici, C., Close, T. J., De Bellis, L., Cattivelli, L., and Roose, M. L. (2011). Expression of the H+-ATPase AHA10 proton pump is associated with citric acid accumulation in lemon juice sac cells. *Funct. Integr. Genomics* 11, 551–563. doi: 10.1007/s10142-011-0226-3

Banerjee, A., Wani, S. H., and Roychoudhury, A. (2017). Epigenetic Control of Plant Cold Responses. *Front. Plant Sci.* 8:1643. doi: 10.3389/fpls.2017.01643

Bernet, G. P., Margaix, C., Jacas, J., Carbonell, E. A., and Asins, M. J. (2005). Genetic analysis of citrus leafminer susceptibility. *Theor. Appl. Genet.* 110, 1393–1400. doi: 10.1007/s00122-005-1943-6

Biere, A., Marak, H. B., and van Damme, J. M. (2004). Plant chemical defense against herbivores and pathogens: Generalized defense or trade-offs? *Oecologia* 140, 430–441. doi: 10.1007/s00442-004-1603-6

Borredá, C., Perez-Roman, E., Talon, M., and Terol, J. (2022). Comparative transcriptomics of wild and commercial Citrus during early ripening reveals how domestication shaped fruit gene expression. *BMC Plant Biol.* 22:123. doi: 10.1186/s12870-022-03509-9

Buck, R., and Flores-Rentería, L. (2022). The Syngameon Enigma. *Plants* 11:895. doi: 10.3390/plants11070895

Butelli, E., Garcia-Lor, A., Licciardello, C., Las Casas, G., Hill, L., Recupero, G. R., et al. (2017). Changes in Anthocyanin Production during Domestication of Citrus. *Plant Physiol.* 173, 2225–2242. doi: 10.1104/pp.16.01701

Butelli, E., Licciardello, C., Ramadugu, C., Durand-Hulak, M., Celant, A., Recupero, G. R., et al. (2019). Noemi controls production of flavonoid pigments and fruit acidity and illustrates the domestication routes of modern citrus varieties. *Curr. Biol.* 29, 158–164. doi: 10.1016/j.cub.2018.11.040

Cercós, M., Soler, G., Iglesias, D. J., Gadea, J., Forment, J., and Talón, M. (2006). Global analysis of gene expression during development and ripening of citrus fruit flesh. A proposed mechanism for citric Acid utilization. *Plant Mol. Biol.* 62, 513–527. doi: 10.1007/s11103-006-9037-7

Chen, J., Li, G., Zhang, H., Yuan, Z., Li, W., Peng, Z., et al. (2021). Primary Bitter Taste of Citrus is Linked to a Functional Allele of the 1,2-Rhamnosyltransferase Gene Originating from Citrus grandis. *J. Agric. Food Chem.* 69, 9869–9882. doi: 10.1021/acs.jafc.1c01211

Degenhardt, J., and Gershenzon, J. (2000). Demonstration and characterization of (E)-nerolidol synthase from maize: A herbivore-inducible terpene synthase participating in (3E)-4,8-dimethyl-1,3,7-nonatriene biosynthesis. *Planta* 210, 815–822. doi: 10.1007/s004250050684

Deng, X., Yang, X., Yamamoto, M., and Biswas, M. K. (2020). "Domestication and history," in *The Genus Citrus*, eds M. Talon, M. Caruso, and F. G. Gmitter (Sawston: Woodhead Publishing), 33–55. doi: 10.1016/B978-0-12-812163-4.00003-6

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dutilleul, C., Chavarria, H., Rézé, N., Sotta, B., Baudouin, E., and Guillas, I. (2015). Evidence for ACD5 ceramide kinase activity involvement in Arabidopsis response to cold stress. *Plant Cell Environ.* 38, 2688–2697. doi: 10.1111/pce.12578

Erasto, P., and Viljoen, A. M. (2008). Limonene-a review: Biosynthetic, ecological and pharmacological relevance. *Nat. Prod. Commun.* 3:1934578X0800300728. doi: 10.1177/1934578X0800300728

Faizal, A., and Geelen, D. (2013). Saponins and their role in biological processes in plants. *Phytochem. Rev.* 12, 877–893. doi: 10.1007/s11101-013-9322-4

Fang, P., Wang, Y., Wang, M., Wang, F., Chi, C., Zhou, Y., et al. (2021). Crosstalk between Brassinosteroid and Redox Signaling Contributes to the Activation of CBF Expression during Cold Responses in Tomato. *Antioxidants* 10:509. doi: 10.3390/antiox10040509

Faraco, M., Spelt, C., Bliek, M., Verweij, W., Hoshino, A., Espen, L., et al. (2014). Hyperacidification of vacuoles by the combined action of two different P-ATPases in the tonoplast determines flower color. *Cell Rep.* 6, 32–43. doi: 10.1016/j.celrep.2013.12.009

Feng, G., Wu, J., Xu, Y., Lu, L., and Yi, H. (2021). High-spatiotemporal-resolution transcriptomes provide insights into fruit development and ripening in Citrus sinensis. *Plant Biotechnol. J.* 19, 1337–1353. doi: 10.1111/pbi.13549

Gonzalez-Ibeas, D., Ibanez, V., Perez-Roman, E., Borredá, C., Terol, J., and Talon, M. (2021a). Shaping the biology of citrus: I. Genomic determinants of evolution. *Plant Genome* 14:e20133. doi: 10.1002/tpg2.20104

Gonzalez-Ibeas, D., Ibanez, V., Perez-Roman, E., Borredá, C., Terol, J., and Talon, M. (2021b). Shaping the biology of citrus: II. Genomic determinants of domestication. *Plant Genome* 14:e20133. doi: 10.1002/tpg2.20133

Gross, J. (1987). "Carotenoids," in *Pigments in Fruits*, ed. B. S. Schweigert (London: Academic Press), 87–186.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812. doi: 10.1093/bioinformatics/btu393

Hamston, T. J., de Vere, N., King, R. A., Pellicer, J., Fay, M. F., Cresswell, J. E., et al. (2018). Apomixis and Hybridization Drives Reticulate Evolution and Phyletic Differentiation in Sorbus L.: Implications for Conservation. *Front. Plant Sci.* 9:1796. doi: 10.3389/fpls.2018.01796

Hasanuzzaman, M., Nahar, K., and Fujita, M. (2013). "Extreme Temperature Responses, Oxidative Stress and Antioxidant Defense in Plants," in *Abiotic Stress - Plant Responses and Applications in Agriculture*, eds K. Vahdati and C. Leslie (London: IntechOpen), doi: 10.5772/54833

Huang, X. Y., Wang, C. K., Zhao, Y. W., Sun, C. H., and Hu, D. G. (2021). Mechanisms and regulation of organic acid accumulation in plant vacuoles. *Hortic. Res.* 8:227. doi: 10.1038/s41438-021-00702-z

Iglesias, D. J., Cercós, M., Colmenero-Flores, J. M., Naranjo, M. A., Ríos, G., Carrera, E., et al. (2007). Physiology of citrus fruiting. *Braz. J. Plant Physiol.* 19, 333–362. doi: 10.1590/S1677-04202007000400006

Jha, R. K., Patel, J., Patel, M. K., Mishra, A., and Jha, B. (2021). Introgression of a novel cold and drought regulatory-protein encoding CORA-like gene, SbCDR, induced osmotic tolerance in transgenic tobacco. *Physiol. Plant* 172, 1170–1188. doi: 10.1111/ppl.13280

Jirschitzka, J., Schmidt, G. W., Reichelt, M., Schneider, B., Gershenzon, J., and D'Auria, J. C. (2012). Plant tropane alkaloid biosynthesis evolved independently in the Solanaceae and Erythroxylaceae. *Proc. Natl. Acad. Sci. U.S.A.* 109, 10304–10309. doi: 10.1073/pnas.1200473109

Kalita, B., Roy, A., Annamalai, A., and Lakshmi, P. T. V. (2021). A molecular perspective on the taxonomy and journey of Citrus domestication. *Perspect. Plant Ecol. Evol. Syst.* 53:125644. doi: 10.1016/j.ppees.2021.125644

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44:D457–D462. doi: 10.1093/nar/gkv1070

Kariola, T., Brader, G., Helenius, E., Li, J., Heino, P., and Palva, E. T. (2006). EARLY RESPONSIVE TO DEHYDRATION 15, a negative regulator of abscisic acid responses in Arabidopsis. *Plant Physiol.* 142, 1559–1573. doi: 10.1104/pp.106.086223

Khoo, H. E., Azlan, A., Tang, S. T., and Lim, S. M. (2017). Anthocyanidins and anthocyanins: Colored pigments as food, pharmaceutical ingredients, and the potential health benefits. *Food Nutr. Res.* 61:1361779. doi: 10.1080/16546628.2017.1361779

Klemens, P. A., Patzke, K., Trentmann, O., Poschet, G., Büttner, M., Schulz, A., et al. (2014). Overexpression of a proton-coupled vacuolar glucose exporter impairs freezing tolerance and seed germination. *New Phytol.* 202, 188–197. doi: 10.1111/nph.12642

Köllner, T. G., Held, M., Lenk, C., Hiltpold, I., Turlings, T. C., Gershenzon, J., et al. (2008). A maize (E)-beta-caryophyllene synthase implicated in indirect defense responses against herbivores is not expressed in most American maize varieties. *Plant Cell* 20, 482–494. doi: 10.1105/tpc.107.051672

Lee, H. L., Kim, S. Y., Kim, E. J., Han, D. Y., Kim, B. G., and Ahn, J. H. (2019). Synthesis of Methylated Anthranilate Derivatives Using Engineered Strains of *Escherichia coli*. *J. Microbiol. Biotechnol.* 29, 839–844. doi: 10.4014/jmb.1904.04022

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint]. 1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, L. J., Tan, W. S., Li, W. J., Zhu, Y. B., Cheng, Y. S., and Ni, H. (2019). Citrus Taste Modification Potentials by Genetic Engineering. *Int. J. Mol. Sci.* 20:6194. doi: 10.3390/ijms20246194

Li, S. J., Yin, X. R., Xie, X. L., Allan, A. C., Ge, H., Shen, S. L., et al. (2016). The Citrus transcription factor, CitERF13, regulates citric acid accumulation *via* a protein-protein interaction with the vacuolar proton pump, CitVHA-c4. *Sci. Rep.* 6:20151. doi: 10.1038/srep20151

Liang, M., Cao, Z., Zhu, A., Liu, Y., Tao, M., Yang, H., et al. (2020). Evolution of self-compatibility by a mutant Sm-RNase in citrus. *Nat. Plants* 6, 131–142. doi: 10.1038/s41477-020-0597-3

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lu, H. P., Luo, T., Fu, H. W., Wang, L., Tan, Y. Y., Huang, J. Z., et al. (2018). Resistance of rice to insect pests mediated by suppression of serotonin biosynthesis. *Nat. Plants* 4, 338–344. doi: 10.1038/s41477-018-0152-7

Luo, W., and Brouwer, C. (2013). Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29, 1830–1831. doi: 10.1093/bioinformatics/btt285

Luo, Z. W., Cho, J. S., and Lee, S. Y. (2019). Microbial production of methyl anthranilate, a grape flavor compound. *Proc. Natl. Acad. Sci. U.S.A.* 116, 10749–10756. doi: 10.1073/pnas.1903875116

Matsuura, H. N., and Fett-Neto, A. G. (2015). "Plant Alkaloids: Main Features, Toxicity, and Mechanisms of Action," in *Plant Toxins. Toxinology*, eds P. Gopalakrishnakone, C. Carlini, and R. Ligabue-Braun (Dordrecht: Springer), doi: 10.1007/978-94-007-6728-7_2-1

Merelo, P., Agustí, J., Arbona, V., Costa, M. L., Estornell, L. H., Gómez-Cadenas, A., et al. (2017). Cell Wall Remodeling in Abscission Zone Cells during Ethylene-Promoted Fruit Abscission in Citrus. *Front. Plant Sci.* 8:126. doi: 10.3389/fpls.2017.00126

Mierziak, J., Kostyn, K., and Kulma, A. (2014). Flavonoids as important molecules of plant interactions with the environment. *Molecules* 19, 16240–16265. doi: 10.3390/molecules191016240

Mithöfer, A., and Boland, W. (2012). Plant defense against herbivores: Chemical aspects. *Annu. Rev. Plant Biol.* 63, 431–450. doi: 10.1146/annurev-arplant-042110-103854

Moreira, X., Abdala-Roberts, L., Gols, R., and Francisco, M. (2018). Plant domestication decreases both constitutive and induced chemical defences by direct selection against defensive traits. *Sci. Rep.* 8:12678. doi: 10.1038/s41598-018-31041-0

Nakano, M., Shimada, T., Endo, T., Fujii, H., Nesumi, H., Kita, M., et al. (2012). Characterization of genomic sequence showing strong association with polyembryony among diverse Citrus species and cultivars, and its synteny with Vitis and Populus. *Plant Sci.* 183, 131–142. doi: 10.1016/j.plantsci.2011.08.002

Nathanson, J. A. (1984). Caffeine and related methylxanthines: Possible naturally occurring pesticides. *Science* 226, 184–187. doi: 10.1126/science.6207592

National Center for Biotechnology Information [NCBI]. (1988). *National Center for Biotechnology Information*. Bethesda: National Center for Biotechnology Information.

Nogata, Y., Sakamoto, K., Shiratsuchi, H., Ishii, T., Yano, M., and Ohta, H. (2006). Flavonoid composition of fruit tissues of citrus species. *Biosci. Biotechnol. Biochem.* 70, 178–192. doi: 10.1271/bbb.70.178

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raab, T., López-Ráez, J. A., Klein, D., Caballero, J. L., Moyano, E., Schwab, W., et al. (2006). FaQR, required for the biosynthesis of the strawberry flavor compound 4-hydroxy-2,5-dimethyl-3(2H)-furanone, encodes an enone oxidoreductase. *Plant Cell* 18, 1023–1037. doi: 10.1105/tpc.105.039784

Rao, M. J., Zuo, H., and Xu, Q. (2021). Genomic insights into citrus domestication and its important agronomic traits. *Plant Commun.* 2:100138. doi: 10.1016/j.xplc.2020.100138

Sadka, A., Shlizerman, L., Kamara, I., and Blumwald, E. (2019). Primary Metabolism in Citrus Fruit as Affected by Its Unique Structure. *Front. Plant Sci.* 10:1167. doi: 10.3389/fpls.2019.01167

Schmelz, E. A., Huffaker, A., Sims, J. W., Christensen, S. A., Lu, X., Okada, K., et al. (2014). Biosynthesis, elicitation and roles of monocot terpenoid phytoalexins. *Plant J.* 79, 659–678. doi: 10.1111/tpj.12436

Shi, C. Y., Hussain, S. B., Yang, H., Bai, Y. X., Khan, M. A., and Liu, Y. Z. (2019). CsPH8, a P-type proton pump gene, plays a key role in the diversity of citric acid accumulation in citrus fruits. *Plant Sci.* 289:110288. doi: 10.1016/j.plantsci.2019.110288

Shi, C. Y., Song, R. Q., Hu, X. M., Liu, X., Jin, L. F., and Liu, Y. Z. (2015). Citrus PH5-like H(+)-ATPase genes: Identification and transcript analysis to investigate their possible relationship with citrate accumulation in fruits. *Front. Plant Sci.* 6:135. doi: 10.3389/fpls.2015.00135

Strazzer, P., Spelt, C. E., Li, S., Bliek, M., Federici, C. T., Roose, M. L., et al. (2019). Hyperacidification of Citrus fruits by a vacuolar proton-pumping P-ATPase complex. *Nat. Commun.* 10:744. doi: 10.1038/s41467-019-08516-3

Su, C. F., Wang, Y. C., Hsieh, T. H., Lu, C. A., Tseng, T. H., and Yu, S. M. (2010). A novel MYBS3-dependent pathway confers cold tolerance in rice. *Plant Physiol.* 153, 145–158. doi: 10.1104/pp.110.153015

Swingle, W. T., and Reece, P. C. (1967). "History, world distribution, botany, and varieties," in *The Citrus Industry, Revised 2nd*, eds W. Reuther, H. J. Webber, and L. D. Batchelor (Berkeley: University of California Press), 190–430.

Tadeo, F., Terol, J., Rodrigo, M. J., Licciardello, C., and Sadka, A. (2020). "Fruit growth and development," in *The Genus Citrus*, eds M. Talon, M. Caruso, and F. G. Gmitter (Sawston: Woodhead Publishing), 245–269. doi: 10.1016/B978-0-12-812163-4.00012-7

Talon, M., Wu, G. A., Gmitter, F. G. Jr., and Rokhsar, D. (2020). "The origin of citrus," in *The Genus Citrus*, eds M. Talon, M. Caruso, and F. G. Gmitter (Sawston: Woodhead Publishing), 9–31. doi: 10.1016/B978-0-12-812163-4.00002-4

Tanaka, T. (1954). *Species Problem in Citrus.* Tokyo: Japanese Society for Promotion of Science.

Terol, J., Nueda, M. J., Ventimilla, D., Tadeo, F., and Talon, M. (2019). Transcriptomic analysis of Citrus clementina mandarin fruits maturation reveals a MADS-box transcription factor that might be involved in the regulation of earliness. *BMC Plant Biol.* 19:47. doi: 10.1186/s12870-019-1651-z

Tholl, D., Sohrabi, R., Huh, J. H., and Lee, S. (2011). The biochemistry of homoterpenes–common constituents of floral and herbivore-induced plant volatile bouquets. *Phytochemistry* 72, 1635–1646. doi: 10.1016/j.phytochem.2011.01.019

UniProt Consortium (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49:D480–D489. doi: 10.1093/nar/gkaa1100

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Wang, J., and De Luca, V. (2005). The biosynthesis and regulation of biosynthesis of Concord grape fruit esters, including 'foxy' methylanthranilate. *Plant J.* 44, 606–619. doi: 10.1111/j.1365-313X.2005.02552.x

Wang, L., He, F., Huang, Y., He, J., Yang, S., Zeng, J., et al. (2018). Genome of wild Mandarin and domestication history of Mandarin. *Mol. Plant* 11, 1024–1037. doi: 10.1016/j.molp.2018.06.001

Wang, L., Huang, Y., Liu, Z., He, J., Jiang, X., He, F., et al. (2021). Somatic variations led to the selection of acidic and acidless orange cultivars. *Nat. Plants* 7, 954–965. doi: 10.1038/s41477-021-00941-x

Wang, X., Xu, Y., Zhang, S., Cao, L., Huang, Y., Cheng, J., et al. (2017). Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nat. Genet.* 49, 765–772. doi: 10.1038/ng.3839

Wei, Q. J., Ma, Q. L., Zhou, G. F., Liu, X., Ma, Z. Z., and Gu, Q. Q. (2021). Identification of genes associated with soluble sugar and organic acid accumulation in 'Huapi' kumquat (Fortunella crassifolia Swingle) *via* transcriptome analysis. *J. Sci. Food Agric.* 101, 4321–4331. doi: 10.1002/jsfa.11072

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Wu, A., Allu, A. D., Garapati, P., Siddiqui, H., Dortay, H., Zanor, M. I., et al. (2012). JUNGBRUNNEN1, a reactive oxygen species-responsive NAC transcription factor, regulates longevity in Arabidopsis. *Plant Cell* 24, 482–506. doi: 10.1105/tpc.111.090894

Wu, G. A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., et al. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* 32, 656–662. doi: 10.1038/nbt.2906

Wu, G. A., Sugimoto, C., Kinjo, H., Azama, C., Mitsube, F., Talon, M., et al. (2021). Diversification of mandarin citrus by hybrid speciation and apomixis. *Nat. Commun.* 12, 1–10. doi: 10.1038/s41467-021-24653-0

Wu, G. A., Terol, J., Ibanez, V., Lopez-Garcia, A., Perez-Roman, E., Borredá, C., et al. (2018). Genomics of the origin, evolution and domestication of citrus. *Nature* 544, 311–316. doi: 10.1038/nature25447

Yactayo-Chang, J. P., Tang, H. V., Mendoza, J., Christensen, S. A., and Block, A. K. (2020). Plant Defense Chemicals against Insect Pests. *Agronomy* 10:1156. doi: 10.3390/agronomy10081156

Yang, X., Li, H., Yu, H., Chai, L., Xu, Q., and Deng, X. (2017). Molecular phylogeography and population evolution analysis of Citrus ichangensis (*Rutaceae*). *Tree Genet. Genomes* 13:29. doi: 10.1007/s11295-017-1113-4

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 17 October 2022
DOI 10.3389/fpls.2022.979742

OPEN ACCESS

EDITED BY
Riccardo Velasco,
Research Centre of Viticulture and Oenology (CREA), Italy

REVIEWED BY
Rosario Muleo,
University of Tuscia, Italy
Kashmir Singh,
Panjab University, India

*CORRESPONDENCE
Laurent G. Deluc
delucl@oregonstate.edu
Satyanarayana Gouthu
gouthus@oregonstate.edu

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 27 June 2022
ACCEPTED 20 September 2022
PUBLISHED 17 October 2022

CITATION
Gouthu S, Mandelli C, Eubanks BA and Deluc LG (2022) Transgene-free genome editing and RNAi ectopic application in fruit trees: Potential and limitations.
Front. Plant Sci. 13:979742.
doi: 10.3389/fpls.2022.979742

Transgene-free genome editing and RNAi ectopic application in fruit trees: Potential and limitations

Satyanarayana Gouthu[1]*, Christian Mandelli[2], Britt A. Eubanks[1] and Laurent G. Deluc[1,2]*

[1]Department of Horticulture, Oregon State University, Corvallis, OR, United States, [2]Oregon Wine Research Institute, Oregon State University, Corvallis, OR, United States

For the past fifteen years, significant research advances in sequencing technology have led to a substantial increase in fruit tree genomic resources and databases with a massive number of OMICS datasets (transcriptomic, proteomics, metabolomics), helping to find associations between gene(s) and performance traits. Meanwhile, new technology tools have emerged for gain- and loss-of-function studies, specifically in gene silencing and developing tractable plant models for genetic transformation. Additionally, innovative and adapted transformation protocols have optimized genetic engineering in most fruit trees. The recent explosion of new gene-editing tools allows for broadening opportunities for functional studies in fruit trees. Yet, the fruit tree research community has not fully embraced these new technologies to provide large-scale genome characterizations as in cereals and other staple food crops. Instead, recent research efforts in the fruit trees appear to focus on two primary translational tools: transgene-free gene editing *via* Ribonucleoprotein (RNP) delivery and the ectopic application of RNA-based products in the field for crop protection. The inherent nature of the propagation system and the long juvenile phase of most fruit trees are significant justifications for the first technology. The second approach might have the public favor regarding sustainability and an eco-friendlier environment for a crop production system that could potentially replace the use of chemicals. Regardless of their potential, both technologies still depend on the foundational knowledge of gene-to-trait relationships generated from basic genetic studies. Therefore, we will discuss the status of gene silencing and DNA-based gene editing techniques for functional studies in fruit trees followed by the potential and limitations of their translational tools (RNP delivery and RNA-based products) in the context of crop production.

KEYWORDS

gene silencing and editing, RNP delivery, RNA-based ectopic application, fruit trees, genetic tools

Introduction

Fruit trees are an essential specialty crop, part of the worldwide food production and economic system, representing at least ~750 M metric tons produced in 2020 (www.fao.org). Conventional breeding has ensured for decades the improvement of consumer-driven traits, including yield, size, nutritional properties, aroma, taste, and the introduction of agronomic characteristics, like tolerance to abiotic and biotic stress. Even with modern molecular approaches, breeding is slow due to the long juvenile phase of most fruit tree species, and the heterozygous nature of the varieties prevents them from maintaining the integrity of their original genetic makeup without several cycles of crosses. Therefore, conventional breeding may not be the most efficient approach to rapidly developing new varieties to meet the challenges of evolving climate, "volatile" consumer preferences, and other changing socio-economic factors such as decreasing labor force and energy costs.

While the recent increase in fruit tree crop genomic resources and database availability is regarded as a significant trigger to improving the understanding of gene function, recent advances in advanced biotechnology tools like RNAi-based gene silencing and gene editing are of paramount importance to accelerating gene function studies beyond the gene-to-traits associations inferred from most "OMICS" technologies. For the past twenty years, significant signs of progress have been made in most fruit trees for reverse and forward genetics programs (Peña et al., 2004; Malnoy et al., 2009; Chaïb et al., 2010; Petri et al., 2018; Savadi et al., 2021). Thanks to more precise and advanced genetic systems, the functional characterization of key genes to essential performance traits in fruit trees is rapidly increasing. Yet, there is still a significant gap in the amount of scientific information generated from fruit trees compared to other major crops that will incite the development of more translational and sustainable technology to respond to immediate needs.

In the first section of this review, we will summarize the most recently advanced tools, RNAi-based gene silencing and gene editing *via* DNA-targeting Cas effectors that could be exploited to advance fundamental knowledge on the gene(s) to trait associations for primary fruit and vine trees (apple, grape, pear, citrus, kiwifruit, and prunus). A few examples from recent literature will showcase the current knowledge of fruit trees. In the following two sections, we will discuss the emerging development of transformative tools that are gaining public and scientific traction: Ribonucleoprotein delivery and ectopic application of RNA in plants. We will cover the recent advances in both technologies, their potential, their limitations, and the major scientific priorities that need to be addressed for these tools to become efficient and transformative in fruit trees for improved crop production.

Current status of gene silencing and DNA-based gene editing tools for fruit trees

Gene silencing involves suppressing gene expression by either repressing its transcription (Transcriptional Gene Silencing or TGS) or influencing the mRNA expression or the protein level, known as Post-Transcriptional Gene silencing (PTGS). Several tools for PTGS and TGS based on hairpin RNAs (hpRNA), trans-acting small Interfering RNAs (tasiRNA), and microRNAs have been developed over the past ten years. Gene silencing based on hairpin remains the most popular, with improved versions like introducing an intron between the RNA arms to enhance the stability of the hpRNA (Wulfert and Krueger, 2018). Viral-Induced Gene Silencing (VIGS), based on a modified virus containing a fragment of a gene to be silenced, leads to the production of double-stranded RNAs (dsRNA) complementary to the target gene. MicroRNA-induced gene silencing (MIGS) introduced the multiplexing approach to target multiple related or unrelated genes (Han et al., 2015). The development of artificial microRNAs (amiRNA)-based gene silencing, based on the expression of custom primary microRNAs, opened new opportunities for a broader targeting (Carbonell et al., 2014). Like MIGS, artificial or synthetic tasiRNAs (atasiRNAs and syn-tasiRNAs) operate through the action of secondary siRNAs that induce selective gene silencing (Figure 1) (Cisneros and Carbonell, 2020). The syn-tasiRNAs construct expressing different syn-tasiRNAs from a single precursor is a potent tool to target multiple viral RNAs. The simultaneous expression of several syn-tasiRNAs against Tomato spotted wilt virus (TSWV), an economically harmful pathogen in tomato crops worldwide, resulted in strong resistance against the virus in all generated transgenic lines (Carbonell, 2019). These last systems were proven effective in monocots and major fruit crops like tomato, but few studies expanded their use to fruit tree (Charrier et al., 2019b). These performant genetic tools, in conjunction with the rapidly increased implementation of machine learning tools, could exponentially increase the identification of efficient siRNA species with greater on-target efficacy and fewer off-target risks in fruit trees models (Wang et al., 2010; Fahlgren et al., 2016; Ahmed et al., 2020). Yet, in recent years, few of these tools have been applied to fruit trees studies, except VIGS, sense-gene-induced post-transcriptional gene silencing (S-PTGS) approaches (Liu et al., 2018; Qi et al., 2019; Singh et al., 2019; Werner Ribeiro et al., 2020) and RNAi-based vector systems that generated a hairpin structure (Pessina et al., 2016; Li et al., 2020; Huang et al., 2021; Wu et al., 2021). To the best of our knowledge, no functional gene studies have explored the advantages of amiRNAs and syn-tasiRNAs studies in fruit trees with the exception of Charrier's study (Charrier et al.,

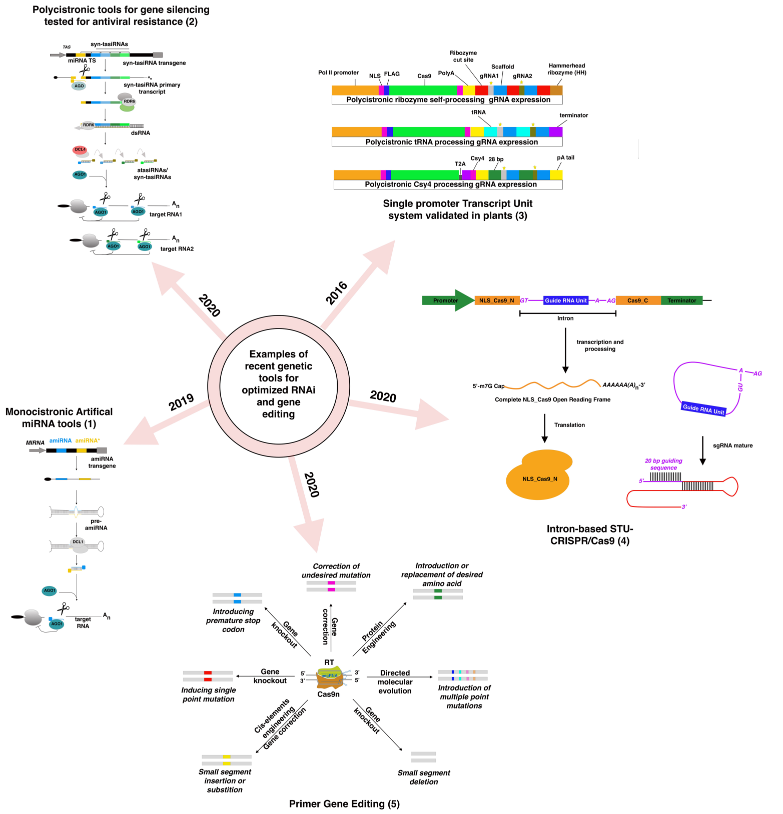

Polycistronic tools for gene silencing tested for antiviral resistance (2)
TAS
syn-tasiRNAs
miRNA TS
syn-tasiRNA transgene
AGO
syn-tasiRNA primary transcript
RDR6
dsRNA
DCL4
atasiRNAs/ syn-tasiRNAs
AGO1
An
target RNA1
target RNA2
Pol II promoter
NLS
FLAG
Cas9
PolyA
Ribozyme cut site
gRNA1
Scaffold
gRNA2
Hammerhead ribozyme (HH)
Polycistronic ribozyme self-processing gRNA expression
tRNA
terminator
Polycistronic tRNA processing gRNA expression
T2A
Csy4
28 bp
pA tail
Polycistronic Csy4 processing gRNA expression
Single promoter Transcript Unit system validated in plants (3)
2020
2016
2019
2020
2020
Examples of recent genetic tools for optimized RNAi and gene editing
Promoter
NLS_Cas9_N
GT
Guide RNA Unit
A
AG
Cas9_C
Terminator
Intron
transcription and processing
5'-m7G Cap
AAAAAA(A)n-3'
Complete NLS_Cas9 Open Reading Frame
Translation
NLS_Cas9_N
GU
sgRNA mature
20 bp guiding sequence
5'
3'
Intron-based STU-CRISPR/Cas9 (4)
Monocistronic Artifical miRNA tools (1)
MIRNA
amiRNA
amiRNA*
amiRNA transgene
pre-amiRNA
DCL1
AGO1
target RNA
An
Correction of undesired mutation
Introduction or replacement of desired amino acid
Introducing premature stop codon
Gene knockout
Gene correction
Protein Engineering
RT
pegRNA
5'
3'
3'
5'
Cas9n
Gene knockout
Inducing single point mutation
Directed molecular evolution
Introduction of multiple point mutations
Cis-elements engineering
Gene correction
Gene knockout
Small segment insertion or substition
Small segment deletion
Primer Gene Editing (5)

FIGURE 1
Examples of recently developed and validated genetic tools in plants for RNAi-silencing and DNA-based gene editing based on figures from previously published figures: (1) (Carbonell, 2019), (2) (Cisneros and Carbonell, 2020), (3) (Lowder et al., 2016), (4) (Zhong et al., 2020), (5) (Hassan et al., 2020).

2019b), while polycistronic amiRNA and syn-tasiRNAs tools have demonstrated their efficiency in creating antiviral resistance (Carbonell et al., 2019; Miao et al., 2021). The lack of robust and tractable genetic systems in fruit trees, in conjunction with the explosion of the gene-editing multifaceted technology, could potentially explain this lack of willingness to adopt more performing and higher throughput RNAi-based technologies for knock-down generation.

The gene-editing technology, as it is, already offers more versatile tools than RNAi-based gene silencing for multiplexed targeting. Implementing sgRNA arrays within the same construct simultaneously targeting up to 12 genes offers more significant opportunities for high throughput screening of mutants (Tang et al., 2016; Stuttmann et al., 2021). If the uncoordinated expression of the two components of the editing system (Cas protein and the guide RNAs) had been a significant drawback at the beginning of the editing era, the progressive adoption of Single Transcript Unit systems has radically improved the editing efficiency and the versatility of the tools regardless the targeted crops (Tang et al., 2016; Tang et al., 2019; Zhong et al., 2020). Identifying new Cas9 and Cas12a with different PAM requirements has expanded the range of DNA recognition for the broader-targeted gene editing (Tang et al., 2017; Wang et al., 2017a; Zhong et al., 2018). Precise genome editing techniques have suffered from conceptual pitfalls for years. Recent technology advances have attempted to address major bottlenecks. The engineering of the genetic cassette enabling for local presence of the donor template near the cut site was found to improve the editing efficiency rate (Ali et al., 2020). Using CRISPR base editors (a modified Cas9 with a cytosine or adenine deaminase domain) is regarded as a promising and exciting alternative to avoid donor templates. Still, the catalog is currently limited to C-to-T and A-to-G base conversion (Komor et al., 2016). To overcome this limitation, a new type of primer Editors was developed, and based on chimeric nCas9 protein fused to an M-MLV reverse transcriptase, a primer-editing guide RNA (pegRNA) designed to mediate site-specific nicking then serves as a template for RT (Anzalone et al., 2019). This system was successfully adopted for monocots and is likely to work for dicots (Butt et al., 2020; Lin et al., 2020; Xu et al., 2020).

To the best of our knowledge, most gene-editing studies were designed to create stable knockout *via Agrobacterium-*

mediated transformation to infer association to major agronomic traits like flower and fruit architecture/ composition, disease resistance, and for improving breeding purposes like in kiwi fruits (Varkonyi-Gasic et al., 2019; Charrier et al., 2019a; Pompili et al., 2020; Wan et al., 2020; Iocco-Corena et al., 2021). All of them used the popular CRISPR-Cas9 system to achieve editing. A special note should be made on gene-editing technology to combat Citrus canker caused by the *Xanthomonas citri* subspecies a significant disease for *citrus* production worldwide (Vojnov et al., 2010). Through several studies (Hu et al., 2014; Jia et al., 2016; Jia et al., 2017), one biallelic mutant of the CsLOB1 promoter region involved in the interaction with the bacterial TALE (transcription activator-like effector) was found to confer to 'Duncan' grapefruits complete immunity to *Xanthomonas* (Jia et al., 2022). In other fruit crops beyond fruit trees, gene editing including in strawberry (Zhou et al., 2018), banana (Kaur et al., 2018), and watermelon (Tian et al., 2018), were all through stable transgenic transformations.

The recent gene-edited crops approved by USDA also used stable expression through *Agrobacterium*-mediated transformation or particle bombardment but were followed by the segregation of transgenes through selfing and crossing (Lacroix and Citovsky, 2020; Gao, 2021) to generate transgene-free plants, which is not possible in clonal fruit crops. Direct modification of crop genomes to introduce economically essential traits without GMO labels has strengthened plant breeding efforts. USDA views the crop varieties developed through genome editing technology as the products of plant breeding as long as no foreign DNA is inserted into the genome ("USDA APHIS | Regulated Article Letters of Inquiry"). This encouraged the development of transgene-free approaches to introduce agronomic traits, and the number of gene-edited crops approved by the USDA jumped from 7 to 70 from 2019 to 2020 (Bomgardner, 2020). Country-wide status of the regulatory, and legislative status towards gene-edited crops has been well discussed in the review (Turnbull et al., 2021). Yet many clonally propagated fruit crops did not benefit from this technology mostly because of the recalcitrance of these crops to have gene-editing reagents delivered into regenerable plant materials without the need for further crossings.

Similarly, technologies based on ectopic RNA application to induce RNAi have also gained significant traction due to the non-GMO nature of the technology (Taning et al., 2021). Currently used for plant protection against fungal pathogens and pests, the technology is an alternative to conventional pesticides for more sustained production systems. RNAi-based products offer multiple advantages compared to their chemical counterparts (Taning et al., 2020). First, the dsRNA active molecules can be designed to target the expression of different genes without changing the sequence-dependent mode of action. Secondly, the availability of increasingly robust *in silico* tools for dsRNA design, in conjunction with the growing access to genomic resources, makes it possible to design species-specific molecules with negligible off-target effects compared with current broad-spectrum pesticides with undesirable side-effects. dsRNA molecules can be rapidly degraded, limiting their long-term environmental persistence. RNAi-based biocontrol delivered by exogenous application of formulated RNA molecules might have the public favor because plants treated with exogenous dsRNA are not considered genetically modified organisms (Shew et al., 2017). Finally, when one compares transgenic approaches, the fast and temporary use of this technology during specific times of the growing season may also offer more leverage, versatility, and reactiveness in the number of applications and the nature of applied materials. There is an increasing number of studies reporting RNA-based product applications in many crops, including fruit trees, but all were performed in a laboratory setting. The recent study conducted by Wise et al., 2022 (see below) is encouraging, but RNAi technology's applicability to field conditions may vary from a given crop production system to another. Multiple limiting factors exist in a field setting, including overcoming the plant's physical barriers for uptake, the effects of environmental factors on the extent of RNA silencing, and achieving systemic silencing to the whole plant. Altogether, this will need to be addressed to maximize the scalability and processibility of the technology in a crop production system.

RNP delivery: Applications and limitations

Besides eliminating transgenes after generating transgenic gene-edited plants through conventional stable transformation methods, direct transgene-free editing can be performed through either transient expression of plasmids or directly using CRISPR elements as ribonucleoproteins. Transient expression of CRISPR elements without integration into the plant genome has been reported in the grapevine using a Geminiviral replicon system (Olivares et al., 2021). But direct gene editing using CRISPR elements as ribonucleoproteins (RNP), without the use of DNA, first demonstrated by (Woo et al., 2015), is the most promising and desirable option to generate transgene-free plants because it avoids DNA insertions and accomplishes gene editing in one generation without unwanted crossings in most clonal crops. However, the success of this technique in fruit crops depends on two main factors: cellular delivery of RNPs and identification of edited material in the absence of selection markers, which are discussed below.

Delivery of CRISPR RNPs is a significant challenge in plants because standard transfection techniques used in animals are typically ineffective in intact plant cells. The genome editing reagents can be delivered into protoplasts without cell walls *via* polyethylene glycol (PEG)-mediated transfection. Therefore, protoplast transfection is commonly used in model organisms and many crops to demonstrate the efficiency of RNP-mediated gene editing. PEG-mediated RNP delivery into protoplasts has been performed in many plant species such as Arabidopsis, rice, lettuce, tobacco (Woo et al., 2015), petunia (Yu et al., 2021), maize (Sant'Ana et al., 2020), wheat (Liang et al., 2017), soybean (Kim et al., 2017), potato (Andersson et al., 2018), cabbage (Murovec et al., 2018), including fruit crops of grapevine, apple (Malnoy et al., 2016), and banana (Wu et al., 2020). These studies reported editing efficiencies in regenerated microcalli, shoots, or plants in the low 11% in petunia to 25% in the potato model. Besides cellular internalization of RNPs and successful editing, the editing efficiency estimates largely depend on protoplast regeneration efficiency, which remains very low in some species or impossible in many fruit crops. Following PEG-mediated protoplast transfection, gene-edited plants were regenerated in some plant models, such as *N. benthamiana* and *B. oleracea* (Woo et al., 2015; Lin et al., 2018; Hsu et al., 2021). In fruit crops, including grapevine, apple, and citrus, these studies were limited to demonstrating RNP-mediated transgene-free gene-editing technique (Malnoy et al., 2016; Zhang et al., 2022b) mainly due to the challenges in regenerating plants from protoplasts. In this regard, the recent establishment of protoplast regeneration protocols in banana, grapevine, guava, and oil palm holds a lot of promise (Reed and Bargmann, 2021). Still, an ideal delivery method for fruit crops would be able to carry CRISPR RNPs and penetrate the cell wall and cell membrane into intact regenerable cells. To overcome the protoplast regeneration, CRISPR RNPs can be transfected *via* PEG transfection into zygotic cells of rice, taking advantage of the immature cell wall during the early zygotic period. The researchers achieved targeted mutations in 14-64% of plants (Toda et al., 2019). However, *in vitro* electro-fusion of isolated gametes is technically challenging for broader application, especially in clonally propagated crop species. On the other hand, transgene-free gene editing has been attempted through biolistic delivery of RNPs in immature wheat embryos and intact tobacco BY2 cells with 3-5% mutagenesis frequencies (Liang et al., 2017; Liu et al., 2020a). Other potential technologies being explored for RNP delivery include nanoparticles and the cell-penetrating peptides (Bilichak et al., 2020).

The cell wall, which makes RNP delivery challenging with current techniques, comprises a complex network of carbohydrates with a negative charge and allows only small molecules through. Studies that estimated the pore size of cell walls found that the cell wall size exclusion limit (SEL) was generally within the 5-20 nm range (Etxeberria et al., 2016; Cunningham et al., 2018). Once in the apoplastic space across the cell wall, the cell membrane has a much larger exclusion limit of 300-500 nm (Wang et al., 2019). So, to pass these two barriers, the RNPs and the carrier should be smaller than the cell wall SEL, and the carrier must have a motif enabling the plasma membrane passage likely *via* endocytosis. While the size of the Cas9 RNP is expected to be 7 to 9 nm, the size of the RNPs with carrier complexes might reach from 25 nm in the case of individually nanocapsuled Cas9 RNPs to 500 nm in case of aggregated nano assemblies that are much larger than the cell wall SEL (Mout et al., 2017a; Mout et al., 2017b; Chen et al., 2019). Biolistic delivery circumvents the cell wall SEL and cell membrane permeability issues through mechanical force. Most plant tissues are amenable to biolistic gene transfer but problems with strong cuticles, lignified cell walls, or hairy surfaces that resist particle penetration can occur. RNP-mediated genome editing using biolistic methods in intact tissues has been demonstrated in rice (Banakar et al., 2020), maize (Svitashev et al., 2016), and wheat (Liang et al., 2017; Liang et al., 2019; Liu et al., 2020b) with editing efficiency usually less than 10%. It was shown that the protein delivery of the Cas9/gRNA RNPs into plant cells had lower off-target cleavage rates when compared with the DNA-based delivery of the Cas9/gRNA complex (Svitashev et al., 2016; Zhang et al., 2016; Liang et al., 2017). The limitation of the biolistic approach to delivering CRISPR RNPs is the need to adapt particle bombardment protocols for each type of target tissue, which necessitates the adjustment of several critical variables such as particle diameter and distance from the target material (Lacroix and Citovsky, 2020) and low transformation frequency due to a small number of cells receiving microprojectiles (Banakar et al., 2019). Despite its low rate of delivery and possible integration of DNA fragments into the genome and genome-scale sequence disruptions (Zhang et al., 2016; Banakar et al., 2019; Liu et al., 2019), CRISPR RNPs delivery through biolistics is still a practical method in fruit crops. Biolistic RNP delivery into intact (Awasthi et al., 2022).

There are several reports of large biomolecules measuring >200 nm delivered across the cell wall in calli and intact plant tissues with the help of nanoparticle carriers and cell-penetrating peptides (CPPs) without forced biolistics (Ng et al., 2016; Guo et al., 2019), which is difficult to explain. Specific nanoparticles interact with the cell wall changing the pore sizes and formation of new pores (Asli and Neumann, 2009; Ma et al., 2010; Palocci et al., 2017) and preincubation with certain pro-endocytotic peptide carriers causing cell wall modifications were also reported (Wang et al., 2021a). Size dynamics of cell wall pores can also vary depending on cell type, degree of development, and physiological stage of the cell. Permeability to nanoparticles might increase in newly synthesized cell walls of actively dividing cells and cultured cells where the wall texture is less dense and less structurally organized (Navarro et al., 2008; Palocci et al., 2017). Assuming interaction occurs between the carrier particles and the biopolymers of the cell wall, the internalization of proteins bigger than the exclusion limit is plausible. Various nanoparticle platforms, including lipid nanoparticles, polymer-based nanoparticles, DNA nanoclews, and gold-based nanoparticles, have successfully delivered CRSPR RNPs across the cell membrane into human cell lines for the genome editing (Duan et al., 2021). However, there is no literature on nanoparticle-mediated RNP delivery for genome editing in walled plant cells, which could be a massive advantage for fruit crops to avoid the need for crossing and protoplast regeneration. Other potential tools for protein delivery into intact plant cells and tissues are cell-penetrating peptides (Ng et al., 2016; Guo et al., 2019; Midorikawa et al., 2019). A recent study by Numata's group screened 55 CPPs, without protein cargo to determine the optimal CPP characteristics for penetration into the intact plant tissues of different species (Numata et al., 2018). The optimal composition of CPPs for the highest penetrating efficiency and nuclear localization differ across the plant species. Still, in general, Lys-containing CPPs seem to be more efficient for plant delivery (Numata et al., 2018) compared to Arg-rich peptides such as Tat peptide favored in animal cells (Taylor and Zahid, 2020; Trofimenko et al., 2021), which could be due to the differences in significant lipid components of their cell membranes. Generally, CPPs, similar to nanocarriers, are believed to internalize by classical endocytic pathways (Xie et al., 2020; Francia et al., 2022), but detailed studies are needed to clarify the differences. In animal systems, simple co-incubation is enough to internalize the proteins delivered through nanocarriers or CPPs. Due to the presence of the cell wall, mild infiltration force such as vacuum or centrifugal force is applied to deliver protein cargoes in plants without loss of regeneration efficiency (Kimura et al., 2019; Watanabe et al., 2021). CPPs have been used in animal cells to deliver genome-editing elements as RNPs non-biolistically, and these short, positively charged peptides were shown to translocate across cell membranes with high delivery efficiencies (Liu et al., 2014; Ramakrishna et al., 2014; Alghuthaymi et al., 2021; Gustafsson et al., 2021). In plants, nanocarriers and CPPs were used to deliver large proteins such as alcohol dehydrogenase (150 KDa) across the cell wall into intact plant tissues (Ng et al., 2016; Wang et al., 2021a). How these proteins attached to CPPs and nanomaterials much larger than cell wall SEL can pass through remains unanswered. Regardless of the mechanism, the ability of CPPs to deliver large proteins across the cell wall highlights their potential for transgene-free genome editing in clonal fruit plants.

Transgene-free gene editing, either through transient expression of plasmid DNAs or RNP delivery, does not afford selection in contrast to stable transformation with selectable marker genes. The reported transgene-free mutation rate in tobacco and wheat through transient expression or RNP biolistic techniques is currently low at 2.5 to 9% using explants and 0.5% cells while using calli (Svitashev et al., 2016; Chen et al., 2018a; Zhang et al., 2019b). When no selection pressure is applied during callus and shoot regeneration following these techniques, most regenerated embryos/shoots should be non-mutant. It should be followed by high throughput screenings such as sequencing, high-resolution melting analysis, etc. Approaches like selectable co-editing followed by Zhang et al., 2019b that confers herbicide tolerance *via* co-editing of acetolactate synthase gene is appealing but may not be applicable for all the crop species.

Ectopic application of siRNA: A real opportunity for lab-to-field transitions in fruit trees?

Several factors like penetration, stability and diffusion of RNA molecules in plants must be considered to evaluate the ectopic application of RNAs as a promising tool in the field, especially for tree crops, due to size and field-practice constraints. dsRNA molecules can be delivered through different methods, including foliar spray, recombinant microbes, nanoparticles, trunk injection, and root soaking, with variable outcomes that rely on the plant model itself (Koch et al., 2016; Wang and Jin, 2017; Liu et al., 2021; Pugsley et al., 2021). To the best of our knowledge, there is currently one study reporting on the systemic effect of trunk injection-mediated delivery of dsRNA in fruit tree crops under field conditions (Wise et al., 2022). The two year-study clearly showed a gradual decline but persistence of dsRNA molecules in the tree canopies over the growing season following the treatment in the spring. The potential of dsRNA delivery has been extensively adopted and reported in many instances with pests and pathogens as a successful method for plant protection. However, with few exceptions showing significant results (Wang et al., 2016) against major pathogens for crop production like *Botrytis cinerea*, most silencing studies through RNA application were validated using transgenic materials that were over-expressing the transgenes (Dalakouras et al., 2016; Schwartz et al., 2020; Hendrix et al., 2021). Foliar application is another promising avenue for tree crops but suffers the same issues as RNP delivery regarding physical barriers to cross. The lipophilic nature of the cuticle hampers the absorption of exogenous hydrophilic and polar molecules like nucleic acids (Schreiber, 2005) that can be penetrated *via* an abrasion, high-pressure spraying, and abaxial stomatal flooding but the applicability of such treatment in a field setting is not realistic (Dalakouras et al., 2016). Beyond the cuticle, the dsRNA molecules' size, length, and shape are essential determining factors in crossing the next physical barrier, the cell wall. It remains a significant hurdle for delivering long RNA molecules and even for a siRNA molecule that does not exceed 26 nucleotide long in many instances. With an averaged pore size exclusion of 6 nm, the size of dsRNA cannot exceed, in theory, more than 16 nm to cross the cell wall (Bennett et al., 2020; Kurczyńska et al., 2021). A recent study in tobacco BY-2 cells suggests a pore SEL for 90 bp long nucleic acids corresponding to 31 nm (Bennett et al., 2020), which seems to suggest that crossing the cell wall is instead a flexible and dynamic process, which can be potentially manipulated to a certain extent and for which the use of nanocarriers may play a critical role in increasing the uptake (Schwartz et al., 2020). The plasma membrane appears to be less problematic to be crossed as several studies have reported the internalization of RNA molecules as long as 500 nm (Wang et al., 2014). Still, it may depend on the shape of the nucleic acids (Zhang et al., 2022a).

The use of nanocarriers to deliver drugs, proteins, and nucleic acids has been widely exploited in the medical sciences (Zhang et al., 2022a). Their use in plant sciences remains highly potential but anecdotal for pest and pathogen control. Nanocarriers serve two purposes for ectopic application of dsRNA: protection and improved uptake (Šečić and Kogel, 2021). Several classes of nanocarriers/nanoparticles (NPs) were shown to protect and extend the integrity of RNA molecules at the leaf surface and inside the cells (Mitter et al., 2017). Numerous carriers have been tested for delivery to various plant models (Wang et al., 2021b; Zhang et al., 2022a). The proper class of NPs used for a given crop primarily depends on the delivery method (spray techniques, root drenching, or trunk injection). Additionally, other criteria need to be specifically evaluated in field conditions because plants may respond differently to a given nanocarrier depending on the plant's architecture and environmental conditions. Some recently tested carriers include but are not limited to carbon nanodots (Schwartz et al., 2020), carbon nanotubes (CN) (Kwak et al., 2019; Demirer et al., 2020), Layered Double Hydroxides of clay (LDH) (Mitter et al., 2017), iron oxide compounds (Cai et al., 2020), multifaceted histidine-based nanocarriers (He et al., 2020), and cationic polymers mimicking Arginine-rich cell-penetrating peptides (CPP) (Parsons et al., 2018). Studies comparing naked RNA versus complexed ones to NPs have ascertained an improved foliar uptake with NPs (Mitter et al., 2017; Schwartz et al., 2020; Delgado-Martín et al., 2022).

Whether this improved uptake is associated with better RNA protection, hence limiting their degradation, or the NPs' physical and chemical properties influencing the RNA's internalization *via* endocytosis mechanisms is debatable (Zhang et al., 2022a). Layered Double Hydroxide molecules were found to increase the uptake of dsRNA molecules by directly maintaining their integrity on leave surfaces (Mitter et al., 2017). Finally, non-engineered and non-metal particles like DNA nanostructures are also capable of delivering exogenous biomolecules like siRNA because of their inherent biocompatibility with plant structural components. The reduced risk of phytotoxicity and traceability compared with conventional NPs renders their use even more attractive for sustainability purposes (Zhang et al., 2019a). Another aspect of the NP's choice is their size. Carbon Dots (CDs), besides the simplistic and advantageous scalability of their synthesis, have an average hydrodynamic diameter of 2.6 nm, which is below the average SEL of the cell wall, even when combined with siRNA (4.7 ± 0.8 nm) (Wang et al., 2014). Interestingly, complexes including dsRNA instead of siRNA showed hydrodynamic diameters in the 160 to 350 nm range with an efficient RNA uptake (Schwartz et al., 2020; Ng et al., 2021; Delgado-Martín et al., 2022). These results again challenge the issue of crossing the cell wall wherein the SEL can be somewhat overcome depending on the employed NP's physical and chemical characteristics.

Most efficient silencing experiments were observed with the ectopic application of 21-22 nucleotide siRNA instead of longer dsRNA regardless of the silencing extent (local or systemic) (Bennett et al., 2020). Thus, the selection of an effective siRNA sequence remains an important prerequisite for potent applications of RNAi-based products in the field as the identification of siRNA with high efficacy *via* genetic engineering remains largely dependent on genetic studies with stable transformations (Mitter et al., 2017; Schwartz et al., 2020; Delgado-Martín et al., 2022) or experiments targeting GFP expression (Dalakouras et al., 2018; Schwartz et al., 2020). The role of 22 nucleotide siRNA species in triggering a systemic RNAi in the whole plant is well documented. It requires the action of the RDR6 polymerase that is likely responsible for the amplification, the transitivity, and the systemic spread of RNAi (McHale et al., 2013; Taochy et al., 2017; Chen et al., 2018b). However, the recruitment of RDR6 can occur with any-sized sRNA that contains an asymmetric bulge in its duplex structure, which leaves open the opportunity to implement a silencing with a siRNA species other than 22 nucleotide long (Manavella et al., 2012; Dalakouras et al., 2016; Dalakouras et al., 2018). Therefore, coupling the increasing genomic resources of fruit trees to developing robust *in silico* tools to predict different classes of siRNAs species with a systemic silencing potential is a significant priority. Algorithms like Support Vector Machines (SVM) can be trained and tested over a sequence dataset, which has already been experimentally validated. Once the best prediction parameters are set up, they can be used to predict siRNAs from long dsRNA sequences. Sequence composition features have historically represented most of the critical features used in the SVM pipeline as di- and tri-nucleotide counts, global and local GC content, duplex flexibility, and thermodynamics stability (Shabalina et al., 2006; Sciabola et al., 2021).

Additionally, target accessibility and the 5' siRNA composition to load into AGO proteins are critical to the silencing characteristics of a siRNA sequence (Gago-Zachert et al., 2019). *In silico* tools that strategize a hybrid-SVM-based prediction approach based on efficacy scores from previously designed models (nonspecific and toxic siRNAs removal, intended versus unintended target transcripts, RISC loading efficacy of the siRNA, target site accessibility, highly specific siRNA) combined with training datasets will generate a list of siRNA to be tested with higher confidence (Ahmed et al., 2020). Through this approach, Ahmed et al. (2020) identified new siRNA candidates targeting highly expressed gene (GFP gene) and endogenous genes in tobacco, Arabidopsis, and periwinkle with relatively high confidence (correlation coefficient greater than 0.7 between the measured predicted efficacy) of the siRNA candidate. Overall, using siRNAs could represent the most efficient way to overcome the plant's physical barriers. Still, it requires significant progress in developing robust predicting tools to identify effective siRNA.

Ultimately, the systemic component of silencing remains essential to assess the technology in a field setting. Optimizing systemic silencing within trees will reduce ectopic application costs to a minimum. The plant's age could also be an important influencing factor. While young plants tend to be more susceptible to pathogens and pests due to less accumulation of waxy cuticles and trichomes, these unprotected portions may be more amenable to absorbing RNA-based products for better protection. When plant maturity is reached, both plant size and mature leaf structure become challenging for efficient absorption of sprayed RNAs, systemic spread, and the resulting silencing. Werner et al., 2020 found that the foliar application of small hairpin dsRNA effectively inhibited *Fusarium graminearum* infection in non-sprayed tissues of barley (*Hordeum vulgare*). By improving the RNA uptake, the NPs also are likely to impact the strength of secondary siRNA production and probably the extent of systemic spread of silencing. Schwartz et al., 2020 demonstrated that carbon nanodots bound to siRNA molecules silence target genes in both locally and newly emerging leaves. Mitter et al. (2017) found a more significant systemic movement of dsRNA-treated with LDH in non-treated parts of the cucumber and tobacco plants. Delgado-Martín et al. (2022) demonstrated greater efficiency of Carbon Dots to induce a systemic spread of dsRNA with even lesser dsRNA molecules applied, which could be advantageous in terms of synthesis cost. Though, very little data have demonstrated systemic silencing of endogenous genes, which brings uncertainty to the translation of the technology, especially to control low expressed endogenous genes (Marcianò et al., 2021; Nerva et al., 2022). There is still a substantial work to determine whether the technology is suitable enough to apply *i)* to the set of plant gene(s) associated with traits and *ii)* to induce a systemic and extended response in the plants, which would limit the production costs not only of the dsRNA but also of the nanocarrier. In conclusion, to achieve a successful Spray Gene Induced Silencing with a foliar application of RNA, several significant milestones need to be completed in conjunction with a solid knowledge of the target crop system to optimize the delivery methods and the potential addition of a nanocarrier (Figure 2).

Discussion and perspectives

The availability of the foundational knowledge related to gene (s)-to-trait associations is a prerequisite for fully translating RNAi and gene editing technologies into sustainable and transformative tools for improved fruit tree production. Like other plant models, this information is mainly generated through genetic engineering studies which still are scarce in fruit trees. As the lag created by the lack of robust and tractable systems for genetic studies tends to abate, the adoption of multiplexing tools through both gene silencing and gene editing for fruit trees studies would help accelerate the acquisition of this foundational information A good example is the development of the microvine model of grapevine, which has been extensively used for reverse genetic and physiological studies (Chaïb et al., 2010; Pellegrino et al., 2019; Torregrosa et al., 2019). Major research programs developed to generate mutant collections that exist in monocots need to expand in fruit trees (Salomé, 2020; Liu et al., 2020a). Loss-of-function studies using RNAi gene silencing remain a popular tool for functional genomic purposes even with the emergence of advanced and versatile gene-editing tools. The recent development of novel RNA-targeting CRISPR/Cas effectors, like Cas13, which could take over RNAi-based silencing tools, is promising. However, collateral cleavage events of non-target RNAs with RNA-targeting Cas effectors are often observed. Its

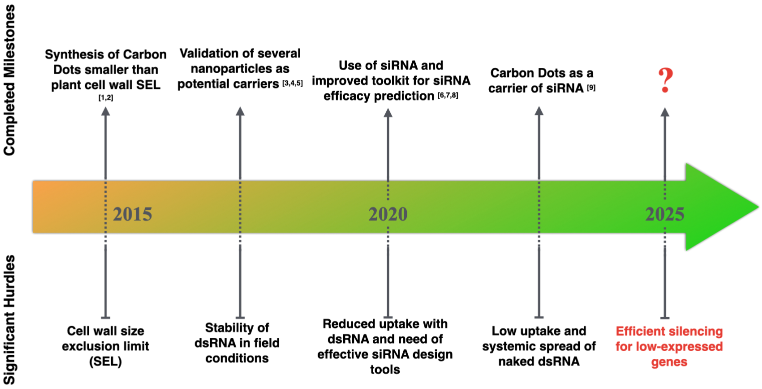

Completed Milestones
Synthesis of Carbon Dots smaller than plant cell wall SEL [1,2]
Validation of several nanoparticles as potential carriers [3,4,5]
Use of siRNA and improved toolkit for siRNA efficacy prediction [6,7,8]
Carbon Dots as a carrier of siRNA [9]
?
2015
2020
2025
Significant Hurdles
Cell wall size exclusion limit (SEL)
Stability of dsRNA in field conditions
Reduced uptake with dsRNA and need of effective siRNA design tools
Low uptake and systemic spread of naked dsRNA
Efficient silencing for low-expressed genes

FIGURE 2
Significant milestones achieved for RNA-based ectopic application. [1, 2]: Wang et al., 2014; Schwartz et al., 2020, [3-5]: Mitter et al., 2017, Zhang et al., 2019a, Demirer et al., 2020, [6-8]: Ahmed et al., 2020, Bennett et al., 2020, Sciabola et al., 2021, [9]: Delgado-Martín et al., 2022.

broad application remains uncertain, leaving room and time for developing more efficient RNAi-based silencing tools (Aman et al., 2018; Ai et al., 2022). Repurposing current gene-editing tools to non-editing applications like transcriptional regulation (CRISPR interference and CRISPR activation) and other aspects of epigenetic regulations is an exciting research avenue for generating mutants. Implementing such tools with new synthetic and inducible promoter systems may also offer alternate routes to prevent the hurdle of lethal mutants and provide more accurate gene-to-trait information in a spatial and temporal context. The NGS technology has enabled the generation of a massive amount of "OMICS" data in fruit trees that can increase the catalog of critical genes and their roles in primary performance traits. If the translational tools discussed in this review reach a certain level of effectiveness along with an effort to communicate the advantages stressing the non-GMO nature of these technologies, they are likely to be better accepted by the public.

Unlike RNA-based ectopic application, RNP delivery still faces major constraints, such as the clonal propagation nature of fruit trees and their recalcitrance to plant regeneration, which renders this approach a challenging alternate avenue. Efforts to develop i) protoplast regeneration protocols and ii) RNP delivery to intact regenerable tissues should be major priorities because they can significantly bolster transgene-free gene editing in fruit tree models. Nanocarriers and CPPs used in animal models must be extensively exploited in plant models for improved RNP delivery through walled regenerable tissues like the isolated examples of biolistic delivery in banana and CPP-mediated delivery in wheat microspores (Bilichak et al., 2020; Awasthi et al., 2022). Even where gene editing is performed through a transient transformation in apple (Chen et al., 2018a) and RNP-delivery in bananas (Awasthi et al., 2022), it is technically challenging to identify mutant cells. Once these milestones are achieved, the selection of edited mutants will remain tedious unless the editing efficiency rate of CRISPR/Cas is dramatically improved, making the regeneration of edited material effortless and rapid.

Pests and pathogens have conserved virulence mechanisms across host species, and the interaction mechanisms with multiple host plants often share commonalities. The knowledge from RNAi studies, such as targeting the fungal effector and RNAi fungal machinery of *Botrytis cinerea* (Wang et al., 2016; Wang et al., 2017b), the tubulin of *Drosophila suzukii* (Taning et al., 2016) and host plant's housekeeping genes in wheat against powdery mildew fungus (Schaefer et al., 2020) should be adopted in fruit trees. Botrytis and powdery mildew are significant pathogens in many fruit trees, such as cherry, apple, and grapevine. The orthologous fungal effectors or host plant genes could be targeted through RNAi application. Though the implementation of this technology remains an issue in fruit tree orchards, a universal delivery method for all the major crops is unlikely to be developed because every crop has its plant architecture with different leaf shapes, which could be problematic for efficient delivery. The current estimates place the price-per-hectare of RNAi-based biopesticide in the range of $20-120, which relies on the cost of dsRNA synthesis and the crop production system. This corresponds to nearly 50% of the average expenses related to purchasing chemical-based products for treating the same area. Then, the broader use of this technology in the field would reduce the production cost. Future improvements in the design and the silencing efficacy of dsRNA molecules, along with the use of new carrier molecules, will favor the uptake and decrease the cost of synthesis of the active ingredients needed. Unlike gene-edited products where yield penalty can be associated with the disruption of the target genes, the RNAi technology applied to the field may offer more opportunities to manipulate traits of interest in a spatial and temporal context with lesser side effects. This technology could strengthen the confidence between producers and consumers if accepted. Overall, due to the increasing need to comply with a set of restrictive but necessary food biosafety rules, the application of biopesticides like RNAi-based products may result in better public acceptance than conventional chemical treatments. Also, by targeting traits related to yield and secondary metabolites, increased production of more nutritional food per square foot could be achieved.

Author contributions

LD and SG contribute to the overall design of the review. SG wrote RNP delivery sections, CM, BE, and LD contributed to the RNAi-based ectopic application, and LD wrote the first section related to the current gene editing and silencing tools. All the authors contribute to the editing of the final draft.

Funding

Part of this work is supported by [Crosscutting Program - Agricultural Innovation through Gene Editing] [grant no. 2021-67013-34555/project accession no. 1026009] from the USDA National Institute of Food and Agriculture and the Northwest Center for Small Fruits Research center (Grant # 000036).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmed, F., Senthil-Kumar, M., Dai, X., Ramu, V. S., Lee, S., Mysore, K. S., et al. (2020). pssRNAit: A web server for designing effective and specific plant siRNAs with genome-wide off-target assessment. *Plant Physiol.* 184, 65–81. doi: 10.1104/pp.20.00293

Ai, Y., Liang, D., and Wilusz, J. E. (2022). CRISPR/Cas13 effectors have differing extents of off-target effects that limit their utility in eukaryotic cells. *Nucleic Acids Res.* 50 (11). doi: 10.1093/nar/gkac159

Alghuthaymi, M. A., Ahmad, A., Khan, Z., Khan, S. H., Ahmed, F. K., Faiz, S., et al. (2021). Exosome/Liposome-like nanoparticles: New carriers for CRISPR genome editing in plants. *Int. J. Mol. Sci.* 22, 7456. doi: 10.3390/ijms22147456

Ali, Z., Shami, A., Sedeek, K., Kamel, R., Alhabsi, A., Tehseen, M., et al. (2020). Fusion of the Cas9 endonuclease and the VirD2 relaxase facilitates homology-directed repair for precise genome engineering in rice. *Commun. Biol.* 3, 44. doi: 10.1038/s42003-020-0768-9

Aman, R., Mahas, A., Butt, H., Ali, Z., Aljedaani, F., and Mahfouz, M. (2018). Engineering RNA virus interference *via* the CRISPR/Cas13 machinery in arabidopsis. *Viruses* 10, 732. doi: 10.3390/v10120732

Andersson, M., Turesson, H., Olsson, N., Fält, A.-S., Ohlsson, P., Gonzalez, M. N., et al. (2018). Genome editing in potato *via* CRISPR-Cas9 ribonucleoprotein delivery. *Physiol. Plant.* 164, 378–384. doi: 10.1111/ppl.12731

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149–157. doi: 10.1038/s41586-019-1711-4

Asli, S., and Neumann, P. M. (2009). Colloidal suspensions of clay or titanium dioxide nanoparticles can inhibit leaf growth and transpiration *via* physical effects on root water transport. *Plant Cell Environ.* 32, 577–584. doi: 10.1111/j.1365-3040.2009.01952.x

Awasthi, P., Khan, S., Lakhani, H., Chaturvedi, S., Shivani,, Kaur, N., et al. (2022). Transgene-free genome editing supports the role of carotenoid cleavage dioxygenase 4 as a negative regulator of β-carotene in banana. *J. Exp. Bot.* 73, 3401–3416. doi: 10.1093/jxb/erac042

Banakar, R., Eggenberger, A. L., Lee, K., Wright, D. A., Murugan, K., Zarecor, S., et al. (2019). High-frequency random DNA insertions upon co-delivery of CRISPR-Cas9 ribonucleoprotein and selectable marker plasmid in rice. *Sci. Rep.* 9, 19902. doi: 10.1038/s41598-019-55681-y

Banakar, R., Schubert, M., Collingwood, M., Vakulskas, C., Eggenberger, A. L., and Wang, K. (2020). Comparison of CRISPR-Cas9/Cas12a ribonucleoprotein complexes for genome editing efficiency in the rice phytoene desaturase (OsPDS) gene. *Rice* 13, 4. doi: 10.1186/s12284-019-0365-z

Bennett, M., Deikman, J., Hendrix, B., and Iandolino, A. (2020). Barriers to efficient foliar uptake of dsRNA and molecular barriers to dsRNA activity in plant cells. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00816

Bilichak, A., Sastry-Dent, L., Sriram, S., Simpson, M., Samuel, P., Webb, S., et al. (2020). Genome editing in wheat microspores and haploid embryos mediated by delivery of ZFN proteins and cell-penetrating peptide complexes. *Plant Biotechnol. J.* 18, 1307–1316. doi: 10.1111/pbi.13296

Bomgardner, M. (2020). Gene-edited crops approved by USDA jumps from 7 to 70 between 2019 and 2020. In: *Genetic literacy project*.

Butt, H., Rao, G. S., Sedeek, K., Aman, R., Kamel, R., and Mahfouz, M. (2020). Engineering herbicide resistance *via* prime editing in rice. *Plant Biotechnol. J.* 18, 2370–2372. doi: 10.1111/pbi.13399

Cai, L., Cai, L., Jia, H., Liu, C., Wang, D., and Sun, X. (2020). Foliar exposure of Fe3O4 nanoparticles on nicotiana benthamiana: Evidence for nanoparticles uptake, plant growth promoter and defense response elicitor against plant virus. *J. Hazard. Mater.* 393, 122415. doi: 10.1016/j.jhazmat.2020.122415

Carbonell, A. (2019). Design and high-throughput generation of artificial small RNA constructs for plants. *Methods Mol. Biol.* 1932, 247–260. doi: 10.1007/978-1-4939-9042-9_19

Carbonell, A., López, C., and Daròs, J.-A. (2019). Fast-forward identification of highly effective artificial small RNAs against different tomato spotted wilt virus isolates. *Mol. Plant Microbe Interact.* 32, 142–156. doi: 10.1094/MPMI-05-18-0117-TA

Carbonell, A., Takeda, A., Fahlgren, N., Johnson, S. C., Cuperus, J. T., and Carrington, J. C. (2014). New generation of artificial MicroRNA and synthetic trans-acting small interfering RNA vectors for efficient gene silencing in arabidopsis. *Plant Physiol.* 165, 15–29. doi: 10.1104/pp.113.234989

Chaïb, J., Torregrosa, L., Mackenzie, D., Corena, P., Bouquet, A., and Thomas, M. R. (2010). The grape microvine - a model system for rapid forward and reverse genetics of grapevines. *Plant J.* 62, 1083–1092. doi: 10.1111/j.1365-313X.2010.04219.x

Charrier, A., Vergne, E., Dousset, N., Richer, A., Petiteau, A., and Chevreau, E. (2019a). Efficient targeted mutagenesis in apple and first time edition of pear using the CRISPR-Cas9 system. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00040

Charrier, A., Vergne, E., Joffrion, C., Richer, A., Dousset, N., and Chevreau, E. (2019b). An artificial miRNA as a new tool to silence and explore gene functions in apple. *Transgenic Res.* 28, 611–626. doi: 10.1007/s11248-019-00170-1

Chen, G., Abdeen, A. A., Wang, Y., Shahi, P. K., Robertson, S., Xie, R., et al. (2019). A biodegradable nanocapsule delivers a Cas9 ribonucleoprotein complex for *in vivo* genome editing. *Nat. Nanotechnol.* 14, 974–980. doi: 10.1038/s41565-019-0539-2

Chen, L., Li, W., Katin-Grazzini, L., Ding, J., Gu, X., Li, Y., et al. (2018a). A method for the production and expedient screening of CRISPR/Cas9-mediated non-transgenic mutant plants. *Hortic. Res.* 5, 13. doi: 10.1038/s41438-018-0023-4

Chen, W., Zhang, X., Fan, Y., Li, B., Ryabov, E., Shi, N., et al. (2018b). A genetic network for systemic RNA silencing in Plants1[OPEN]. *Plant Physiol.* 176, 2700–2719. doi: 10.1104/pp.17.01828

Cisneros, A. E., and Carbonell, A. (2020). Artificial small RNA-based silencing tools for antiviral resistance in plants. *Plants* 9, 669. doi: 10.3390/plants9060669

Cunningham, F. J., Goh, N. S., Demirer, G. S., Matos, J. L., and Landry, M. P. (2018). Nanoparticle-mediated delivery towards advancing plant genetic engineering. *Trends Biotechnol.* 36, 882–897. doi: 10.1016/j.tibtech.2018.03.009

Dalakouras, A., Jarausch, W., Buchholz, G., Bassler, A., Braun, M., Manthey, T., et al. (2018). Delivery of hairpin RNAs and small RNAs into woody and herbaceous plants by trunk injection and petiole absorption. *Frontiers in Plant Science* 9

Dalakouras, A., Wassenegger, M., McMillan, J. N., Cardoza, V., Maegele, I., Dadami, E., et al. (2016). Induction of silencing in plants by high-pressure spraying of *In vitro*-synthesized small RNAs. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01327

Delgado-Martín, J., Delgado-Olidén, A., and Velasco, L. (2022). Carbon dots boost dsRNA delivery in plants and increase local and systemic siRNA production. *Int J Mol Sci* 23 (10), 484602. doi: 10.1101/2022.03.16.484602

Demirer, G. S., Zhang, H., Goh, N. S., Pinals, R. L., Chang, R., and Landry, M. P. (2020). Carbon nanocarriers deliver siRNA to intact plant cells for efficient gene knockdown. *Sci. Adv.* 6, eaaz0495. doi: 10.1126/sciadv.aaz0495

Duan, L., Ouyang, K., Xu, X., Xu, L., Wen, C., Zhou, X., et al. (2021). Nanoparticle delivery of CRISPR/Cas9 for genome editing. *Front. Genet.* 12. doi: 10.3389/fgene.2021.673286

Etxeberria, E., Gonzalez, P., Bhattacharya, P., Sharma, P., and Ke, P. C. (2016). Determining the size exclusion for nanoparticles in citrus leaves. *HortScience* 51, 732–737. doi: 10.21273/HORTSCI.51.6.732

Fahlgren, N., Hill, S. T., Carrington, J. C., and Carbonell, A. (2016). P-SAMS: a web site for plant artificial microRNA and synthetic trans-acting small interfering RNA design. *Bioinformatics* 32, 157–158. doi: 10.1093/bioinformatics/btv534

Francia, V., Reker-Smit, C., and Salvati, A. (2022). Mechanisms of uptake and membrane curvature generation for the internalization of silica nanoparticles by cells. *Nano Lett.* 22, 3118–3124. doi: 10.1021/acs.nanolett.2c00537

Gago-Zachert, S., Schuck, J., Weinholdt, C., Knoblich, M., Pantaleo, V., Grosse, I., et al. (2019). Highly efficacious antiviral protection of plants by small interfering RNAs identified *in vitro*. *Nucleic Acids Res.* 47, 9343–9357. doi: 10.1093/nar/gkz678

Gao, C. (2021). Genome engineering for crop improvement and future agriculture. *Cell* 184, 1621–1635. doi: 10.1016/j.cell.2021.01.005

Guo, B., Itami, J., Oikawa, K., Motoda, Y., Kigawa, T., and Numata, K. (2019). Native protein delivery into rice callus using ionic complexes of protein and cell-penetrating peptides. *PloS One* 14, e0214033. doi: 10.1371/journal.pone.0214033

Gustafsson, O., Rädler, J., Roudi, S., Lehto, T., Hällbrink, M., Lehto, T., et al. (2021). Efficient peptide-mediated *In vitro* delivery of Cas9 RNP. *Pharmaceutics* 13, 878. doi: 10.3390/pharmaceutics13060878

Han, Y., Zhang, B., Qin, X., Li, M., and Guo, Y. (2015). Investigation of a miRNA-induced gene silencing technique in petunia reveals alterations in miR173 precursor processing and the accumulation of secondary siRNAs from endogenous genes. *PloS One* 10, e0144909. doi: 10.1371/journal.pone.0144909

Hassan, M. M., Yuan, G., Chen, J.-G., Tuskan, G. A., and Yang, X. (2020). Prime editing technology and its prospects for future applications in plant biology research. *BioDesign Res.* 2020, 1-14. doi: 10.34133/2020/9350905

Hendrix, B., Zheng, W., Bauer, M. J., Havecker, E. R., Mai, J. T., Hoffer, P. H., et al. (2021). Topically delivered 22 nt siRNAs enhance RNAi silencing of endogenous genes in two species. *Planta* 254, 60. doi: 10.1007/s00425-021-03708-y

He, J., Xu, S., and Mixson, A. J. (2020). The multifaceted histidine-based carriers for nucleic acid delivery: Advances and challenges. *Pharmaceutics* 12, E774. doi: 10.3390/pharmaceutics12080774

Hsu, C.-T., Yuan, Y.-H., Lin, Y.-C., Lin, S., Cheng, Q.-W., Wu, F.-H., et al. (2021). Efficient and economical targeted insertion in plant genomes *via* protoplast regeneration. *CRISPR J.* 4, 752–760. doi: 10.1089/crispr.2021.0045

Huang, D., Wang, Q., Zhang, Z., Jing, G., Ma, M., Ma, F., et al. (2021). Silencing MdGH3-2/12 in apple reduces drought resistance by regulating AM colonization. *Horticulture Res.* 8, 84. doi: 10.1038/s41438-021-00524-z

Hu, Y., Zhang, J., Jia, H., Sosso, D., Li, T., Frommer, W. B., et al. (2014). Lateral organ boundaries 1 is a disease susceptibility gene for citrus bacterial canker disease. *PNAS* 111, E521–E529. doi: 10.1073/pnas.1313271111

Iocco-Corena, P., Chaïb, J., Torregrosa, L., Mackenzie, D., Thomas, M. R., and Smith, H. M. (2021). VviPLATZ1 is a major factor that controls female flower morphology determination in grapevine. *Nat. Commun.* 12, 6995. doi: 10.1038/s41467-021-27259-8

Jia, H., Omar, A. A., Orbović, V., and Wang, N. (2022). Biallelic editing of the LOB1 promoter *via* CRISPR/Cas9 creates canker-resistant 'Duncan' grapefruit. *Phytopathology*® 112, 308–314. doi: 10.1094/PHYTO-04-21-0144-R

Jia, H., Orbovic, V., Jones, J. B., and Wang, N. (2016). Modification of the PthA4 effector binding elements in type I CsLOB1 promoter using Cas9/sgRNA to produce transgenic Duncan grapefruit alleviating XccΔpthA4:dCsLOB1.3 infection. *Plant Biotechnol. J.* 14, 1291–1301. doi: 10.1111/pbi.12495

Jia, H., Zhang, Y., Orbović, V., Xu, J., White, F. F., Jones, J. B., et al. (2017). Genome editing of the disease susceptibility gene CsLOB1 in citrus confers resistance to citrus canker. *Plant Biotechnol. J.* 15, 817–823. doi: 10.1111/pbi.12677

Kaur, N., Alok, A., Shivani,, Kaur, N., Pandey, P., Awasthi, P., et al. (2018). CRISPR/Cas9-mediated efficient editing in phytoene desaturase (PDS) demonstrates precise manipulation in banana cv. rasthali genome. *Funct. Integr. Genomics* 18, 89–99. doi: 10.1007/s10142-017-0577-5

Kim, E., Koo, T., Park, S. W., Kim, D., Kim, K., Cho, H.-Y., et al. (2017). *In vivo* genome editing with a small Cas9 orthologue derived from campylobacter jejuni. *Nat. Commun.* 8, 14500. doi: 10.1038/ncomms14500

Kimura, M., Yoshizumi, T., and Numata, K. (2019). A centrifugation-assisted peptide-mediated gene transfer method for high-throughput analyses. *Plant Biotechnol. (Tokyo)* 36, 49–52. doi: 10.5511/plantbiotechnology.18.1115a

Koch, A., Biedenkopf, D., Furch, A., Weber, L., Rossbach, O., Abdellatef, E., et al. (2016). An RNAi-based control of fusarium graminearum infections through spraying of long dsRNAs involves a plant passage and is controlled by the fungal silencing machinery. *PloS Pathog.* 12, e1005901. doi: 10.1371/journal.ppat.1005901

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424. doi: 10.1038/nature17946

Kurczyńska, E., Godel-Jędrychowska, K., Sala, K., and Milewska-Hendel, A. (2021). Nanoparticles–plant interaction: What we know, where we are? *Appl. Sci.* 11, 5473. doi: 10.3390/app11125473

Kwak, S.-Y., Lew, T. T. S., Sweeney, C. J., Koman, V. B., Wong, M. H., Bohmert-Tatarev, K., et al. (2019). Chloroplast-selective gene delivery and expression in planta using chitosan-complexed single-walled carbon nanotube carriers. *Nat. Nanotechnol.* 14, 447–455. doi: 10.1038/s41565-019-0375-4

Lacroix, B., and Citovsky, V. (2020). "Biolistic approach for transient gene expression studies in plants," in *Biolistic DNA delivery in plants: Methods and protocols*. Eds. S. Rustgi and H. Luo (New York, NY: Springer US), 125–139. Methods in Molecular Biology. doi: 10.1007/978-1-0716-0356-7_6

Liang, Z., Chen, K., and Gao, C. (2019). Biolistic delivery of CRISPR/Cas9 with ribonucleoprotein complex in wheat. *Plant Genome Editing CRISPR Syst.* 1917, 327–335. doi: 10.1007/978-1-4939-8991-1_24

Liang, Z., Chen, K., Li, T., Zhang, Y., Wang, Y., Zhao, Q., et al. (2017). Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. *Nat. Commun.* 8, 14261. doi: 10.1038/ncomms14261

Li, Z.-X., Lan, J.-B., Liu, Y.-Q., Qi, L.-W., and Tang, J.-M. (2020). Investigation of the role of AcTPR2 in kiwifruit and its response to botrytis cinerea infection. *BMC Plant Biol.* 20, 557. doi: 10.1186/s12870-020-02773-x

Lin, C., Hsu, C., Yang, L., Lee, L., Fu, J., Cheng, Q., et al. (2018). Application of protoplast technology to CRISPR/Cas9 mutagenesis: from single-cell mutation detection to mutant plant regeneration. *Plant Biotechnol. J.* 16, 1295–1310. doi: 10.1111/pbi.12870

Lin, Q., Zong, Y., Xue, C., Wang, S., Jin, S., Zhu, Z., et al. (2020). Prime genome editing in rice and wheat. *Nat. Biotechnol.* 38, 582–585. doi: 10.1038/s41587-020-0455-x

Liu, J., Gaj, T., Patterson, J. T., Sirk, S. J., and Iii, C. F. B. (2014). Cell-penetrating peptide-mediated delivery of TALEN proteins *via* bioconjugation for genome engineering. *PloS One* 9, e85755. doi: 10.1371/journal.pone.0085755

Liu, S., Geng, S., Li, A., Mao, Y., and Mao, L. (2021). RNAi technology for plant protection and its application in wheat. *aBIOTECH* 2, 365–374. doi: 10.1007/s42994-021-00036-3

Liu, H.-J., Jian, L., Xu, J., Zhang, Q., Zhang, M., Jin, M., et al. (2020a). High-throughput CRISPR/Cas9 mutagenesis streamlines trait gene identification in Maize[OPEN]. *Plant Cell* 32, 1397–1413. doi: 10.1105/tpc.19.00934

Liu, J., Nannas, N. J., Fu, F., Shi, J., Aspinwall, B., Parrott, W. A., et al. (2019). Genome-scale sequence disruption following biolistic transformation in rice and maize. *Plant Cell* 31, 368–383. doi: 10.1105/tpc.18.00613

Liu, H., Qian, M., Song, C., Li, J., Zhao, C., Li, G., et al. (2018). Down-regulation of PpBGAL10 and PpBGAL16 delays fruit softening in peach by reducing polygalacturonase and pectin methylesterase activity. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01015

Liu, W., Rudis, M. R., Cheplick, M. H., Millwood, R. J., Yang, J.-P., Ondzighi-Assoume, C. A., et al. (2020b). Lipofection-mediated genome editing using DNA-free delivery of the Cas9/gRNA ribonucleoprotein into plant cells. *Plant Cell Rep.* 39, 245–257. doi: 10.1007/s00299-019-02488-w

Lowder, L., Malzahn, A., and Qi, Y. (2016). Rapid evolution of manifold CRISPR systems for plant genome editing. *Front Plant Sci*. 7. doi: 10.3389/fpls.2016.01683

Ma, X., Geiser-Lee, J., Deng, Y., and Kolmakov, A. (2010). Interactions between engineered nanoparticles (ENPs) and plants: Phytotoxicity, uptake and accumulation. *Sci. Total Environ.* 408, 3053–3061. doi: 10.1016/j.scitotenv.2010.03.031

Malnoy, M. A., Korban, S., Boresjza-Wysocka, E., and Aldwinckle, H. C. (2009). "Apple," in *Compendium of transgenic crop plants (John Wiley & Sons, Ltd)* (Oxford UK: Blackwell Publishing), 1–52. doi: 10.1002/9781405181099.k0401

Malnoy, M., Viola, R., Jung, M.-H., Koo, O.-J., Kim, S., Kim, J.-S., et al. (2016). DNA-Free genetically edited grapevine and apple protoplast using CRISPR/Cas9 ribonucleoproteins. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01904

Manavella, P. A., Koenig, D., and Weigel, D. (2012). Plant secondary siRNA production determined by microRNA-duplex structure. *PNAS* 109, 2461–2466. doi: 10.1073/pnas.1200169109

Marcianò, D., Ricciardi, V., Marone Fassolo, E., Passera, A., Bianco, P. A., Failla, O., et al. (2021). RNAi of a putative grapevine susceptibility gene as a possible downy mildew control strategy. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.667319

McHale, M., Eamens, A. L., Finnegan, E. J., and Waterhouse, P. M. (2013). A 22-nt artificial microRNA mediates widespread RNA silencing in arabidopsis. *Plant J.* 76, 519–529. doi: 10.1111/tpj.12306

Miao, S., Liang, C., Li, J., Baker, B., and Luo, L. (2021). Polycistronic artificial microRNA-mediated resistance to cucumber green mottle mosaic virus in cucumber. *Int. J. Mol. Sci.* 22, 12237. doi: 10.3390/ijms222212237

Midorikawa, K., Kodama, Y., and Numata, K. (2019). Vacuum/Compression infiltration-mediated permeation pathway of a peptide-pDNA complex as a non-viral carrier for gene delivery in planta. *Sci. Rep.* 9, 271. doi: 10.1038/s41598-018-36466-1

Mitter, N., Worrall, E. A., Robinson, K. E., Li, P., Jain, R. G., Taochy, C., et al. (2017). Clay nanosheets for topical delivery of RNAi for sustained protection against plant viruses. *Nat. Plants* 3, 16207. doi: 10.1038/nplants.2016.207

Mout, R., Ray, M., Lee, Y.-W., Scaletti, F., and Rotello, V. M. (2017a). *In vivo* delivery of CRISPR/Cas9 for therapeutic gene editing: Progress and challenges. *Bioconjugate Chem.* 28, 880–884. doi: 10.1021/acs.bioconjchem.7b00057

Mout, R., Ray, M., Yesilbag Tonga, G., Lee, Y.-W., Tay, T., Sasaki, K., et al. (2017b). Direct cytosolic delivery of CRISPR/Cas9-ribonucleoprotein for efficient gene editing. *ACS Nano* 11, 2452–2458. doi: 10.1021/acsnano.6b07600

Murovec, J., Guček, K., Bohanec, B., Avbelj, M., and Jerala, R. (2018). DNA-Free genome editing of brassica oleracea and b. rapa protoplasts using CRISPR-Cas9 ribonucleoprotein complexes. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01594

Navarro, E., Baun, A., Behra, R., Hartmann, N. B., Filser, J., Miao, A.-J., et al. (2008). Environmental behavior and ecotoxicity of engineered nanoparticles to algae, plants, and fungi. *Ecotoxicology* 17, 372–386. doi: 10.1007/s10646-008-0214-0

Nerva, L., Guaschino, M., Pagliarani, C., De Rosso, M., Lovisolo, C., and Chitarra, W. (2022). Spray-induced gene silencing targeting a glutathione s-transferase gene improves resilience to drought in grapevine. *Plant Cell Environ.* 45, 347–361. doi: 10.1111/pce.14228

Ng, H. K. M., Lim, G. K., and Leo, C. P. (2021). Comparison between hydrothermal and microwave-assisted synthesis of carbon dots from biowaste and chemical for heavy metal detection: A review. *Microchem. J.* 165, 106116. doi: 10.1016/j.microc.2021.106116

Ng, K. K., Motoda, Y., Watanabe, S., Sofiman Othman, A., Kigawa, T., Kodama, Y., et al. (2016). Intracellular delivery of proteins *via* fusion peptides in intact plants. *PloS One* 11, e0154081. doi: 10.1371/journal.pone.0154081

Numata, K., Horii, Y., Oikawa, K., Miyagi, Y., Demura, T., and Ohtani, M. (2018). Library screening of cell-penetrating peptide for BY-2 cells, leaves of arabidopsis, tobacco, tomato, poplar, and rice callus. *Sci. Rep.* 8, 10966. doi: 10.1038/s41598-018-29298-6

Olivares, F., Loyola, R., Olmedo, B., de los Angele Miccono, M., Aguirre, C., Vergara, R., et al. (2021). CRISPR/CAs9 targeted Editing of Genes Associated with Fungal Susceptibility in Vitis vinifera L., cv. Thomspson Seedless Using Geminivirus-Derived Replicons. *Front Plant Sci* 12. doi 10.3389/fpls.2021.791030

Palocci, C., Valletta, A., Chronopoulou, L., Donati, L., Bramosanti, M., Brasili, E., et al. (2017). Endocytic pathways involved in PLGA nanoparticle uptake by grapevine cells and role of cell wall and membrane in size selection. *Plant Cell Rep.* 36, 1917–1928. doi: 10.1007/s00299-017-2206-0

Parsons, K. H., Mondal, M. H., McCormick, C. L., and Flynt, A. S. (2018). Guanidinium-functionalized interpolyelectrolyte complexes enabling RNAi in resistant insect pests. *Biomacromolecules* 19, 1111–1117. doi: 10.1021/acs.biomac.7b01717

Pellegrino, A., Romieu, C., Rienth, M., and Torregrosa, L. (2019). "The microvine: A versatile plant model to boost grapevine studies in physiology and genetics" in *Advances in grape and wine biotechnology*. 1–13, Eds. A. Morata and I. Loira (IntechOpen).

Peña, L., Cervera, M., Fagoaga, C., Pérez, R., Romero, J., Juárez, J., et al. (2004). "Agrobacterum-mediated transformation of citrus," in *Transgenic crops of the world: Essential protocols*. Ed. I. S. Curtis (Springer Netherlands: Dordrecht), 145–156. doi: 10.1007/978-1-4020-2333-0_11

Pessina, S., Lenzi, L., Perazzoli, M., Campa, M., Dalla Costa, L., Urso, S., et al. (2016). Knockdown of MLO genes reduces susceptibility to powdery mildew in grapevine. *Hortic. Res.* 3, 16016. doi: 10.1038/hortres.2016.16

Petri, C., Alburquerque, N., Faize, M., Scorza, R., and Dardick, C. (2018). Current achievements and future directions in genetic engineering of European plum (Prunus domestica l.). *Transgenic Res.* 27, 225–240. doi: 10.1007/s11248-018-0072-3

Pompili, V., Dalla Costa, L., Piazza, S., Pindo, M., and Malnoy, M. (2020). Reduced fire blight susceptibility in apple cultivars using a high-efficiency CRISPR/Cas9-FLP/FRT-based gene editing system. *Plant Biotechnol. J.* 18, 845–858. doi: 10.1111/pbi.13253

Pugsley, C. E., Isaac, R. E., Warren, N. J., and Cayre, O. J. (2021). Recent advances in engineered nanoparticles for RNAi-mediated crop protection against insect pests. *Front. Agron.* 3. doi: 10.3389/fagro.2021.652981

Qi, X., Liu, C., Song, L., and Li, M. (2019). Arabidopsis EOD3 homologue PaCYP78A6 affects fruit size and is involved in sweet cherry (Prunus avium l.) fruit ripening. *Scientia Hortic.* 246, 57–67. doi: 10.1016/j.scienta.2018.10.041

Ramakrishna, S., Kwaku Dad, A.-B., Beloor, J., Gopalappa, R., Lee, S.-K., and Kim, H. (2014). Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res.* 24, 1020–1027. doi: 10.1101/gr.171264.113

Reed, K. M., and Bargmann, B. O. R. (2021). Protoplast regeneration and its use in new plant breeding technologies. *Front. Genome Ed* 3. doi: 10.3389/fgeed.2021.734951

Salomé, P. A. (2020). A roadmap toward Large-scale genome editing in crops. *Plant Cell* 32, 1340–1341. doi: 10.1105/tpc.20.00144

Sant'Ana, R. R. A., Caprestano, C. A., Nodari, R. O., and Agapito-Tenfen, S. Z. (2020). PEG-delivered CRISPR-Cas9 ribonucleoproteins system for gene-editing screening of maize protoplasts. *Genes* 11, 1029. doi: 10.3390/genes11091029

Savadi, S., Mangalassery, S., and Sandesh, M. S. (2021). Advances in genomics and genome editing for breeding next generation of fruit and nut crops. *Genomics* 113, 3718–3734. doi: 10.1016/j.ygeno.2021.09.001

Schaefer, L. K., Parlange, F., Buchmann, G., Jung, E., Wehrli, A., Herren, G., et al. (2020). Cross-kingdom RNAi of pathogen effectors leads to quantitative adult plant resistance in wheat. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00253

Schreiber, L. (2005). Polar paths of diffusion across plant cuticles: New evidence for an old hypothesis. *Ann. Bot.* 95, 1069–1073. doi: 10.1093/aob/mci122

Schwartz, S. H., Hendrix, B., Hoffer, P., Sanders, R. A., and Zheng, W. (2020). Carbon dots for efficient small interfering RNA delivery and gene silencing in plants. *Plant Physiol.* 184, 647–657. doi: 10.1104/pp.20.00733

Sciabola, S., Xi, H., Cruz, D., Cao, Q., Lawrence, C., Zhang, T., et al. (2021). PFRED: A computational platform for siRNA and antisense oligonucleotides design. *PloS One* 16, e0238753. doi: 10.1371/journal.pone.0238753

Šečić, E., and Kogel, K.-H. (2021). Requirements for fungal uptake of dsRNA and gene silencing in RNAi-based crop protection strategies. *Curr. Opin. Biotechnol.* 70, 136–142. doi: 10.1016/j.copbio.2021.04.001

Shabalina, S. A., Spiridonov, A. N., and Ogurtsov, A. Y. (2006). Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinf.* 7, 65. doi: 10.1186/1471-2105-7-65

Shew, A. M., Danforth, D. M., Nalley, L. L., Nayga, R. M., Tsiboe, F., and Dixon, B. L. (2017). New innovations in agricultural biotech: Consumer acceptance of topical RNAi in rice production. *Food Control* 81, 189–195. doi: 10.1016/j.foodcont.2017.05.047

Singh, K., Dardick, C., and Kumar Kundu, J. (2019). RNAi-mediated resistance against viruses in perennial fruit plants. *Plants* 8, 359. doi: 10.3390/plants8100359

Stuttmann, J., Barthel, K., Martin, P., Ordon, J., Erickson, J. L., Herr, R., et al. (2021). Highly efficient multiplex editing: one-shot generation of 8× nicotiana benthamiana and 12× arabidopsis mutants. *Plant J.* 106, 8–22. doi: 10.1111/tpj.15197

Svitashev, S., Schwartz, C., Lenderts, B., Young, J. K., and Mark Cigan, A. (2016). Genome editing in maize directed by CRISPR–Cas9 ribonucleoprotein complexes. *Nat. Commun.* 7, 13274. doi: 10.1038/ncomms13274

Tang, X., Lowder, L. G., Zhang, T., Malzahn, A. A., Zheng, X., Voytas, D. F., et al. (2017). A CRISPR-Cpf1 system for efficient genome editing and transcriptional repression in plants. *Nat. Plants* 3, 17103. doi: 10.1038/nplants.2017.103

Tang, X., Ren, Q., Yang, L., Bao, Y., Zhong, Z., He, Y., et al. (2019). Single transcript unit CRISPR 2.0 systems for robust Cas9 and Cas12a mediated plant genome editing. *Plant Biotechnol. J.* 17, 1431–1445. doi: 10.1111/pbi.13068

Tang, X., Zheng, X., Qi, Y., Zhang, D., Cheng, Y., Tang, A., et al. (2016). A single transcript CRISPR-Cas9 system for efficient genome editing in plants. *Mol. Plant* 9, 1088–1091. doi: 10.1016/j.molp.2016.05.001

Taning, C. N., Arpaia, S., Christiaens, O., Dietz-Pfeilstetter, A., Jones, H., Mezzetti, B., et al. (2020). RNA-Based biocontrol compounds: current status and perspectives to reach the market. *Pest Manage. Sci.* 76, 841–845. doi: 10.1002/ps.5686

Taning, C. N. T., Christiaens, O., Berkvens, N., Casteels, H., Maes, M., and Smagghe, G. (2016). Oral RNAi to control drosophila suzukii: laboratory testing against larval and adult stages. *J. Pest Sci.* 89, 803–814. doi: 10.1007/s10340-016-0736-9

Taning, C. N. T., Mezzetti, B., Kleter, G., Smagghe, G., and Baraldi, E. (2021). Does RNAi-based technology fit within EU sustainability goals? *Trends Biotechnol.* 39, 644–647. doi: 10.1016/j.tibtech.2020.11.008

Taochy, C., Gursanscky, N. R., Cao, J., Fletcher, S. J., Dressel, U., Mitter, N., et al. (2017). A genetic screen for impaired systemic RNAi highlights the crucial role of DICER-LIKE 2. *Plant Physiol.* 175, 1424–1437. doi: 10.1104/pp.17.01181

Taylor, R. E., and Zahid, M. (2020). Cell penetrating peptides, novel vectors for gene therapy. *Pharmaceutics* 12, 225. doi: 10.3390/pharmaceutics12030225

Tian, S., Jiang, L., Cui, X., Zhang, J., Guo, S., Li, M., et al. (2018). Engineering herbicide-resistant watermelon variety through CRISPR/Cas9-mediated base-editing. *Plant Cell Rep.* 37, 1353–1356. doi: 10.1007/s00299-018-2299-0

Toda, E., Koiso, N., Takebayashi, A., Ichikawa, M., Kiba, T., Osakabe, K., et al. (2019). An efficient DNA- and selectable-marker-free genome-editing system using zygotes in rice. *Nat. Plants* 5, 363–368. doi: 10.1038/s41477-019-0386-z

Torregrosa, L., Rienth, M., Romieu, C., and Pellegrino, A. (2019). The microvine, a model for studies in grapevine physiology and genetics. *OENO One* 53(3). doi: 10.20870/oeno-one.2019.53.3.2409

Trofimenko, E., Grasso, G., Heulot, M., Chevalier, N., Deriu, M. A., Dubuis, G., et al. (2021). Genetic, cellular, and structural characterization of the membrane potential-dependent cell-penetrating peptide translocation pore. *eLife* 10, e69832. doi: 10.7554/eLife.69832

Turnbull, C., Lillemo, M., and Hvoslef-Eide, T. A. K. (2021). Global regulation of genetically modified crops amid the gene edited crop boom – a review. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.630396

Varkonyi-Gasic, E., Wang, T., Voogd, C., Jeon, S., Drummond, R. S. M., Gleave, A. P., et al. (2019). Mutagenesis of kiwifruit CENTRORADIALIS-like genes transforms a climbing woody perennial with long juvenility and axillary flowering into a compact plant with rapid terminal flowering. *Plant Biotechnol. J.* 17, 869–880. doi: 10.1111/pbi.13021

Vojnov, A. A., Morais do Amaral, A., Dow, J. M., Castagnaro, A. P., and Marano, M. R. (2010). Bacteria causing important diseases of citrus utilise distinct modes of pathogenesis to attack a common host. *Appl. Microbiol. Biotechnol.* 87, 467–477. doi: 10.1007/s00253-010-2631-2

Wang, J. W., Cunningham, F. J., Goh, N. S., Boozarpour, N. N., Pham, M., and Landry, M. P. (2021a). Nanoparticles for protein delivery in planta. *Curr. Opin. Plant Biol.* 60, 102052. doi: 10.1016/j.pbi.2021.102052

Wang, J. W., Grandio, E. G., Newkirk, G. M., Demirer, G. S., Butrus, S., Giraldo, J. P., et al. (2019). Nanoparticle-mediated genetic engineering of plants. *Mol. Plant* 12, 1037–1040. doi: 10.1016/j.molp.2019.06.010

Wang, L., Huang, C., and Yang, J. Y. (2010). Predicting siRNA potency with random forests and support vector machines. *BMC Genomics* 11, S2. doi: 10.1186/1471-2164-11-S3-S2

Wang, M., and Jin, H. (2017). Spray-induced gene silencing: a powerful innovative strategy for crop protection. *Trends Microbiol.* 25, 4–6. doi: 10.1016/j.tim.2016.11.011

Wang, X., Li, L., Li, L., Song, F., and Song, F. (2021b). Interplay of nanoparticle properties during endocytosis. *Crystals* 11, 728. doi: 10.3390/cryst11070728

Wang, M., Mao, Y., Lu, Y., Tao, X., and Zhu, J. (2017a). Multiplex gene editing in rice using the CRISPR-Cpf1 system. *Mol. Plant* 10, 1011–1013. doi: 10.1016/j.molp.2017.03.001

Wan, D.-Y., Guo, Y., Cheng, Y., Hu, Y., Xiao, S., Wang, Y., et al. (2020). CRISPR/Cas9-mediated mutagenesis of VvMLO3 results in enhanced resistance to powdery mildew in grapevine (Vitis vinifera). *Horticulture Res.* 7, 116. doi: 10.1038/s41438-020-0339-8

Wang, M., Weiberg, A., Dellota, E., Yamane, D., and Jin, H. (2017b). Botrytis small RNA bc-siR37 suppresses plant defense genes by cross-kingdom RNAi. *RNA Biol.* 14, 421–428. doi: 10.1080/15476286.2017.1291112

Wang, M., Weiberg, A., Lin, F.-M., Thomma, B. P. H. J., Huang, H.-D., and Jin, H. (2016). Bidirectional cross-kingdom RNAi and fungal uptake of external RNAs confer plant protection. *Nat. Plants* 2, 16151. doi: 10.1038/nplants.2016.151

Wang, Q., Zhang, C., Shen, G., Liu, H., Fu, H., and Cui, D. (2014). Fluorescent carbon dots as an efficient siRNA nanocarrier for its interference therapy in gastric cancer cells. *J. Nanobiotechnol.* 12, 58. doi: 10.1186/s12951-014-0058-0

Watanabe, K., Odahara, M., Miyamoto, T., and Numata, K. (2021). Fusion peptide-based biomacromolecule delivery system for plant cells. *ACS Biomater. Sci. Eng.* 7, 2246–2254. doi: 10.1021/acsbiomaterials.1c00227

Werner, B. T., Gaffar, F. Y., Schuemann, J., Biedenkopf, D., and Koch, A. M. (2020). RNA-Spray-Mediated silencing of fusarium graminearum AGO and DCL genes improve barley disease resistance. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00476

Werner Ribeiro, C., Dugé de Bernonville, T., Glévarec, G., Lanoue, A., Oudin, A., Pichon, O., et al. (2020). "ALSV-based virus-induced gene silencing in apple tree (Malus × domestica l.)," in *Virus-induced gene silencing in plants: Methods and protocols*. Eds. V. Courdavault and S. Besseau (New York, NY: Springer US), 183–197. Methods in Molecular Biology. doi: 10.1007/978-1-0716-0751-0_14

Wise, J. C., Wise, A. G., Rakotondravelo, M., Vandervoort, C., Seeve, C., and Fabbri, B. (2022). Trunk injection delivery of dsRNA for RNAi-based pest control in apple trees. *Pest Manag Sci.* 78, 3528–3533. doi: 10.1002/ps.6993

Woo, J. W., Kim, J., Kwon, S. I., Corvalán, C., Cho, S. W., Kim, H., et al. (2015). DNA-Free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nat. Biotechnol.* 33, 1162–1164. doi: 10.1038/nbt.3389

Wu, R., Cooney, J., Tomes, S., Rebstock, R., Karunairetnam, S., Allan, A. C., et al. (2021). RNAi-mediated repression of dormancy-related genes results in evergrowing apple trees. *Tree Physiol.* 41, 1510–1523. doi: 10.1093/treephys/tpab007

Wulfert, S., and Krueger, S. (2018). Phosphoserine Aminotransferase1 is part of the phosphorylated pathways for serine biosynthesis and essential for light and sugar-dependent growth promotion. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01712

Wu, S., Zhu, H., Liu, J., Yang, Q., Shao, X., Bi, F., et al. (2020). Establishment of a PEG-mediated protoplast transformation system based on DNA and CRISPR/Cas9 ribonucleoprotein complexes for banana. *BMC Plant Biol.* 20, 425. doi: 10.1186/s12870-020-02609-8

Xie, J., Bi, Y., Zhang, H., Dong, S., Teng, L., Lee, R. J., et al. (2020). Cell-penetrating peptides in diagnosis and treatment of human diseases: From preclinical research to clinical application. *Front. Pharmacol.* 11. doi: 10.3389/fphar.2020.00697

Xu, R., Li, J., Liu, X., Shan, T., Qin, R., and Wei, P. (2020). Development of plant prime-editing systems for precise genome editing. *Plant Commun.* 1, 100043. doi: 10.1016/j.xplc.2020.100043

Yu, J., Tu, L., Subburaj, S., Bae, S., and Lee, G.-J. (2021). Simultaneous targeting of duplicated genes in petunia protoplasts for flower color modification *via* CRISPR-Cas9 ribonucleoproteins. *Plant Cell Rep.* 40, 1037–1045. doi: 10.1007/s00299-020-02593-1

Zhang, Y., Cheng, Y., Fang, H., Roberts, N., Zhang, L., Vakulskas, C. A., et al. (2022b). Highly efficient genome editing in plant protoplasts by ribonucleoprotein delivery of CRISPR-Cas12a nucleases. *Front. Genome Ed.* 4. doi: 10.3389/fgeed.2022.780238

Zhang, H., Demirer, G. S., Zhang, H., Ye, T., Goh, N. S., Aditham, A. J., et al. (2019a). DNA Nanostructures coordinate gene silencing in mature plants. *Proc. Natl. Acad. Sci.* 116, 7543–7548. doi: 10.1073/pnas.1818290116

Zhang, H., Goh, N. S., Wang, J. W., Pinals, R. L., González-Grandío, E., Demirer, G. S., et al. (2022a). Nanoparticle cellular internalization is not required for RNA delivery to mature plant leaves. *Nat. Nanotechnol.* 17, 197–205. doi: 10.1038/s41565-021-01018-8

Zhang, Y., Liang, Z., Zong, Y., Wang, Y., Liu, J., Chen, K., et al. (2016). Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. *Nat. Commun.* 7, 1–8. doi: 10.1038/ncomms12617

Zhang, R., Liu, J., Chai, Z., Chen, S., Bai, Y., Zong, Y., et al. (2019b). Generation of herbicide tolerance traits and a new selectable marker in wheat using base editing. *Nat. Plants* 5, 480–485. doi: 10.1038/s41477-019-0405-0

Zhong, Z., Liu, S., Liu, X., Liu, B., Tang, X., Ren, Q., et al. (2020). Intron-based single transcript unit CRISPR systems for plant genome editing. *Rice* 13, 8. doi: 10.1186/s12284-020-0369-8

Zhong, Z., Zhang, Y., You, Q., Tang, X., Ren, Q., Liu, S., et al. (2018). Plant genome editing using FnCpf1 and LbCpf1 nucleases at redefined and altered PAM sites. *Mol. Plant* 11, 999–1002. doi: 10.1016/j.molp.2018.03.008

Zhou, J., Wang, G., and Liu, Z. (2018). Efficient genome editing of wild strawberry genes, vector development and validation. *Plant Biotechnol. J.* 16, 1868–1877. doi: 10.1111/pbi.12922

TYPE Original Research
PUBLISHED 27 October 2022
DOI 10.3389/fpls.2022.1007266

Check for updates

Macadamia germplasm and genomic database (MacadamiaGGD): A comprehensive platform for germplasm innovation and functional genomics in *Macadamia*

Pan Wang[1,2], Yi Mo[1,2], Yi Wang[1,2], Yuchong Fei[1,2], Jianting Huang[1,2], Jun Ni[1,2]* and Zeng-Fu Xu[1,2]*

[1]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Forestry, Guangxi University, Nanning, China, [2]Key Laboratory of National Forestry and Grassland Administration for Fast-Growing Tree Breeding and Cultivation in Central and Southern China, College of Forestry, Guangxi University, Nanning, China

As an important nut crop species, macadamia continues to gain increased amounts of attention worldwide. Nevertheless, with the vast increase in macadamia omic data, it is becoming difficult for researchers to effectively process and utilize the information. In this work, we developed the first integrated germplasm and genomic database for macadamia (MacadamiaGGD), which includes five genomes of four species; three chloroplast and mitochondrial genomes; genome annotations; transcriptomic data for three macadamia varieties, germplasm data for four species and 262 main varieties; nine genetic linkage maps; and 35 single-nucleotide polymorphisms (SNPs). The database serves as a valuable collection of simple sequence repeat (SSR) markers, including both markers that are based on macadamia genomic sequences and developed in this study and markers developed previously. MacadamiaGGD is also integrated with multiple bioinformatic tools, such as search, JBrowse, BLAST, primer designer, sequence fetch, enrichment analysis, multiple sequence alignment, genome alignment, and gene homology annotation, which allows users to conveniently analyze their data of interest. MacadamiaGGD is freely available online (http://MacadamiaGGD.net). We believe that the database and additional information of the SSR markers can help scientists better understand the genomic sequence information of macadamia and further facilitate molecular breeding efforts of this species.

KEYWORDS

***Macadamia*, germplasm, genome, SSR, SNP, MacadamiaGGD, genomic database, molecular breeding**

OPEN ACCESS

EDITED BY
MANUEL TALÓN,
Instituto Valenciano de Investigaciones Agrarias, Spain

REVIEWED BY
Priyanka Sharma,
University of Queensland, Australia
Alberto García,
Universitat Politècnica de València, Spain

*CORRESPONDENCE
Jun Ni
nijun@gxu.edu.cn
Zeng-Fu Xu
zfxu@gxu.edu.cn

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 30 July 2022
ACCEPTED 11 October 2022
PUBLISHED 27 October 2022

CITATION
Wang P, Mo Y, Wang Y, Fei Y, Huang J, Ni J and Xu Z-F (2022) Macadamia germplasm and genomic database (MacadamiaGGD): A comprehensive platform for germplasm innovation and functional genomics in *Macadamia*.
Front. Plant Sci. 13:1007266.
doi: 10.3389/fpls.2022.1007266

Introduction

Macadamia (*Macadamia* spp.), which belongs to the Proteaceae family (Urata, 1954), is an evergreen perennial flowering plant species (Storey and Hamilton, 1953) originating from southern Queensland and northern New South Wales in Australia (Moncur et al., 1985). Macadamia has already become one of the most important economic oil crop species worldwide (Sedgley, 1983; Aradhya et al., 1998; Topp et al., 2019) due to the high level of monounsaturated fatty acid-palmitoleic acid (omega-7) in its nuts, which can effectively lower blood total cholesterol and benefit human health (Nagao et al., 1992; Moodley et al., 2007; Arroyo-Caro et al., 2016). To date, four macadamia species, namely, *Macadamia integrifolia* (Maiden & Betche), *M. tetraphylla* (L. A. S. Johnson), *M. ternifolia* (F. Muell), and *M. jansenii* (C.L. Gross & P.H. Weston), have been identified (Mast et al., 2008), among which only *M. integrifolia*, *M. tetraphylla*, and their hybrids are most widely planted worldwide (SAMAC, 2020). The other two species, *M. ternifolia* and *M. jansenii*, have not yet been used for any commercial purpose because they produce only small, unpalatable, bitter, inedible nuts, the mature nuts of which contain high cyanogenic glycoside levels (Trueman, 2013; Mai et al., 2020).

Macadamia plants are diploid (2n = 28) (Peace et al., 2003) and their genome size ranges from 758 to 896 megabase (Mb) (Nock et al., 2020; Niu et al., 2022a). In recent years, several *de novo*-assembled macadamia genomes have been reported, providing new insight for genetic breeding. In 2016, the first assembled draft genome of macadamia (*M. integrifolia* cultivar HAES 741) was finished and released by Nock's lab, the staff of whom used the short-read Illumina sequence platform (193493 scaffolds, N50 = 4745 bp, 518 Mb) (Nock et al., 2016). In 2020, the first sequence-based genetic linkage maps of macadamia were constructed (Langdon et al., 2020). In 2020, an improved chromosome-scale genome assembly of *M. integrifolia* cultivar HAES 741 was completed by the use of the short-read Illumina and long-read Pacific Biosciences (PacBio) sequencing platforms (4094 scaffolds, N50 = 413 kb, 745 Mb) (Nock et al., 2020). Furthermore, in 2020, by using the third-generation sequencing (TGS) platforms Oxford Nanopore (PromethION), PacBio (Sequel I), and BGI (Single-tube Long Fragment Read), researchers assembled the genome of *M. jansenii* (Murigneux et al., 2020). In addition, the genomes of *M. integrifolia* (249 contigs, N50 = 5.3 Mb, 738 Mb), *M. tetraphylla* (153 contigs, N50 = 10.0 Mb, 707 Mb), *M. ternifolia* (211 contigs, N50 = 6.4 Mb, 716 Mb), and *M. jansenii* (284 contigs, N50 = 4.5 Mb, 738 Mb) were assembled by use of the PacBio HiFi TGS platform (Sharma et al., 2021a). The genome of *M. jansenii* has been improved by Hi-C assembly (219 scaffolds, N50 = 52 Mb, 758 Mb) (Sharma et al., 2021c) and was further updated by the latest hifiasm assembly (779 contigs, N50 = 46 Mb, 826 Mb) (Sharma et al., 2021b). Recently, the genome of the cultivar HAES 344 was sequenced and assembled into 14 pseudochromosomes by the use of Illumina NovaSeq and PacBio Sequel II sequencing (5387 contigs, N50 = 281 kb, 794 Mb) (Lin et al., 2022). A chromosome-scale genome assembly of *M. tetraphylla* has also been constructed from long-read Oxford Nanopore Technologies (ONT) sequencing data (1059 scaffolds, N50 = 51 Mb, 751 Mb) (Niu et al., 2022a). Moreover, in recent years, the chloroplast and mitochondrion genomes of *M. integrifolia*, *M. tetraphylla*, and *M. ternifolia* have been assembled and thoroughly annotated (Niu et al., 2022b).

As inbreeding decline occurs in macadamia, it is vitally important to understand the genetic distances between individuals (Steiger et al., 2003). The morphological characteristics of macadamia could be greatly influenced by the environment; thus, it is sometimes difficult to identify genetic relationships through phenotypic observations (Hardner, 2016). The use of DNA marker systems has become one of the most efficient strategies to evaluate genetic distance and genetic foundation (Ranketse et al., 2022). DNA marker systems, including isozyme (Vithanage and Winks, 1992; Aradhya et al., 1998), randomly amplified DNA fingerprinting (RAF) (Peace et al., 2002; Peace et al., 2004; Peace et al., 2005), amplified fragment length polymorphism (AFLP) (Steiger et al., 2003), sequence tagged site (STS) (Vithanage et al., 1998), random amplified polymorphic DNA (RAPD) (Vithanage et al., 1998), randomly amplified microsatellite fingerprinting (RAMiFi) (Peace et al., 2004), simple sequence repeat (SSR) (Schmidt et al., 2006; Nock et al., 2014b; Langdon et al., 2019; Ranketse et al., 2022), diversity array technology (DArT) and single-nucleotide polymorphism (SNP) markers (Alam et al., 2018; O'Connor et al., 2019b), have been developed for the genetic and molecular breeding of macadamia. Genome-wide association studies (GWASs) have also greatly facilitated the identification of new molecular markers associated with yield traits (O'Connor et al., 2019a; O'Connor et al., 2020). As codominant, highly reproducible, highly polymorphic and cost-efficient DNA markers, SSRs have been preferred for use in studies of genetic identification and diversity analysis. To date, although the sequencing of the whole genomes of different macadamia species has been completed, genome-based development of SSR markers has not been reported.

With the rapidly developed sequencing technologies, the genomes of dozens of plant species have been sequenced each year. Nevertheless, how to integrate and well manage the large amount of omics data is still a task. In recent years, the genomic databases of some economic crops were well constructed and greatly facilitated the researchers to use the genome, transcriptome, or phenotype data. Citrus Genome Database (CGD, https://www.citrusgenomedb.org/) integrates genomes, maps, markers, phenotype data, and quantitative trait loci of agronomic traits of 25 citrus species. The Rice Genome Hub (RGH, https://rice-genome-hub.southgreen.fr), which is part of the South Green Bioinformatics platform, also integrates large amount of rice omics data with a large number of powerful in-house tools (Droc et al., 2019). Rice Annotation Project Database

(RAP-DB, https://rapdb.dna.affrc.go.jp/) is consisted of updated genome annotation and focuses on the comprehensive analysis of genome structure and function of rice genes (Project, 2007). Gossypium Resource and Network Database (GRAND, http://grand.cricaas.com.cn) contains the genomic, transcriptomic, phenotypic, and integrative analysis tools for cotton (Zhang et al., 2022). With the inspirations from these databases, in this study we developed the first integrated germplasm and functional genomic database for macadamia (MacadamiaGGD).

Currently, large amounts of macadamia omics data lack centralized management. These data are distributed across multiple repositories or personal websites, with the same data from the same source in different repositories. In addition, many macadamia omics data lack the management of versions. The same data has different versions and accession numbers in different repositories, which can make it difficult for users to find the most updated dataset. The main purpose of the MacadamiaGGD described in this article is to provide the germplasm data, genome resources, transcriptome (RNA-seq) data, molecular marker information and genetic linkage map information to assist in the scientific research and molecular breeding of macadamia. And several commonly used bioinformatics tools are also integrated with MacadamiaGGD, which can help the researchers better utilize the database.

Materials and methods

Data sources and processing

In MacadamiaGGD, we integrated the genetic information data, including that of five genomes of four species, the chloroplast and mitochondrion genomes of three species and genome annotations, which were previously released in public databases, including the National Center for Biotechnology Information (NCBI) Assembly database, the *GigaScience* database (*Giga*DB), and the China National Center for Bioinformation (CNCB) Genome Warehouse (GWH) database. In addition, transcriptomic data for three macadamia varieties were downloaded from the NCBI Sequence Read Archive (SRA) database. The germplasm, genetic linkage map, SNP and SSR marker data were retrieved from the NCBI PubMed database and other databases, as summarized in Table 1. The components of data integration mainly include the data source, the data transform, and the data sink in the database. Extract, transform, and load (ETL) architecture was applied to data integration. In data integration process, raw data were collected, transformed, sorted, cleaned, aggregated, and stored *via* using PostgreSQL 9.5.25, Scala 2.13.1, AKKA 2.6.5, and SBT 1.3.5 (Figures 1A, B). Processed raw data were applied for variation calling and data visualization though using HTML5, CSS3, Java Script, Slick 3.3.2, Bootstrap 3.3.0 and Play Framework 2.8.2 (Figure 1B).

Development of the database

MacadamiaGGD was deployed in the Ubuntu 16.04 operation system using AKKA 2.6.5 (https://akka.io) as the web server, PostgreSQL 9.5.25 (https://www.postgresql.org) as the database server, Scala 2.13.1 (https://www.scala-lang.org) as the programming language and SBT 1.3.5 (https://www.scala-sbt.org) as the interactive building tool. All the data were managed and stored in the PostgreSQL Database. The website interface was generated *via* Bootstrap 3.3.0 (https://getbootstrap.com) and Play Framework 2.8.2 (https://www.playframework.com/). The web interface of MacadamiaGGD was developed using HTML5, CSS3, Java Script. The query function was enforced based on the Slick 3.3.2 middleware tier. JBrowse 1.16.6 (https://www.jbrowse.org) was used for genome visualization.

Sample collection and DNA isolation

Leaf samples of 21 macadamia accessions for DNA isolation were collected from the macadamia plantation in Chongzuo, Guangxi, China (Table S1). The DNA was isolated following a previously described method (Doyle, 1991), with slight modifications. To avoid problems of low efficiency and insufficient grinding due to manual grinding, young leaves were ground in a Tissuelyser-192 (Shanghai Jingxin Industrial Development Co., Ltd., China) and extracted with a 2% cetyltrimethylammonium bromide (CTAB) buffer. Nucleic acids were isolated with a chloroform: isoamyl alcohol (24:1) solution. DNA was purified with ethanol and resuspended in sterile distilled water. The DNA quality and concentration were assessed using ultraviolet spectrometry *via* a Nanodrop 2000c (Thermo Fisher Scientific, MA, USA) and agarose gel electrophoresis. The purified DNA was stored at -20°C until use.

Genome-wide SSR screening and characterization

New microsatellite markers were screened in the *M. integrifolia* HAES 741 reference genome (https://www.ncbi.nlm.nih.gov/bioproject/748012) by using SSRHunter 1.3 (http://www.bio2soft.net) (Li and Wan, 2005). The search criteria were set as 2, 3, and 4 nucleotides, corresponding to at least 4 repetitions. Afterward, the SSRs, comprising no fewer than 30 repeated motifs and being evenly distributed on each chromosome, were preferentially selected. To further confirm the quality of the SSR markers, each sequence was again queried *via* BLAST within MacadamiaGGD and tested *via* polymerase chain reaction (PCR).

Primer 3 (https://primer3.org) was used to design primer pairs flanking the sequences of the screened SSR motifs. The primer design parameters were as follows: primer length, 17-25

TABLE 1 Summary of all datasets in MacadamiaGGD.

Dataset	Species	References	Repository/ Accession number	URL
Germplasm	*M. integriflia* *M. ternifolia* *M. tetraphylla* *M. jansenii*	Vithanage and Winks (1992); Aradhya et al. (1998); Peace et al. (2002); Peace et al. (2005); Allan (2007); He (2008); Gitonga et al. (2009); Hardner et al. (2009); Machado Neto and Moryia (2010); Zhang (2011); Hardner (2016); Zeng and Du (2017); Alam et al. (2018); Tang et al. (2018); Toft et al. (2018); Langdon et al. (2019); O'Connor et al. (2019b); Tan et al. (2019); Tan et al. (2020); Mai et al. (2021); Tan et al. (2021); Lin et al. (2022)		http://MacadamiaGGD.net/nut/toRef
Genome assembly	*M. integriflia* HAES 741	Nock et al. (2020)	NCBI/PRJNA748012	https://www.ncbi.nlm.nih.gov/bioproject/748012
	M. integriflia HAES 344	Lin et al. (2022)	CNCB/PRJCA004595	https://ngdc.cncb.ac.cn/gwh/Assembly/23196/show
	M. tetraphylla	Sharma et al. (2021a)	GigaDB/100906; NCBI/PRJNA694456	http://gigadb.org/dataset/view/id/100906/
	M. ternifolia	Sharma et al. (2021a)	GigaDB/100906; NCBI/PRJNA694456	http://gigadb.org/dataset/view/id/100906/
	M. jansenii	Sharma et al. (2021a)	GigaDB/100906; NCBI/PRJNA694456	http://gigadb.org/dataset/view/id/100906/
Genome annotation	*M. integriflia* HAES 741	Nock et al. (2020)	NCBI/PRJNA748012	https://www.ncbi.nlm.nih.gov/bioproject/748012
	M. integriflia HAES 344	Lin et al. (2022)	CNCB/PRJCA004595	https://ngdc.cncb.ac.cn/gwh/Assembly/23196/show
Chloroplast assembly and annotation	*M. integriflia*	Nock et al. (2014a)	NCBI/PRJNA264682	https://www.ncbi.nlm.nih.gov/genome/?term=txid60698
	M. ternifolia	Liu et al. (2017)	NCBI/PRJNA421511	https://www.ncbi.nlm.nih.gov/genome/browse/#!/organelles/66349/
	M. tetraphylla	Liu et al. (2018)	NCBI/MH778649	https://www.ncbi.nlm.nih.gov/nuccore/MH778649
Mitochondrion assembly and annotation	*M. integriflia*	Niu et al. (2022b)	NCBI/MW566570	https://www.ncbi.nlm.nih.gov/nuccore/MW566570
	M. ternifolia	Niu et al. (2022b)	NCBI/MW566571	https://www.ncbi.nlm.nih.gov/nuccore/MW566571
	M. tetraphylla	Niu et al. (2022b)	NCBI/MW566572	https://www.ncbi.nlm.nih.gov/nuccore/MW566572
Transcriptome	*M. integriflia* HAES 741	Nock et al. (2020)	NCBI/PRJNA593881	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA593881
	M. integriflia HAES 344	Lin et al. (2022)	NCBI/PRJNA706119	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA706119
	M. integriflia H2	Lin et al. (2022)	NCBI/PRJNA706119	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA706119
Genetic linkage maps	*M. integriflia*	Langdon et al. (2020)		https://researchportal.scu.edu.au/esploro/outputs/dataset/991012821025202368
SSRs	*M. integriflia*	Schmidt et al. (2006); Nock et al. (2014b); Langdon et al. (2019)		http://MacadamiaGGD.net/nut/toRef
SNPs	*M. integriflia*	O'Connor et al. (2019a); O'Connor et al. (2020)		http://MacadamiaGGD.net/nut/toRef
	M. integriflia *M. ternifolia* *M. integriflia* × *M. tetraphylla* *M. jansenii*	(Alam et al., 2018)		

The datasets were deposited in the repositories of the China National Center for Bioinformation (CNCB), the National Center for Biotechnology Information (NCBI), and the GigaDB.

bp; melting temperature (Tm), 53 °C; amplicon size, 350-500 bp; and GC content, 40-60%.

Marker analysis, data analysis and map construction

The SSR PCR mixture (10 μL) comprised 1 μL of DNA, 0.4 μL of each primer (10 μM), 5 μL of Rapid Taq Master Mix (Vazyme, China) and 3.2 μL of double-distilled water. The amplification reaction program was as follows: 5 min at 95°C; 36 cycles of (30 s at 95°C, 53°C and 72°C); and a final extension of 5 min at 72 °C. Afterward, the mixture was held at 16°C. The PCR products were examined by electrophoresis on a 7% nondenaturing polyacrylamide gel run at 220 V for 40 min and visualized by silver staining. The density distribution map of polymorphic SSR markers on chromosomes was generated using MG2C software (http://mg2c.iask.in/mg2c_v2.1/).

Results

Overview of MacadamiaGGD

MacadamiaGGD contains the most comprehensive bioinformatics datasets of macadamia (including five genomes, a total of 89.28 Gb of transcriptomic data, three chloroplast and mitochondrion genomes, germplasm data for four species and 262 main varieties, nine genetic linkage maps, 35 SNPs and 657 SSR markers), which provides convenient access to the large amount of germplasm and genomic information of macadamia (Figure 1A). MacadamiaGGD is composed of 11 main functional modules: Home, Germplasm, Genomes, Expression, BLAST, Markers, Maps, Tools, References, Download and Help (Figure 1C). MacadamiaGGD can be used to search and visualize genomic information by using various tools, including search, JBrowse, BLAST, primer designer, sequence fetch, enrichment analysis, multiple sequence alignment, genome alignment, and gene homology annotation (Figure 1). MacadamiaGGD also provides information about macadamia germplasm and genome-related references. In summary, researchers can use the above functional modules of the database to quickly acquire the germplasm and genomic information of macadamia.

Germplasm

In the Germplasm module of MacadamiaGGD, 23 agronomic traits of four species and 16 agronomic traits of 262 main varieties were carefully described, including tree vigor, leaf type, fruit shape, flower color, the early-bloom stage and full-boom stage, and others. Users can easily obtain information on the morphological characteristics of four macadamia species and 262 varieties in the germplasm module. In addition, a phylogenetic analysis tool based on the results of Alam et al. (2018), which shows genetic distances between individuals genotypes, is provided in this module.

Genome browse and search

The MacadamiaGGD database provides public information on the assembled genomes of the *M. integrifolia*, *M. tetraphylla*, *M. ternifolia*, and *M. jansenii*, which are available in different public databases. For example, when "Genomes" is clicked on, the column header label appears, showing the suboptions as in Figure 2A. We can choose any label to access the sublinks and search for the needed information. When the user enters a gene "*LOC122078696*" in *Macadamia integrifolia* HAES 741 genome Browse, it will get the structure and function annotation information of all transcript of the gene (Figure 2B). Moreover, when the user clicks "Search", a new layer appears with four options: "Keyword", "Gene ID", "Gene Name", and "Region" (Figure 2C). Then, if one clicks "Gene ID", the interface appears as a blank box (Figure 2C). The user can enter the gene "*LOC122078696*" in the box and click the Search button; then, the requested information is displayed (Figure 2C).

Genome JBrowse

Gene annotations in MacadamiaGGD are displayed graphically in the genome JBrowse, which includes the information of the gene location, nucleotide sequences, amino acid sequences, and other features. For example, if a user selects the genomic region from 192751 bp to 203445 bp on Chromosome 14 (NC_056557.1) for browsing, all genes located within this zone are displayed properly (Figure 2D). Further, when the mRNA XM_042644823.1 is clicked on, detailed information on its mRNA, coding sequence (CDS), and other features are displayed (Figure 2E).

Transcriptomes of macadamia from different tissues

In the expression module of MacadamiaGGD, a total of 89.28 Gb of raw RNA-seq data were collected from tissues of young leaves, shoots, and flowers from the cultivar 'Mauka' (Nock et al., 2020); tissues of leaves, stems, flowers, and roots from the cultivar 'Kau'; and shells and kernels at five different development stages from cultivar 'Hinde' (Lin et al., 2022). By mapping the transcriptome data to the reference genome and using transcripts per million (TPM) for calculation, we acquired the expression matrix of the annotated genes of macadamia.

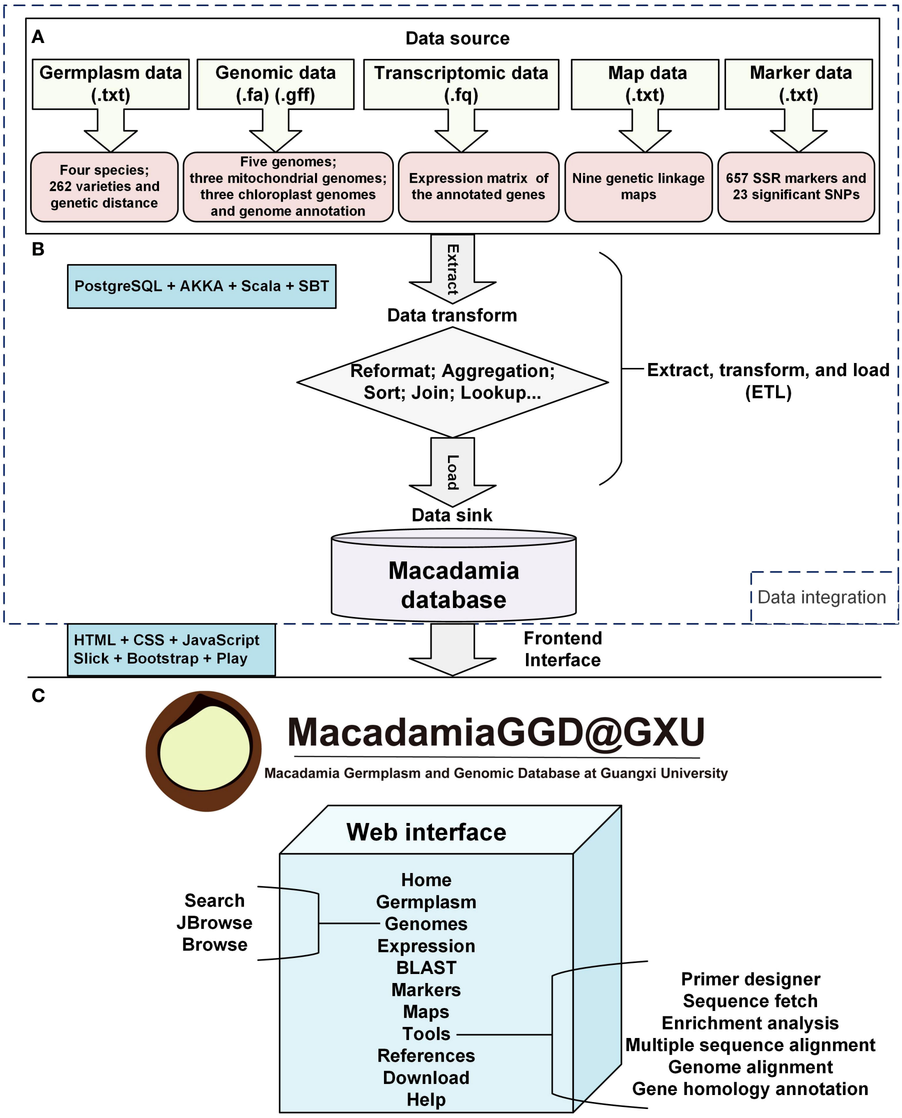

A
Data source
Germplasm data (.txt)
Genomic data (.fa) (.gff)
Transcriptomic data (.fq)
Map data (.txt)
Marker data (.txt)
Four species; 262 varieties and genetic distance
Five genomes; three mitochondrial genomes; three chloroplast genomes and genome annotation
Expression matrix of the annotated genes
Nine genetic linkage maps
657 SSR markers and 23 significant SNPs
B
PostgreSQL + AKKA + Scala + SBT
Extract
Data transform
Reformat; Aggregation; Sort; Join; Lookup...
Extract, transform, and load (ETL)
Load
Data sink
Macadamia database
Data integration
HTML + CSS + JavaScript Slick + Bootstrap + Play
Frontend Interface
C
MacadamiaGGD@GXU
Macadamia Germplasm and Genomic Database at Guangxi University
Web interface
Home
Germplasm
Genomes
Expression
BLAST
Markers
Maps
Tools
References
Download
Help
Search
JBrowse
Browse
Primer designer
Sequence fetch
Enrichment analysis
Multiple sequence alignment
Genome alignment
Gene homology annotation

FIGURE 1
Feature diagram of MacadamiaGGD. MacadamiaGGD is a collection of germplasm, genomic, transcriptomic, maps, and molecular marker data of macadamia, and multiple bioinformatic tools. All the data are stored and managed in a PostgreSQL database. **(A)**, Data source layer. **(B)**, Middleware layer. **(C)**, Application layer.

BLAST

BLAST is the most commonly used tool and is included as a separate module in the MacadamiaGGD database. It allows users to perform both BLASTp and BLASTn searches to rapidly align sequences to the database. In the BLAST module, pasting the DNA/protein sequences in the query box or uploading a FASTA file is acceptable. For example, the users can enter “Example 1” sequence in the blank box and select the against database type, e-value, and max target sequence number and then click the “Run” button to obtain the comparison results *via* the “BLASTp” function (Figure 3A). In addition, when pulling down the search result interface, a user is presented with all the comparison results (Figure 3B), including the description information of the candidate subject sequences alignment parameters (Figure 3C) and the matching information between the query sequence and each subject sequence (Figure 3D).

Markers

In the “markers” module, we included 657 SSR markers and 35 SNPs. Macadamia trees have a relatively long juvenile period (commonly four to five years); thus, it would take a great deal of time to select high-yielding cultivars for breeding. Molecular markers that are associated with key yield traits are extremely important for developing rapid cycle breeding programs in macadamia (O'Connor et al., 2020). To verify the polymorphism of SSR markers from previous research (Schmidt et al., 2006; Nock et al., 2014b; Langdon et al., 2019), we randomly selected 8 primer pairs from MacadamiaGGD (Table S2) and identified polymorphisms of these SSRs *via* electrophoresis. The results showed that the selected primer pairs were polymorphic.

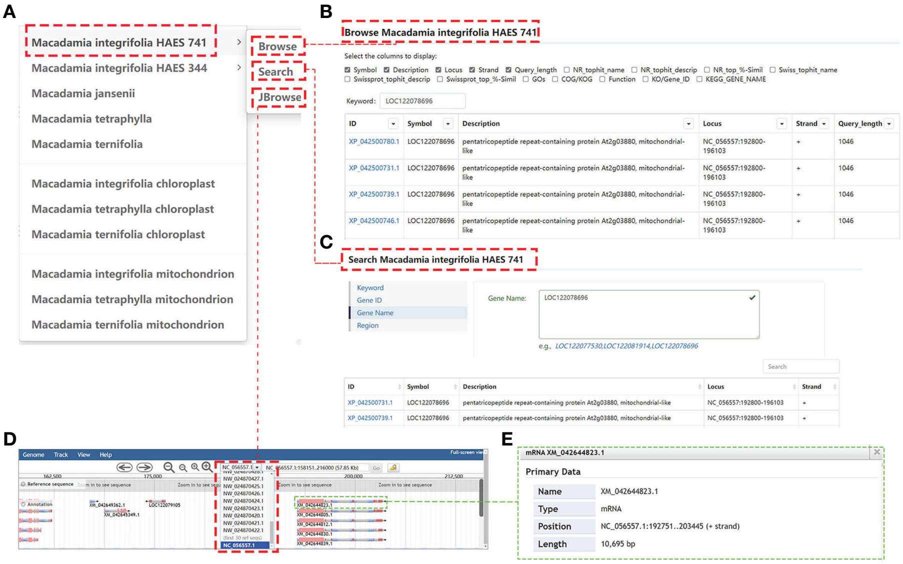
In this study, a total of 145593 SSR loci were obtained from *M. integrifolia* HAES 741 genomic sequences (Nock et al., 2020). They were evenly distributed on 14 chromosomes, with an average density of 10400 loci per chromosome (Table 2). SSR motifs exist as one of three main types: dinucleotide repeats (DNRs), trinucleotide repeats (TNRs) and tetranucleotide repeats (TTRs). Among these SSRs, DNRs were the most abundant (115139), followed by TNRs (26400) and TTRs (4054), which accounted for 79%, 18% and 3%, respectively (Table 2). A total of 927 primer pairs were designed by the selection of the SSR loci with repeat numbers ≥30 from the total SSR loci (Table S3). Out of 927 amplified products, 605 primer pairs were polymorphic, with an average of 1.17 SSR markers per Mb on 14 chromosomes. According to the SSR density distribution map, chromosome 5 had the highest number of SSRs (81), but chromosome 12 had only 13 SSRs (Figure 4). In addition, a total of 35 SNPs were included in the "markers" module, which were significantly associated with the yield component traits identified by genome-wide association studies (GWASs) (O'Connor et al., 2019a; O'Connor et al., 2020).

Maps

The map module contains nine genetic linkage maps derived from three macadamia cultivars, HAES 741, HVP A268 and HVP A4. In each map, there were 14 linkage groups (LGs), which correspond to the number of haploid chromosomes in macadamia. When the users open this module, the features of the maps are displayed, including the description and number of maps. The images of the maps appear at the lower left of the module, while the detailed information of the LG location, the marker numbers, the largest and smallest gap, the total length and the average length between markers is displayed at the lower right.

Tools

The tools module contains several utilities, including "Primer designer", "Sequence Fetch", "Enrichment analysis", "Multiple sequence alignment", "Genome alignment", and "Gene homology annotation", which allow a relatively complete bioinformatics analysis. The user can click the "Primer designer" button, input the nucleic acid sequence or


A
Macadamia integrifolia HAES 741
Macadamia integrifolia HAES 344
Macadamia jansenii
Macadamia tetraphylla
Macadamia ternifolia
Macadamia integrifolia chloroplast
Macadamia tetraphylla chloroplast
Macadamia ternifolia chloroplast
Macadamia integrifolia mitochondrion
Macadamia tetraphylla mitochondrion
Macadamia ternifolia mitochondrion
Browse
Search
JBrowse
B
C
D
Genome
Track
View
Help
Full-screen view
200,000
212,500
Zoom in to see sequence
Go
E

Browse Macadamia integrifolia HAES 741

Select the columns to display:

☑ Symbol ☑ Description ☑ Locus ☑ Strand ☑ Query_length ☐ NR_tophit_name ☐ NR_tophit_descrip ☐ NR_top_%-Simil ☐ Swiss_tophit_name ☐ Swissprot_tophit_descrip ☐ Swissprot_top_%-Simil ☐ GOs ☐ COG/KOG ☐ Function ☐ KO/Gene_ID ☐ KEGG_GENE_NAME

Keyword: LOC122078696

ID	Symbol	Description	Locus	Strand	Query_length
XP_042500780.1	LOC122078696	pentatricopeptide repeat-containing protein At2g03880, mitochondrial-like	NC_056557:192800-196103	+	1046
XP_042500731.1	LOC122078696	pentatricopeptide repeat-containing protein At2g03880, mitochondrial-like	NC_056557:192800-196103	+	1046
XP_042500739.1	LOC122078696	pentatricopeptide repeat-containing protein At2g03880, mitochondrial-like	NC_056557:192800-196103	+	1046
XP_042500746.1	LOC122078696	pentatricopeptide repeat-containing protein At2g03880, mitochondrial-like	NC_056557:192800-196103	+	1046

Search Macadamia integrifolia HAES 741

Keyword
Gene ID
Gene Name
Region

Gene Name: LOC122078696 ✓

e.g., LOC122077530,LOC122081914,LOC122078696

Search

ID	Symbol	Description	Locus	Strand
XP_042500731.1	LOC122078696	pentatricopeptide repeat-containing protein At2g03880, mitochondrial-like	NC_056557:192800-196103	+
XP_042500739.1	LOC122078696	pentatricopeptide repeat-containing protein At2g03880, mitochondrial-like	NC_056557:192800-196103	+

mRNA XM_042644823.1

Primary Data

Name	XM_042644823.1
Type	mRNA
Position	NC_056557.1:192751..203445 (+ strand)
Length	10,695 bp

FIGURE 2
General view of the "Genomes" module. **(A)**, The genome module includes "11 macadamia genomes", and three tools including "Browse", "Search", and "JBrowse". **(B)**, The Browse information of gene "*LOC122078696*" in *Macadamia integrifolia* HAES 741 genome. **(C)**, Showing the Search result of gene "*LOC122078696*". **(D)**, The JBrowse information of gene "*LOC122078696*". **(E)**, Detailed description interface of mRNA XM_042644823.1.

select a scaffold range, adjust the appropriate parameters, and click the "Run" button to obtain a satisfactory pair of primers. Users can screen functional genes of interest (GOIs) based on the data of the *M. integrifolia* transcriptome, click the "Enrichment analysis" button, input the gene ID in the dialog box and select Kyoto Encyclopedia of Genes and Genomes (KEGG) or Gene Ontology (GO) for functional clustering analysis. "Sequence Fetch" can be used to efficiently obtain the sequence of GOI from the *M. integrifolia* genome, which can acquire either a certain or multiple gene sequences at the same time. "Muscle" is a multisequence alignment tool that not only can be used to obtain homology between genes but also can be used to build an intuitive diagram. The "primer designer" tool can be used to design specific primers to clone GOIs for functional research. In addition, by using the "LASTZ" and "GeneWise" tools, users can complete genome alignment and gene homology annotation, respectively.

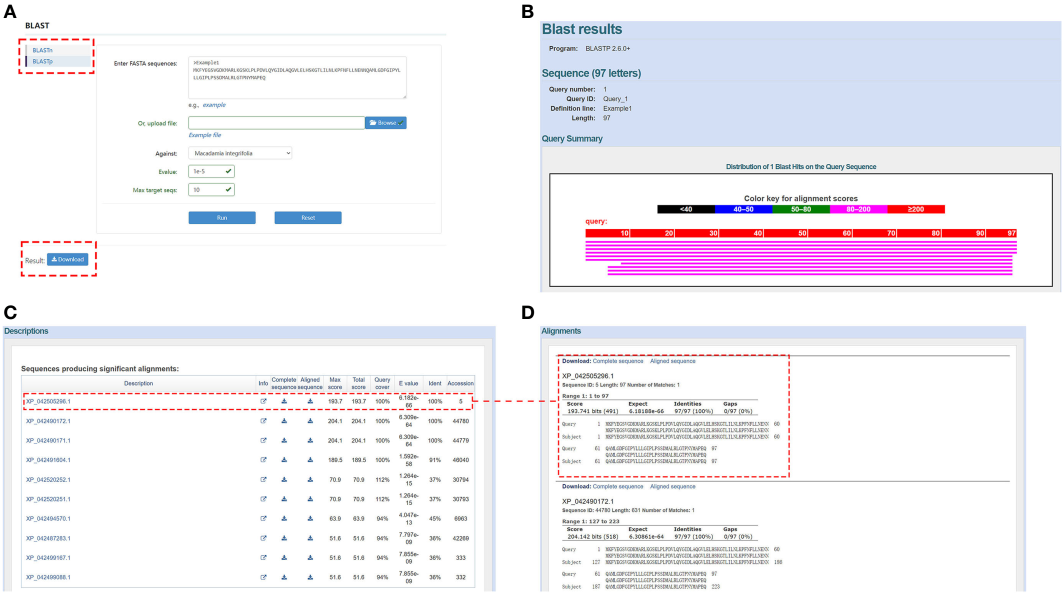
References

Currently, the "Reference" module contains the macadamia germplasm and genome-related references, which allows users to query approximately 40 articles information related to the data contained in MacadamiaGGD. The completion and optimization of macadamia genome sequencing results among these publications contribute to the study of macadamia functional genomics and comparative genomics and are convenient for molecular plant breeding efforts.

A case study involving the use of MacadamiaGGD

MacadamiaGGD integrates BLAST, enrichment analysis, and other tools for functional genomic research of Macadamia. Acyltransferases are the potential molecular targets for genetic engineering to increase the oil content and alter the fatty acid composition in the oil crops (Zhang et al., 2021). Here, we provide a case study on the *diacylglycerol acyltransferases* (*DGATs*) of *M. integrifolia* by using the "BLAST", "GO enrichment", "JBrowse", and "Gene Expression" function of MacadamiaGGD. By using BLAST in MacadamiaGGD, the Conserved Domains Database (CDD) of the NCBI database, the SMART database (https://smart.embl.de/) and MEGA 11 software (https://megasoftware.net/), we obtain one DGAT1 (MiDGAT1), three DGAT2 (MiDGAT2-1, MiDGAT2-2, MiDGAT2-3), and one DGAT3 (MiDGAT3) (Figure 5A). Of the five *MiDGAT* genes, two genes (*MiDGAT2-1*, *MiDGAT2-3*) were mapped to chromosome 14, and their physical positions were very close (Figure 5B).

To verify the expression features of *MiDGATs* during triacylglycerol (TAG) biosynthesis, we downloaded the transcriptome expression data of *M. integrifolia* kernel



BLAST

Enter FASTA sequences:

```
>Example1
MKFYEGSVGDKMARLKGSKLPLPDVLQYGIDLAQGVLELHSKGTLILNLKPFNFLLNENNQAMLGDFGIPYL
LLGIPLPSSDMALRLGTPNYMAPEQ
```

e.g., example

Or, upload file:

Example file

Against: Macadamia integrifolia

Evalue: 1e-5

Max target seqs: 10

Result:

Blast results

Program: BLASTP 2.6.0+

Sequence (97 letters)

Query number: 1
Query ID: Query_1
Definition line: Example1
Length: 97

Query Summary

Distribution of 1 Blast Hits on the Query Sequence

Descriptions

Sequences producing significant alignments:

Description	Info	Complete sequence	Aligned sequence	Max score	Total score	Query cover	E value	Ident	Accession
XP_042505296.1				193.7	193.7	100%	6.182e-66	100%	5
XP_042490172.1				204.1	204.1	100%	6.309e-64	100%	44780
XP_042490171.1				204.1	204.1	100%	6.309e-64	100%	44779
XP_042491604.1				189.5	189.5	100%	1.592e-58	91%	46040
XP_042520252.1				70.9	70.9	112%	1.264e-15	37%	30794
XP_042520251.1				70.9	70.9	112%	1.264e-15	37%	30793
XP_042494570.1				63.9	63.9	94%	4.047e-13	45%	6963
XP_042487283.1				51.6	51.6	94%	7.797e-09	36%	42269
XP_042499167.1				51.6	51.6	94%	7.855e-09	36%	333
XP_042499088.1				51.6	51.6	94%	7.855e-09	36%	332

Alignments

```
Download: Complete sequence   Aligned sequence

XP_042505296.1
Sequence ID: 5 Length: 97 Number of Matches: 1

Range 1: 1 to 97
  Score              Expect       Identities     Gaps
  193.741 bits (491)  6.18188e-66  97/97 (100%)  0/97 (0%)

Query      1  MKFYEGSVGDKMARLKGSKLPLPDVLQYGIDLAQGVLELHSKGTLILNLKPFNFLLNENN  60
              MKFYEGSVGDKMARLKGSKLPLPDVLQYGIDLAQGVLELHSKGTLILNLKPFNFLLNENN
Subject    1  MKFYEGSVGDKMARLKGSKLPLPDVLQYGIDLAQGVLELHSKGTLILNLKPFNFLLNENN  60

Query     61  QAMLGDFGIPYLLLGIPLPSSDMALRLGTPNYMAPEQ  97
              QAMLGDFGIPYLLLGIPLPSSDMALRLGTPNYMAPEQ
Subject   61  QAMLGDFGIPYLLLGIPLPSSDMALRLGTPNYMAPEQ  97

Download: Complete sequence   Aligned sequence

XP_042490172.1
Sequence ID: 44780 Length: 631 Number of Matches: 1

Range 1: 127 to 223
  Score              Expect       Identities     Gaps
  204.142 bits (518)  6.30861e-64  97/97 (100%)  0/97 (0%)

Query      1  MKFYEGSVGDKMARLKGSKLPLPDVLQYGIDLAQGVLELHSKGTLILNLKPFNFLLNENN  60
              MKFYEGSVGDKMARLKGSKLPLPDVLQYGIDLAQGVLELHSKGTLILNLKPFNFLLNENN
Subject  127  MKFYEGSVGDKMARLKGSKLPLPDVLQYGIDLAQGVLELHSKGTLILNLKPFNFLLNENN  186

Query     61  QAMLGDFGIPYLLLGIPLPSSDMALRLGTPNYMAPEQ  97
              QAMLGDFGIPYLLLGIPLPSSDMALRLGTPNYMAPEQ
Subject  187  QAMLGDFGIPYLLLGIPLPSSDMALRLGTPNYMAPEQ  223
```

FIGURE 3
View of the "BLAST" module. **(A)**, Demonstration of the "BLASTp" box. **(B)**, Example of the search result after a sequence was input. **(C)**, Descriptions of the alignment result. **(D)**, Match information between the query sequence and subject sequences.

TABLE 2 Characterization of the screened SSRs in *Macadamia integrifolia*.

Chromosome	DNR	TNR	TTR	All SSR loci	Proportion to all SSR loci (%)	All SSRs	SSRs densitydistribution on chromosome(1/Mb)
Chr1	6453	1276	196	7925	5.44	21	0.58
Chr2	10092	2386	352	12830	8.81	28	0.64
Chr3	9217	2048	349	11614	7.98	25	0.66
Chr4	8935	2142	318	11395	7.83	67	1.79
Chr5	11249	2537	400	14187	9.74	81	1.72
Chr6	8208	1823	291	10322	7.1	63	1.54
Chr7	8730	2065	289	11084	7.61	24	0.65
Chr8	8878	1991	288	11157	7.66	73	2.09
Chr9	6890	1566	249	8705	5.98	22	0.52
Chr10	7443	1833	269	9547	6.56	44	1.29
Chr11	7625	1856	271	9752	6.7	59	1.74
Chr12	7315	1666	250	9231	6.34	13	0.41
Chr13	6562	1516	257	8335	5.72	57	1.95
Chr14	7542	1695	275	9512	6.53	29	0.88
Total	115139	26400	4054	145593	100	606	

DNR, dinucleotide repeat; TNR, trinucleotide repeat; TTR, tetranucleotide repeat; Mb, megabase.

development from MacadamiaGGD. By using gene ontology (GO) annotation information available from MacadamiaGGD, we conducted the GO enrichment analysis of the five *MiDGAT* genes from *M. integrifolia*. The results showed the five *MiDGAT*s were enriched in more than 30 GO terms, which are involved in fatty acid and TAG biosynthesis in plants (Figure 5C). Further, we also investigated the expression profile of *MiDGAT*s at five stages of kernel development. *MiDGAT2-1* and *MiDGAT2-3* were highly expressed in stages I and II (Figure 5D). *MiDGAT2-2* exhibited low expression levels in stages I and II, whereas it was highly expressed in stages III, IV, and V. Consistent with these results, The expression pattern of *MiDGAT2* was recently found to be mainly correlated with fatty acid biosynthesis at different stages of developing kernels (Gao et al., 2021).

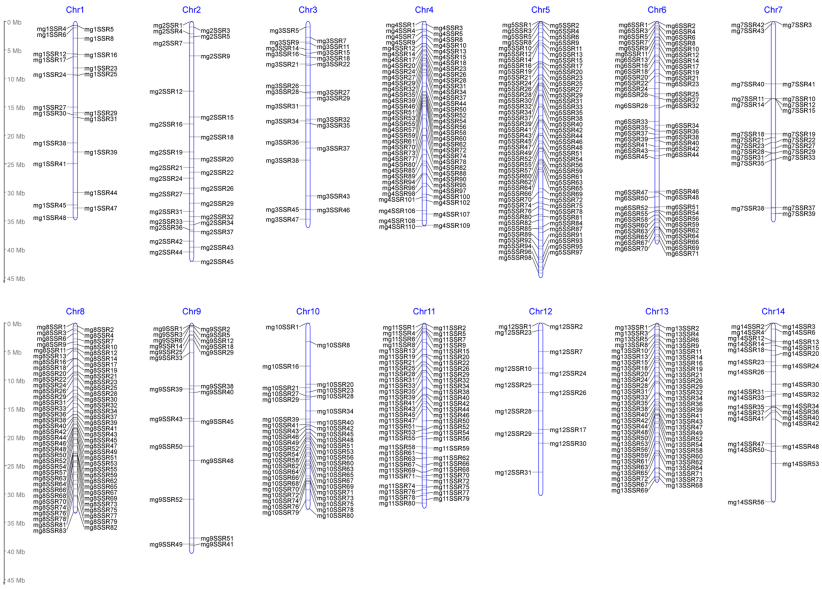


FIGURE 4
Density distribution map of polymorphic SSR markers on chromosomes in *Macadamia integrifolia*.

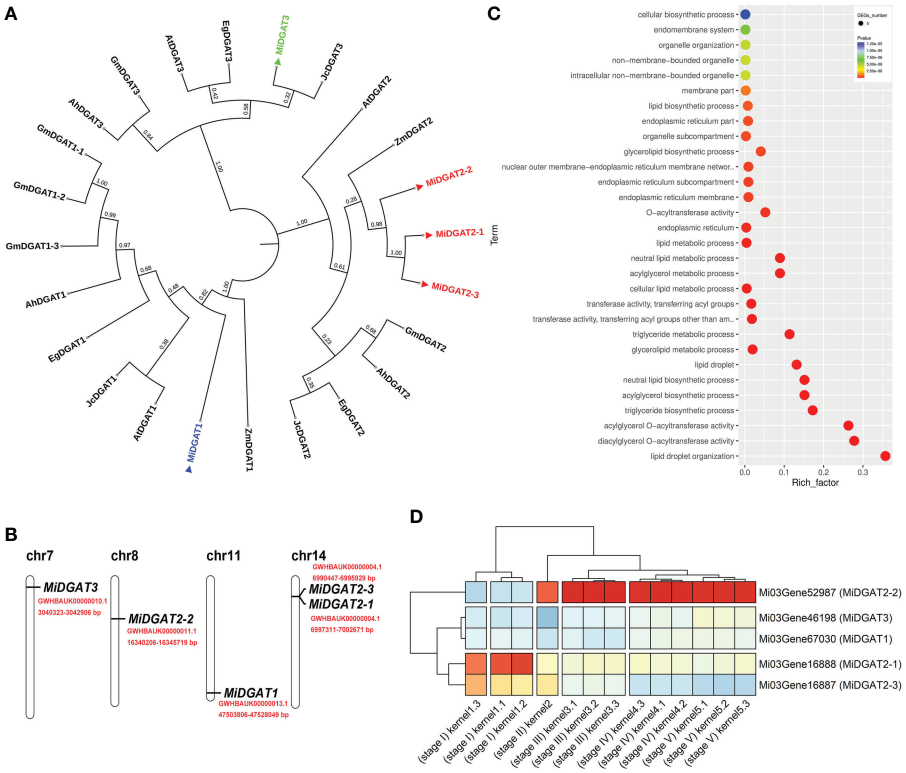


FIGURE 5
A case study for the application of MacadamiaGGD. **(A)**, Phylogenetic analysis of macadamia MiDGATs and DGATs from other plants. The phylogenetic tree was constructed *via* the neighbor-joining method and 1000 bootstraps by the software MEGA 11 (https://megasoftware.net/). The tree was visualized by iTOL (https://itol.embl.de/). Macadamia MiDGAT proteins and their sequence accessions are MiDGAT1 (Mi03Gene67030), MiDGAT2-1 (Mi03Gene16888), MiDGAT2-2 (Mi03Gene52987), MiDGAT2-3 (Mi03Gene16887) and MiDGAT3 (Mi03Gene46198) from *Macadamia integrifolia*. The proteins and their sequence accessions from other plants are AtDGAT1 (NP_179535), AtDGAT2 (AEE78802) and AtDGAT3 (Q9C5W0.2) from *Arabidopsis thaliana*, GmDGAT1-1 (NP_001237289), GmDGAT1-2 (NP_001237684.2), GmDGAT1-3 (NP_001242457.1), GmDGAT2 (NP_001299586.1) and GmDGAT3 (XP_003542403.1) from *Glycine max*, AhDGAT1 (AGT57761.1), AhDGAT2 (AEO11788.1) and AhDGAT3 (AAX62735.1) from *Arachis hypogaea*, JcDGAT1 (NP_001292926), JcDGAT2 (NP_001292973) and JcDGAT3 (XP_012083005.1) from *Jatropha curcas*, ZmDGAT1 (NP_001349157.1) and ZmDGAT2 (AQL03438.1) from *Zea mays*, and EgDGAT1 (XP_039165824.1), EgDGAT2 (XP_010033619.2) and EgDGAT3 (XP_010024878.2) from *Eucalyptus grandis*. **(B)**, Distribution of *MiDGAT* genes within the macadamia genome. The chromosome number is indicated at the top of each chromosome. The red font indicates the specific physical position of the genes. **(C)**, GO enrichment of *MiDGATs*. **(D)**, Expression pattern of *MiDGATs* at different developmental stages of macadamia kernels. The transcripts per million (TPM) values of expression levels are graphically represented by the Pheatmap package (R 4.2.0).

Discussion

The macadamia database MacadamiaGGD serves as an integrated germplasm and genomic research platform that can facilitate the genomic research and molecular breeding of macadamia. MacadamiaGGD integrates the currently published macadamia datasets of genomes, genetic maps, molecular markers, and morphological data of four macadamia species. MacadamiaGGD consists of 11 functional modules: Home, Germplasm, Genomes, Expression, BLAST, Markers, Maps, Tools, References, Download and Help.

Compared to other existing genome databases, the MacadamiaGGD provides a more comprehensive database and tools to characterize germplasms and genes of macadamia species. For example, “Phylogenetic Analysis”, which is integrated in the Germplasm module of MacadamiaGGD, was not included in the Citrus Genome Database (CGD, https://www.citrusgenomedb.org/), the Rice Genome Hub (RGH, https://rice-genome-hub.southgreen.fr) (Droc et al., 2019), the Kiwifruit Genome Database (KGD; http://kiwifruitgenome.org/) (Yue et al., 2020), and the functional genomics database for cannabis (CannabisGDB, https://gdb.supercann.net) (Cai et al., 2021). Databases of two kinds of molecular markers, SSR and SNP, are included in MacadamiaGGD, but not available in RGH, KGD, CannabisGDB, and the Gossypium Resource and Network Database (GRAND, http://grand.cricaas.com.cn) (Zhang et al., 2022). And MacadamiaGGD provides genetic linkage maps of nine genotypes, whereas genetic linkage maps are not available in KGD, CannabisGDB, and GRAND. Given the comprehensive information, interactive nature, and user-friendly database, MacadamiaGGD makes it easy to retrieve genomic information of macadamia. Thus, MacadamiaGGD not only provides a convenient way for researchers to understand and acquire basic germplasm and genomic information but also can largely help advance the molecular breeding of macadamia in the future.

The macadamia genome was used for the exploration of the SSR motifs, which were found to be evenly distributed across all 14 chromosomes. However, the percentage of the three SSR motifs was different, among which DNRs accounted for 79%, TNRs accounted for 18%, and TTRs accounted for 3%. This pattern is consistent with that in *Myrica rubra* (Jiao et al., 2012), in which DNRs were dominant. In this study, 927 primer pairs were designed for the verification of SSR locus polymorphisms, among which 605 primer pairs were found to be polymorphic. The density of microsatellite distribution was approximately 1.17 SSRs/Mb on 14 chromosomes, which was much higher than that in previous studies (Nock et al., 2014b). The main reason for this discrepancy may be due to the differences in genome quality and the SSR prediction method. In summary, we developed the first database of macadamia germplasm, genome, and genome-based SSR marker information, which will facilitate the molecular breeding of macadamia.

Conclusion

In conclusion, we developed the first comprehensive macadamia germplasm and genomic database MacadamiaGGD, which could serve as a central portal for macadamia species. MacadamiaGGD integrates data from germplasm, genomes, transcriptomes, genetic linkage maps, and SSR markers from various macadamia species. MacadamiaGGD also provides a group of user-friendly modules that enable users worldwide to efficiently retrieve and analyze genomic data. At present, MacadamiaGGD is in its first version but will be updated in a timely manner when new macadamia germplasm and omics data are available or published. We believe that MacadamiaGGD not only will broaden the understanding of the germplasm, genetics and genomics of macadamia species but also will facilitate the molecular breeding of macadamia.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Author contributions

Z-FX and JN designed the research. PW, YM, YW, YF, and JH collected and processed genomic and germplasm data. YM and PW developed the SSRs. PW, Z-FX and JN wrote the first draft of this manuscript. All authors contributed to the edit of this manuscript and the construction of MacadamiaGGD. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by start-up research funds from the Guangxi University.

Acknowledgments

We would like to thank all the macadamia researchers who have created valuable data resources collected in McadamiaGGD. Thanks to Mr. Zheng Yin and Mr. Zequn Zheng (VGsoft Team, China) for assisting in the website construction.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1007266/full#supplementary-material

References

Alam, M., Neal, J., O'connor, K., Kilian, A., and Topp, B. (2018). Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. *PLoS One* 13 (8), e0203465. doi: 10.1371/journal.pone.0203465

Allan, P. (2007). Evaluation and identification of old and new macadamia cultivars and selections at pietermaritzburg. *South Afr. J. Plant Soil* 24 (2), 124–129. doi: 10.1080/02571862.2007.10634793

Aradhya, M. K., Yee, L. K., Zee, F. T., and Manshardt, R. M. (1998). Genetic variability in macadamia. *Genet. Resour. Crop Evol.* 45 (1), 19–32. doi: 10.1023/A:1008634103954

Arroyo-Caro, J. M., Manas-Fernandez, A., Alonso, D. L., and Garcia-Maroto, F. (2016). Type I diacylglycerol acyltransferase (MtDGAT1) from macadamia tetraphylla: cloning, characterization, and impact of iits heterologous expression on triacylglycerol composition in yeast. *J. Agric. Food Chem.* 64 (1), 277–285. doi: 10.1021/acs.jafc.5b04805

Cai, S., Zhang, Z., Huang, S., Bai, X., Huang, Z., Zhang, Y. J., et al. (2021). CannabisGDB: a comprehensive genomic database for *Cannabis sativa* L. *Plant Biotechnol. J.* 19 (5), 857–859. doi: 10.1111/pbi.13548

Doyle, J. (1991). "DNA Protocols for plants," in *Molecular techniques in taxonomy*. Eds. G. M. Hewitt, A. W. B. Johnston and J. P. W. Young (Berlin, Heidelberg: Springer Berlin Heidelberg) 57, 283–293. doi: 10.1007/978-3-642-83962-7_18

Droc, G., Dereeper, A., Ruiz, M., Antoine, C., Barca, M., and Tranchant-Dubreuil, C. (2019). The South green rice genome hub. *Plant and Animal Genome* 17, 1. doi: agritrop.cirad.fr/590957/

Gao, Y., Sun, Y., Gao, H., Chen, Y., Wang, X., Xue, J., et al. (2021). Ectopic overexpression of a type-II DGAT (CeDGAT2-2) derived from oil-rich tuber of *Cyperus esculentus* enhances accumulation of oil and oleic acid in tobacco leaves. *Bioproducts* 14 (1), 76. doi: 10.1186/s13068-021-01928-8

Gitonga, L. N., Muigai, A. W. T., Kahangi, E. M., and Ngamau, K. (2009). Status of macadamia production in Kenya and the potential of biotechnology in enhancing its genetic improvement. *J. Plant Breed. Crop Sci.* 1 (3), 049–059. doi: 10.5897/jpbcs.9000122

Hardner, C. M. (2016). Macadamia domestication in hawai'i. *Genet. Resour. Crop Evol.* 63 (8), 1411–1430. doi: 10.1007/s10722-015-0328-1

Hardner, C. M., Peace, C., Lowe, A. J., Neal, J., Pisanu, P., Powell, M., et al. (2009). "Genetic resources and domestication of macadamia," in J. Janick (Ed.) *Horticultural Reviews*, vol. 35, (Hoboken, New Jersey: John Wiley & Sons, Inc) pp. 1–125. doi: 10.1002/9780470593776.ch1

He, X. (2008). *Studies on genetic diversity of macadamia (Macadamia spp.) germplasm resources* (Southwest University: Master).

Jiao, Y., Jia, H., Li, X., Chai, M., Jia, H., Chen, Z., et al. (2012). Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (Myrica rubra). *BMC Genomics* 13 (1), 201. doi: 10.1186/1471-2164-13-201

Langdon, K. S., King, G. J., Baten, A., Mauleon, R., Bundock, P. C., Topp, B. L., et al. (2020). Maximising recombination across macadamia populations to generate linkage maps for genome anchoring. *Sci. Rep.* 10 (1), 5048. doi: 10.1038/s41598-020-61708-6

Langdon, K. S., King, G. J., and Nock, C. J. (2019). DNA Paternity testing indicates unexpectedly high levels of self-fertilisation in macadamia. *Tree Genet. Genomes* 15 (2), 29. doi: 10.1007/s11295-019-1336-7

Lin, J., Zhang, W., Zhang, X., Ma, X., Zhang, S., Chen, S., et al. (2022). Signatures of selection in recently domesticated macadamia. *Nat. Commun.* 13 (1), 242. doi: 10.1038/s41467-021-27937-7

Liu, J., Niu, Y. F., Ni, S. B., He, X. Y., and Shi, C. (2017). Complete chloroplast genome of a subtropical fruit tree *Macadamia ternifolia* (Proteaceae). *Mitochondrial DNA Part B Resour.* 2 (2), 738–739. doi: 10.1080/23802359.2017.1390401

Liu, J., Niu, Y. F., Ni, S. B., He, X. Y., Zheng, C., Liu, Z. Y., et al. (2018). The whole chloroplast genome sequence of *Macadamia tetraphylla* (Proteaceae). *Mitochondrial DNA B Part Resour.* 3 (2), 1276–1277. doi: 10.1080/23802359.2018.1532836

Li, Q., and Wan, J. M. (2005). SSRHunter: development of a local searching software for SSR sites. *Yi Chuan = Hereditas* 27 (5), 808–810. doi: 10.16288/j.yczz.2005.05.024

Machado Neto, N. B., and Moryia, A. T. (2010). Variability in *Macadamia integrifolia* by RAPD markers. *Crop Breed. Appl. Biotechnol.* 10 (3), 266–270. doi: 10.1590/s1984-70332010000300013

Mai, T. T. P., Alam, M. M., Hardner, C. M., Henry, R. J., and Topp, B. L. (2020). Genetic structure of wild germplasm of macadamia: species assignment, diversity and phylogeographic relationships. *Plants (Basel)* 9 (6), 741. doi: 10.3390/plants9060714

Mai, T. T. P., Hardner, C. M., Alam, M. M., Henry, R. J., and Topp, B. L. (2021). Phenotypic characterisation for growth and nut characteristics revealed the extent of genetic diversity in wild macadamia germplasm. *Agriculture* 11 (7), 680. doi: 10.3390/agriculture11070680

Mast, A. R., Willis, C. L., Jones, E. H., Downs, K. M., and Weston, P. H. (2008). A smaller macadamia from a more vagile tribe: inference of phylogenetic relationships, divergence times, and diaspore evolution in macadamia and relatives (tribe macadamieae; proteaceae). *Am. J. Bot.* 95 (7), 843–870. doi: 10.3732/ajb.0700006

Moncur, M. W., Stephenson, R. A., and Trochoulias, T. (1985). Floral development of *Macadamia integrifolia* maiden & betche under Australian conditions. *Scientia Hortic.* 27 (1), 87–96. doi: 10.1016/0304-4238(85)90058-5

Moodley, R., Kindness, A., and Jonnalagadda, S. B. (2007). Elemental composition and chemical characteristics of five edible nuts (almond, Brazil, pecan, macadamia and walnut) consumed in Southern Africa. *J. Environ. Sci. Health Part B* 42, 585–591. doi: 10.1080/03601230701391591

Murigneux, V., Rai, S. K., Furtado, A., Bruxner, T. J. C., Tian, W., Harliwong, I., et al. (2020). Comparison of long-read methods for sequencing and assembly of a plant genome. *Gigascience* 9 (12), 1–11. doi: 10.1093/gigascience/giaa146

Nagao, M. A., Hirae, H. H., and Stephenson, R. A. (1992). Macadamia: cultivation and physiology. *Crit. Rev. Plant Sci.* 10 (5), 441–470. doi: 10.1080/07352689209382321

Niu, Y., Li, G., Ni, S., He, X., Zheng, C., Liu, Z., et al. (2022a). The chromosome-scale reference genome of *Macadamia tetraphylla* provides insights into fatty acid biosynthesis. *Front. Genet.* 13, 835363. doi: 10.3389/fgene.2022.835363

Niu, Y., Lu, Y., Song, W., He, X., Liu, Z., Zheng, C., et al. (2022b). Assembly and comparative analysis of the complete mitochondrial genome of three macadamia species *(M. integrifolia, M. ternifolia* and *M. tetraphylla). PloS One* 17 (5), e0263545. doi: 10.1371/journal.pone.0263545

Nock, C. J., Baten, A., Barkla, B. J., Furtado, A., Henry, R. J., and King, G. J. (2016). Genome and transcriptome sequencing characterises the gene space of *Macadamia integrifolia* (Proteaceae). *BMC Genomics* 17 (1), 937. doi: 10.1186/s12864-016-3272-3

Nock, C. J., Baten, A., and King, G. J. (2014a). Complete chloroplast genome of *Macadamia integrifolia* confirms the position of the gondwanan early-diverging eudicot family proteaceae. *BMC Genomics* 15 (9), S13. doi: 10.1186/1471-2164-15-S9-S13

Nock, C. J., Baten, A., Mauleon, R., Langdon, K. S., Topp, B., Hardner, C., et al. (2020). Chromosome-scale assembly and annotation of the macadamia genome (*Macadamia integrifolia* HAES 741). *G3-Genes Genomes Genet.* 10 (10), 3497–3504. doi: 10.1534/g3.120.401326

Nock, C. J., Elphinstone, M. S., Ablett, G., Kawamata, A., Hancock, W., Hardner, C. M., et al. (2014b). Whole genome shotgun sequences for microsatellite discovery and application in cultivated and wild macadamia (Proteaceae). *Appl. Plant Sci.* 2 (4), 1300089. doi: 10.3732/apps.1300089

O'Connor, K., Hayes, B., Hardner, C., Alam, M., and Topp, B. (2019a). Selecting for nut characteristics in macadamia using a genome-wide association study. *Hortic. Sci.* 54 (4), 629–632. doi: 10.21273/hortsci13297-18

O'Connor, K., Hayes, B., Hardner, C., Nock, C., Baten, A., Alam, M., et al. (2020). Genome-wide association studies for yield component traits in a macadamia breeding population. *BMC Genomics* 21 (1), 199. doi: 10.1186/s12864-020-6575-3

O'Connor, K., Kilian, A., Hayes, B., Hardner, C., Nock, C., Baten, A., et al. (2019b). Population structure, genetic diversity and linkage disequilibrium in a macadamia breeding population using SNP and silicoDArT markers. *Tree Genet. Genomes* 15 (2), 24. doi: 10.1007/s11295-019-1331-z

Peace, C. P., Allan, P., Vithanage, V., Turnbull, C. N., and Carroll, B. J. (2005). Genetic relationships amongst macadamia varieties grown in South Africa as assessed by RAF markers. *South Afr. J. Plant Soil* 22 (2), 71–75. doi: 10.1080/02571862.2005.10634684

Peace, C. P., Vithanage, V., Neal, J., Turnbull, C. G. N., and Carroll, B. J. (2004). A comparison of molecular markers for genetic analysis of macadamia. *J. Hortic. Sci. Biotechnol.* 79 (6), 965–970. doi: 10.1080/14620316.2004.11511874

Peace, C. P., Vithanage, V., Turnbull, C. G. N., and Carroll, B. J. (2002). Characterising macadamia germplasm with codominant radiolabelled DNA amplification fingerprinting (RAF) markers. *Acta Hortic.* 575, 371–380. doi: 10.17660/ActaHortic.2002.575.42

Peace, C. P., Vithanage, V., Turnbull, C. G. N., and Carroll, B. J. (2003). A genetic map of macadamia based on randomly amplified DNA fingerprinting (RAF) markers. *Euphytica* 134 (1), 17–26. doi: 10.1023/A:1026190529568

Project, R. A. (2007). The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36 (suppl_1), D1028–D1033. doi: 10.1093/nar/gkm978

Ranketse, M., Hefer, C. A., Pierneef, R., Fourie, G., and Myburg, A. A. (2022). Genetic diversity and population structure analysis reveals the unique genetic composition of South African selected macadamia accessions. *Tree Genet. Genomes* 18 (2), 15. doi: 10.1007/s11295-022-01543-0

SAMAC (2020)Statistics of the Southern African macadamia industry. In: *Macadamias South Africa*. Available at: https://www.samac.org.za/industry-statistics/ (Accessed 11/05 2020).

Schmidt, A. L., Scott, L., and Lowe, A. J. (2006). Isolation and characterization of microsatellite loci from macadamia. *Mol. Ecol. Notes* 6 (4), 1060–1063. doi: 10.1111/j.1471-8286.2006.01434.x

Sedgley, M. (1983). Pollen tube growth in macadamia. *Scientia Hortic.* 18 (4), 333–341. doi: 10.1016/0304-4238(83)90015-8

Sharma, P., Al-Dossary, O., Alsubaie, B., Al-Mssallem, I., Nath, O., Mitter, N., et al. (2021a). Improvements in the sequencing and assembly of plant genomes. *Gigabyte* 2021, 1–10. doi: 10.46471/gigabyte.24

Sharma, P., Masouleh, A. K., Topp, B., Furtado, A., and Henry, R. J. (2021b). De novo chromosome level assembly of a plant genome from long read sequence data. *Plant J.* 109 (3), 727–736. doi: 10.1111/tpj.15583

Sharma, P., Murigneux, V., Haimovitz, J., Nock, C. J., Tian, W., Kharabian Masouleh, A., et al. (2021c). The genome of the endangered *Macadamia jansenii* displays little diversity but represents an important genetic resource for plant breeding. *Plant Direct* 5 (12), e364. doi: 10.1002/pld3.364

Steiger, D. L., Moore, P. H., Zee, F., Liu, Z., and Ming, R. (2003). Genetic relationships of macadamia cultivars and species revealed by AFLP markers. *Euphytica* 132 (3), 269–277. doi: 10.1023/A:1025025522276

Storey, W. B., and Hamilton, R. A. (1953). "The macadamia nut industry in Hawaii," in *California Avocado Society 1953-54 Yearbook* 38, 63–67. Available at: http://avocadosource.com/CAS_Yearbooks/CAS_38_1953-54/CAS_1953-54_PG_063-067.pdf

Tang, Y., Yang, X., Cai, Y., Li, M., Zeng, L., Zheng, W., et al. (2018). Optimization and application of SSR-PCR reaction system for macadamia. *Fujian J. Agric. Sci.* 33 (2), 154–158. doi: 10.19303/j.issn.1008-0384.2018.02.009

Tan, Q., Wang, W., Chen, H., Wei, Y., Zheng, S., Huang, X., et al. (2020). Genetic diversity analysis of macadamia varieties based on single-nucleotide polymorphism. *Mol. Plant Breed.* 18 (21), 7246–7253. doi: 10.13271/j.mpb.018.007246

Tan, Q., Wang, W., Wei, Y., Zheng, S., Huang, X., He, X., et al. (2019). Diversity analysis of fruit traits related to yield in macadamia germplasms. *J. Fruit Sci.* 36 (12), 1630–1637. doi: 10.13925/j.cnki.gsxb.20190087

Tan, Q., Wei, Y., Huang, X., Zhang, T., Xu, P., Song, H., et al. (2021). Analysis of fruit characteristics and nutrients of 10 accessions of *Macadamia integrifolia*. *J. Fruit Sci.* 38 (5), 672–680. doi: 10.13925/j.cnki.gsxb.20200372

Toft, B. D., Alam, M., and Topp, B. (2018). Estimating genetic parameters of architectural and reproductive traits in young macadamia cultivars. *Tree Genet. Genomes* 14 (4), 50. doi: 10.1007/s11295-018-1265-x

Topp, B. L., Nock, C. J., Hardner, C. M., Alam, M., and O'connor, K. M. (2019). "Macadamia (*Macadamia* spp.) breeding," in *Advances in plant breeding strategies: Nut and beverage crops: Volume 4*. Eds. J. M. Al-Khayri, S. M. Jain and D. V. Johnson(Cham: Springer Nature Switzerland AG), 221–251. doi: 10.1007/978-3-030-23112-5_7

Trueman, S. J. (2013). The reproductive biology of macadamia. *Scientia Hortic.* 150, 354–359. doi: 10.1016/j.scienta.2012.11.032

Urata, U. (1954). *Pollination requirements of macadamia* (Honolulu (HI: Technical Bulletin), 1–40.

Vithanage, V., Hardner, C., Anderson, K. L., Meyers, N., Mcconchie, C., and Peace, C. (1998). Progress made with molecular markers for genetic improvement of macadamia. *Acta Hortic* 461, 199–208. doi: 10.17660/ActaHortic.1998.461.20H

Vithanage, V., and Winks, C. W. (1992). Isozymes as genetic markers for macadamia. *Scientia Hortic.* 49 (1), 103–115. doi: 10.1016/0304-4238(92)90147-5

Yue, J., Liu, J., Tang, W., Wu, Y. Q., Tang, X., Li, W., et al. (2020). Kiwifruit genome database (KGD): a comprehensive resource for kiwifruit genomics. *Horticult. Res.* 7, 117. doi: 10.1038/s41438-020-0338-9

Zeng, H., and Du, L. (2017). *Illustrated guide to identication of macadamia cultivars* (Beijing: China Agriculture Press).

Zhang, M. (2011). *Study on the Fruit Main Ingredients of the Twenty-eight Macadamia Germplasms* (Hainan University, China: Master Thesis).

Zhang, Z., Chai, M., Yang, Z., Yang, Z., and Fan, L. (2022). GRAND: an integrated genome, transcriptome resources, and gene network database for gossypium. *Front. Plant Sci.* 13, 773107. doi: 10.3389/fpls.2022.773107

Zhang, T. T., He, H., Xu, C. J., Fu, Q., Tao, Y. B., Xu, R., et al. (2021). Overexpression of type 1 and 2 diacylglycerol acyltransferase genes (JcDGAT1 and JcDGAT2) enhances oil production in the woody perennial biofuel plant *Jatropha curcas*. *Plants (Basel)* 10 (4), 699. doi: 10.3390/plants10040699

TYPE Original Research
PUBLISHED 14 November 2022
DOI 10.3389/fpls.2022.975942

Building a foundation for gene family analysis in Rosaceae genomes with a novel workflow: A case study in *Pyrus* architecture genes

Huiting Zhang[1,2], Eric K. Wafula[3], Jon Eilers[1], Alex E. Harkess[4,5], Paula E. Ralph[3], Prakash Raj Timilsena[3], Claude W. dePamphilis[3], Jessica M. Waite[1] and Loren A. Honaas[1]*

[1]Tree Fruit Research Laboratory, Agricultural Research Service (ARS), United States Department of Agriculture (USDA), Wenatchee, WA, United States, [2]Department of Horticulture, Washington State University, Pullman, WA, United States, [3]Department of Biology, The Pennsylvania State University, University Park, PA, United States, [4]College of Agriculture, Auburn University, Auburn, AL, United States, [5]HudsonAlpha Institute for Biotechnology, Huntsville, AL, United States

OPEN ACCESS

EDITED BY
Riccardo Velasco,
Research Centre of Viticulture and Oenology (CREA), Italy

REVIEWED BY
Andrea Zuccolo,
King Abdullah University of Science and Technology, Saudi Arabia
Alessandro Cestaro,
Fondazione Edmund Mach, Italy
Douglas W. Bryant
NewLeaf Symbiotics, St. Louis, United States

*CORRESPONDENCE
Loren A. Honaas
loren.honaas@usda.gov

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 22 June 2022
ACCEPTED 21 September 2022
PUBLISHED 14 November 2022

CITATION
Zhang H, Wafula EK, Eilers J, Harkess AE, Ralph PE, Timilsena PR, dePamphilis CW, Waite JM and Honaas LA (2022) Building a foundation for gene family analysis in Rosaceae genomes with a novel workflow: A case study in *Pyrus* architecture genes.
Front. Plant Sci. 13:975942.
doi: 10.3389/fpls.2022.975942

The rapid development of sequencing technologies has led to a deeper understanding of plant genomes. However, direct experimental evidence connecting genes to important agronomic traits is still lacking in most non-model plants. For instance, the genetic mechanisms underlying plant architecture are poorly understood in pome fruit trees, creating a major hurdle in developing new cultivars with desirable architecture, such as dwarfing rootstocks in European pear (*Pyrus communis*). An efficient way to identify genetic factors for important traits in non-model organisms can be to transfer knowledge across genomes. However, major obstacles exist, including complex evolutionary histories and variable quality and content of publicly available plant genomes. As researchers aim to link genes to traits of interest, these challenges can impede the transfer of experimental evidence across plant species, namely in the curation of high-quality, high-confidence gene models in an evolutionary context. Here we present a workflow using a collection of bioinformatic tools for the curation of deeply conserved gene families of interest across plant genomes. To study gene families involved in tree architecture in European pear and other rosaceous species, we used our workflow, plus a draft genome assembly and high-quality annotation of a second *P. communis* cultivar, 'd'Anjou.' Our comparative gene family approach revealed significant issues with the most recent 'Bartlett' genome - primarily thousands of missing genes due to methodological bias. After correcting assembly errors on a global scale in the 'Bartlett' genome, we used our workflow for targeted improvement of our genes of interest in both *P. communis* genomes, thus laying the groundwork for future functional studies in pear tree architecture. Further, our global gene family classification of 15 genomes across 6 genera provides a valuable and previously unavailable resource for the Rosaceae research community. With it, orthologs and other gene family members can be easily identified across any of the classified genomes. Importantly, our workflow can be easily adopted for any other plant genomes and gene families of interest.

KEYWORDS

tree architecture gene, gene family, comparative genomics, targeted genome re-annotation, European pear genome, Rosaceae, PlantTribes2

1 Introduction

Advancements in plant genome sequencing and assembly have vigorously promoted research in non-model organisms. In horticultural species, new genome sequences are being released every month (Chen et al., 2021a; Chen et al., 2021b; Wang et al., 2021a, Wang et al., 2021b; Xu et al., 2021). These genomes have broadened our understanding of targeted cultivars and provided fundamental genomic resources for molecular breeding and more in-depth studies of economically important crop traits such as those involved in plant architecture. Although many gene families have been identified as important for architectural traits, such as dwarfing, weeping, and columnar growth (Hill and Hollender, 2019), the study of these genes and their functionality in new species is still hampered by inaccurate information about their gene models or domain structures, and the frequent lack of 1:1 orthology between related genes of different species. Sequencing and annotating a diversity of related genomes are crucial steps for obtaining this level of information.

Crops, most of which have gone through more than ten thousand years of domestication to meet human requirements, have a wide diversity in forms, sometimes even within the same species (Stansell and Björkman, 2020). One such example is in the *Brassica* species, where *B. rapa* encompasses morphologically diverse vegetables such as Chinese cabbage, turnips, and mizuna; and cabbage, stem kale, and Brussels sprouts are the same biological species, *B. oleracea*. Therefore, a single reference genome does not represent the complex genome landscape, or pan-genome, for a single crop species. To understand the genetic basis of the diverse *Brassica* morphotypes, many attempts have been made to explore the genomes of *Brassica* (Cheng et al., 2016a, Cheng et al., 2016b; Stansell et al., 2018; Stansell and Björkman, 2020; Mabry et al., 2021). In one of those attempts, genomes from 199 *B. rapa* and 119 *B. oleracea* accessions were sequenced and analyzed using a comparative genomic framework (Cheng et al., 2016a, Cheng et al., 2016b). Genomic selection signals and candidate genes were identified for traits associated with leaf-heading and tuber-forming morphotypes. Compared to *Brassica*, pome fruits may not appear to have as much diversity in their vegetative appearance, but they do have great diversity in terms of fruit quality, rootstock growth and performance, and post-harvest physiology. However, genome studies and pan-genome scale investigations in pome fruits are still in their infancy. In cultivated apple (*Malus domestica*), genomes of four different cultivars (Velasco et al., 2010; Daccord et al., 2017; Zhang et al., 2019; Sun et al., 2020b; Khan et al., 2022) have been published, providing resources to study: (1) small (SNPs and small InDels) and large scale (chromosome rearrangements) differences that can help explain cultivar diversity, and (2) gene content differences that may contribute to cultivar specific traits. However, genomic resources for European pear (*Pyrus communis*) cultivars are limited to just two published genomes (Chagné et al., 2014; Linsmith et al., 2019) from a single cultivar, ‘Bartlett’. More European pear genomes will afford new perspectives that help us understand shared and unique traits for important cultivars in *Pyrus*, as well as other Rosaceae.

Besides understanding large scale genomic characteristics, new genomes also provide rich resources for reverse genetic studies (Tollenaere et al., 2012; Wu et al., 2012). To obtain the actual sequence of a target gene, reverse genetic approaches in the pre-genome era relied on sequence and domain homology and technologies such as RACE PCR (Takos et al., 2006), which could be challenging and time consuming. Alternatively, in species with high-quality reference genomes, the annotation is generally considered to contain all the genes and target genes that could ideally be identified with a sequence similarity search (*i.e.*, BLAST). However, reports of annotation errors, such as imperfect gene models and missing functional genes are very common (Marx et al., 2016; Pertea et al., 2018; Pilkington et al., 2018). Another complicating factor is that duplication events (*i.e.*, whole genome duplication, regional tandem duplication) and polyploidy occur in the majority of flowering plants, including most crop species, posing substantial challenges to genome assembly and annotation (Kyriakidou et al., 2018). Moreover, instances of neofunctionalization and subfunctionalization occur frequently following duplication events (Hughes et al., 2014), which sometimes will result in large and complex gene families (Yang et al., 2015; Yoshida et al., 2019). Therefore, a one-to-one relationship between a gene in a model organism and its ortholog in other plant species, or even between closely related species and varieties, is rare (Xiao et al., 2013). Without understanding the orthology and paralogy between members of a given gene family, it is difficult to translate knowledge of a gene in a model organism to another species of interest.

In the present study, we assembled a draft genome for the European pear cultivar 'd'Anjou', improved the current 'Bartlett' assembly (*i.e.*, Bartlett.DH_V2), and developed a workflow that allows highly efficient target gene identification in any plant genome of interest. We used our workflow which iteratively curated and improved gene models for architecture-related genes from both the polished Bartlett.DH_v2 and the d'Anjou genomes. Importantly, we recovered many genes that were missing from gene families of interest (50 genes in the cultivar 'Bartlett') and corrected errors in others across the genus *Pyrus*. This work demonstrates that the integration of comparative genomics and phylogenomics can facilitate and enhance gene annotation, and thus gene discovery, in important plant reference genomes.

2 Materials and methods

2.1 Plant materials and sequencing

The 'd'Anjou' plants were purchased from Van Well's nursery in East Wenatchee, WA, USA and grown in the USDA ARS greenhouse #6 at Wenatchee, WA, USA. Fresh leaves (~1.5g) from one 'd'Anjou' plant were flash frozen and used for DNA extraction. A CTAB isolation protocol (Michiels et al., 2003) was used to generate high-molecular-weight genomic DNA with the following modifications: the extraction was performed at large-scale with 100 ml of extraction buffer in a 250 ml Nalgene centrifuge bottle; the isopropanol precipitation was performed at room temperature (~ 5 minutes) followed immediately by centrifugation; after a 15-minute incubation in the first pellet wash solution, the pellet was transferred to a 50 ml centrifugation tube *via* sterile glass hook before performing the second pellet wash; following the second pellet wash, centrifugation, and air drying, the pellet was resuspended in 2 ml TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0) and allowed to resuspend at 4°C overnight. The concentration of the DNA was measured by a Qubit 2.0 fluorometer (Invitrogen) and 50 ug DNA was digested with RNase A (Qiagen, final concentration 10 ug/ml, 37°C for 30 minutes) and then further cleaned up using the PacBio recommended, user-shared gDNA clean-up protocol (https://www.pacb.com/search/?q=user+shared+protocols) performed at large-scale with the DNA sample brought up to 2 ml with TE and all other volumes scaled up accordingly. The final pellet was resuspended in 100 ul TE. The final DNA concentration was measured by Qubit fluorometer, and 500 ng was loaded onto a PFG (Bio-Rad CHEF) to check the size range. The DNA ranged in size from 15 Kb to 100 Kb with a mean fragment size around 50 Kb. The purity of the DNA as measured by the NanoDrop spectrophotometer (ThermoFisher) was 260/280 nm: 1.91; 260/230 nm: 2.51. Cleaned-up gDNA was sent to the Penn State Genomics Core facility (University Park, PA, USA) for Pacbio and Illumina library construction and sequencing. A total of 10 ug gDNA was used to construct PacBio SMRTbell libraries and sequenced on a PacBio Sequel system. A small subset of the same gDNA was used to make Illumina TruSeq library and was sequenced on an Illumina HiSeq 2500 platform. In addition, 4 ug of the same gDNA was sent to the DNA technologies and Expression Analysis Core Laboratory at UC Davis (Davis, CA, USA) to construct an Illumina 10X Chromium library, which was sequenced on an Illumina NovaSeq 6000 sequencer.

2.2 Genome assembly and post-assembly processing

To create the initial backbone assembly of d'Anjou, Canu assembler v2.1.1 (Koren et al., 2017) was used to correct and trim PacBio continuous long reads (CLR) followed by a hybrid assembly of Illumina short reads and PacBio CLR with MaSuRCA assembler v3.3.2 (Zimin et al., 2013). Next, Supernova v2.1.1, the 10x Genomics *de novo* assembler (Weisenfeld et al., 2017), was used to assemble linked-reads at five different raw read coverage depths of approximately 50x, 59x, 67x, 78x, and 83x based on the kmer estimated genome size, and the resulting phased assembly graph was translated to produce two parallel pseudo-haplotype sequence representations of the genome. The Supernova assembler can only handle raw data between 30- to 85-fold coverage of the estimated genome size. Therefore, the muti-coverage assemblies provide an opportunity to capture most of the genome represented in the ~234-fold coverage sequenced 10x Chromium read data. One of the pseudo-haplotypes at each of the five coverages was used for subsequent meta-assembly construction to improve the backbone assembly, and the quality of which was assessed using a combination of assembly metrics, including (1) contig and scaffold contiguity (L50), (2) completeness of conserved land plants (embryophyta_odb10) benchmarking universal single-copy orthologs (BUSCO v5.2.2) (Manni et al., 2021), and (3) an assembly size closer to the expected d'Anjou haploid genome size. The backbone assembly was incrementally improved by bridging gaps and joining contigs with the Quickmerge program (Chakraborty et al., 2016) using contigs from the five primary Supernova assemblies in decreasing order of assembly quality. The resulting meta-assembly at each merging step was only retained if improvement in contiguity, completeness, and assembly size was observed.

Next, the long-distance information of DNA molecules provided in linked-reads was used to correct errors introduced in the meta-assembly during both the *de novo* and merging steps of the assembly process with Tigmint (Jackman et al., 2018) and ARCS (Yeo et al., 2017). Tigmint aligns linked-reads to an assembly to identify and correct potential errors and breaks. The improved assembly is then re-scaffolded into highly contiguous sequences with ARCS using the long-distance information contained in the linked-reads. To further improve the d'Anjou meta-assembly, trimmed paired-reads from both the short insert Illumina and 10x Chromium libraries were used to iteratively fill gaps between contigs using GapFiller v1.10 (Boetzer and Pirovano, 2012), and correct base errors and local misassemblies with Pilon v1.23 (Walker et al., 2014). The genome assembly process is illustrated in Supplementary Figure 1.

2.3 Pseudomolecule construction

Before constructing the chromosomal-scale pseudomolecules, extraneous DNA sequences present in meta-assembly were identified and excluded (Supplementary Figure 1). Megablast searches with e-value < 1e-10 was performed against the NCBI nucleotide collection database (nt), and then the best matching Megablast hits (max_target_seqs = 100) against the NCBI taxonomy database were queried to determine their taxonomic attributions. Assembly sequences with all their best-matching sequences not classified as embryophytes (land plants) were considered contaminants and discarded. A second iteration of Megablast searches of all the remaining sequences (embryophytes) was performed against the NCBI RefSeq plant organelles database to identify chloroplast and mitochondrion sequences. Assembly sequences with high similarity (> 80% identity; > 50% coverage) to plant organelle sequences were discarded (Yoshida et al., 2019; Hämälä et al., 2021). Finally, the remaining meta-assembly contigs and scaffolds were ordered and oriented into chromosomal-scale pseudomolecules with RaGOO (Alonge et al., 2019) using the *Pyrus communis* Bartlett.DH_v2 (Linsmith et al., 2019) reference chromosomes (Supplementary Figure 1).

2.4 Assembly validation

The completeness of both the contig and scaffold assembly were evaluated by searching against the land plants (embryophyta_odb10) gene set with BUSCO v4 (Manni et al., 2021) (Supplementary Table 1). Synteny comparison between Bartlett.DH_v2 and d'Anjou meta-assembly were evaluated with D-GENIES (Cabanettes and Klopp, 2018) using repeat masked (http://www.repeatmasker.org) DNA alignments generated by minimap2 (Li et al., 2016b). Synteny results of the whole genome and each of the 17 *Pyrus communis* chromosomes are shown in Figure 1 and Supplementary Figure 2, respectively.

2.5 Gene prediction

Prior to protein-coding gene annotation, we first estimated and masked the repetitive sequences in the d'Anjou meta-assembly following the protocol described by (Campbell et al., 2014). The meta-assembly was first searched using MITE-Hunter (Han and Wessler, 2010) and LTRharvest/LTRdigest (Ellinghaus et al., 2008; Steinbiss et al., 2009) to collect consensus

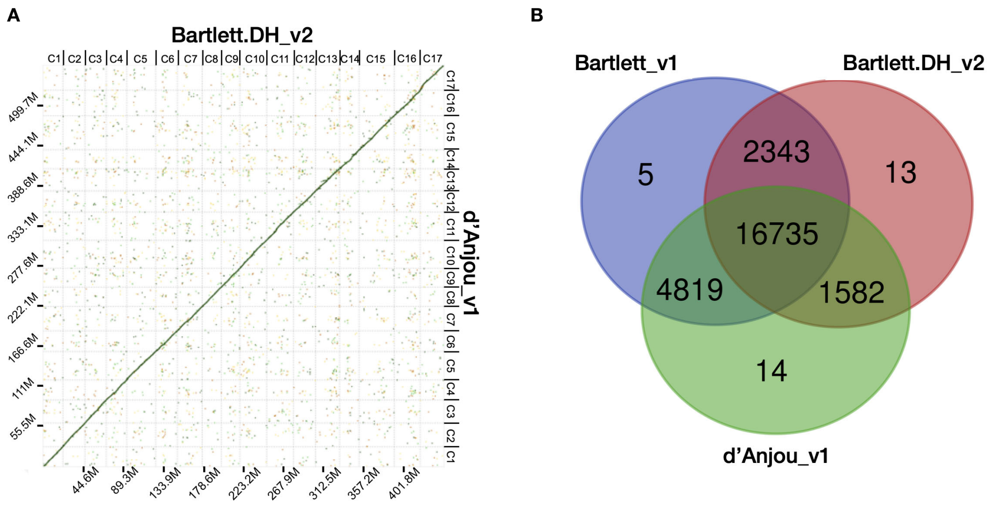


FIGURE 1
Characterization of the d'Anjou genome and protein orthology among European pears. **(A)** Dot plot of genome alignment of Bartlett.DH_v2 (x axis) and d'Anjou (y axis). **(B)** Overlap and distinctiveness of gene annotations among three *Pyrus communis* genotypes, Bartlett_v1, Bartlett.DH_v2, and d'Anjou.

miniature inverted-repeat transposable elements (MITEs) and long terminal repeat retrotransposons (LTRs), respectively. LTRs were filtered to remove false positives and elements with nested insertions. The cleaned LTRs were then used together with the MITEs to mask the genomes. The unmasked regions of the genomes were then annotated with RepeatModeler (http://www.repeatmasker.org/RepeatModeler) to predict additional *de novo* repetitive sequences. All collected repetitive sequences were compared to a BLAST database of plant proteins from SwissProt and RefSeq, and sequences with significant hits were excluded from the repeat masking library.

To supplement *ab initio* gene predictions, extensive extrinsic gene annotation homology evidence was collected, including (1) d'Anjou RNA-seq data from our previous study (Honaas et al., 2021); (2) homologous protein evidence of closely related species: *Malus domestica*, *Prunus persica*, *Pyrus betulifolia*, *Pyrus communis* 'Bartlett', *Pyrus* x *bretschneideri*, *Rosa chinensis*, and *Rubus occidentalis* retrieved from the Genome Database for Rosaceae (GDR) (Jung et al., 2018), and (3) protein sequences from the plant model species, *Arabidopsis thaliana* (Cheng et al., 2017).

Protein-coding gene annotations from the *Pyrus communis* reference genomes of Bartlett_v1 and Bartlett.DH_v2 were separately transferred (liftovers) to pseudomolecules of d'Anjou meta-assembly using the FLO (Pracana et al., 2017) (https://github.com/wurmlab/flo) pipeline based on the UCSC Genome Browser Kent-Toolkit (Kuhn et al., 2013). Next, repetitive and low complexity regions of the pseudomolecules were masked with RepeatMasker in the MAKER pipeline (release 3.01.02) (Cantarel et al., 2008) using the previously described d'Anjou-specific repeat library. Then, the MAKER pipeline updated the transferred annotations with gene annotation homology evidence (described above) and predicted additional protein coding genes with Augustus (Stanke et al., 2004; Hoff and Stanke, 2019) and SNAP (Korf, 2004). Only predicted gene models supported by annotation evidence, encode a Pfam domain, or both, were retained.

2.6 Computation of pear orthogroups

To compare the gene content of the three *Pyrus communis* genomes, Bartlett_v1, Bartlett.DH_v2, and d'Anjou, orthologous and paralogous protein clusters were estimated with OrthoFinder v1.1.8 (Emms and Kelly, 2015) from annotated proteins in all the genomes.

2.7 Bartlett.DH_v2 genome polishing

To improve the base quality of the publicly available pear reference genome, the *Pyrus communis* Bartlett.DH_v2 assembly was iteratively polished with two rounds of Pilon v1.24 (Walker et al., 2014) using the raw Illumina shotgun reads from the Bartlett.DH_v2 genome projects obtained from the NCBI Short Read Archive (SRA accessions: SRR10030340, SRR10030308), and completeness and accuracy assessed with the BUSCO v5.2.2 (Manni et al., 2021) embryophyta_odb10 database.

2.8 Gene family identification

Protein sequences of tree architecture candidate genes gleaned from published literature were sorted into pre-computed orthologous gene family clusters of 26 representative land-plant genomes (26Gv2.0) using the both BLASTp (Camacho et al., 2009) and HMMER hmmscan (Eddy, 2011) sequence search option of the *GeneFamilyClassifier* tool implemented in the PlantTribes 2 pipeline (https://github.com/dePamphilis/PlantTribes). Classification results of these architecture genes, including orthogroup taxa gene counts, corresponding superclusters (super orthogroups) at multiple clustering stringencies, and orthogroup-level annotations from multiple public biological functional databases are reported in Supplementary Table 2.

2.9 Gene family analysis

All the tools used in this process are modules from the command line version of PlantTribes 2 pipeline and are processed on SCINet (https://scinet.usda.gov/) with customized scripts Supplementary File 8. Protein coding genes from 14 Rosaceae genomes (*Fragaria vesca*, *Rosa chinensis*, *Rubus occidentalis*, *Prunus avium*, *Malus domestica* HFTH, *M. domestica* GDDH13, *M. domestica* Gala, *M. sieversii*, *M. sylvestris*, *Pyrus communis* Bartlett_v1, *Pyrus communis* Bartlett.DH, *Pyrus ussuriensis* x *communis*, *Pyrus bretschneideri*, *Pyrus communis* d'Anjou. Source of data and corresponding publications listed in Supplementary Table 3) were sorted into orthologous groups (26Gv2.0) with the *GeneFamilyClassifier* tool as previously described, after a quality control filtration using the *AssemblyPostProcessor* tool. A detailed summary of the Rosaceae gene family classification results are in Supplementary Table 3. Sequences classified into the orthogroups of interest (with candidate genes in this study) were integrated with scaffold backbone gene models using the *GeneFamilyIntegrator* tool. Gene names were modified as shown in Supplementary Table 4 for easier recognition of the species and cultivar. Amino acid multiple sequence alignments and their corresponding DNA codon alignments were generated by the *GeneFamilyAligner* tool with the L-INS-i algorithm implemented in MAFFT (Katoh et al., 2002). Sites present in less than 10% of the aligned DNA sequences were removed with trimAL (Capella-Gutiérrez et al., 2009). Maximum likelihood (ML) phylogenetic trees were estimated from the trimmed DNA alignments using the RAxML algorithm (Stamatakis, 2014) option in the *GeneFamilyPhylogenyBuilder* tool. One hundred bootstrap replicates (unless otherwise indicated) were conducted for each tree to estimate the reliability of the branches. The multiple sequence alignments were visualized in the Geneious R9 software (Kearse et al., 2012) with Clustal color scheme. The phylogeny was colored with a custom script and visualized with Dendroscope version 3.7.5 (Huson and Scornavacca, 2012). Gene sequences, alignments, and phylogenies are available in Supplementary Files 1–3.

2.10 Domain prediction

To estimate domain structures of proteins in each orthogroup, the predicted amino acid sequences (either obtained from public databases or generated by the PlantTribes *AssemblyPostProcessor* tool) were submitted to interproscan v5.44-79.0 (Jones et al., 2014) on SCINet and searched against all the databases.

2.11 Targeted gene family annotation

The following approaches were used in parallel to annotate candidate genes from the original Bartlett.DH_v2, the polished Bartlett.DH_v2, and the d'Anjou genome assemblies:

2.11.1 TGFam-finder

The 'RESOURCE.config' and 'PROGRAM_PATH.config' files were generated according to the author's instruction. The three targeted genome assemblies mentioned above were used as the *target genomes*. Complete protein sequences from apples and pears in the same orthogroup were used as *protein for domain identification*. Complete protein sequences from other Rosaceae species and *Arabidopsis thaliana* in the same orthogroup were used as *resource proteins* for each annotation step. For each orthogroup, Pfam annotations from the InterProScan results were used as *TSV for domain identification*. For orthogroups without Pfam descriptions, MobiDBLite information was used as *TSV for domain identification* (Kim et al., 2020).

2.11.2 *Bitacora*

Arabidopsis genes from targeted gene families (orthogroups of interest) were used to generate a multiple sequence alignment and HMM profile using MAFFT (Katoh et al., 2002) and hmmbuild (Eddy, 2011). The resulting files were then used as input for Bitacora v1.3, (Vizueta et al., 2020) running in both genome mode and full mode to identify genes of interest in the genome assemblies mentioned above.

2.12 Manual curation and gene model verification

In cases where both TGFam-Finder and Bitacora failed to predict a full-length gene, the gene model was curated manually.

2.12.1 Curation with orthologous gene models

First, the genomic region containing the target sequence was determined either by the general feature format file (gff) or a BLASTn search using the coding sequence of the target gene or a closely related gene as a query. Next, a genomic fragment containing the target sequence and 3kb upstream and downstream of the targeted region was extracted. Then, the incomplete transcript(s), predicted exons, and complete gene models from a closely related species were mapped to the extracted genomic region using Geneious R9 (Kearse et al., 2012) with the *Map to Reference* function. The final gene model was determined by using the full-length coding sequence of a closely related gene as a reference.

2.12.2 Curation with RNA-seq read mapping

The gff3 files obtained from Bitacora were loaded into an Apollo docker container v2.6.3 (Dunn et al., 2019) for verification of the predicted gene models using expression data. Publicly available RNA-seq data (Nham et al., 2015; Nham et al., 2017; Gabay et al., 2018; Zhang et al., 2018, Zhang et al., 2020; Hewitt et al., 2020) for *Pyrus* were used as inputs of an RNA-seq aligner, STAR v2.7.8a (Dobin et al., 2013), and alignments were performed with maximum intron size set to 5kb and default settings. Intron-exon structure was compared to the aligned expression data. If there was insufficient RNA-seq coverage from the targeted cultivar, data from other cultivars and *Pyrus* species were used as supporting evidence. Summaries of read mapping results are available in Supplementary Files 4, 5. Curated gene models from the original Bartlett.DH_v2 were transferred to the polished genome for validation.

Gene model cartoons were generated using the *visualize gene structure* function in TBtools v1.09854 (Chen et al., 2020). Final gene models and their corresponding chromosomal locations are available in Supplementary Files 6, 7.

3 Results

3.1 The draft d'Anjou genome

3.1.1 Genome assembly

We generated approximately 134 million paired-end reads from Illumina HiSeq and a total of 1,054,992 PacBio continuous long reads (CLR) with a read length N50 of 20 Kb, providing an estimated 67-fold and 21-fold coverage respectively of the expected 600 Mb *Pyrus communis* genome (Chagné et al., 2014). Additionally, approximately 468 million 2 x 150 bp paired reads (~234-fold coverage) with an estimated mean molecule length (linked-reads) of 20 kb were generated using 10x Genomics Chromium Technology (Supplementary Table 5). The final meta-assembly, generated with a combination of the three datasets, contains 5,800 scaffolds with a N50 of 358 Kb (Table 1). The cleaned contigs and scaffolds were ordered and oriented into 17 pseudochromosomes guided by the reference genome, *Pyrus communis* 'Bartlett.DH_v2' (Linsmith et al., 2019).

Next, we compared the d'Anjou meta-assembly to two published reference assemblies of Bartlett (Chagné et al., 2014; Linsmith et al., 2019) to assess assembly contiguity, completeness, and structural accuracy. The Benchmarking Universal Single-Copy Ortholog (BUSCO) (Manni et al., 2021) analysis showed that the d'Anjou genome captured 97.4% complete genes in the embryophyta_odb10 gene sets, comparable to the reference genomes (Table 1; Supplementary Table 6). Furthermore, synteny comparisons between the draft d'Anjou genome and the reference Bartlett.DH_v2 genome showed high collinearities at both whole-genome and chromosomal levels (Figure 1A; Supplementary Figure 2).

3.1.2 Annotation

Combining information such as *de novo* transcriptome assembly, homologous proteins of closely related species, and protein-coding gene annotations from the two 'Bartlett' genomes, we identified a total of 45,981 protein coding genes in d'Anjou (Table 1). Of those putative genes 76.63% were annotated with functional domains from Pfam (Mistry et al., 2020) and the remaining are supported by annotation evidence, primarily d'Anjou RNA-Seq reconstructed transcripts (Honaas et al., 2021). These results indicate that we captured a large majority of the gene space in the d'Anjou genome. This affords a range of analyses including gene and gene family characterization, plus global-scale comparisons with other Rosaceae species including the 'Bartlett' cultivar.

3.2 Comparison among three European pear genomes

To study the shared and genotype-specific genes among the three European pear genomes (Bartlett version1, Bartlett double haploid version 2, and d'Anjou version 1), we constructed 25,511 protein clusters (orthogroups), comprising 77.71% of all the genes. While numbers of predicted genes from the Bartlett_v1 and d'Anjou genomes may be overestimated due to the presence of alternative haplotype segments in the assembly caused by high heterozygosity (Linsmith et al., 2019), this should have very little effect on orthogroup circumscription. Further, the process of creating a double haploid reduces genome heterozygosity, but should retain estimates of orthogroup content. Hence, we formulated the following hypotheses: (1) a large majority of gene families are shared by all three genotypes; (2) few genotype-specific gene families are present in each genome; (3) the commercial 'Bartlett' genotype and the double haploid 'Bartlett' genotype (roughly version 1.0 and 2.0 of this genome, respectively) should have virtually identical gene family circumscriptions; and (4) we should detect very few gene families that are unique to either 'Bartlett' genome and shared with 'd'Anjou'. The protein clustering analysis results (Table 1; Figure 1B) support our hypotheses 1 and 2: 65.60% of the orthogroups contain genes from all three genotypes and only 0.12% of the orthogroups are species-specific. However, among the 8,744 orthogroups containing genes from two genotypes, more than half (55.11%) are shared

TABLE 1 Comparison of genome assembly and annotation, and orthogroups among *Pyrus communis* genotypes.

Characteristics	Bartlett_v1	Bartlett.DH_v2	d'Anjou
Assembly			
Assembly size (Mb)	600	507.7	600
Number of scaffolds	142,083	592	5800
Scaffold N50	88 Kb	8.1 Mb	358.88 Kb
Pseudochromosomes	17	17	17
Complete BUSCOs	96.3%	98.3%	97.4%
Annotation			
Predicted gene number	43,419	37,445	45,981
Complete BUSCOs	93.1%	81.8%	92.9%
Mean CDS length	1209	1120	1343
Gene family classification			
Percentage of genes classified into pear orthogroups	76.2	76.2	80.4
Percentage of pear orthogroups containing genes	93.7	81	90.7
Number of 26Gv2 orthogroups containing genes	9878	9668	9837

between d'Anjou and Bartlett_v1, 18.10% are shared by d'Anjou and Bartlett.DH_v2, and only 26.80% are shared between the two Bartlett genomes, which does not support hypotheses 3 and 4.

To better understand why these hypotheses lacked support, we took a broader look at gene family content by comparing a collection of Rosaceae genomes, including the pear genomes in question. We assigned all the predicted protein coding genes from 14 Rosaceae genomes of interest (Chagné et al., 2014; Daccord et al., 2017; Li et al., 2017; Shirasawa et al., 2017; Raymond et al., 2018; VanBuren et al., 2018; Xue et al., 2018; Linsmith et al., 2019; Ou et al., 2019; Zhang et al., 2019; Sun et al., 2020b) to orthogroups constructed with a 26-genome scaffold, covering most of the major lineages of land plants (Supplementary Figure 3). Out of the 18,110 orthogroups from this database, *Prunus persica*, a rosaceous species included in the genome scaffold, has representative genes in 10,290 orthogroups. Genes from most apple and pear genomes (Bartlett_v1, d'Anjou, *Malus domestica* HFTH_v1.0, *M. domestica* GDDH13_v1.1, *M. domestica* Gala_v1.0, *M. sieversii*_v1.0, *M. sylvestris*_v1.0) are present in more than 9,800 orthogroups, however, genes from Bartlett.DH_v2 were only found in 9,688 orthogroups (Table 1; Supplementary Table 3). These results suggest there are many genes not annotated in the Bartlett.DH_v2 genome.

3.3 Genome-wide identification of selected architecture genes

3.3.1 A selection of architecture genes

With this new comparative genomic information, our next steps were two-fold: first, to leverage information from the three European pear genomes and other available Rosaceae genomes, to identify and improve a set of tree architecture-related gene models of interest, and second, to use these architecture gene families as a test case to investigate potential issues in the Bartlett.DH_v2 genome.

Many aspects of tree architecture are important for improving pear growth and maintenance, harvest, ripening, tree size and orchard modernization, disease resistance, and soil microbiome interaction. Traits of interest include dwarfing and dwarfism, root system architecture traits, and branching and branch growth. We selected key gene families known to be involved, particularly those that have been previously shown to influence architectural traits in fruit trees (Supplementary Table 7). The identification of genes within these families, as well as their genomic locations, correct gene models, and domain conservation, is an important early step in testing and understanding their relationships and functions.

3.3.2 Overview of the gene dentification workflow

Here, we developed a high throughput workflow (Figure 2), leveraging a subset of the best Rosaceae plant genomes and a phylogenomic perspective, to efficiently and accurately generate lists of genes in gene families of interest and phylogenetic relationships of genes from different plant lineages. Our workflow, consisting of three main steps, implemented various functions from PlantTribes2 (Wafula, 2019; https://github.com/dePamphilis/PlantTribes) and other software (Kim et al., 2020; Vizueta et al., 2020) for targeted gene annotation.

3.3.3 Step 1 - An initial gene list and preliminary phylogenies

In Step 1, representative plant architecture genes obtained from the literature were assigned into orthogroups based on sequence similarity, giving us 22 orthogroups of interest (Supplementary Tables 2, 8). Note that OG12636 is a monocot-specific orthogroup, thus not included in the downstream analysis of this section). We then leveraged the gene classification results of the aforementioned 14 Rosaceae genomes (Supplementary Table 3) and identified genes assigned to the 21 orthogroups of interest at a plant family level. Next, these Rosaceae genes were integrated with sequences from the 26 scaffolding species in the targeted 21 orthogroups for multiple sequence alignments, which were used to infer phylogeny. At the end of this step, we obtained our initial list of genes in each orthogroup and the phylogenetic relationship of genes in each gene family.

After examining the 21 orthogroups, we identified 64, 105, 94, and 53 genes from *Prunus persica*, Gala_v1, d'Anjou, and Bartlett.DH_v2, respectively (Supplementary Table 9). A whole genome duplication (WGD) event occurred in the common ancestor of *Malus* and *Pyrus* (Sun et al., 2020b), but was not shared with *Prunus.* Therefore, we expect to see an approximate 1:2 ratio in gene numbers in many cases, which explains fewer genes in *Prunus* compared to Gala_v1 and d'Anjou. However, the low gene count in Bartlett.DH_v2 was unexpected. For instance, we observed a clade within a PIN orthogroup (OG1145) comprised of short *PIN* genes (Křeček et al., 2009), which seemed to lack genes from the Bartlett.DH_v2 genome altogether (Figure 3A). One gene copy is found in *Prunus* and Rosoideae species, and two copies are found in most of the Maleae genomes, but none were identified in Bartlett.DH_v2. In addition, in the four genomes mentioned above, we found a number of problematic genes (Supplementary Table 9), for example genes that appeared shorter than all other orthologs or contained unexpected indels likely due to assembly or annotation errors.

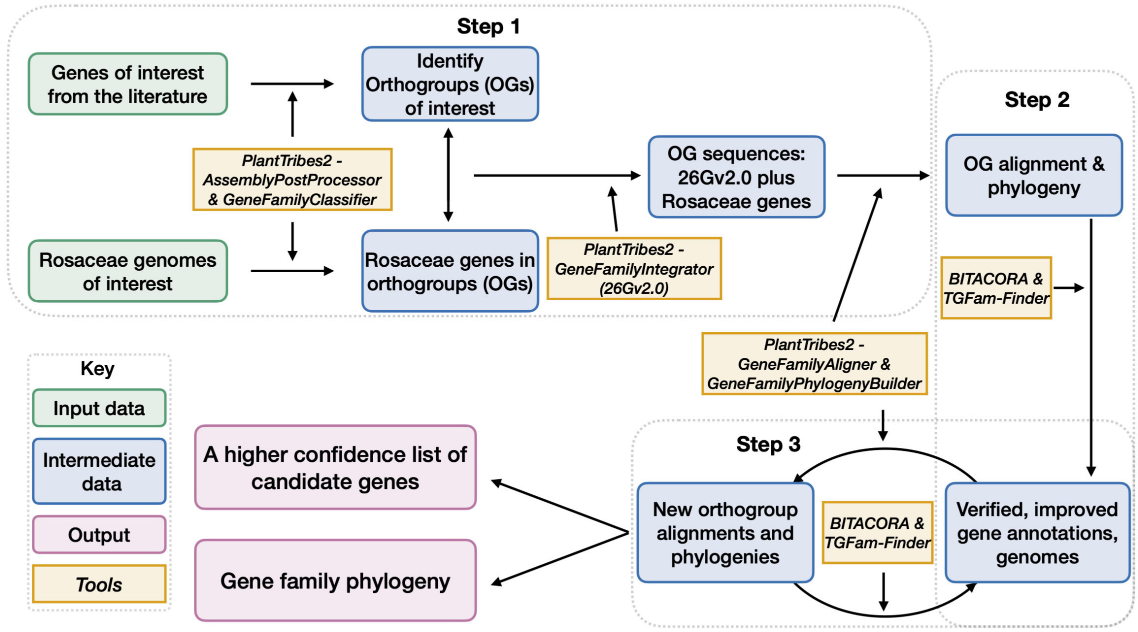

Step 1
Genes of interest from the literature
Identify Orthogroups (OGs) of interest
PlantTribes2 - AssemblyPostProcessor & GeneFamilyClassifier
OG sequences: 26Gv2.0 plus Rosaceae genes
Rosaceae genomes of interest
Rosaceae genes in orthogroups (OGs)
PlantTribes2 - GeneFamilyIntegrator (26Gv2.0)
Step 2
OG alignment & phylogeny
BITACORA & TGFam-Finder
PlantTribes2 - GeneFamilyAligner & GeneFamilyPhylogenyBuilder
Key
Input data
Intermediate data
Output
Tools
A higher confidence list of candidate genes
Gene family phylogeny
Step 3
New orthogroup alignments and phylogenies
BITACORA & TGFam-Finder
Verified, improved gene annotations, genomes

FIGURE 2
A workflow for candidate gene identification, curation, and gene family construction. Gray dotted boxes outlined the three steps of this workflow. Boxes with green outlines are input information. Boxes with blue outlines are intermediate outputs and boxes with purple outlines are final outputs. Contents in boxes with orange outlines are software used for generating the outputs.

3.3.4 Step 2 and 3 - Iterative reannotation of problematic gene models

Inaccurate and missing gene models are common in any genome, especially in the early annotation versions (Marx et al., 2016; Pilkington et al., 2018). In model organisms, such as human, mouse (https://www.gencodegenes.org/), and *Arabidopsis* (https://www.arabidopsis.org/), gene annotations are continuously being improved using experimental evidence, improved data types (*e.g.* full-length RNA molecule sequencing), and both manual and computational curation. Building a better genome assembly is another way to detect additional genes. For instance, the BUSCO completeness score increased from 86.7% in the initial 'Golden Delicious' apple genome (Velasco et al., 2010) to 94.9% in the higher-quality GDDH13 genome (Daccord et al., 2017), indicating that the latter genome captured approximately 120 more conserved single-copy genes. Hence, we hypothesized that the potentially missing and problematic gene models we observed in the two European pears could be improved by: (1) using additional gene annotation approaches; and (2) searching against improved genome assemblies.

To test whether further gene annotation would improve problematic gene models, we moved forward to Step 2 of our workflow, using results from Step 1 as inputs. For each orthogroup containing problematic European pear genes (Supplementary Table 9), we used a subset of high-quality gene models from Rosids identified in Step 1 as inputs and re-annotated these gene families in the two pear genomes. After using a combination of annotation software and manual curation, we found a total of 98 genes from the d'Anjou genome, and reduced the number of problematic or incomplete genes from 34 to 3. In Bartlett.DH_v2, we identified 20 complete genes that were not annotated in the original genome and improved the sequences of 7 previously problematic genes. However, the total number of the selected architecture genes in Bartlett.DH_v2 (73 genes among which 15 were problematic or incomplete) was still notably lower than that of d'Anjou (98 with 3 incomplete genes) or Gala (105 with 15 being incomplete, see Supplementary Table 9). In Step 3, which involves iterative steps of phylogenetic analysis and targeted gene re-annotation, we added additional information such as the improved d'Anjou genes and RNA-seq datasets as new resources to annotate Bartlett.DH_v2 genes, but found no improvements in identifying unannotated genes or improving problematic models.

Results gathered after the first iteration of Step 3 supported our hypothesis that extra annotation steps could help improve imperfect gene models and identify missing genes in the two targeted European pear genomes. However, there were still about 30 genes potentially missing in Bartlett.DH_v2, which led us to test whether polishing the genome assembly would further improve problematic or missing gene models.

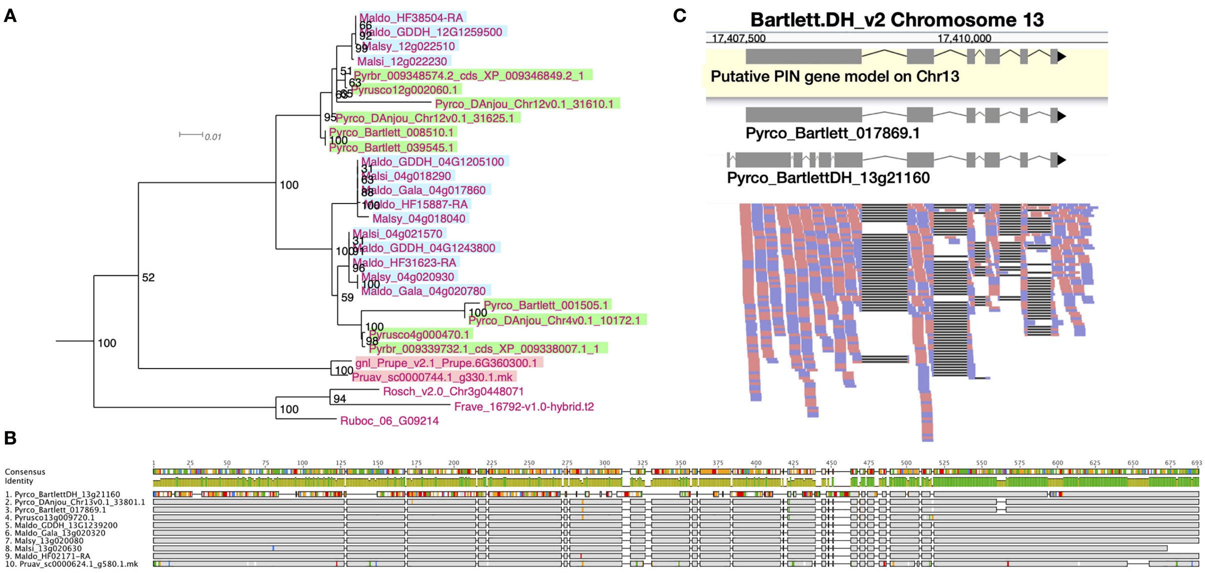


FIGURE 3
Phylogeny, amino acid sequence comparison, and RNAseq read mapping of *PIN* genes. **(A)** One clade of short *PINs* from OG1145 phylogeny. *Malus* genes are indicated with a blue background, *Pyrus* with a green background, and *Prunus* with a pink background. **(B)** Amino acid sequence alignment of orthologous genes from 10 Amygdaloideae species in the long *PIN* gene family (OG438). Sites identical to the consensus are shown in gray and sites different from the consensus are shown with a color following the Clustal color scheme in Geneious R9. Green color in the identity row indicates 100% identical across all sequences and greeny-brown color indicates identity from > 30% to < 100% identity. Gaps in the alignment are shown with a straight line. **(C)** RNAseq reads (forward: red; reverse: blue) mapped to a fragment of chromosome 13 in the Bartlett.DH_v2 genome, where a long *PIN* gene, *Pyrco_BartlettDH_13g21160*, was annotated. The gene model in the yellow box is a putative gene model predicted with RNAseq reads mapped to this region. The two gene models above the read mapping are retrieved from the original annotations of Bartlett_v1 (*Pyrco_Bartlett_017869.1*) and Bartlett.DH_v2 (*Pyrco_BartlettDH_13g21160*).

3.3.5 Step 3 - adding Bartlett.DH_v2 genome polishing

The quality of genome assembly is affected by many factors, including sequencing depth, contig contiguity, and post-assembly polishing. Attempts to improve a presumably high-quality genome are time consuming, and may prove useless if the genome is already in good condition. To initially determine whether polishing the genome assembly would be useful, we first investigated the orthogroups with problematic Bartlett.DH_v2 genes to seek for evidence of assembly derived annotation issues. Indeed, in most cases where we failed to annotate a gene from presumably the correct genomic region, we observed unexpected indels while comparing the Bartlett.DH_v2 genome assembly to other pears (Supplementary Figure 4; Supplementary Table 10). Unexpected indels in the Bartlett.DH_v2 genome were associated with incorrect gene models as well. For example, Figure 3B shows a subset of amino acid sequence alignments for a specific member (*Pyrco_BartlettDH_13g21160*) of a PIN orthogroup (OG438) comprised of the long *PIN* genes (Křeček et al., 2009), in which the Bartlett.DH_v2 gene model shared low sequence identity with orthologs from other Maleae species and *Prunus*. To validate the identity of the problematic gene models, we leveraged RNAseq data from various resources (Nham et al., 2015; Nham et al., 2017; Gabay et al., 2018; Zhang et al., 2018; Hewitt et al., 2020; Zhang et al., 2020) and mapped them to the Bartlett.DH_v2 gene models. In most cases where a conflict was present between the pear consensus, for a given gene of interest, and the Bartlett.DH_v2 gene model, the reads supported the consensus (Figure 3C). The frequent occurrence of truncated and missing genes in the Bartlett.DH_v2 genome may be caused by assembly errors (*e.g.*, base call errors, adapter contamination) that create erroneous open reading frames. This observation provided us with the first piece of evidence that the differences in gene family content observed in the Bartlett.DH_v2 genome may not only be caused by misannotations, but also assembly issues.

To further test whether improvement to the genome assembly would allow us to capture the problematic and missing genes, we polished the Bartlett.DH_v2 genome with Illumina reads from the original publication (Linsmith et al., 2019). We identified 98.40% complete BUSCOs in the polished genome assembly, very similar to the original assembly (Supplementary Table 6), indicating that polishing did not remove BUSCO genes. Using the polished genome, we reiterated Step 3 of our workflow and annotated a total of 103 genes in our gene families of interest, with only two gene models being incomplete (Supplementary Table 9). This new result doubled the number of genes we identified from the original genome annotation and brought the expected gene number into parity with other pome fruit genomes. This supports our hypothesis that genes were missing due to methodological reasons, and in this case, due to assembly errors.

3.4 Curation of a challenging gene family: The IGT family

Some gene families are more complex than others. For example, it is more difficult to study the evolution of resistance (R) genes than most BUSCO genes because the former comprises fast-evolving multigene families while the latter are universally conserved single-copy gene families. Within the architecture gene families we studied, the IGT family is more challenging than many others because members of this family have relatively low levels of sequence conservation outside of a few conserved domains (Yoshihara et al., 2013). Previous reports identified four major clades (LAZY1-like, DRO1-like, TAC1-like, and LAZY5-like) in this gene family (Waite and Dardick, 2021). Study of LAZY1 in model species identified 5 conserved regions (Yoshihara et al., 2013) (Figure 4). The same domains are also present in other LAZY1-like and DRO1-like proteins and the first 4 domains are found in TAC1-like proteins across land plants (Yoshihara and Spalding, 2017). LAZY5-like, the function of which is largely unknown, has only domains I and V. Early research of the *TAC1-like* and *LAZY1-like IGT* genes identified these genes as grass-specific (Li et al., 2007; Yu et al., 2007), as BLAST searches failed to find homologs in other plant lineages.

Using Arabidopsis and rice *IGT* genes as queries, our workflow identified five orthogroups (Supplementary Table 2), containing all the pre-characterized *IGT* genes in angiosperms. The phylogeny constructed with these five orthogroups largely supported previous classification of the four clades (Waite and Dardick, 2021), and provided more information regarding the evolutionary history of this gene family (Figure 5; Supplementary Figure 5). The TAC1-like clade, which is sister to the others, is divided into two monophyletic groups; one contains only monocots while the other has representatives from all the other angiosperm lineages. The LAZY1-like and LAZY5-like clades form one large monophyletic group, which is sister to the DRO1-like clade. Within Rosaceae, a near 1:2 ratio of gene number was expected between peach and pear due to the WGD in the common ancestor of the Maleae. Compared to the six known peach *IGT* genes (Waite and Dardick, 2021), we found 11 orthologs in Bartlett.DH_v2 (including 1 short gene, *Pycro_BartlettDH_LAZY.Chr10*, caused by an unexpected premature stop codon) and 9 in d'Anjou (*Pycro_Danjou_DRO.Chr2* and *Pycro_Danjou_LAZY.Chr10* failed to be annotated due to missing information in the genome). The resulting phylogeny (Figure 5) shows that we have now identified most of the expected *IGT* genes in European pears.

Besides low sequence similarity, *IGT* genes also have unique intron-exon arrangements, which are conserved across

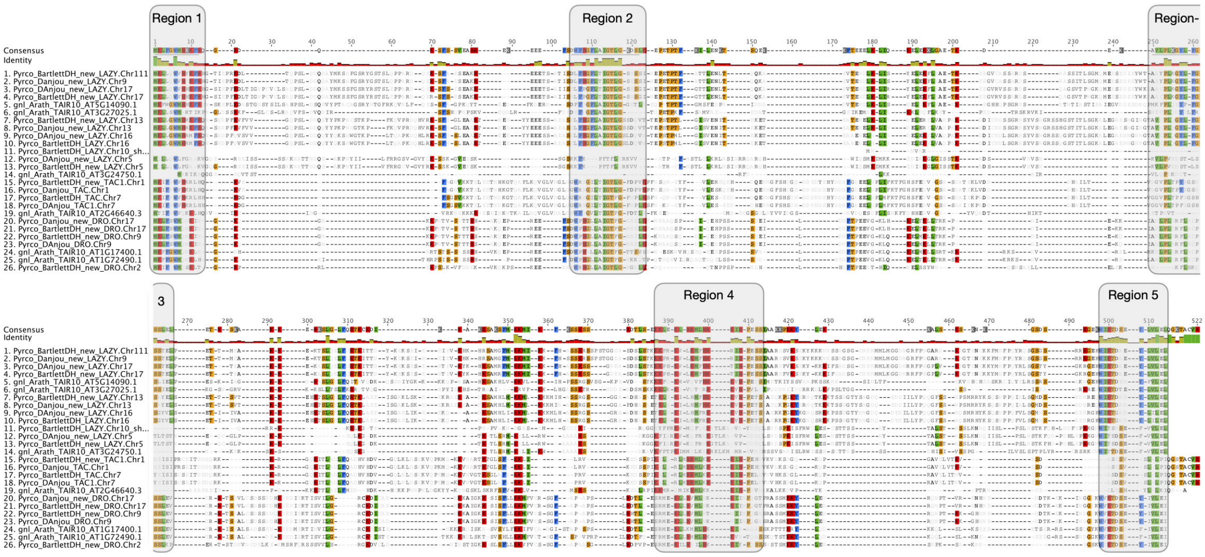

Region 1
Region 2
Region-
3
Region 4
Region 5
Consensus
Identity
1. Pyrco_BartlettDH_new_LAZY.Chr111
2. Pyrco_Danjou_new_LAZY.Chr9
3. Pyrco_DAnjou_new_LAZY.Chr17
4. Pyrco_BartlettDH_new_LAZY.Chr17
5. gnl_Arath_TAIR10_AT5G14090.1
6. gnl_Arath_TAIR10_AT3G27025.1
7. Pyrco_BartlettDH_new_LAZY.Chr13
8. Pyrco_Danjou_new_LAZY.Chr13
9. Pyrco_DAnjou_new_LAZY.Chr16
10. Pyrco_BartlettDH_LAZY.Chr16
11. Pyrco_BartlettDH_LAZY.Chr10_sh...
12. Pyrco_DAnjou_new_LAZY.Chr5
13. Pyrco_BartlettDH_new_LAZY.Chr5
14. gnl_Arath_TAIR10_AT3G24750.1
15. Pyrco_BartlettDH_new_TAC1.Chr1
16. Pyrco_Danjou_TAC.Chr1
17. Pyrco_BartlettDH_TAC.Chr7
18. Pyrco_DAnjou_TAC1.Chr7
19. gnl_Arath_TAIR10_AT2G46640.3
20. Pyrco_Danjou_new_DRO.Chr17
21. Pyrco_BartlettDH_new_DRO.Chr17
22. Pyrco_BartlettDH_new_DRO.Chr9
23. Pyrco_DAnjou_DRO.Chr9
24. gnl_Arath_TAIR10_AT1G17400.1
25. gnl_Arath_TAIR10_AT1G72490.1
26. Pyrco_BartlettDH_new_DRO.Chr2

FIGURE 4
Amino acid comparison of *IGT* genes. Amino acid alignment of *IGT* genes from *Arabidopsis thaliana*, Bartlett.DH_v2, and d'Anjou. Sites consisting of a similar amino acid type as the consensus were highlighted with a background color following the Clustal color scheme in Geneious R9. Green color in the identity row indicates 100% identical across all sequences, greeny-brown color indicates identity from > 30% to < 100% identity, and red color indicates identity < 30%. Five conserved regions (domains) were highlighted with shaded boxes. Note: Region 3 is in both the upper and lower parts of the figure.

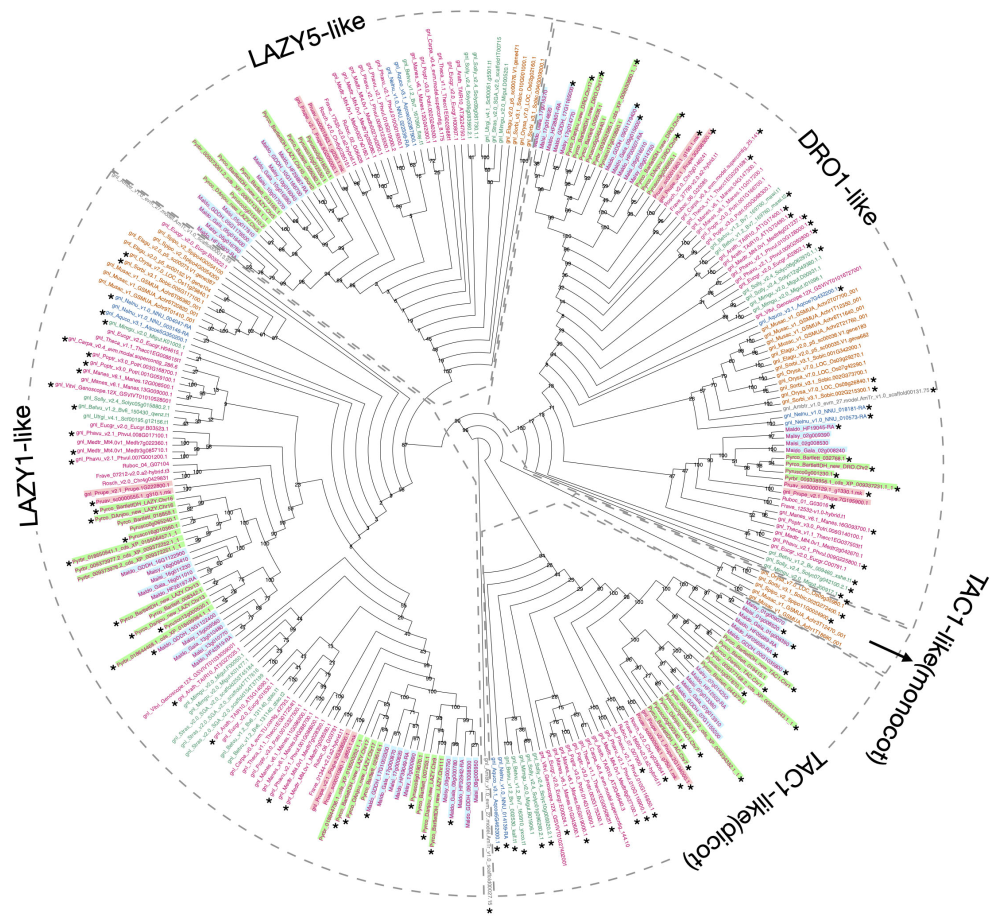


FIGURE 5
Phylogeny comparison of IGT genes. Cladogram of the IGT gene family (including LAZY1-like, LAZY5-like, TAC1-like, and DRO-like, separated by gray dotted boxes). The TAC1-like group was further divided into two monophyletic groups, with one containing only monocot genes and the other containing dicot genes. Genes are colored and highlighted as shown in Supplementary Figure 3. 1000 bootstrap replicates were conducted to estimate reliability and the numbers on the node indicate bootstrap support. Gene models with the expected domain structure and intron-exon structure were marked with *.

Arabidopsis and a few other plant species (Uga et al., 2013; Yoshihara et al., 2013; Waite and Dardick, 2021). These genes all contain 5 exons, but unlike most genes, the first exon only comprises six nucleotides and the last exon contains ~20 nucleotides. Annotation of short exons, especially when transcriptome evidence is limited, can be very challenging and skipping such exons could cause problems in gene discovery (Mount, 2000; Guo and Liu, 2015; Sharma et al., 2018). For instance, the annotation of *AtAPC11* (*At3g05870*) was inaccurate until Guo and Liu identified a single-nucleotide exon in this gene (Guo and Liu, 2015).

To determine whether we captured the correct *IGT* gene models in the targeted genomes, we investigated the protein sequence alignments and gene features. In the original annotation, only three gene models (*Pyrco_BartlettDH_16g10510*, *Pyrco_BartlettDH_07g15250*, *Pyrco_DAnjou_Chr7v0.1_17442.1*) have the correct intron-exon combination and the expected domains. In the iterative re-annotation steps of our workflow, we identified 6 additional accurate gene models leveraging sequence orthology and transcriptome evidence. We further investigated all the sequences we identified as *IGT* genes, seeking the presence or absence of the expected domain features. However, even among gene models from the best annotated genomes used to construct the 26Gv2.0 database, only 45.16% (56/124) have the expected domain features (indicated with an * next to gene names in Figure 5. LAZY5-like was not taken into consideration due to its unique structure). In most cases, although the signature IGT domain (II) is correctly identified in the genes, domains I and V are usually missing or incorrect, likely due to misannotation of the first and last short exons. In Rosaceae, besides Bartlett.DH_v2 and d'Anjou, only 34.38% (33/96) had the expected domains (Figure 5). This finding motivated us to manually investigate the targeted genomes to

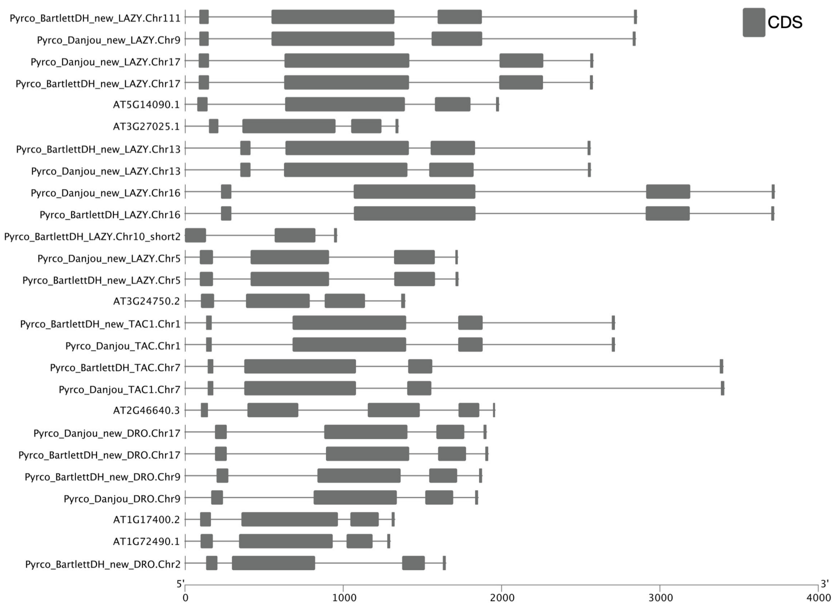

Pyrco_BartlettDH_new_LAZY.Chr111
Pyrco_Danjou_new_LAZY.Chr9
Pyrco_Danjou_new_LAZY.Chr17
Pyrco_BartlettDH_new_LAZY.Chr17
AT5G14090.1
AT3G27025.1
Pyrco_BartlettDH_new_LAZY.Chr13
Pyrco_Danjou_new_LAZY.Chr13
Pyrco_Danjou_new_LAZY.Chr16
Pyrco_BartlettDH_LAZY.Chr16
Pyrco_BartlettDH_LAZY.Chr10_short2
Pyrco_Danjou_new_LAZY.Chr5
Pyrco_BartlettDH_new_LAZY.Chr5
AT3G24750.2
Pyrco_BartlettDH_new_TAC1.Chr1
Pyrco_Danjou_TAC.Chr1
Pyrco_BartlettDH_TAC.Chr7
Pyrco_Danjou_TAC1.Chr7
AT2G46640.3
Pyrco_Danjou_new_DRO.Chr17
Pyrco_BartlettDH_new_DRO.Chr17
Pyrco_BartlettDH_new_DRO.Chr9
Pyrco_Danjou_DRO.Chr9
AT1G17400.2
AT1G72490.1
Pyrco_BartlettDH_new_DRO.Chr2
CDS
5'
3'
0
1000
2000
3000
4000

FIGURE 6
Intron-exon structure comparison of IGT genes. Cartoon illustrating intron-exon structures of IGT genes from *Arabidopsis thaliana* (Araport11), Bartlett.DH_v2, and d'Anjou. Boxes indicate exons, lines indicate introns. UTR regions are not shown in this figure.

annotate the IGT genes. Using the correct gene models as reference, plus a careful manual curation, we were able to annotate 19 complete gene models of the 20 expected IGT genes from the two targeted pear genomes (Figures 4, 6).

4 Discussion

A second European pear cultivar genome from 'd'Anjou' provided additional insights into gene families across Rosaceae. By leveraging perspectives from comparative genomics and phylogenomics, we developed a high-throughput workflow using a collection of bioinformatic tools that takes a list of genes of interest from the literature and genomes of interest as input, and produces a curated list of the targeted genes in the query genomes.

In the case study presented here, candidate genes from 16 plant architecture-related gene families were identified from 15 Rosaceae genomes. The study of gene families consists primarily of two initial parts: first, identification of all the members in these families, and second, investigation of their phylogenetic relationships. Many attempts (Feng et al., 2019; Cancino-García et al., 2020; Zheng et al., 2020) to identify genes of interest from a genome have relied solely on a BLAST search querying a homolog from a model organism, which may be distantly related. However, such a method is insufficient in identifying all members of a large complex gene family or a fast-evolving and highly-divergent family, such as the *IGT* genes. They may also incorrectly include genes in a gene family based only on one or a few highly conserved regions that are insufficient for gene family membership. Compared to a BLAST-only approach, the gene classification process in our workflow used a combination of BLAST and HMMER search against an objectively pre-classified gene family scaffold, which provides a better result by taking into consideration both sensitivity and specificity (Wafula, 2019). This allowed us to efficiently identify even very challenging genes. Moreover, instead of selecting homologs based on simple statistics such as identity or bitscore, we took a phylogenetic approach and a sample dataset with references from a wide range of land plants to increase the accuracy of identifying orthologs and paralogs. Phylogenetic relationships revealed by a small number of taxa, for instance using only one species of interest and one model organism, can be inaccurate. For example, in our phylogenetic analysis with rich taxon sampling, *PIN5-1* and *PIN5-2* from *Pyrus bretschneideri* are sisters to all other *PINs* (Supplementary Figure 6), challenging the phylogenetic relationship inferred with *PINs* only from *P. bretschneideri* and *Arabidopsis thaliana* (Qi et al., 2020).

The iterative quality control steps in the workflow helped identify problems that existed in certain gene models and provided hints about where to make targeted improvements to important *Pyrus* genomic resources. The highly contiguous assembly of Bartlett.DH_v2 provided a valuable reference to anchor the shorter scaffolds from d'Anjou, which is essential for a good annotation. On the other hand, the perspective afforded by the d'Anjou genome led us to examine the Bartlett.DH_v2 genome assembly further. We developed and tested hypotheses regarding unexpected gene annotation patterns in the two targeted European pear genomes among various Maleae species and cultivars. This led to a polished assembly and improved annotations that allowed us to curate a high confidence list of candidate genes and gene models for downstream analyses. By adding targeted iterations of genome assembly and annotation, we now have a better starting point for reverse genetic analyses and understanding functionality of architecture-related genes in pears.

The challenges we encountered as we laid the groundwork for reverse genetics studies to understand pear architecture genes, and the approaches we took to evaluate and tackle these challenges, reinforce the idea that genome assembly and annotation are iterative processes. We found that relating gene accession IDs and inconsistent gene names back to gene sequences in various databases was often difficult and time consuming. Objective, global-scale gene classification, as we used here *via* PlantTribes2 (Wafula, 2019), can help researchers work across genomes and among various genome resources. Further, guidance from consortia such as AgBioData (Harper et al., 2018) is helping facilitate work such as we have described here that includes the acquisition and analysis of genome-scale data. Our starting point for understanding putative architecture genes in pear was with genes of interest from several plant species - an approach that many researchers will find familiar. With genes of interest in hand, our workflow provides a comparative genome approach to efficiently identify, investigate, and then improve and/or validate genes of interest across genomes and genome resources.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA762155.

Author contributions

HZ, JW, LH conceived and designed the research. PR prepared gDNA for sequencing. HZ, EW, PT, JE, JW, CD, and AH performed the genome assembly and gene family analysis. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Washington Tree Fruit Research Commission project PR-17-104, the Agricultural Research Service in the US Department of Agriculture, and the Dottie and Lloyd Huck Endowment.

Acknowledgments

The authors would like to acknowledge Heidi Hargarten for maintaining the d'Anjou plant and collecting leaf tissue for sequencing. They also thank Craig Praul at Penn State and Diana Burkart-Waco and Lutz Froenicke at UC Davis for sequencing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.975942/full#supplementary-material

References

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S. , Sedlazeck, F. J., et al. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20 (1), 224. doi: 10.1186/s13059-019-1829-6

Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biol.* 13 (6), R56. doi: 10.1186/gb-2012-13-6-r56

Cabanettes, F., and Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6, e4958. doi: 10.7717/peerj.4958

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinf.* 10 (1), 421. doi: 10.1186/1471-2105-10-421

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-p. *Curr. Protoc. Bioinf.* 48 (1), 4.11.1–4.11.39. doi: 10.1002/0471250953.bi0411s48

Cancino-García, V. J., Ramírez-Prado, J. H., and De-la-Peña, C. (2020). Auxin perception in agave is dependent on the species' auxin response factors. *Sci. Rep.* 10 (1), 3860. doi: 10.1038/s41598-020-60865-y

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18 (1), 188–196. doi: 10.1101/gr.6743907

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25 (15), 1972–1973. doi: 10.1093/bioinformatics/btp348

Chagné, D., Crowhurst, R. N., Pindo, M., Thrimawithana, A., Deng, C., Ireland, H., et al. (2014). The draft genome sequence of European pear (Pyrus communis l. 'Bartlett'). *PloS One* 9 (4), e92644. doi: 10.1371/journal.pone.0092644

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., and Emerson, J. J. (2016). Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* 44 (19), e147–e147. doi: 10.1093/nar/gkw654

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13 (8), 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, D. X., Pan, Y., Wang, Y., Cui, Y. Z., Zhang, Y. J., Mo, R. Y., et al. (2021a). The chromosome-level reference genome of coptischinensis provides insights into genomic evolution and berberine biosynthesis. *Horticult. Res.* 8 (1), 121. doi: 10.1038/s41438-021-00559-2

Chen, J., Xie, F. F., Cui, Y. Z., Chen, C. B., Lu, W. J., Hu, X. D., et al. (2021b). A chromosome-scale genome sequence of pitaya (Hylocereus undatus) provides novel insights into the genome evolution and regulation of betalain biosynthesis. *Horticult. Res.* 8 (1), 164. doi: 10.1038/s41438-021-00612-0

Cheng, C., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: A complete reannotation of the arabidopsis thaliana reference genome. *Plant J.* 89 (4), 789–804. doi: 10.1111/tpj.13415

Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y., et al. (2016a). Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in brassica rapa and brassica oleracea. *Nat. Genet.* 48 (10), 1218–1224. doi: 10.1038/ng.3634

Cheng, F., Wu, J., Cai, C., Fu, L., Liang, J., Borm, T., et al. (2016b). Genome resequencing and comparative variome analysis in a brassica rapa and brassica oleracea collection. *Sci. Data* 3 (1), 160119. doi: 10.1038/sdata.2016.119

Daccord, N., Celton, J-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., et al. (2017). High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* 49 (7), 1099–1106. doi: 10.1038/ng.3886

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29 (1), 15–21. doi: 10.1093/bioinformatics/bts635

Dunn, N. A., Unni, D. R., Diesh, C., Munoz-Torres, M., Harris, N. L., Yao, E., et al. (2019). Apollo: Democratizing genome annotation. *PloS Comput. Biol.* 15 (2), e1006790. doi: 10.1371/journal.pcbi.1006790

Eddy, S. R. (2011). Accelerated profile HMM searches. *PloS Comput. Biol.* 7 (10), e1002195. doi: 10.1371/journal.pcbi.1002195

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

Emms, D. M., and Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157. doi: 10.1186/s13059-015-0721-2

Feng, Y., Sun, Q., Zhang, G., Wu, T., Zhang, X., Xu, X., et al. (2019). Genome-wide identification and characterization of ABC transporters in nine rosaceae species identifying MdABCG28 as a possible cytokinin transporter linked to dwarfing. *Int. J. Mol. Sci.* 20 (22), 5783. doi: 10.3390/ijms20225783

Gabay, G., Faigenboim, A., Dahan, Y., Izhaki, Y., and Itkin, M. (2018). "Transcriptome analysis and metabolic profiling reveal the key role of α-linolenic acid in dormancy regulation of European pear,". *J. Exp. Bot.* 70 (3), 1017–1031. doi: 10.1093/jxb/ery405

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29 (7), 644–652. doi: 10.1038/nbt.1883

Guo, L., and Liu, C.-M. (2015). A single-nucleotide exon found in arabidopsis. *Sci. Rep.* 5 (1), 18087. doi: 10.1038/srep18087

Hämälä, T., et al. (2021). "Genomic structural variants constrain and facilitate adaptation in natural populations of theobroma cacao, the chocolate tree," in *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.2102914118

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. doi: 10.1093/nar/gkq862

Harper, L., Campbell, J., Cannon, E. K.S., Jung, S., Poelchau, M., Walls, R., et al. (2018). AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database* 2018, 1–32. doi: 10.1093/database/bay088

Hewitt, S. L., Hendrickson, C. A., and Dhingra, A. (2020). Evidence for the involvement of vernalization-related genes in the regulation of cold-induced ripening in 'D'Anjou' and 'Bartlett' pear fruit. *Sci. Rep.* 10 (1), 8478. doi: 10.1038/s41598-020-65275-8

Hill, J. L., and Hollender, C. A. (2019). Branching out: New insights into the genetic regulation of shoot architecture in trees. *Curr. Opin. Plant Biol.* 47, 73–80. doi: 10.1016/j.pbi.2018.09.010

Hoff, K. J., and Stanke, M. (2019). "Predicting genes in single genomes with AUGUSTUS,". *Curr. Protoc. Bioinf.* 65 (1), e57. doi: 10.1002/cpbi.57

Honaas, L., Hargarten, H., Hadish, J., Ficklin, S. P., Serra, S., Musacchi, S., et al. (2021). Transcriptomics of differential ripening in 'd'Anjou' pear (Pyrus communis l.). *Front. Plant Sci.* 12, 609684. doi: 10.1038/s41438-021-00505-2

Hughes, T. E., Langdale, J. A., and Kelly, S. (2014). The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.* 24 (8), 1348–1355. doi: 10.1101/gr.172684.114

Huson, D. H., and Scornavacca, C. (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systemat. Biol.* 61 (6), 1061–1067. doi: 10.1093/sysbio/sys062

Jackman, S. D., Coombe, L., Chu, J., Warren, R. L., Vandervalk, B. P., Yeo, S., et al. (2018). Tigmint: Correcting assembly errors using linked reads from large molecules. *BMC Bioinf.* 19 (1), 393. doi: 10.1186/s12859-018-2425-6

Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 30 (9), 1236–1240. doi: 10.1093/bioinformatics/btu031

Jung, S., Lee, T., Cheng, C-H., Buble, K., Zheng, P., Yu, J., et al. (2018). 15 years of GDR: New data and functionality in the genome database for rosaceae. *Nucleic Acids Res.* 47 (D1), D1137–D1145. doi: 10.1093/nar/gky1000

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30 (14), 3059–3066. doi: 10.1093/nar/gkf436

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28 (12), 1647–1649. doi: 10.1093/bioinformatics/bts199

Khan, A., Carey, S., Serrano, A., Zhang, H., Hargarten, H., Hale, H., et al. (2022), A phased, chromosome-scale genome of 'Honeycrisp' apple (Malus domestica). *Gigabyte* 2022, 1–15. doi: 10.46471/gigabyte.69

Kim, S., Cheong, K., Park, J., Kim M-S., , Kim, J., Seo M-K., , et al. (2020). TGFam-finder: A novel solution for target-gene family annotation in plants. *N. Phytol.* 227 (5), 1568–1581. doi: 10.1111/nph.16645

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., Phillippy, A. M., et al. (2017). Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27 (5), 722–736. doi: 10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinf.* 5 (1), 59. doi: 10.1186/1471-2105-5-59

Křeček, P., Skůpa, P., Libus, J., Naramoto, S., Tejos, R., Friml, J., et al. (2009). The PIN-FORMED (PIN) protein family of auxin transporters. *Genome Biol.* 10 (12), 249. doi: 10.1186/gb-2009-10-12-249

Kuhn, R. M., Haussler, D., and Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings Bioinf.* 14 (2), 14x4–1161. doi: 10.1093/bib/bbs038

Kyriakidou, M., Tai, H. H., Anglin, N.L., Ellis, D., and Strömvik, M. V. (2018). Current strategies of polyploid plant genome sequence assembly. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01660

Li, H. (2016b). Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 32 (14), 2103–2110. doi: 10.1093/bioinformatics/btw152

Li, P., Wang, Y., Qian, Q., Fu, Z., Wang, M., Zeng, D., et al. (2007). LAZY1 controls rice shoot gravitropism through regulating polar auxin transport. *Cell Res.* 17 (5), 402–410. doi: 10.1038/cr.2007.38

Li, Y., Wei, W., Feng, J., Luo, H., Pi, M., Liu, Z., et al. (2017). Genome re-annotation of the wild strawberry fragaria vesca using extensive illumina- and SMRT-based RNA-seq datasets. *DNA Res.* 25 (1), dsx038. doi: 10.1093/dnares/dsx038

Linsmith, G., Rombauts, S., Montanari, S., Deng, C.H., Celton, J-M., Guérif, P., et al. (2019). Pseudo-chromosome–length genome assembly of a double haploid 'Bartlett' pear (Pyrus communis l.). *GigaScience* 8 (12). doi: 10.1093/gigascience/giz138

Mabry, M. E., Turner-Hissong, S. D., Gallagher, E. Y., McAlvay, A. C., An, H., Edger, P. P., et al. (2021). The evolutionary history of wild, domesticated, and feral brassica oleracea (Brassicaceae). *Mol. Biol. Evol* 38:4419–34. doi: 10.1093/molbev/msab183

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, msab199-. doi: 10.1093/molbev/msab199

Marx, H., Minogue, C. E., Jayaraman, D., Richards, A. L., Kwiecien, N. W., Siahpirani, A. F., et al. (2016). A proteomic atlas of the legume medicago truncatula and its nitrogen-fixing endosymbiont sinorhizobium meliloti. *Nat. Biotechnol.* 34 (11), 1198–1205. doi: 10.1038/nbt.3681

Michiels, A., Ende, W. V., Tucker, M., Liesbet, V., and Laere, A. V. (2003). Extraction of high-quality genomic DNA from latex-containing plants. *Analytical. Biochem.* 315 (1), 85–89. doi: 10.1016/s0003-2697(02)00665-6

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49 (D1), gkaa913-. doi: 10.1093/nar/gkaa913

Mount, S. M. (2000). Genomic sequence, splicing, and gene annotation. *Am. J. Hum. Genet.* 67 (4), 788–792. doi: 10.1086/303098

Nham, N. T., Freitas, S. T. de, Macnish, A. J., Carr, K. M., Kietikul, T., Guilatco, A., et al. (2015). A transcriptome approach towards understanding the development of ripening capacity in 'Bartlett' pears (Pyrus communis l.). *BMC Genomics* 16 (1), 762. doi: 10.1186/s12864-015-1939-9

Nham, N. T., Macnish, A. J., Zakharov, F., and Mitcham, E. J. (2017). 'Bartlett' pear fruit (Pyrus communis l.) ripening regulation by low temperatures involves genes associated with jasmonic acid, cold response, and transcription factors. *Plant Sci.* 260, 8–18. doi: 10.1016/j.plantsci.2017.03.008

Ou, C., Wang, F., Wang, J., Li, S., Zhang, Y., Fang, M., et al. (2019). A *de novo* genome assembly of the dwarfing pear rootstock zhongai 1. *Sci. Data* 6 (1), 281. doi: 10.1038/s41597-019-0291-3

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y-C., et al. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* 19 (1), 208. doi: 10.1186/s13059-018-1590-2

Pilkington, S. M., Crowhurst, R., Hilario, E., Nardozza, S., Fraser, L., Peng, Y., et al. (2018). A manually annotated actinidia chinensis var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. *BMC Genomics* 19 (1), 257. doi: 10.1186/s12864-018-4656-3

Pracana, R., Priyam, A., Levantis, I., Nichols, R.A., and Wurm, Y. (2017). The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. *Molecular Ecology*. 26, 2864–2879. doi: 10.1111/mec.14054

Qi, L., Chen, L., Wang, C., Zhang, S., Yang, Y., Liu, J., et al. (2020). Characterization of the auxin efflux transporter PIN proteins in pear. *Plants* 9 (3), 349. doi: 10.3390/plants9030349

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., et al. (2018). The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.* 50 (6), 772–777. doi: 10.1038/s41588-018-0110-3

Sharma, S., Sharma, S. N., and Saxena, R. (2018). Identification of short exons disunited by a short intron in eukaryotic DNA regions. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 17 (5), 1660–1670. doi: 10.1109/tcbb.2019.2900040

Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H., et al. (2017). The genome sequence of sweet cherry (Prunus avium) for use in genomics-assisted breeding. *DNA Res.* 24 (5), dsx020-. doi: 10.1093/dnares/dsx020

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30 (9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 32 (suppl_2), W309–W312. doi: 10.1093/nar/gkh379

Stansell, Z., and Björkman, T. (2020). From landrace to modern hybrid broccoli: the genomic and morphological domestication syndrome within a diverse b. oleracea collection. *Horticult. Res.* 7 (1), 159. doi: 10.1038/s41438-020-00375-0

Stansell, Z., Hyma, K., Fresnedo-Ramírez, J., Sun, Q., Mitchell, S., Björkman, T., et al. (2018). Genotyping-by-sequencing of brassica oleracea vegetables reveals unique phylogenetic patterns, population structure and domestication footprints. *Horticult. Res.* 5 (1), 38. doi: 10.1038/s41438-018-0040-3

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. doi: 10.1093/nar/gkp759

Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., et al. (2020b). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* 52 (12), 1423–1432. doi: 10.1038/s41588-020-00723-9

Takos, A. M., Jaffé, F. W., Jacob, S.R., Bogs, J., Robinson, S. P., and Walker, A. R. (2006). Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples. *Plant Physiol.* 142 (3), 1216–1232. doi: 10.1104/pp.106.088104

Tollenaere, R., Hayward, A., Dalton-Morgan, J., Campbell, E., Lee, J. R. M., Lorenc, M. T., et al. (2012). Identification and characterization of candidate Rlm4 blackleg resistance genes in brassica napus using next-generation sequencing. *Plant Biotechnol. J.* 10 (6), 709–715. doi: 10.1111/j.1467-7652.2012.00716.x

Uga, Y., Sugimoto, K., Ogawa, S., Rane, J., Ishitani, M., Hara, N., et al. (2013). Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. *Nat. Genet.* 45 (9), 1097–1102. doi: 10.1038/ng.2725

VanBuren, R., Wai, C. M., Colle, M., Wang, J., Sullivan, S., Bushakra, J. M., et al. (2018). A near complete, chromosome-scale assembly of the black raspberry (Rubus occidentalis) genome. *GigaScience* 7 (8), giy094-. doi: 10.1093/gigascience/giy094

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (Malus × domestica borkh.). *Nat. Genet.* 42 (10), 833–839. doi: 10.1038/ng.654

Vizueta, J., Sánchez-Gracia, A., and Rozas, J. (2020). Bitacora: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Mol. Ecol. Resour.* 20 (5), 1445–1452. doi: 10.1111/1755-0998.13202

Wafula, E. K. (2019). *Computational methods for comparative genomics of non-model species: A case study in the parasitic plant family Orobanchaceae.* [dissertation]. [University Park (PA)]: The Pennsylvania State University

Waite, J. M., and Dardick, C. (2021). The roles of the IGT gene family in plant architecture: past, present, and future. *Curr. Opin. Plant Biol.* 59, 101983. doi: 10.1016/j.pbi.2020.101983

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* 9 (11), e112963. doi: 10.1371/journal.pone.0112963

Wall, P. K., Leebens-Mack, J., Müller, K. F., Field, D., Altman, N. S., and dePamphilis, C. W. (2008). PlantTribes: A gene and gene family resource for comparative genomics in plants. *Nucleic Acids Res.* 36 (suppl_1), D970–D976. doi: 10.1093/nar/gkm972

Wang, J., Xu, S., Mei, Y., Cai, S., Gu, Y., Sun, M., et al. (2021a). A high-quality genome assembly of morinda officinalis, a famous native southern herb in the lingnan region of southern China. *Horticult. Res.* 8 (1), 135. doi: 10.1038/s41438-021-00551-w

Wang, P., Yu, J., Jin, S., Chen, S., Yue, C., Wang, W., et al. (2021b). Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. *Horticult. Res.* 8 (1), 107. doi: 10.1038/s41438-021-00542-x

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Res.* 27 (5), 757–767. doi: 10.1101/gr.214874.116

Wu, J., Wei, K., Cheng, F., Li, S., Wang, Q., Zhao, J., et al. (2012). A naturally occurring InDel variation in BraA.FLC.b (BrFLC2) associated with flowering time variation in brassica rapa. *BMC Plant Biol.* 12 (1), 151. doi: 10.1186/1471-2229-12-151

Xiao, D., Zhao, J. J., Hou, X. L., Basnet, R. K., Carpio, D. P.D., Zhang, N. W., et al. (2013). The brassica rapa FLC homologue FLC2 is a key regulator of flowering time, identified through transcriptional co-expression networks. *J. Exp. Bot.* 64 (14), 4503–4516. doi: 10.1093/jxb/ert264

Xue, H., Wang, S., Yao, J-L., Deng, C. H., Wang, L., Su, Y., et al. (2018). Chromosome level high-density integrated genetic maps improve the pyrus bretschneideri 'DangshanSuli' v1.0 genome. *BMC Genomics* 19 (1), 833. doi: 10.1186/s12864-018-5224-6

Xu, X., Yuan, H., Yu, X., Huang, S., Sun, Y., Zhang, T., et al. (2021). The chromosome-level stevia genome provides insights into steviol glycoside biosynthesis. *Horticult. Res.* 8 (1), 129. doi: 10.1038/s41438-021-00565-4

Yang, Z., Wafula, E. K., Honaas, L. A., Zhang, H., Das, M., Fernandez-Aparicio, M., et al. (2015). Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. *Mol. Biol. Evol.* 32 (3), 767–790. doi: 10.1093/molbev/msu343

Yeo, S., Coombe, L., Warren, R.L., Chu, J., and Birol, I. (2017). ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* 34 (5), 725–731. doi: 10.1093/bioinformatics/btx675

Yoshida, S., Kim, S., Wafula, E. K., Tanskanen, J., Kim, Y-M., Honaas, L., et al. (2019). “Genome sequence of striga asiatica provides insight into the evolution of plant parasitism,”. *Curr. Biol.* 29 (18), 3041–3052.e4. doi: 10.1016/j.cub.2019.07.086

Yoshihara, T., and Spalding, E. P. (2017). LAZY genes mediate the effects of gravity on auxin gradients and plant architecture. *Plant Physiol.* 175 (2), 959–969. doi: 10.1104/pp.17.00942

Yoshihara, T., Spalding, E. P., and Iino, M. (2013). AtLAZY1 is a signaling component required for gravitropism of the arabidopsis thaliana inflorescence. *Plant J.* 74 (2), 267–279. doi: 10.1111/tpj.12118

Yu, B., Lin, Z., Li, H., Li, X., Li, J., Wang, Y., et al. (2007). TAC1, a major quantitative trait locus controlling tiller angle in rice. *Plant J.* 52 (5), 891–898. doi: 10.1111/j.1365-313x.2007.03284.x

Zhang, H., Yang, Y., Li, D., Song, J., Ma, C., and Wang, R. (2018). RNA-Seq analysis of the tissue-specific expressed genes of pyrus betulaefolia in root, stem and leaf. *Acta Hortic. Sin.* 45 (10), 1881–1894. doi: 10.16420/j.issn.0513-353x.2017-0783

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., et al. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat. Commun.* 10 (1), 1494. doi: 10.1038/s41467-019-09518-x

Zhang, Z., Tian, C., Zhang, Y., Li, C., Li, X., Yu, Q., et al. (2020). Transcriptomic and metabolomic analysis provides insights into anthocyanin and procyanidin accumulation in pear. *BMC Plant Biol.* 20 (1), 129. doi: 10.1186/s12870-020-02344-0

Zheng, X., Xiao, Y., Tian, Y., Yang, S., and Wang, C. (2020). PcDWF1, a pear brassinosteroid biosynthetic gene homologous to AtDWARF1, affected the vegetative and reproductive growth of plants. *BMC Plant Biol.* 20 (1), 109. doi: 10.1186/s12870-020-2323-8

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29 (21), 2669–2677. doi: 10.1093/bioinformatics/btt476

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 December 2022
DOI 10.3389/fpls.2022.975917

OPEN ACCESS

EDITED BY
Ksenija Gasic,
Clemson University, United States

REVIEWED BY
Hongliang Zhu,
China Agricultural University, China
Sachin Rustgi,
Clemson University, United States

*CORRESPONDENCE
Concetta Licciardello
concetta.licciardello@crea.gov.it

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 22 June 2022
ACCEPTED 13 October 2022
PUBLISHED 13 December 2022

CITATION
Salonia F, Ciacciulli A, Pappalardo HD, Poles L, Pindo M, Larger S, Caruso P, Caruso M and Licciardello C (2022) A dual sgRNA-directed CRISPR/Cas9 construct for editing the fruit-specific β-cyclase 2 gene in pigmented citrus fruits. *Front. Plant Sci.* 13:975917. doi: 10.3389/fpls.2022.975917

A dual sgRNA-directed CRISPR/Cas9 construct for editing the fruit-specific β-cyclase 2 gene in pigmented citrus fruits

Fabrizio Salonia[1,2†], Angelo Ciacciulli[1†], Helena Domenica Pappalardo[1], Lara Poles[1,2], Massimo Pindo[3], Simone Larger[3], Paola Caruso[1], Marco Caruso[1] and Concetta Licciardello[1*]

[1]Council for Agricultural Research and Economics (CREA) - Research Centre for Olive, Fruit and Citrus Crops, Acireale, Italy, [2]Department of Agriculture, Food and Environment (Di3A), University of Catania, Catania, Italy, [3]Research and Innovation Centre, Trento with S. Michele all' Adige, Trento, Italy

CRISPR/Cas9 genome editing is a modern biotechnological approach used to improve plant varieties, modifying only one or a few traits of a specific variety. However, this technology cannot be easily used to improve fruit quality traits in citrus, due to the lack of knowledge of key genes, long juvenile stage, and the difficulty regenerating whole plants of specific varieties. Here, we introduce a genome editing approach with the aim of producing citrus plantlets whose fruits contain both lycopene and anthocyanins. Our method employs a dual single guide RNA (sgRNA)-directed genome editing approach to knockout the fruit-specific β-cyclase 2 gene, responsible for the conversion of lycopene to beta-carotene. The gene is targeted by two sgRNAs simultaneously to create a large deletion, as well as to induce point mutations in both sgRNA targets. The EHA105 strain of *Agrobacterium tumefaciens* was used to transform five different anthocyanin-pigmented sweet oranges, belonging to the Tarocco and Sanguigno varietal groups, and 'Carrizo' citrange, a citrus rootstock as a model for citrus transformation. Among 58 plantlets sequenced in the target region, 86% of them were successfully edited. The most frequent mutations were deletions (from -1 to -74 nucleotides) and insertions (+1 nucleotide). Moreover, a novel event was identified in six plantlets, consisting of the inversion of the region between the two sgRNAs. For 20 plantlets in which a single mutation occurred, we excluded chimeric events. Plantlets did not show an altered phenotype in vegetative tissues. To the best of our knowledge, this work represents the first example of the use of a genome editing approach to potentially improve qualitative traits of citrus fruit.

KEYWORDS

anthocyanins, β-LCY2, genome editing, antioxidant compounds, regeneration, blood oranges, GoldenBraid 3.0

Introduction

Citrus represents a group of species that is highly appreciated by consumers. They have high levels of bioactive compounds (Duarte et al., 2016), such as anthocyanins and carotenoids (e.g., lycopene), which are main sources of antioxidants (Rao and Agarwal, 1999; Krinsky and Johnson, 2005; Smeriglio et al., 2016; Pojer et al., 2020). Anthocyanins reduce inflammation, protect against cardiovascular disease (Mazza, 2007; He and Giusti, 2010), prevent cancer and inhibit its growth, and help fight obesity and type 2 diabetes associated with insulin resistance (Anderson et al., 2004; Cooke et al., 2005; Wang and Stoner, 2008). Lycopene and other carotenoids inhibit lipoprotein oxidation, thus reducing the negative effects of cancer and cardiovascular diseases, inflammatory processes, and Parkinson's disease (Gerster, 1997; Bramley, 2000). Demand for anthocyanin- or lycopene-pigmented citrus fruits has been increasing, and new varieties have recently been released (Barry et al., 2020).

Anthocyanins are water-soluble compounds synthesized in the phenylpropanoid pathway (Winkel-Shirley, 2001). The biosynthetic pathway and regulatory complex have been comprehensively described in plants (Tanaka et al., 2008; Petroni and Tonelli, 2011). In citrus, *Ruby* (a MYB-like member of the MBW complex) is the crucial transcription factor responsible for the control of anthocyanin production (Butelli et al., 2012; Butelli et al., 2017; Butelli et al., 2019). It is switched on mainly under cold conditions (Lo Piero et al., 2005; Crifò et al., 2011; Carmona et al., 2017) and is light-dependent (Huang et al., 2019). The crucial role of *Ruby* has been demonstrated by its overexpression in non-pigmented citrus fruits, which turns such fruits purple (Dutt et al., 2016). Among citrus fruits, the presence of anthocyanins in so-called blood oranges results in a range of colors from red to purple, characterizing the flesh and the rind of varieties such as Moro' Tarocco, Sanguigno, and Sanguinello (Caruso et al., 2016). The compound may also change the color of other parts of plant tissues, including young leaves, petals, stamens, styles, and stigmas (Fabroni et al., 2016; Catalano et al., 2020).

The carotenoid pathway in plants is also extensively researched. Biosynthetic genes and their expression profiles, as well as corresponding enzymes, have been described (Kato et al., 2004; Rodrigo et al., 2004; Kato et al., 2006; Tanaka et al., 2008; Chen et al., 2010), although the regulatory genes controlling the pathways are largely unknown. Carotenoid biosynthesis starts from the condensation of two geranylgeranyl diphosphates to produce 15-cis-pythoene, and then a series of desaturations induces the production of lycopene. At this crucial point, the pathway splits into two branches catalyzed by the lycopene cyclase family (*LCY*) genes, in particular epsilon- (ϵ-*LCY*) and beta-cyclase (β-*LCY*); they code for the corresponding enzymes lycopene ϵ-cyclase (*LCY*ϵ) and lycopene β-cyclase (*LCY*β). Both of these are responsible for cyclization and thus degradation of lycopene (Sandmann, 2002). LCYϵ forms monocyclic δ-carotene, which is a substrate used by LCYβ to produce α-carotene. LCYβ can also add two β-rings to lycopene, leading to the biosynthesis of β, β-carotenoids. In citrus, the expression of *LCYs* is different during the transition from chloroplasts to chromoplasts, corresponding to the fruit color break stage, characterized by downregulation of ϵ-*LCY* and overexpression of β-*LCY* (Kato et al., 2004; Alquézar et al., 2008a). Two β-*LCY* genes (β-*LCY1* and β-*LCY2*) have been isolated from sweet orange and grapefruit (Alquézar et al., 2009; Mendes et al., 2011). The β-*LCY1* is expressed in green tissues (leaves, roots, petals, and fruit), while the β-*LCY2* is strongly induced in fruit tissues, particularly in flavedo and pulp (Alquézar et al., 2009). Gene expression and transcriptomic studies on lycopene-pigmented mutants compared to non-pigmented fruits have revealed that lycopene accumulation in pulp is attributable to downregulation of β-*LCY2* gene expression (Alquézar et al., 2008b; Alquézar et al., 2009; Mendes et al., 2011; Alquézar et al., 2013; Liu et al., 2016; Jiang et al., 2019; Lana et al., 2020; Promkaew et al., 2020; Tatmala et al., 2020; Zhang H. et al., 2021).

Some sweet orange varieties (Cara Cara, Hong Anliu, Vaniglia sanguigno, Kirkwood Red), a lemon (Pink fleshed), several pummelos (Chandler, Tubtim Siam, Siam Red Ruby, Thong Dee, Pomelit), and grapefruits (Star Ruby, Flame, Ruby Red, Thompson) are characterized by lycopene accumulation in the pulp, and in some cases also in the rind. The intensity of lycopene pigmentation is rather variable in the different varieties.

To date, no reported study has developed citrus varieties with both anthocyanins and lycopene in their pulp, likely because these traits are difficult to combine through traditional breeding approaches (Salonia et al., 2020). In tomato, the crop with the highest content of lycopene, the activation of the anthocyanin biosynthetic pathway has successfully been performed *via* genetic engineering (Butelli et al., 2008) and traditional breeding (Hazra et al., 2018) approaches. In citrus, traditional breeding takes a long time and requires substantial resources to obtain progeny and evaluate their traits. Moreover, the generation of citrus hybrids that accumulate both compounds is not always feasible because some cultivars are sexually incompatible, sterile, polyembrionic (Talon and Gmitter, 2008), or chimeric (Caruso et al., 2020). One way around this may be to use so-called new plant breeding techniques, which are advanced technologies of genetic engineering that can induce DNA modifications that may be indistinguishable from naturally evolved ones. Genome editing *via* clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein 9 (Cas9) represents a promising strategy to induce a target mutation into a gene of interest controlling a specific trait without modifying the rest of the genome.

Genome editing *via* CRISPR/Cas9 has been widely applied to fruit crops, for example, to induce resistance against *Botrytis cinerea* in grape (Wang et al., 2018) and *Plasmopara viticola* in grapevine (Li et al., 2020), to produce apple varieties resistant to fire blight caused by *Erwinia amylovora* (Pompili et al., 2020), and to produce early-flowering genotypes by knockout Terminal Flower 1 in pear (Charrier et al., 2019) and CENTRORADIALIS in kiwifruit (Varkonyi-Gasic et al., 2019). In citrus, CRISPR/Cas9 has been used exclusively to introduce resistance against citrus canker disease in grapefruit (Jia et al., 2016; Jia et al., 2019) and sweet orange (Wang et al., 2019). Several studies have used genome editing to improve fruit quality (Li et al., 2018a; Xing et al., 2020; Li et al., 2022), and some of them have focused on the lycopene accumulation pathway. Li et al., (2018b) promoted the biosynthesis of lycopene, inhibiting the conversion from lycopene into β- and α-carotene in tomato. Similarly, genome editing of β-cyclase was used to develop a β-carotene-enriched banana variety (Kaur et al., 2020). Following these examples and pursuing an approach that has already been used in tomato (Zsögön et al., 2018; Natalini et al., 2021), we targeted *β-LCY2* using a dual sgRNA approach to produce loss-of-function mutants that stimulate lycopene accumulation in anthocyanin-rich sweet oranges.

Material and methods

Plant material

Mature fruits of 'Carrizo' citrange (*Citrus sinensis* L Osbeck × *Poncirus trifoliata* L. Raf.) and the 'Valencia', 'Doppio sanguigno', 'Vaccaro', 'Tarocco TDV', 'Tarocco Lempso', 'Bud Blood' sweet orange (*C. sinensis* L. Osbeck) were collected from December to February (2019-2020) from the CREA citrus germplasm of Palazzelli (Lentini, Siracusa, Italy; 37°20'22" N, 14°53'31" E).

Optimization of regeneration protocol

Polyembrionic seeds of 'Valencia', 'Doppio Sanguigno' and 'Tarocco TDV' were sterilized with 1% hypochlorite solution and washed three times with sterile water. Then they were sown in tubes containing basal Murashige and Skoog (MS) medium (25 g/L sucrose, 4.4 g/L MS basal medium including vitamins, 7 g/L python agar) and incubated at 25°C under dark conditions for 4–5 weeks. Nucellar seedlings of about 10 cm were recovered from each seed and used for regeneration and transformation experiments. Internodal stem segments were cut and used in regeneration tests. Four different MS media with different hormone concentrations, RDM1, RDM2, RSM1, and RMS2 (Supplementary Table 1), were tested to evaluate regeneration efficiency measured after 1 month of culture with a 16 h photoperiod at 25 ± 1°C.

Citrus transformation protocol

Agrobacterium-mediated transformation was carried out for all varieties using internodal stem segments from nucellar seedlings as explants, except in the case of 'Tarocco Lempso' for which we used embryogenic calli (obtained from unfertilized ovules cultivated in liquid basal MS medium), and 'Bud Blood', for which we used cotyledons (started from sterilized seeds). Transformation experiments were performed using a previously reported method (Orbović and Grosser, 2015) with minor modifications. After *Agrobacterium* infection and co-cultivation, explants were cultured in RDM1 regeneration medium (25 g/L sucrose, 4.4 g/L MS basal medium including vitamins, 1 mg/L BAP) with antibiotics (70 mg/L kanamycin and 400 mg/L cefotaxime) at 25 ± 1°C for 2 weeks under dark conditions and then transferred to a 16 h photoperiod. Explants were transferred to fresh medium every 4 weeks to stimulate the production of transgenic shoots. After 8–10 weeks, shoots were separated from explants and cultured in RDM1 medium with antibiotics to enhance the effect of selection medium. *NptII*-kanamycin-resistant 'Doppio Sanguigno', 'Vaccaro', 'Tarocco TDV', 'Tarocco Lempso' and 'Bud Blood' shoots were cultured in MS basal medium including vitamins for 2–3 weeks to stimulate plant growth and enhance the efficiency of mini grafts on 'Carrizo' rootstock. 'Carrizo' shoots, used as transformation control, were cultured in MS medium with 0.5 mg/L NAA for 4–5 weeks to induce rooting.

Design of sgRNAs

The plasmid for genome editing (pDGB3_alpha1) has been constructed using the GoldenBraid 3.0 system (Supplementary Table 2) (Vazquez-Vilar et al., 2017). Two sgRNAs were designed with the coding sequence of *β-LCY2* (FJ516404) and chosen among the most suitable according to the parameters indicated in the Benchling (www.benchling.com) and CRISPR-P 2.0 (Lei et al., 2014; Liu et al., 2017) tools (Supplementary Table 3). The criteria used for selection of the sgRNAs were the presence of the same sgRNA in both tools, a distance of no more than 300 bp between the sgRNAs, and an on-target and off-target score higher than 50%. In this study, the sgRNAs were spaced 231 bp from each other, and the on-target score was 71% for sgRNA1 and 55% for sgRNA2 in CRISPR-P 2.0 and 69.8% (sgRNA1) and 50% (sgRNA2) in Benchling. The presence of off-target was verified in both tools, and none of the off-target were found with less than two mismatches in 'seed region' of the sgRNA. Although the sgRNAs were designed based on sweet orange (the species that will be edited), we confirmed that no mutations occurred in the 'Carrizo' during this process; to do this, a *β-LCY2* sequence was blasted against the *P. trifoliata* genome using the Phtyozome blast tool (https://phytozome-next.jgi.doe.gov/blast-search).

Vector construction

The sgRNAs were domesticated and were used to assemble a CRISPR/Cas9 construct, following the iterative cloning strategy of GoldenBraid 3.0 (Vazquez-Vilar et al., 2017; www.gbcloning.upv.es). Each assembled vector was validated through enzymatic digestion (Supplementary Figure 1). The final pDGB3_alpha1 plasmid contained both sgRNAs with a U6 promoter and sgRNA scaffold; the *Cas9* with a 35S promoter and a NOS terminator; and a *nptII* selectable marker gene with a NOS promoter and terminator (Figure 1A). The correct assembly of the final vector was checked by Sanger sequencing, using a set of primers that included both sgRNAs (Supplementary Table 3). Sanger sequencing was performed using the 3130 Genetic Analyzer (Applied Biosystem) according to a previous study (Arlotta et al., 2020). The vector was used for the transformation of *A. tumefaciens* strain EHA105.

Identification of transformed plants

Resistant shoots were analyzed to detect the presence of *nptII* selectable marker and *Cas9* genes. DNA was extracted using the CTAB protocol with a few modifications (Arlotta et al., 2020). PCR was performed using Taq DNA Polymerase, following the manufacturer's instructions (VWR, Life Science, Haasrode, Belgium). Amplification conditions were 1 cycle at 95°C for 2 min; 35 cycles at 95°C for 25 s, 57°C for 30 s, and 72°C for 70 s; and a final cycle at 72°C for 5 min. All products were visualized *via* gel electrophoresis (1.5% agarose and 2.5 μL GelRed in 100 mL TAE 1X). Positive 'Carrizo' shoots were transferred in jiffy substrate under growth chamber conditions (25 ± 1°C, a 16 h photoperiod, 90% relative humidity), while 'Doppio Sanguigno', 'Vaccaro', 'Tarocco TDV', 'Tarocco Lempso', 'Bud Blood' shoots were mini grafted on 'Carrizo' rootstock (Supplementary Figure 2). PCR amplification was repeated with DNA extracted from leaves of mini grafted plantlets after 5 months. The primer sequences are listed in Supplementary Table 3.

Detection of *β-LCY2*-edited plantlets *via* PCR

The genomic DNA of 66 'Doppio Sanguigno', 7 'Vaccaro', 15 'Tarocco TDV', 10 'Tarocco Lempso', 13 'Bud Blood', and 149 'Carrizo' transformed plants was extracted from lateral leaves and apical shoots. The DNA was used as template to amplify a 380 bp fragment using a primers set (Supplementary Table 3) designed to amplify the region containing both sgRNAs on *β-LCY2* (Figure 1B). The amplification protocol was the same as described above. All PCR products were visualized *via* gel

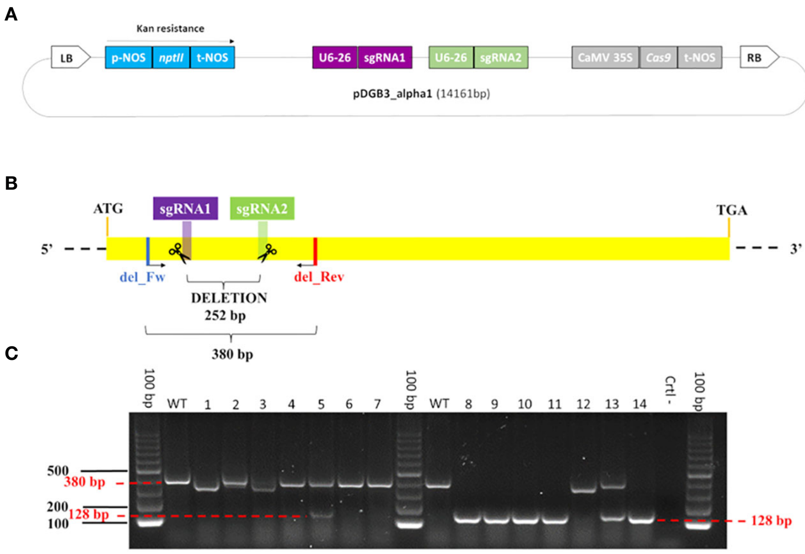


FIGURE 1
Genome editing construct for the *β-LCY2* gene and validation of edited plantlets. **(A)** Map (not at scale) of the final vector p-DGB3_alpha1 (14.161 bp). From the left border (LB) to the right border (RB), the main portions of the construct are: the *neomycin phosphotransferase II* gene (*nptII*, light blue) with the NOS promoter (p-NOS) and terminator (t-NOS); the two sgRNAs (purple and green) each with the U6-26 promoter; the *Cas9* gene (silver) with the CaMV 35S promoter and NOS terminator. **(B)** Schema of sgRNAs into *β-LCY2*. The size of the deletion between the sgRNAs is 252 bp; the length of the amplicon using the del_Fw and del_Rev primers is 380 bp. **(C)** Gel electrophoresis showing the PCR products of edited plantlets. Samples #4, #6, #7, and #12 show profile I, in which one or both sgRNAs are putatively edited; samples #8, #9, #10, #11, and #14 show profile II, consisting of a large deletion of 252 bp, with an amplicon of 128 bp; samples #5 and #13 show profile III, consisting of two amplicons of 380 bp and 128 bp; samples #1, #2, and #3 show amplicons whose lengths are a little bit different than expected, due to a diverse editing event, and described as profile IV; 100 bp is the ladder; WT is the wild type; Ctrl- is the negative control.

electrophoresis (1.5% agarose and 2.5 μL GelRed (Biotium, Fremont, California, USA) in 100 mL TAE 1X).

Detection of *β-LCY2* editing events *via* high-throughput sequencing

The *β-LCY2* CRISPR/Cas9-targeted region of 58 transformed citrus lines (36 for 'Doppio Sanguigno', 2 for 'Vaccaro', 5 for 'Tarocco TDV', 2 for 'Tarocco Lempso', 5 for 'Bud Blood', and 8 for 'Carrizo') was screened *via* high-throughput sequencing (HTS). The *β-LCY2* region containing both target sites was amplified using specific primers with overhang Illumina adapters (Supplementary Table 3). PCR was performed with 12.5 μL PCRBIO HS Taq Mix Red (PCRBiosystems, London, UK), 0.4 μM each primer, and 25 ng DNA template in a final volume of 25 μL. Amplification conditions were 1 cycle at 95°C for 5 min; 33 cycles at 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and a final cycle at 72°C for 5 min.

After purification and quantification, the pooled amplicon libraries were sequenced on an Illumina MiSeq platform (MiSeq Control Software 2.0.5) as reported by Quail et al. (2012). The CRISPResso2 pipeline (https://crispresso.pinellolab.partners.org/submission; Pinello et al., 2016) was used to process the raw paired-end reads (PE300), saved into 'fastq' files, to visualize the mutation profiles of the *β-LCY2* target sequence. CRISPResso2 default parameters were used, except for *double guide* mode and *in BATCH* too, and in standalone environment. Moreover, for plantlets coded 11Dk, 15Dk, 32Da, 34Da, 521A, and 7Dk the CRISPResso2 settings were modified using 30% minimum homology for alignment instead of the 60% under default parameter.

For three samples, the CRISPResso2 analysis failed due to the presence of a presumable inversion, even though this event was correctly visualized through basic functions of STAR (Dobin et al., 2013; http://code.google.com/p/rna-star/) and SAMtools. Two primers were designed to confirm the reliability of the inversion (Supplementary Table 3) and amplification was performed using Taq DNA polymerase (VWR, Life Science, Haasrode, Belgium), as described above.

Results

Optimization of regeneration protocol and identification of the suitable medium for transformation experiments

A minimum of 30 explants of 'Valencia', 'Doppio Sanguigno', 'Tarocco TDV' were separately cultivated in four MS media (RDM1, RDM2, RSM1, and RMS2) (Supplementary Table 1). 'Valencia' was used as a control because this variety can be efficiently regenerated using available protocols (Boscariol et al., 2003). The regenerant shoots were obtained from indirect organogenesis through callus stage. RDM1 medium showed the highest percentage of explants producing shoots (PEPS) for all varieties (Supplementary Table 4); 'Doppio Sanguigno' had a PEPS of 90% in both RDM1 and RMS1, higher than that of 'Valencia' (80%). RDM1 medium also induced the highest regeneration efficiency in 'Tarocco TDV' (Supplementary Table 4). Based on these results, we used RDM1 as the most efficient medium for *Agrobacterium*-mediated transformations of all blood orange varieties and 'Carrizo' citrange.

Generation of dual sgRNA construct for the editing of *β-LCY2*

As already mentioned, the genome editing vector was assembled following the interactive cloning strategy of GoldenBraid 3.0 (https://gbcloning.upv.es/), one of the most flexible and feasible approaches for designing such vectors (Figure 1A; Supplementary Table 2). The assembly of two vectors, each with one of the sgRNAs (Level 1, Supplementary Figure 1), required careful evaluation of several colonies; in fact, the integration efficiency of the sgRNAs was low (1 in 30 screened white colonies was positive). The two vectors of Level 2 (Supplementary Figure 1), one containing the sgRNA2 cassette and *nptII* (omega_1R) and the other containing sgRNA1 and *Cas9* (omega 2), showed an integration of almost 100%. In our experience, the integration of sgRNAs with U6 promoter and RNA scaffold was more complicated than the assembly of the sgRNAs with *Cas9* and *nptII*. The production of the final vector, consisting of *nptII*, *Cas9*, and both sgRNAs, was validated through PCR using 10 positive colonies, 2 of which were sequenced, producing a 1,220 bp size amplicon (Supplementary Figure 3).

Production of *nptII*- and *Cas9*-positive blood orange varieties

A total of 510 'Doppio Sanguigno', 260 'Vaccaro', 300 'Tarocco TDV', 160 'Tarocco Lempso', 400 'Bud Blood', and 210 'Carrizo' explants were infected with *A. tumefaciens*. After about 12 weeks of culturing on selected medium (RDM1 with 70 mg/L kanamycin and 400 mg/L cefotaxime) explants produced kanamycin-resistant shoots as follows: 92 'Doppio Sanguigno', 43 'Vaccaro', 101 'Tarocco TDV', 22 'Tarocco Lempso', 24 'Bud Blood', and 413 'Carrizo' (Table 1). The presence of *nptII* and *Cas9* was verified through PCR (Supplementary Figure 4). All of the 92 'Doppio Sanguigno' were tested twice, as shoots (coming from the explants) and as plantlets (testing the leaves that resulted from mini grafting). For the other citrus varieties, we were unable to screen all kanamycin-resistant shoots, because some were not recovered due to contamination issues. Therefore, only

TABLE 1 Transformation efficiency of ‘Doppio sanguigno’, ‘Vaccaro’, ‘Tarocco TDV’, ‘Tarocco Lempso’, ‘Bud Blood’ sweet oranges, and ‘Carrizo’ citrange.

Variety	Infected explants	*nptII* resistant regenerants	Regenerants tested by PCR	*nptII/Cas9* PCR positive	TE*
‘Doppio sanguigno’	510	92	92	66	12%
‘Vaccaro’	260	43	7	5	2%
‘Bud blood’	400	24	13	9	2.5%
‘Tarocco TDV’	300	101	15	8	2.7%
‘Tarocco Lempso’	160	22	10	2	1.3%
‘Carrizo’	210	413	219	149	70%

*The transformation efficiency was calculated dividing the number of regenerants positive for nptII and Cas9 by the number of infected explants.

noncontaminated shoots were mini grafted and screened to confirm to the presence of *nptII* and *Cas9*. Integration of both genes was confirmed in 66 out of 92 ‘Doppio Sanguigno’, 5 out of 7 ‘Vaccaro’, 8 out of 15 ‘Tarocco TDV’, 2 out of 10 ‘Tarocco Lempso’, 9 out of 15 ‘Bud Blood’ and 149 out of 219 ‘Carrizo’, resulting in a transformation efficiency (TE) of 12%, 2%, 2.7%, 1.3%, 2.5%, and 70%, respectively (Table 1).

Analysis of plantlets with the *β-LCY2* gene

The *nptII*- and *Cas9*-positive plantlets were screened through PCR amplification to detect the large deletion between the two sgRNAs. Overall, we obtained four main profiles (Figure 1C; Supplementary Table 5): (I) plantlets showing the amplification of one band of 380 bp, presumably having point mutations or short indels within one or both sgRNAs; (II) plantlets in which amplification produced one band of 128 bp, due to the deletion of the region between the sgRNAs; (III) plantlets with two amplicons, whose length were 128 bp and 380 bp; and (IV) other samples displaying amplicons different from 128 bp and 380 bp (Supplementary Table 6). Overall, most plantlets (180 out of 239) showed a single amplicon of 380 bp, resulting presumably edited. Moreover, we found 10 plantlets of ‘Doppio Sanguigno’ and 18 of ‘Carrizo’ with the deletion of the region between the two sgRNAs, suggesting a loss-of-function of the *β-LCY2* target gene. Furthermore, 6 ‘Doppio sanguigno’, 1 ‘Tarocco TDV’ and 12 ‘Carrizo’ plantlets showed both amplicons, indicating that they could be heterozygous or chimeric for the editing events. Finally, 12 samples had a profile that could not be classified (see profile IV).

The *β-LCY2* target region was screened from a subset of 50 transformed lines (36 ‘Doppio Sanguigno’, 2 ‘Vaccaro’, 5 ‘Tarocco TDV’, 2 ‘Tarocco Lempso’, 5 ‘Bud Blood’) *via* HTS. One wild-type plant (not subjected to transformation) of each variety was also sequenced and used as a negative (not edited) control. Moreover, we added 8 plantlets of ‘Carrizo’, simply to compare the editing events in the model accession. On average, 24,000 raw sequence reads aligned for anthocyanin-pigmented sweet oranges and 18,000 for ‘Carrizo’ were obtained for each of the analyzed plantlets. This comprehensive coverage supports the accuracy of the editing events produced.

Among the anthocyanin-pigmented sweet oranges, 86% of ‘Doppio Sanguigno’, 100% of ‘Vaccaro’, 100% of ‘Tarocco TDV’, and 80% of ‘Bud Blood’ of successfully transformed plantlets resulted edited and the mutations occurred in the target site of both sgRNAs; The remaining 14% of ‘Doppio sanguigno’ and 20% of ‘Bud Blood’ were not mutated, as well as two plants of ‘Tarocco Lempso’. All of the ‘Carrizo’ plantlets were successfully edited, and mutations were observed in both sgRNAs too (Supplementary Figure 5A). The dual sgRNA strategy clearly maximized the modification of citrus plants.

Detection of point mutations and indels in *β-LCY2*

Most of the analyzed samples showed full knockout, where 100% of the reads were edited. A limited number of mutated plantlets (4 ‘Doppio Sanguigno’, 3 ‘Tarocco TDV’, 1 ‘Bud Blood’, 1 ‘Carrizo’) showed from 11-98% mutated reads, and could be considered heterozygous or more likely chimeric, while the rest were non-mutated reads, identical to the profile of the wild type (Supplementary Figure 5B).

Several types of mutations were identified and classified as follows: point insertions (+1 nt), small deletions (-1, -2, -3, -4, -5, -7, -8, -12, -15; -23, -25, -26, -29, -51, -56, -74 nts), and substitution (transition of T into C) (Supplementary Table 7). The most frequent type of mutation was an insertion in sgRNA1 and deletion in sgRNA2 (Supplementary Table 7; Supplementary Figure 6). Therefore, we expect that these indels could produce a frameshift and thus incorrect translation of the protein. Supplementary Figure 5C and Supplementary Table 6 show the mutations in the protein sequence, specifically those of homozygous plantlets with a single editing event (Supplementary Figure 5B). Overall, the different editing events caused the introduction of a premature stop codon. A truncated protein is also the result of the deletion of 252 bp, the region between the two sgRNAs (Supplementary Figure 5C).

Inversion, a novel editing event

In addition to point mutations and indels, an unexpected result was observed in two plantlets of 'Doppio sanguigno' (32Da, 34Da) and one of 'Bud Blood' (521A). For these samples, CRISPResso2 failed to align the reads, classifying the editing event as a putative large insertion, although the amplicon size was of 380 bp. To clarify this unexpected result, we realigned the reads against sweet orange genome 3.0 through SAMtools and STAR and visualized the results using the IGV tool. This is because STAR software, which uses the spliced alignment strategy, can handle large deletion in the amplicon as it does with introns. Surprisingly, it was clear that the region between the sgRNAs was inverted compared to the original sequence. Furthermore, to confirm the *in silico* analysis, we designed two primers within the region subjected to the potential inversion, each one coupled with the primers located upstream of sgRNA1 and downstream from sgRNA2 (Figure 2A). The PCR amplification was performed in four different combinations, either considering the absence or the presence of the probable inversion. The wild-type 'Doppio sanguigno' exclusively showed the amplification of the noninverted region (73B). In contrast, the three samples with the suspected inversion showed a profile including the wild type and the inversion. Specifically, the presumable wild-type profile of the 521A 'Bud blood' could have been due to the very short mutations that occurred in about 35% of the mutated reads (Supplementary Table 8); the presence of the inversion was confirmed by the third and fourth primers (Figure 2B). A distinct profile occurred in the 32Da and 34Da 'Doppio sanguigno' plantlets, in which the presumable wild-type (second primer combination) could be the one typical of plantlets with the length of the amplicon different from 380 bp, as described above (Supplementary Table 6).

Logic procedure behind the identification of additional inversions and large deletions

Based on our experience, optimized transformation and regeneration protocols should be followed by a proper bioinformatic analysis to correctly identify induced mutations. Simple detection of editing events is desirable, and the double-guides approach allowed rapid screening through PCR with visualization *via* gel electrophoresis. However, the PCR results contrasted with the first bioinformatic analysis by HTS. The CRISPResso2 tool failed to identify the inversion and the large deletion between the sgRNAs. The insertion visualized by CRISPResso2 contrasted with the band sizes shown in

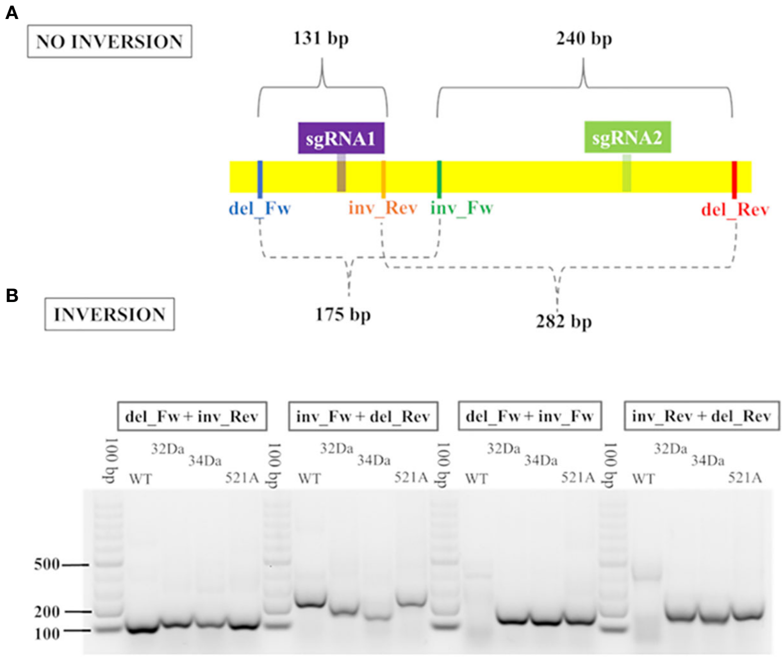


FIGURE 2
Validation of potential inversion *via* PCR. **(A)** Graphical representation of primers incorporated into the region subjected to potential inversion. Each primer del_Fw and del_Rev was coupled with inv_Fw and inv_Rev, for a total of four different combinations. The length of the amplicons produced by each combination is indicated. **(B)** Gel electrophoresis displays the PCR amplicons obtained through the four primer combinations, used to confirm the potential inversion. WT corresponds to 'Doppio sanguigno' wild type, 32Da and 34Da are the codes of 'Doppio sanguigno' plantlets, 521A is the code of the 'Bud Blood' plantlet.

electrophoresis and resulted in the alignment of less than 10 out of 30,000 reads produced by sequencing. We were able to correctly identify the inversion by comparing the PCR/gel results and the CRISPResso2 results using STAR, an additional tool for bioinformatics analysis, which clarified that the large insertion associated with several SNPs (previously visualized by CRISPResso2) corresponded to an inversion. The rationale is that the STAR software, using the spliced alignment strategy, can handle the large deletion in the amplicon as it can with introns. In particular, the amplification of sample 32Da represented the third profile, consisting of a large deletion and a second band similar in size to the one of the wild type. CRISPResso2 analysis displayed an unclear profile for this sample, which looked like a reverse complement of part of the sequence between the two single guides. Approaching the raw reads using STAR produced two alleles, consisting of the deletion represented by splice junctions and the inversion represented as soft-clipped reads between the cutting sites of the guides. In this way, we were able to confirm that the clipped parts were inverted. A schematic of the logic process used to solve the putative inversions and large deletions is shown in Figure 3.

CRISPResso2 showed similar results using 30% minimum homology for alignment to the amplicon. Then about 100% of processed reads were correctly aligned for the large deletion; for the inversion, the fragments external to the cutting sites resulted in a mutated profile. PCR confirmed the inversion in all of the suspected samples.

Sequencing of leaves and apical shoots confirms editing events

Due to the surprisingly high number of homozygous T_0 citrus plantlets (Supplementary Figures 5A, B), we decided to verify if those plants were entirely mutated. To this end, we selected three samples (28Dc and 4DK 'Doppio sanguigno', 292A 'Vaccaro') previously sequenced, and analyzed them again using about 10 mm apical shoots (a more homogenous sample containing all the meristem layers (Sugawara et al., 2002)) using HTS. Moreover, we added a heterozygous sample (15DK 'Doppio sanguigno') as a control, and one of the plantlets showing the inversion (521A 'Bud Blood'), which gave us the opportunity to verify if the Cas9 continued to cut 7 months after the DNA extraction and amplicon sequencing. The results of the amplicon sequencing confirmed the homozygous profile for the three plantlets (28Dc, 4DK, 292A), as well as the heterozygosity of 15DK, although they indicated a slightly different percentage of editing events. The unexpected complete absences of the wild type *b-LCY2* gene sequence in some of the edited plants could allow to observe the induced phenotype in T_0. Regarding sample 521A, in the second round of sequencing, the inversion of the region between the two sgRNAs targets was maintained, although the percentage of mutated reads was lower than the previous sequencing. These data are reported in Supplementary Table 8.

Discussion

Optimization of blood orange regeneration

The application of genome editing technology is possible only if the variety of interest is able to efficiently regenerate after editing. Therefore, much work and resources are needed for the optimization of regeneration and transformation protocols of specific varieties. The unexpectedly high number of escapes may be attributable to a screening performed on acclimated plants,

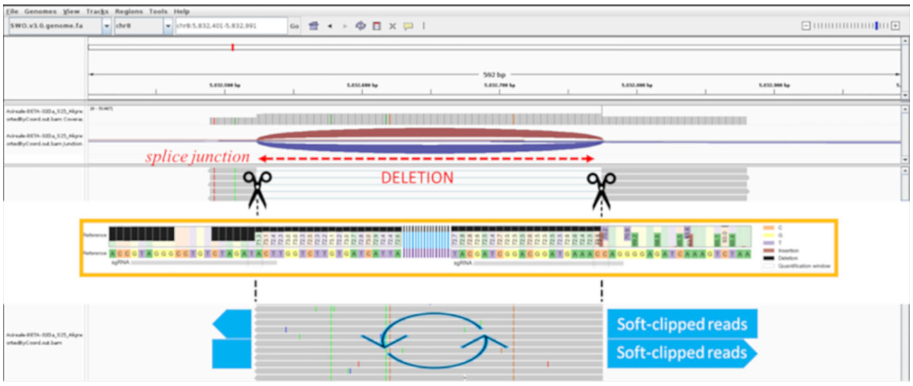


FIGURE 3
IGV screenshot of the 32Da sample realigned by STAR. The scissors represent the cleavage sites of Cas9, the splice junction shows the deletion (indicated in red), light blue large arrows show the soft-clipped portion of the inverted reads. The orange box refers tois the CRISPResso2 output (not to scale with IGV) of the 32Da with a 30% minimum alignment score; the fragments external to the cleavage sites result in a mutated profile with deletions on the left, and insertions and SNPs on the right (as reported in the legend).

that were not exposed to kanamycin pressure for several weeks; therefore, the resulting positive plants displayed a stable integration of the T-DNA. In citrus, *Agrobacterium*-mediated transformation is the most widely used approach to obtain transgenic plants. In fact, about 90% of transformation experiments are performed using this methodology (Gong and Liu, 2013). However, regeneration of transformed plants represents a bottleneck in citrus, which does not readily regenerate (Cervera et al., 2008). The composition of basal medium and the concentration of hormones positively influence organogenesis response and regeneration efficiency (Cervera et al., 2005; Rodríguez et al., 2008). In our study, the RDM1 medium was the most efficient.

As reviewed in Poles et al. (2020), citrus transformation is accession-specific, easy to perform for 'Carrizo' citrange, 'Duncan' grapefruit, and 'Valencia', 'Pineapple' and 'Jincheng' sweet oranges. Other species, such as Clementine and sour orange, are considered recalcitrant. Our preliminarily observations suggested that among sweet oranges, 'Doppio Sanguigno' was the best genotype to use for transformation. Its transformation and regeneration efficiency were lower than 'Carrizo', but still relatively high for editing experiments compared to other varieties tested in the present work. Two previous studies have reported regeneration and transformation protocols for anthocyanin-pigmented citrus varieties. One described transformation of mature tissues of 'Tarocco', showing 72.9% and 9.1% regeneration and transformation efficiencies, respectively (Peng et al., 2019). Other study described transformation of young tissues of 'Maltese half-blood', in which the regeneration and transformation efficiency was 44.6% and 21.4%, respectively (Jardak et al., 2020). In general, our results are comparable to previous results of other sweet orange genotypes, such as 'Valencia' (23.8% TE), Hamlin' (12.8% TE), 'Pineapple' (6.1% TE), and 'Jincheng' (4.7% TE) (Cervera et al., 1998; Boscariol et al., 2003; Zou et al., 2013; Orbović and Grosser, 2015). Our work contributes to the optimization of regeneration and transformation protocols for pigmented blood oranges.

The dual-sgRNA approach, an efficient method to knockout *β-LCY2*

In the last 9 years, since the first study of genome editing technology was published (Gaj et al., 2013), many cloning strategies, adaptable to most applications and plant species, were developed. The generation of genome editing constructs relies on the possibility of using CRISPR/Cas9 plasmids that can be designed using different Cas protein types (Cas9 or Cas12a) or Cas-like and cloning methods (e.g., restriction enzyme ligation, Gateway cloning, Golden Gate assembly). Moreover, different approaches that are able to disrupt gene function could be performed, using one or more sgRNAs (e.g., genome and base editing) to determine base conversions, deletions, insertions, and combination edits introduced into target genomic sites (i.e., prime editing) (Jin et al., 2021). In citrus, the few studies that have applied CRISPR/Cas9 technology used the Golden Gate strategy. This approach has been adopted to introduce resistance to *Xanthomonas citri* (the causal agent of citrus canker disease) to susceptible citrus species (Jia et al., 2016; Jia et al., 2017), to produce a double thorn phenotype in 'Carrizo' citrange (Zhang et al., 2021b), and to induce resistance against the herbicide imazapyr (Alquézar et al., 2022).

In our study, we used knockout of the *β-LCY2* gene to combine anthocyanin and lycopene accumulation in the same fruit. The Benchling and CRISPR-P tools were used to design the sgRNAs; therefore, to maximize as much as possible the efficiency of the sgRNAs to select, we chose those that were the same with both tools. Moreover, the same two sgRNAs were sufficiently separated in the target gene and lacked potential off-targets. The dual-sgRNA approach is not always feasible because it is typically only possible when the distance between two sgRNAs is about 300–500 bp. In our case, however, the distance was 231 bp. Generally, the efficiency of mutation increases when more than one sgRNA is designed in a single gene. The nuclease method is the most efficient for causing loss-of-function, leading to the deletion of a large essential part of the gene (Bortesi and Fischer, 2015). The dual-sgRNA approach also has the advantage of easy visual genotyping of mutants based on amplicon length. In our case, 19% of the *nptII*- and *Cas9*-positive plantlets exhibited the deletion between the sgRNAs, producing a truncated protein. Generally, the DNA cleavage efficiency is different among sgRNAs, as are off-target effects. In our study, the mutations in both sgRNAs occurred in 100% of the edited plantlets, leading to efficient inactivation of the gene. Our data indicate that the use of two sgRNAs maximizes the efficiency of gene knockout, similarly to what has been previously reported in pomegranate (Chang et al., 2019) and in citrus using triple guides (Zhang et al., 2020; Zhang F. et al., 2021).

Furthermore, the DNA cleavage by Cas9 at two sites of the gene may induce rearrangements in the gene sequence, as demonstrated by the six independent events of inversion that we observed in the Doppio sanguigno and Bud blood varieties. Inversions mediated by a dual sgRNA CRISPR/Cas9 targeting system have frequently been induced in animal cell lines (Choi and Meyerson, 2014; Essletzbichler et al., 2014; Li et al., 2015), suggesting that this method is a valid way to create targeted inversion mutations and gene deletions. A combined dual sgRNA/Cas9 system was developed for the creation of targeted deletions and gene replacement (Zhao et al., 2016), as well as inversion mutations (Zhang et al., 2016; Blayney et al., 2020) in *Arabidopsis* and in *Oryza* (Liang et al., 2016). In our case, we can speculate that the inversions were a consequence of plasmid design. In fact, the distance between the two sgRNAs in the plasmid and in the *β-LCY2* were similar (260 bp and 252 bp, respectively). A better understanding of this mechanism could lead to a new application of the CRISPR/Cas system. Therefore, we suppose that both sgRNAs, including the scaffold, could be transcribed as unique mRNA. This hypothesis is supported by the fact that the guides were oriented in the same direction. In all of the cases in which the inversion occurred (and probably in other cases), both guides had bonded with the Cas protein to create a unique complex, able to cut portions of gene of around 260 bp. In this way, the complex creates a nick in both sgRNAs, determining three probable options to the DNA repair through the production of small indels, large deletions, and inversions. In particular, the creation of the inversion simultaneously eliminates PAM sites and a portion of the sgRNAs, releasing the cleavage sites.

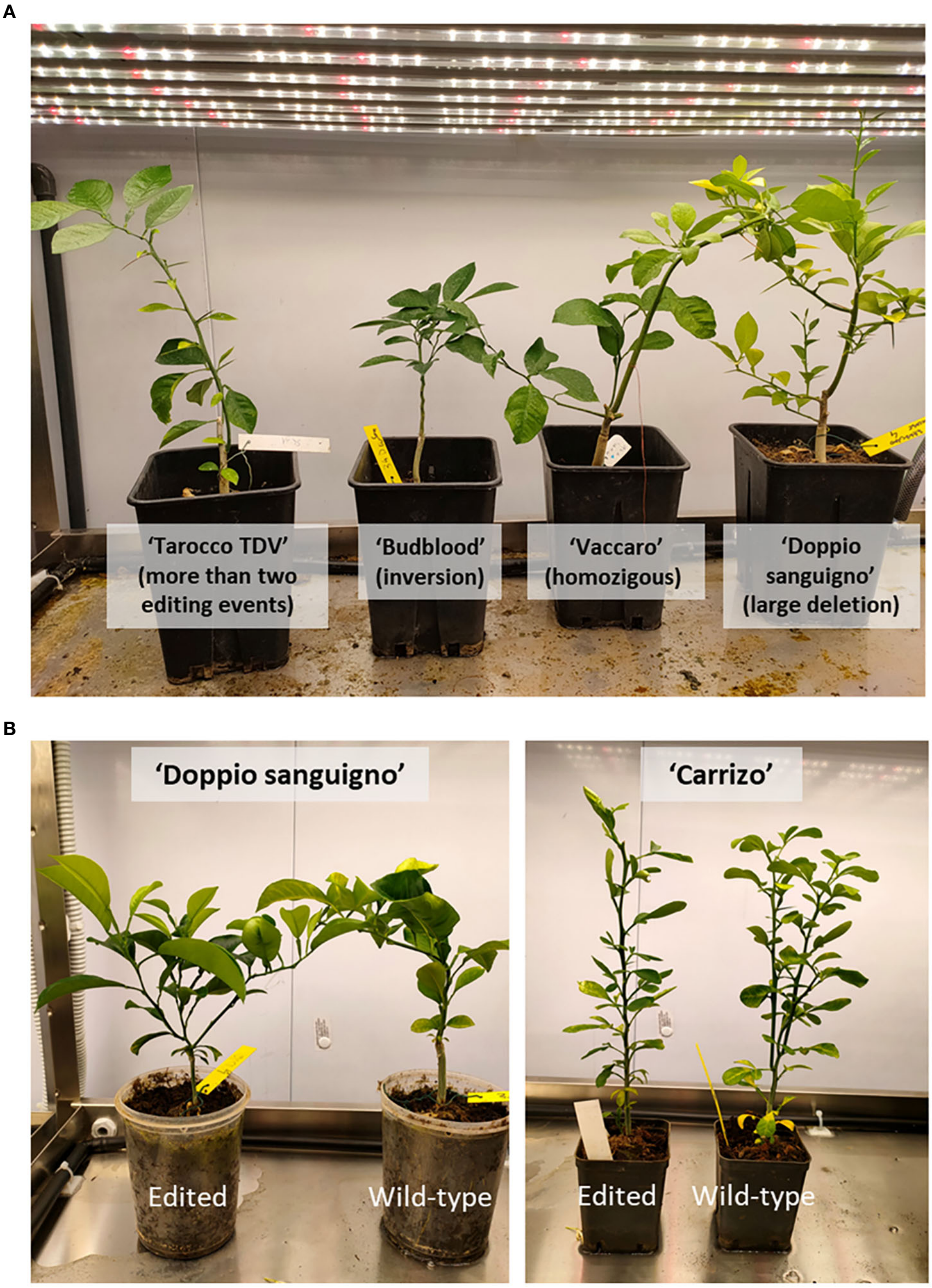


FIGURE 4
Citrus plantlets edited with the β-LCY2 gene. **(A)** Four editing events in four anthocyanin-rich varieties. Differences in the shapes of plantlets are independent of the mutation type. **(B)** Edited 'Doppio Sanguigno' and 'Carrizo' plantlets and related wild type. No differences in the phenotype are seen.

Promising fully edited citrus plants

The CRISPR/Cas9 construct was successfully produced, and 49 anthocyanin-pigmented sweet orange edited plantlets were generated. Moreover, 86% of regenerated and transformed plantlets were fully edited; among them, no heterozygous or chimeric events were observed in 43% of transformed 'Doppio Sanguigno', 50% of 'Vaccaro', 50% of 'Bud Blood', and 69% of 'Carrizo' plantlets. In these cases, we can hypothesize that the entire plant was mutated, as confirmed by the sequencing of different parts of the plantlets (leaves and apical shoots).

Among fully edited plantlets, we place particular attention on those in which a deletion of 252 bp occurred; this produced a specific knockout of *β-LCY2*, leading to the production of a putative truncated protein. A similar result was obtained in plantlets bearing the inversion. Because *β-LCY2* is a fruit-specific chromoplastic gene, we could expect that fruits produced from these plants could be phenotypically mutated. Those plantlets showed a normal phenotype (Figure 4), demonstrating that the knockout of *β-LCY2* did not alter the carotenoid metabolism of vegetative tissues.

Theoretically, the Cas9 protein will continue to cut because the recognition site is present in the DNA target (LeBlanc et al., 2018); suggesting that plantlets that are not edited in the first round of validation may be modified after several months. This is what we observed 5 months after the first PCR screening.

Conclusions

The application of genome editing, mainly if it addressed improving fruit traits of woody plants, is particularly challenging. The data showed in this study add new knowledge for citrus improvement, because genome editing was efficiently adopted to improve qualitative traits of different citrus cultivars. Our strategy generated several edited anthocyanin-pigmented sweet oranges, specifically 38 plantlets of 'Doppio sanguigno', 2 of 'Vaccaro', 5 of 'Tarocco TDV' and 4 of 'Bud blood', including non-chimeric genotypes, as supported by HTS analysis. The complete ablation of the functionally active *β-LCY2* gene in T_0 plants paves the road to the application of this technology to highly heterozygous and vegetatively propagated elite cultivars. For the first time this study reports the transformation of seeds in sweet orange. The transformation protocol for a series of anthocyanin-rich sweet oranges was never tested before. Moreover, the use of the mini grafting in non-sterile environment represents a novelty because it allows acclimation of transformed shoots. The long juvenile stage of citrus plants is a limiting step for the phenotypic evaluation of the edited plants. Grafting the obtained plantlets on early-flowering rootstocks could speed-up fruit production.

Data availability statement

The data presented in the study are depositated in the NCBI repository, accession number PRJNA853727.

Author contributions

FS did the constructs and wrote the manuscript; AC supported in all the bioinformatics analysis of editing events, managed the plants and contributed in the writing of the manuscript; HP and LP performed the optimization of transformation and regeneration protocols; MP and SL did the high throughput sequencing; PC reviewed the manuscript; MC grafted the plants and reviewed the manuscript; CL conceived the work and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work has been funded by Italian Ministry of Agriculture Food and Forestry through the project «BIOTECH Biotecnologie sostenibili per l'agricoltura italiana», subproject CITRUS Citrus improvement by sustainable biotechnologies (DM 15930/7305/2018).

Acknowledgments

Authors are grateful to Diego Orzaez and his laboratory for kindly providing Golden Braid 3.0 plasmids, and to Cinzia Comino who assisted in the preliminary steps of the construct assembly. The participation of FS to this work has been also supported by the "Incarico di associatura" with CREA Research Centre for Olive, Fruit and Citrus crops during 2022. The contribution of AC has been supported also by Fondazione Umberto Veronesi Post-doctoral Fellowships 2022.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.975917/full#supplementary-material

References

Alquézar, B., Bennici, S., Carmona, L., Gentile, A., and Peña, L. (2022). Generation of transfer-DNA-Free base-edited citrus plants. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.835282

Alquézar, B., Rodrigo, M. J., Lado, J., and Zacarías, L. (2013). A comparative physiological and transcriptional study of carotenoid biosynthesis in white and red grapefruit (Citrus paradisi macf.). *Tree Genet. Genomes* 9, 1257–1269. doi: 10.1007/s11295-013-0635-7

Alquézar, B., Rodrigo, M. J., and Zacarías, L. (2008a). Carotenoid biosynthesis and their regulation in citrus fruits. *Tree For. Sci. Biotechnol.* 2, 23–37.

Alquézar, B., Rodrigo, M. J., and Zacarías, L. (2008b). Regulation of carotenoid biosynthesis during fruit maturation in the red-fleshed orange mutant cara cara. *Phytochemistry* 69, 1997–2007. doi: 10.1016/j.phytochem.2008.04.020

Alquézar, B., Zacarías, L., and Rodrigo, M. J. (2009). Molecular and functional characterization of a novel chromoplast-specific lycopene β-cyclase from citrus and its relation to lycopene accumulation. *J. Exp. Bot.* 60, 1783–1797. doi: 10.1093/jxb/erp048

Anderson, J. W., Randles, K. M., Kendall, C. W., and Jenkins, D. J. (2004). Carbohydrate and fiber recommendations for individuals with diabetes: a quantitative assessment and meta-analysis of the evidence. *J. Am. Coll. Nutr.* 23 (1), 5–17. doi: 10.1080/07315724.2004.10719338

Arlotta, C., Ciacciulli, A., Strano, M. C., Cafaro, V., Salonia, F., Caruso, P., et al. (2020). Disease resistant citrus breeding using newly developed high resolution melting and CAPS protocols for alternaria brown spot marker assisted selection. *Agronomy* 10, 1368. doi: 10.3390/agronomy10091368

Barry, G. H., Caruso, M., and Gmitter, F. G.Jr. (2020). "Commercial scion varieties," in *The genus citrus*, *1st edition*. Eds. M. Talon, M. Caruso and F. G. Gmitter (Cambridge, UK: Elsevier) 83–104. doi: 10.1016/B978-0-12-812163-4.00005-X

Blayney, J., Foster, E. M., Jagielowicz, M., Kreuzer, M., Morotti, M., Reglinski, K., et al. (2020). Unexpectedly high levels of inverted re-insertions using paired sgRNAs for genomic deletions. *Methods Protoc.* 3, 53. doi: 10.3390/mps3030053

Bortesi, L., and Fischer, R. (2015). The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol. Adv.* 33, 41–52. doi: 10.1016/j.biotechadv.2014.12.006

Boscariol, R. L., Almeida, W. A. B., Derbyshire, M. T. V. C., Mourão Filho, F. A. A., and Mendes, B. M. J. (2003). The use of the PMI/mannose selection system to recover transgenic sweet orange plants (Citrus sinensis l. osbeck). *Plant Cell Rep.* 22, 122–128. doi: 10.1007/s00299-003-0654-1

Bramley, P. M. (2000). Is lycopene beneficial to human health? *Phytochemistry* 54, 233–236. doi: 10.1016/S0031-9422(00)00103-5

Butelli, E., Garcia-Lor, A., Licciardello, C., Las Casas, G., Hill, L., Recupero, G. R., et al. (2017). Changes in anthocyanin production during domestication of citrus. *Plant Physiol.* 173, 2225–2242. doi: 10.1104/pp.16.01701

Butelli, E., Licciardello, C., Ramadugu, C., Durand-Hulak, M., Celant, A., Reforgiato Recupero, G., et al. (2019). Noemi controls production of flavonoid pigments and fruit acidity and illustrates the domestication routes of modern citrus varieties. *Curr. Biol.* 29 (1), 158–164. doi: 10.1016/j.cub.2018.11.040

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., et al. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* 24, 1242–1255. doi: 10.1105/tpc.111.095232

Butelli, E., Titta, L., Giorgio, M., Mock, H. P., Matros, A., Peterek, S., et al. (2008). Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. *Nat. Biotechnol.* 26, 1301–1308. doi: 10.1038/nbt.1506

Carmona, L., Alquézar, B., Marques, V. V., and Peña, L. (2017). Anthocyanin biosynthesis and accumulation in blood oranges during postharvest storage at different low temperatures. *Food Chem.* 237, 7–14. doi: 10.1016/j.foodchem.2017.05.076

Caruso, M., Ferlito, F., Licciardello, C., Allegra, M., Strano, M. C., Di Silvestro, S., et al. (2016). Pomological diversity of the Italian blood orange germplasm. *Sci. Hortic.* 213, 331–339. doi: 10.1016/j.scienta.2016.10.044

Caruso, M., Smith, M. W., Froelicher, Y., Russo, G., and Gmitter, F.G. (2020). "Traditional breeding," in *The genus citrus*, *1st edition*. Eds. M. Talon, M. Caruso and F. G. Gmitter (Cambridge, UK: Elsevier), 129–148.

Catalano, C., Ciacciulli, A., Salonia, F., Russo, M. P., Caruso, P., Caruso, M., et al. (2020). Target-genes reveal species and genotypic specificity of anthocyanin pigmentation in citrus and related genera. *Genes* 11 (7), 807. doi: 10.3390/genes11070807

Cervera, M., Juárez, J., Navarro, L., and Peña, L. (2005). Genetic transformation of mature citrus plants. *Methods Mol. Biol.* 286, 177–188. doi: 10.1385/1-59259-827-7:177

Cervera, M., Navarro, A., Navarro, L., and Pena, L. (2008). Production of transgenic adult plants from clementine mandarin by enhancing cell competence for transformation and regeneration. *Tree Physiol.* 28, 55–66. doi: 10.1093/treephys/28.1.55

Cervera, M., Pina, J. A., Juárez, J., Navarro, L., and Peña, L. (1998). Agrobacterium-mediated transformation of citrange: Factors affecting transformation and regeneration. *Plant Cell Rep.* 18, 271–278. doi: 10.1007/s002990050570

Chang, L., Wu, S., and Tian, L. (2019). Effective genome editing and identification of a regiospecific gallic acid 4-o-glycosyltransferase in pomegranate (Punica granatum l.). *Hortic. Res.* 6, 123. doi: 10.1038/s41438-019-0206-7

Charrier, A., Vergne, E., Dousset, N., Richer, A., Petiteau, A., and Chevreau, E. (2019). Efficient targeted mutagenesis in apple and first time edition of pear using the CRISPR-Cas9 system. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00040

Chen, C., Costa, M. G. C., Yu, Q., Moore, G. A., and Gmitter, F. G. (2010). Identification of novel members in sweet orange carotenoid biosynthesis gene families. *Tree Genet. Genomes* 6, 905–914. doi: 10.1007/s11295-010-0300-3

Choi, P. S., and Meyerson, M. (2014). Targeted genomic rearrangements using CRISPR/Cas technology. *Nat. Commun.* 5, 1–6. doi: 10.1038/ncomms4728

Cooke, D., Steward, W. P., Gescher, A. J., and Marczylo, T. (2005). Anthocyans from fruits and vegetables–does bright colour signal cancer chemopreventive activity? *Eur. J. Cancer* 41 (13), 1931–1940.

Crifò, T., Puglisi, I., Petrone, G., Recupero, G. R., and Lo Piero, A. R. (2011). Expression analysis in response to low temperature stress in blood oranges: Implication of the flavonoid biosynthetic pathway. *Gene* 476, 1–9. doi: 10.1016/j.gene.2011.02.005

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29 (1), 15–21. doi: 10.1093/bioinformatics/bts635

Duarte, A., Carvalho, C., and Miguel, M. G. (2016). Bioactive compounds of citrus as health promoters. *Natural Bioactive Compounds Fruits Vegetables as Health Promoters* 1, 29–97.

Dutt, M., Erpen, L., and Grosser, J. W. (2018). Genetic transformation of the 'W murcott'tangor: comparison between different techniques. *Scientia Hortic.* 242, 90–94. doi: 10.1016/j.scienta.2018.07.026

Dutt, M., and Grosser, J. W. (2009). Evaluation of parameters affecting agrobacterium-mediated transformation of citrus. *Plant Cell Tiss. Organ Cult.* 98, 331–340. doi: 10.1007/s11240-009-9567-1

Dutt, M., Stanton, D., and Grosser, J. W. (2016). Ornacitrus: Development of genetically modified anthocyanin-expressing citrus with both ornamental and fresh fruit potential. *J. Am. Soc. Hortic. Sci.* 141, 54–61. doi: 10.21273/jashs.141.1.54

Erpen, L., Tavano, E. C. R., Harakava, R., Dutt, M., Grosser, J. W., Piedade, S. M. S., et al. (2018). Isolation, characterization, and evaluation of three citrus sinensis-derived constitutive gene promoters. *Plant Cell Rep.* 37 (8), 1113–1125. doi: 10.1007/s00299-018-2298-1

Essletzbichler, P., Konopka, T., Santoro, F., Chen, D., Gapp, B. V., Kralovics, R., et al. (2014). Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. *Genome Res.* 24, 2059–2065. doi: 10.1101/gr.177220.114

Fabroni, S., Ballistreri, G., Amenta, M., and Rapisarda, P. (2016). Anthocyanins in different citrus species: An UHPLC-PDA-ESI/MSn-assisted qualitative and quantitative investigation. *J. Sci. Food Agric.* 96, 4797–4808. doi: 10.1002/jsfa.7916

Gaj, T., Gersbach, C. A., and Barbas, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* 31, 397–405. doi: 10.1016/j.tibtech.2013.04.004

Gerster, H. (1997). The potential role of lycopene for human health. *J. Am. Coll. Nutr.* 16, 109–126. doi: 10.1080/07315724.1997.10718661

Gong, X.-Q., and Liu, J.-H. (2013). Genetic transformation and genes for resistance to abiotic and biotic stresses in citrus and its related genera. *Plant Cell Tissue Organ Cult.* 113, 137–147. doi: 10.1007/s11240-012-0267-x

Hazra, P., Longjam, M., and Chattopadhyay, A. (2018). Stacking of mutant genes in the development of "purple tomato" rich in both lycopene and anthocyanin contents. *Scientia Hortic.* 239, 253–258. doi: 10.1016/j.scienta.2018.05.039

He, J., and Giusti, M. M. (2010). Anthocyanins: natural colorants with health-promoting properties. *Annu. Rev. Food Sci. Technol.* 1 (1), 163–187. doi: 10.1146/annurev.food.080708.100754

Huang, D., Yuan, Y., Tang, Z., Huang, Y., Kang, C., Deng, X., et al. (2019). Retrotransposon promoter of Ruby1 controls both light- and cold-induced accumulation of anthocyanins in blood orange. *Plant Cell Environ.* 42, 3092–3104. doi: 10.1111/pce.13609

Jardak, R., Boubakri, H., Zemni, H., Gandoura, S., Mejri, S., Mliki, A., et al. (2020). Establishment of an in vitro regeneration system and genetic transformation of the Tunisian'Maltese half-blood'(Citrus sinensis): an agro-economically important variety. *3 Biotech.* 10, 99. doi: 10.1007/s13205-020-2097-6

Jiang, C. C., Zhang, Y. F., Lin, Y. J., Chen, Y., and Lu, X. K. (2019). Illumina® sequencing reveals candidate genes of carotenoid metabolism in three pummelo cultivars (Citrus maxima) with different pulp color. *Int. J. Mol. Sci.* 20 (9), 2246. doi: 10.3390/ijms20092246

Jia, H., Orbović, V., and Wang, N. (2019). CRISPR-LbCas12a-mediated modification of citrus. *Plant Biotechnol. J.* 17, 1928–1937. doi: 10.1111/pbi.13109

Jia, H., Orbovic, V., Jones, J. B., and Wang, N. (2016). Modification of the PthA4 effector binding elements in type I CsLOB1 promoter using Cas9/sgRNA to produce transgenic Duncan grapefruit alleviating XccΔpthA4: DCsLOB1.3 infection. *Plant Biotechnol. J.* 14, 1291–1301. doi: 10.1111/pbi.12495

Jia, H., Xu, J., Orbović, V., Zhang, Y., and Wang, N. (2017). Editing citrus genome *via* SaCas9/sgRNA system. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.02135

Jin, S., Lin, Q., Luo, Y., Zhu, Z., Liu, G., Li, Y., et al. (2021). Genome-wide specificity of prime editors in plants. *Nat. Biotechnol.* 39, 1292–1299. doi: 10.1038/s41587-021-00891-x

Kato, M., Ikoma, Y., Matsumoto, H., Sugiura, M., Hyodo, H., and Yano, M. (2004). Accumulation of carotenoids and expression of carotenoid biosynthetic genes during maturation in citrus fruit. *Plant Physiol.* 134, 824–837. doi: 10.1104/pp.103.031104

Kato, M., Matsumoto, H., Ikoma, Y., Okuda, H., and Yano, M. (2006). The role of carotenoid cleavage dioxygenases in the regulation of carotenoid profiles during maturation in citrus fruit. *J. Exp. Bot.* 57, 2153–2164. doi: 10.1093/jxb/erj172

Kaur, N., Alok, A., Kumar, P., Kaur, N., Awasthi, P., Chaturvedi, S., et al. (2020). CRISPR/Cas9 directed editing of lycopene epsilon-cyclase modulates metabolic flux for β-carotene biosynthesis in banana fruit. *Metab. Eng.* 59, 76–86. doi: 10.1016/j.ymben.2020.01.008

Krinsky, N. I., and Johnson, E. J. (2005). Carotenoid actions and their relation to health and disease. *Mol. Aspects Med.* 26, 459–516. doi: 10.1016/j.mam.2005.10.001

Lana, G., Zacarias-Garcia, J., Distefano, G., Gentile, A., Rodrigo, M. J., and Zacarias, L. (2020). Transcriptional analysis of carotenoids accumulation and metabolism in a pink-fleshed lemon mutant. *Genes* 11 (11), 1294. doi: 10.3390/genes11111294

LeBlanc, C., Zhang, F., Mendez, J., Lozano, Y., Chatpar, K., Irish, V. F., et al. (2018). Increased efficiency of targeted mutagenesis by CRISPR/Cas9 in plants using heat stress. *Plant J.* 93, 377–386. doi: 10.1111/tpj.13782

Lei, Y., Lu, L., Liu, H. Y., Li, S., Xing, F., and Ling-Ling, C. (2014). CRISPR-P : A web tool for synthetic single-guide RNA design of CRISPR-system in plants. *Mol. Plant* 7, 1494–1496. doi: 10.1093/mp/ssu044

Liang, G., Zhang, H., Lou, D., and Yu, D. (2016). Selection of highly efficient sgRNAs for CRISPR/Cas9-based plant genome editing. *Sci. Rep.* 6, 21451. doi: 10.1038/srep21451

Li, R., Fu, D., Zhu, B., Luo, Y., and Zhu, H. (2018b). CRISPR/Cas9-mediated mutagenesis of lncRNA1459 alters tomato fruit ripening. *Plant J.* 94, 513–524. doi: 10.1111/tpj.13872

Li, M. Y., Jiao, Y. T., Wang, Y. T., Zhang, N., Wang, B. B., Liu, R. Q., et al. (2020). CRISPR/Cas9-mediated VvPR4b editing decreases downy mildew resistance in grapevine (*Vitis vinifera l.*). *Horticulture Res.* 7, 149. doi: 10.1038/s41438-020-00371-4

Li, R., Li, R., Li, X., Fu, D., Zhu, B., Tian, H., et al. (2018a). Multiplexed CRISPR/Cas9-mediated metabolic engineering of γ-aminobutyric acid levels in solanum lycopersicum. *Plant Biotechnol. J.* 16 (2), 415–427.

Li, Y., Park, A. I., Mou, H., Colpan, C., Bizhanova, A., Akama-Garren, E., et al. (2015). A versatile reporter system for CRISPR-mediated chromosomal rearrangements. *Genome Biol.* 16, 111. doi: 10.1186/s13059-015-0680-7

Li, J., Scarano, A., Gonzalez, N. M., D'Orso, F., Yue, Y., Nemeth, K., et al. (2022). Biofortified tomatoes provide a new route to vitamin d sufficiency. *Nat. Plants* 8, 611–616.

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., and Chen, L. L. (2017). CRISPR-p 2.0 : An improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant* 10, 530–532. doi: 10.1016/j.molp.2017.01.003

Liu, W., Ye, Q., Jin, X., Han, F., Huang, X., Cai, S., et al. (2016). A spontaneous bud mutant that causes lycopene and β-carotene accumulation in the juice sacs of the parental guanxi pummelo fruits (Citrus grandis (L.) osbeck). *Sci. Hortic.* 198, 379–384. doi: 10.1016/j.scienta.2015.09.050

Lo Piero, A. R., Puglisi, I., Rapisarda, P., and Petrone, G. (2005). Anthocyanins accumulation and related gene expression in red orange fruit induced by low temperature storage. *J. Agric. Food Chem.* 53, 9083–9088. doi: 10.1021/jf051609s

Mazza, G. (2007). Anthocyanins and heart health. *Ann. Ist. Super. Sanità* 43 (4), 369–374.

Mendes, A. F. S., Chen, C., Gmitter, F. G., Moore, G. A., and Costa, M. G. C. (2011). Expression and phylogenetic analysis of two new lycopene β-cyclases from citrus paradisi. *Physiol. Plant* 141 (1), 1–10. doi: 10.1111/j.1399-3054.2010.01415.x

Natalini, A., Acciarri, N., and Cardi, T. (2021). Breeding for nutritional and organoleptic quality in vegetable crops: The case of tomato and cauliflower. *Agriculture* 11 (7), 606. doi: 10.3390/agriculture11070606

Orbović, V., and Grosser, J. W. (2015). "Citrus transformation using juvenile tissue explants," in *Agrobacterium protocols*. Ed. K. Wang (New york, NY: Springer), 245–257. doi: 10.1007/978-1-4939-1658-0_20

Peña, L., Martín-Trillo, M., Juárez, J., Pina, J. A., Navarro, L., and Martínez-Zapater, J. M. (2001). Constitutive expression of arabidopsis LEAFY or APETALA1 genes in citrus reduces their generation time. *Nat. Biotechnol.* 19 (3), 263–267. doi: 10.1038/85719

Peng, A., Zou, X., Xu, L., He, Y., Lei, T., Yao, L., et al. (2019). Improved protocol for the transformation of adult citrus sinensis osbeck 'Tarocco' blood orange tissues. *Vitr. Cell. Dev. Biol. - Plant* 55, 659–667. doi: 10.1007/s11627-019-10011-9

Petroni, K., and Tonelli, C. (2011). Recent advances on the regulation of anthocyanin synthesis in reproductive organs. *Plant Sci.* 181, 219–229. doi: 10.1016/j.plantsci.2011.05.009

Pinello, L., Canver, M. C., Hoban, M. D., Orkin, S. H., Kohn, D. B., Bauer, D. E., et al. (2016). Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.* 34 (7), 695–697. doi: 10.1038/nbt.3583

Pojer, E., Mattivi, F., Johnson, D., Stockley, C. S., Agati, G., Brunetti, C., et al. (2020). Are flavonoids effective antioxidants in plants? twenty years of our investigation. *Antioxidants* 12, 483–508. doi: 10.1111/1541-4337.12024

Poles, L., Licciardello, C., Distefano, G., Nicolosi, E., Gentile, A., and La Malfa, S. (2020). Recent advances of *in vitro* culture for the application of new breeding techniques in citrus. *Plants* 9 (8), 938. doi: 10.3390/plants9080938

Pompili, V., Dalla Costa, L., Piazza, S., Pindo, M., and Malnoy, M. (2020). Reduced fire blight susceptibility in apple cultivars using a high-efficiency CRISPR/Cas9-FLP/FRT-based gene editing system. *Plant Biotechnol. J.* 18, 845–858. doi: 10.1111/pbi.13253

Promkaew, P., Pongprasert, N., Wongs-Aree, C., Kaewsuksaeng, S., Opio, P., Kondo, S., et al. (2020). Carotenoids accumulation and carotenoids biosynthesis gene expression during fruit development in pulp of tubtim-Siam pummelo fruit. *Sci. Hortic.* 260, 108870. doi: 10.1016/j.scienta.2019.108870

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three NGS sequencing platforms. *BMC Genomics* 13, 341. doi: 10.1186/1471-2164-13-341

Rao, A. V., and Agarwal, S. (1999). Role of lycopene as antioxidant carotenoid in the prevention of chronic diseases: a review. *Nutr. Res.* 19 (2), 305–323. doi: 10.1016/S0271-5317(98)00193-6

Rodrigo, M. J., Marcos, J. F., and Zacarías, L. (2004). Biochemical and molecular analysis of carotenoid biosynthesis in flavedo of orange (Citrus sinensis l.) during fruit development and maturation. *J. Agric. Food Chem.* 52, 6724–6731. doi: 10.1021/jf049607f

Rodríguez, A., Cervera, M., Peris, J. E., and Peña, L. (2008). The same treatment for transgenic shoot regeneration elicits the opposite effect in mature explants from two closely related sweet orange (Citrus sinensis (L.) osb.) genotypes. *Plant Cell. Tissue Organ Cult.* 93, 97–106. doi: 10.1007/s11240-008-9347-3

Salonia, F., Ciacciulli, A., Poles, L., Pappalardo, H. D., La Malfa, S., and Licciardello, C. (2020). New plant breeding techniques in citrus for the improvement of important agronomic traits. a review. *Front. Plant Sci.* 11, 1234. doi: 10.3389/fpls.2020.01234

Sandmann, G. (2002). Molecular evolution of carotenoid biosynthesis from bacteria to plants. *Physiol. Plant* 116, 431–440. doi: 10.1034/j.1399-3054.2002.1160401.x

Smeriglio, A., Barreca, D., Bellocco, E., and Trombetta, D. (2016). Chemistry, pharmacology and health benefits of anthocyanins. *Phyther. Res.* 1286, 1265–1286. doi: 10.1002/ptr.5642

Sugawara, K., Wakizuka, T., Oowada, A., Moriguchi, T., and Omura, M. (2002). Histogenic identification by RAPD analysis of leaves and fruit of newly synthesized chimeric citrus. *J. Amer. Soc Hortic. Sci.* 127 (1), 104–107. doi: 10.21273/JASHS.127.1.104

Talon, M., and Gmitter, F. G. (2008). Citrus genomics. *Int. J. Plant Genomics* 2008, 528361. doi: 10.1155/2008/528361

Tanaka, Y., Sasaki, N., and Ohmiya, A. (2008). Biosynthesis of plant pigments: Anthocyanins, betalains and carotenoids. *Plant J.* 54, 733–749. doi: 10.1111/j.1365-313X.2008.03447.x

Tatmala, N., Ma, G., Zhang, L., Kato, M., and Kaewsuksaeng, S. (2020). Characterization of carotenoid accumulation and carotenogenic gene expression during fruit ripening in red colored pulp of 'siam red ruby' pumelo (Citrus grandis) cultivated in thailand. *Hortic. J.* 89, 237–243. doi: 10.2503/hortj.UTD-147

Varkonyi-Gasic, E., Wang, T., Voogd, C., Jeon, S., Drummond, R. S. M., Gleave, A. P., et al. (2019). Mutagenesis of kiwifruit CENTRORADIALIS-like genes transforms a climbing woody perennial with long juvenility and axillary flowering into a compact plant with rapid terminal flowering. *Plant Biotechnol. J.* 17, 869–880. doi: 10.1111/pbi.13021

Vazquez-Vilar, M., Quijano-Rubio, A., Fernandez-Del-Carmen, A., Sarrion-Perdigones, A., Ochoa-Fernandez, R., Ziarsolo, P., et al. (2017). GB3.0: a platform for plant bio-design that connects functional DNA elements with associated biological data. *Nucleic Acids Res.* 45, 2196–2209. doi: 10.1093/nar/gkw1326

Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., et al. (2019). The architecture of intra-organism mutation rate variation in plants. *PloS Biol.* 17, e3000191. doi: 10.1371/journal.pbio.3000191

Wang, L. S., and Stoner, G. D. (2008). Anthocyanins and their role in cancer prevention. *Cancer Lett.* 269 (2), 281–290. doi: 10.1016/j.canlet.2008.05.020

Wang, Z., Wang, S., Li, D., Zhang, Q., Li, L., Zhong, C., et al. (2018). Optimized paired-sgRNA/Cas9 cloning and expression cassette triggers high-efficiency multiplex genome editing in kiwifruit. *Plant Biotechnol. J.* 16, 1424–1433. doi: 10.1111/pbi.12884

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. a colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* 126, 485–493. doi: 10.1104/pp.126.2.485

Xing, S., Chen, K., Zhu, H., Zhang, R., Zhang, H., Li, B., et al. (2020). Fine-tuning sugar content in strawberry. *Genome Biol.* 21 (1), 1–14.

Zhang, H., Chen, J., Peng, Z., Shi, M., Liu, X., Wen, H., et al. (2021a). Integrated transcriptomic and metabolomic analysis reveals a transcriptional regulation network for the biosynthesis of carotenoids and flavonoids in 'Cara cara' navel orange. *BMC Plant Biol.* 21, 1–14. doi: 10.1186/s12870-020-02808-3

Zhang, Z., Mao, Y., Ha, S., Liu, W., Botella, J. R., and Zhu, J.-K. (2016). A multiplex CRISPR/Cas9 platform for fast and efficient editing of multiple genes in arabidopsis. *Plant Cell Rep.* 35, 1519–1533. doi: 10.1007/s00299-015-1900-z

Zhang, F., Rossignol, P., Huang, T., Wang, Y., May, A., Dupont, C., et al. (2020). Reprogramming of stem cell activity to convert thorns into branches. *Curr. Biol.* 30 (15), 2951–2961. doi: 10.1016/j.cub.2020.05.068

Zhang, F., Wang, Y., and Irish, V. F. (2021b). CENTRORADIALIS maintains shoot meristem indeterminacy by antagonizing THORN IDENTITY1 in citrus. *Curr. Biol.* 31 (10), 2237–2242. doi: 10.1016/j.cub.2021.02.051

Zhao, Y., Zhang, C., Liu, W., Gao, W., Liu, C., Song, G., et al. (2016). An alternative strategy for targeted gene replacement in plants using a dual-sgRNA/Cas9 design. *Sci. Rep.* 6, 23890. doi: 10.1038/srep23890

Zou, X., Peng, A., Xu, L., Liu, X., Lei, T., Yao, L., et al. (2013). Efficient auto-excision of a selectable marker gene from transgenic citrus by combining the Cre/loxP system and ipt selection. *Plant Cell Rep.* 32, 1601–1613. doi: 10.1007/s00299-013-1470-x

Zsögön, A., Čermák, T., Naves, E. R., Notini, M. M., Edel, K. H., Weinl, S., et al. (2018). *De novo* domestication of wild tomato using genome editing. *Nat. Biotechnol.* 36, 1211–1216. doi: 10.1038/nbt.4272

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 31 January 2023
DOI 10.3389/fpls.2022.1011199

OPEN ACCESS

EDITED BY
Giorgio Gambino,
National Research Council (CNR), Italy

REVIEWED BY
Aureliano Bombarely,
Polytechnic University of Valencia, Spain
Jia-Ming Song,
Guangxi University, China

*CORRESPONDENCE
Claude W. dePamphilis
cwd3@psu.edu

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 03 August 2022
ACCEPTED 02 December 2022
PUBLISHED 31 January 2023

CITATION
Wafula EK, Zhang H, Von Kuster G, Leebens-Mack JH, Honaas LA and dePamphilis CW (2023) PlantTribes2: Tools for comparative gene family analysis in plant genomics. *Front. Plant Sci.* 13:1011199. doi: 10.3389/fpls.2022.1011199

PlantTribes2: Tools for comparative gene family analysis in plant genomics

Eric K. Wafula[1†], Huiting Zhang[2,3†], Gregory Von Kuster[4], James H. Leebens-Mack[5], Loren A. Honaas[2] and Claude W. dePamphilis[1,4*]

[1]Department of Biology, The Pennsylvania State University, University Park, PA, United States, [2]Tree Fruit Research Laboratory, United States Department of Agriculture (USDA), Agricultural Research Service (ARS), Wenatchee, WA, United States, [3]Department of Horticulture, Washington State University, Pullman, WA, United States, [4]Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, United States, [5]Department of Plant Biology, University of Georgia, Athens, GA, United States

Plant genome-scale resources are being generated at an increasing rate as sequencing technologies continue to improve and raw data costs continue to fall; however, the cost of downstream analyses remains large. This has resulted in a considerable range of genome assembly and annotation qualities across plant genomes due to their varying sizes, complexity, and the technology used for the assembly and annotation. To effectively work across genomes, researchers increasingly rely on comparative genomic approaches that integrate across plant community resources and data types. Such efforts have aided the genome annotation process and yielded novel insights into the evolutionary history of genomes and gene families, including complex non-model organisms. The essential tools to achieve these insights rely on gene family analysis at a genome-scale, but they are not well integrated for rapid analysis of new data, and the learning curve can be steep. Here we present PlantTribes2, a scalable, easily accessible, highly customizable, and broadly applicable gene family analysis framework with multiple entry points including user provided data. It uses objective classifications of annotated protein sequences from existing, high-quality plant genomes for comparative and evolutionary studies. PlantTribes2 can improve transcript models and then sort them, either genome-scale annotations or individual gene coding sequences, into pre-computed orthologous gene family clusters with rich functional annotation information. Then, for gene families of interest, PlantTribes2 performs downstream analyses and customizable visualizations including, (1) multiple sequence alignment, (2) gene family phylogeny, (3) estimation of synonymous and non-synonymous substitution rates among homologous sequences, and (4) inference of large-scale duplication events. We give examples of PlantTribes2 applications in functional genomic studies of economically important plant families, namely transcriptomics in the weedy Orobanchaceae and a core orthogroup analysis (CROG) in Rosaceae.

PlantTribes2 is freely available for use within the main public Galaxy instance and can be downloaded from GitHub or Bioconda. Importantly, PlantTribes2 can be readily adapted for use with genomic and transcriptomic data from any kind of organism.

KEYWORDS

gene family phylogenetics, multiple sequence alignment, genome duplication, galaxy, modular tools, applied agriculture, comparative genomics, CROG analysis

1 Introduction

A rapid and continuing decline in sequencing costs over the last 30 years has contributed to the generation of massive amounts of transcriptome and genome data for non-model plant species (Barrett et al., 2013; Matasci et al., 2014; Sayers et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019; Marks et al., 2021). Integrating new genomic data from diverse plant lineages in phylogenetic studies can provide the evolutionary context necessary for understanding the evolution of gene function (Williams et al., 2014; Pabón-Mora et al., 2014; Yang et al., 2015b; Zhang et al., 2015; Carvalho et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019; Mi et al., 2020; Nagy et al., 2020), resolving species relationships (Timme et al., 2012; Rothfels et al., 2013; Wickett et al., 2014; Zeng et al., 2014; Yang et al., 2015a; Huang et al., 2016; Xiang et al., 2017; One Thousand Plant Transcriptomes Initiative, 2019; Hodel et al., 2022), accurate identification of orthologous and paralogous genes among species (Sonnhammer and Koonin, 2002; Gabaldón, 2008; Schreiber et al., 2014; Emms and Kelly, 2019; Derelle et al., 2020; Fuentes et al., 2021), and unraveling gene and genome duplications (Bowers et al., 2003; Jiao et al., 2011; Jiao et al., 2012; The Amborella Genome Project, 2013; Li et al., 2015; Ren et al., 2018; Zwaenepoelde Peer, 2019; Viruel et al., 2019). However, comparative genomic and phylogenomic analyses typically requires a level of bioinformatic expertise and a scale of computational resources that are inaccessible to many researchers. For instance, a large-scale phylogenomic study may require objective circumscription of representative protein sequences into gene families using a carefully selected set of most appropriate reference genomes. This requires knowledge and skill to assess the quality of available genomic resources as well as an evolutionary perspective to avoid pitfalls that lead to distorted conclusions, such as using a biased selection of reference species or outgroups. In addition, to execute these analytical pipelines, command line skills and the expertise to navigate through and properly set parameters, select appropriate algorithms, and solve potential computation environment conflicts are needed. Although some software (Chen et al., 2020; Tello-Ruiz et al., 2020; Valentin et al., 2020; Bel et al., 2021; Oliveira et al., 2021; Emms and Kelly, 2022) are more user-friendly (*i.e.*, incorporate a graphical user interface, containerized tools, *etc.*) and have pre-defined parameters suitable for plant research, most others still require custom optimization or are mainly applied to species with small genomes (*i.e.*, prokaryotes), or non-plant systems (Dunn et al., 2013; Blom et al., 2016; Lanza et al., 2016; Altenhoff et al., 2019; Pucker et al., 2020; Ebmeyer et al., 2021; Perrin and Rocha, 2021; Pucker, 2022).

With the goal to improve data accessibility, databases have been created to host curated plant-specific genomic information at different scales, ranging from those including sequenced genomes from diverse plant species (i.e., PLAZA 5.0, Bel et al., 2021 and Gramene, Tello-Ruiz et al., 2020) to ones focusing on specific plant groups, such as the Genome Database for Rosaceae (GDR, Jung et al., 2019). Major plant databases are reviewed and described by various authors (Chen et al., 2006; Lyons and Freeling, 2008; Wall et al., 2008; Goodstein et al., 2012; Schreiber et al., 2014; Martinez, 2016; Huerta-Cepas et al., 2016; Kriventseva et al., 2018; Mi et al., 2020; Tello-Ruiz et al., 2020; Bel et al., 2021). Some databases also provide gene homology information and computational tools for comparative genomic analysis (Martinez, 2016). However, analysis tools implemented in such databases are typically limited, static, and can only be used to analyze existing data (Tomcal et al., 2013; Sundell et al., 2015; Spannagl et al., 2016; Nakaya et al., 2017; Tello-Ruiz et al., 2020). Some more recent databases contain flexible tools (*i.e.*, users can select different algorithms), but these are often not scalable (*i.e.*, many have limitations on data size and number of input sequences). For example, the PLAZA 5.0 database contains 134 carefully selected high-quality plant genomes and provides gene family circumscriptions with rich gene homology and annotation information (Bel et al., 2021). However, users can only upload up to 300 new sequences for the BLAST based gene family search function, and add a maximum of 50 external sequences while running a gene family phylogeny on their webserver (https://bioinformatics.psb.ugent.be/plaza/). Limitation on data input make it infeasible to use these databases to perform genome-scale analyses on new datasets brought by the user.

Other new developments aiming to make complicated bioinformatic analyses accessible to more users are workflow management systems which integrate analytic pipelines and complementary software into readily executable packages, such as SnakeMake (Mölder et al., 2021), Nextflow (Tommaso et al., 2017), Pegasus (Deelman et al., 2015), Galaxy Workbench (The Galaxy Community, 2022), and others. Of those, the Galaxy Workbench is an open-source web-based software framework that aims to make command-line tools accessible to users without informatic expertise (The Galaxy Community, 2022), and is popular among biologists. Galaxy implements several comparative genomic tools developed by the bioinformatics community (Darling et al., 2010; Thanki et al., 2018). Such a web-based framework provides a simplified way to execute standardized analyses and workflows. They can also eliminate the complex administrative and programming tasks inherent in performing big data analyses *via* batch processing on the command line, and greatly simplify record keeping and re-implementation of complex analytical processes. Often, scientists can perform analyses with either existing or user implemented tools from a web browser. Additionally, individual institutions can link these web-based platforms to their own high-performance computing resources, allowing computationally intensive analysis not always possible on a purely web-based platform.

In an effort to address these accessibility and computational challenges in genome-scale research and to take advantage of the Galaxy environment, we developed PlantTribes2, a gene family analysis framework that uses objective classifications of annotated protein sequences from genomes or transcriptomes for comparative and evolutionary analyses of gene families from any type of organism, including fungi, microbes, animals, and plants. An initial version of PlantTribes was developed by Wall et al. (2008), but has become outdated due to several of the previously mentioned limitations. In PlantTribes2, we have completely revamped PlantTribes from a static relational database to a flexible analytical pipeline with all new code, new features, and extensive testing. We have developed a well-documented analytic framework complete with training materials including tutorials and sample datasets. Finally, we worked with the Galaxy community to develop Galaxy wrappers for all of the PlantTribes2 tools (Blankenberg et al., 2014. Supplemental Table 1), so they are available on the public server at usegalaxy.org, and can be installed into any Galaxy instance. Finally, we demonstrate genome-scale evolutionary analysis of gene families using PlantTribes2, starting with *de novo* assembled transcriptomes and gene models from whole genome data. Although our examples, sample datasets, and gene family scaffolds are for plants, the pipeline is system agnostic and can be readily used with genome-scale information from any set of related organisms.

2 Pipeline implementation

The PlantTribes2 toolkit is a collection of self-contained modular analysis pipelines that use objective classifications of annotated protein sequences from sequenced genomes for comparative and evolutionary analyses of genome-scale gene families. At the core of PlantTribes2 analyses are the gene family scaffolds, which are clusters of orthologous and paralogous sequences from specified sets of inferred protein sequences. The tools interact with these scaffolds, as described below, to deliver the following outputs: (1) predicted coding sequences and their corresponding translations, (2) a table of pairwise synonymous/non-synonymous substitution rates for either orthologous or paralogous transcript pairs, (3) results of significant duplication components in the distribution of synonymous substitutions rates (Ks), (4) a summary table for transcripts classified into orthologous gene family clusters with their corresponding functional annotations, (5) gene family amino acid and nucleotide fasta sequences, (6) multiple sequence alignments, and (7) inferred maximum likelihood phylogenies (Figure 1)

2.1 Gene family scaffolds

The current release of PlantTribes2 (v1.0.4) provides several plant gene family scaffolds (Supplemental Table 2) used in previously published and ongoing phylogenomic studies (The Amborella Genome Project, 2013; Wickett et al., 2014; Yang et al., 2015b; Li et al., 2018; Shahid et al., 2018; Yang et al., 2019; Timilsena et al., 2022; Timilsena et al., in press; Zhang et al., 2022), the companion paper in this issue), including one Monocot focused scaffold (12Gv1.0) and four iterations of generic Angiosperm focused scaffolds (22Gv1.1, 26Gv1.0, 26Gv2.0, and 37Gv1.0). Complete sets of inferred protein-coding genes from plant genomes represented in each of the PlantTribes2 scaffolds were clustered into gene families (*i.e.*, orthogroups) using at least one of the following protein clustering methods: GFam (clusters of consensus domain architecture) (Sasidharan et al., 2012), OrthoMCL (narrowly defined clusters) (Li et al., 2003; Chen et al., 2006), or OrthoFinder (more broadly defined clusters) (Emms and Kelly, 2015; Emms and Kelly, 2019). Additional clustering of primary gene families was performed using the MCL algorithm (Enright et al., 2002) at 10 stringencies with inflation values from 1.2 to 5.0 to connect distantly, but potentially related orthogroups into larger hierarchical gene families (*i.e.*, super-orthogroups), as described in Wall et al. (2008). We then annotated each orthogroup with gene function information from biological databases, including Gene Ontology (GO) (Ashburner et al., 2000; Carbon et al., 2019), InterPro/Pfam protein domains (Jones et al., 2014; Blum et al., 2020; Mistry

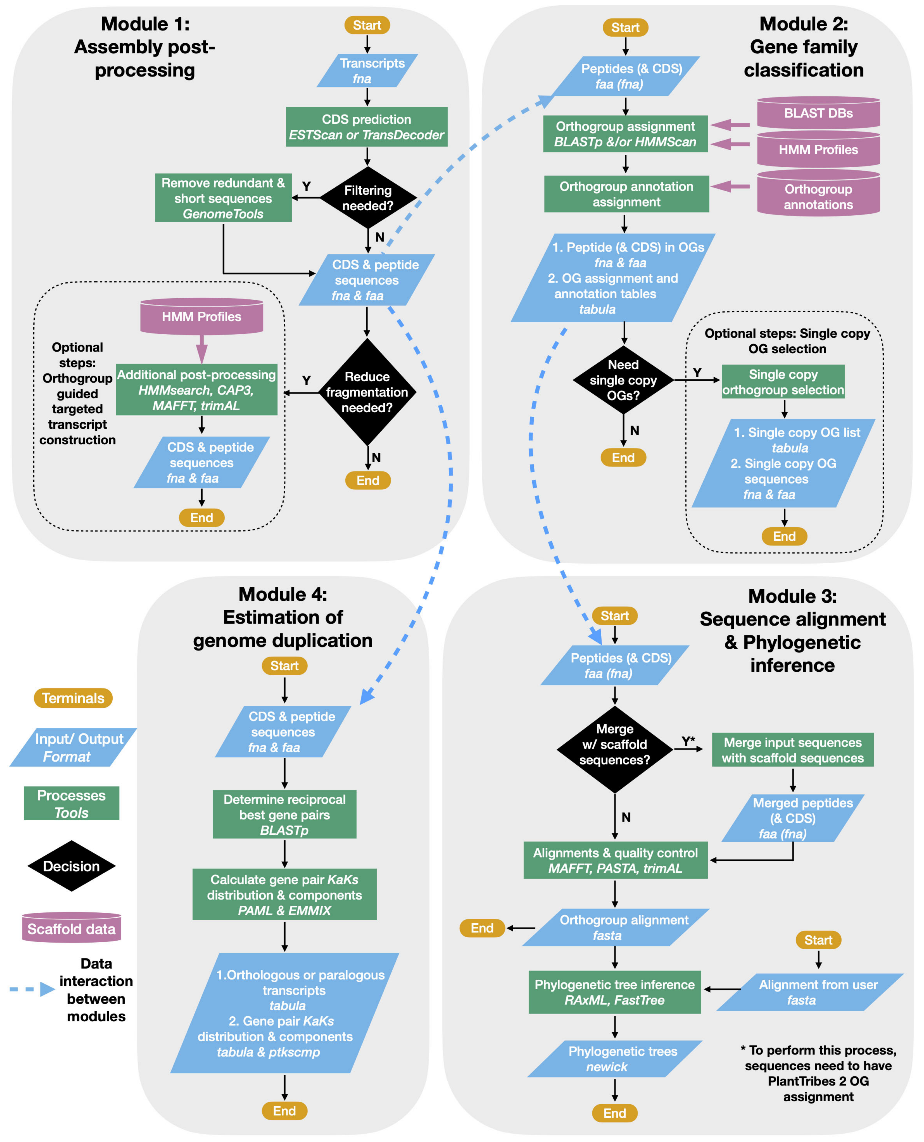

Module 1: Assembly post-processing
Start
Transcripts fna
CDS prediction ESTScan or TransDecoder
Filtering needed?
Y
Remove redundant & short sequences GenomeTools
N
CDS & peptide sequences fna & faa
Reduce fragmentation needed?
Y
N
End
Optional steps: Orthogroup guided targeted transcript construction
HMM Profiles
Additional post-processing HMMsearch, CAP3, MAFFT, trimAL
CDS & peptide sequences fna & faa
End
Module 2: Gene family classification
Start
Peptides (& CDS) faa (fna)
BLAST DBs
HMM Profiles
Orthogroup annotations
Orthogroup assignment BLASTp &/or HMMScan
Orthogroup annotation assignment
1. Peptide (& CDS) in OGs fna & faa
2. OG assignment and annotation tables tabula
Need single copy OGs?
Y
N
End
Optional steps: Single copy OG selection
Single copy orthogroup selection
1. Single copy OG list tabula
2. Single copy OG sequences fna & faa
End
Module 4: Estimation of genome duplication
Start
CDS & peptide sequences fna & faa
Determine reciprocal best gene pairs BLASTp
Calculate gene pair KaKs distribution & components PAML & EMMIX
1.Orthologous or paralogous transcripts tabula
2. Gene pair KaKs distribution & components tabula & ptkscmp
End
Module 3: Sequence alignment & Phylogenetic inference
Start
Peptides (& CDS) faa (fna)
Merge w/ scaffold sequences?
Y*
Merge input sequences with scaffold sequences
Merged peptides (& CDS) faa (fna)
N
Alignments & quality control MAFFT, PASTA, trimAL
Orthogroup alignment fasta
End
Start
Alignment from user fasta
Phylogenetic tree inference RAxML, FastTree
Phylogenetic trees newick
End
* To perform this process, sequences need to have PlantTribes 2 OG assignment
Terminals
Input/ Output Format
Processes Tools
Decision
Scaffold data
Data interaction between modules

FIGURE 1
PlantTribes2 analysis workflow. A schematic diagram illustrating the PlantTribes2 modular analysis workflow. (1) A user provides transcripts for post-processing, resulting in a non-redundant set of predicted coding sequences and their corresponding translations (Module 1). (2) The post-processed transcripts (or user provided sequences) are searched against a gene family scaffold blast and/or hmm database(s), and transcripts are assigned into their putative orthogroups with corresponding metadata (Module 2). (3) Classified transcripts are integrated with their corresponding scaffold gene models to estimate orthogroup multiple sequence alignments and corresponding phylogenetic trees (Module 3). Similarly, sequence alignments and phylogeny can be constructed from user provided data. (4) Synonymous substitution rate (Ks) and nonsynonymous substitution rate (Ka) of paralogs from either the post-processed assembly or inferred from the phylogenetic trees are estimated. The Ks results are used to detect large-scale duplication events and many other evolutionary hypotheses (Module 4).

et al., 2020), The Arabidopsis Information Resource (TAIR) (Berardini et al., 2015), UniProtKB/TrEMBL (The UniProt Consortium, 2021), and UniProtKB/Swiss-Prot (The UniProt Consortium, 2021). The final PlantTribes2 scaffold data sets include (1) orthogroups protein coding sequence fasta, (2) orthogroups protein multiple sequence alignments, (3) orthogroups protein HMM profiles, (4) a scaffold protein BLAST database, (5) a scaffold protein HMM profiles database, and (6) templates for analysis pipelines with scaffold metadata.

For custom applications with any focal group of organisms, a detailed description is available on the GitHub repository (https://github.com/dePamphilis/PlantTribes) for how to build a customized PlantTribes2 gene family scaffold. Building custom gene family scaffolds in PlantTribes2 begins with providing unclassified genome-scale gene sets or converting an existing gene family circumscription and corresponding metadata to a format that is compatible with the PlantTribes2 tools. If running on the command line, such externally circumscribed scaffolds can be directly integrated into PlantTribes2 for user-specific gene family analyses. If running on Galaxy, Galaxy administration tools (Blankenberg et al., 2014, Supplemental Table 1) are available for installing and maintaining these external scaffolds within a Galaxy instance that provides the PlantTribes2 tools.

2.2 Illustrated examples of PlantTribes2 tools

Here we describe the use of each PlantTribes2 tool and provide examples of outputs using a test dataset containing transcripts from two plant species (details can be found in Supplemental Table 3). Detailed step-by-step tutorials using the test data to perform analyses are available for both the Galaxy and the command-line versions of the pipeline.

2.2.1 Assembly post-processing

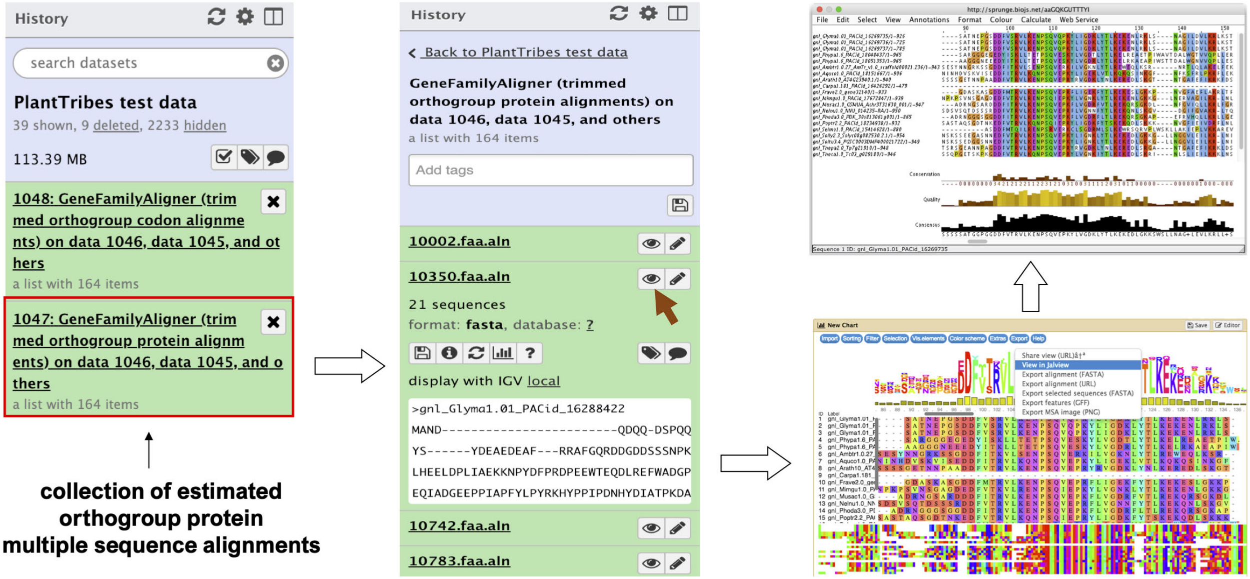
The *AssemblyPostProcessor* tool is an entry point of a PlantTribes2 analysis when the input data is *de novo* transcripts or gene models in some poorly annotated genomes where predicted coding sequences and corresponding peptides do not match. The *AssemblyPostProcessor* pipeline uses either ESTScan (Iseli et al., 1999) or TransDecoder (Haas et al., 2013) to transform transcripts into putative CDSs and their corresponding amino acid translations. Optionally, the resulting predicted coding regions can be filtered to remove duplicated and exact subsequences using GenomeTools (Gremme et al., 2013). The pipeline is implemented with an additional assembly post-processing method that uses scaffold orthogroups to reduce fragmentation in a *de novo* assembly. Homology searches of post-processed transcripts against HMM-profiles (Eddy, 2011) of targeted orthogroups are conducted using HMMER hmmsearch (Eddy, 2011). After assignment of transcripts to targeted orthogroups, orthogroup-specific gene assembly of overlapping primary contigs is performed using CAP3 (Huang and Madan, 1999), an overlap-layout-consensus assembler. Finally, protein multiple sequence alignments of orthogroups are estimated and trimmed using MAFFT (Katoh and Standley, 2013) and trimAL (Capella-Gutiérrez et al., 2009) respectively, to aid in identifying targeted assembled transcripts that are orthologous to the scaffold reference gene models based on the global sequence alignment coverage. A list of *AssemblyPostProcessor* use cases include: (1) processing *de novo* transcriptome assemblies to improve transcript qualities for downstream analyses (Honaas et al., 2016; Yang et al., 2019; Whittle et al., 2021; and example in section 3.2.1); (2) generating matching coding sequences (CDSs) and peptide sequences in genomes with only mRNA sequences (e.g., the *Malus domestica* GDDH13 annotation provided only mRNA sequences but not CDSs, Daccord et al., 2017) and gene information gathered from databases lacking a uniformed naming system and processing protocols - for instance, the numbers of CDSs and peptides do not match in the *Pyrus pyrifolia* 'Cuiguan' genome, and the peptides are named differently from the CDSs (Gao et al., 2021). The *AssemblyPostProcessor*-generated matching CDS and peptide sequences from the aforementioned *Malus* and *Pyrus* genomes among others provided a good starting point for the comparative genomic analyses described in section 3.2.2 and 3.2.3.

2.2.2 Gene family classification

The *GeneFamilyClassifier* tool classifies gene coding sequences either produced by the *AssemblyPostProcessor* tool or from an external source using BLASTp (Camacho et al., 2009) and HMMER (Eddy, 2011) hmmscan (or both classifiers) into pre-computed orthologous gene family clusters (orthogroups) of a PlantTribes2 scaffold. Classified sequences are then assigned with the corresponding orthogroups' metadata, which includes gene counts of scaffold taxa, superclusters (super orthogroups) at multiple clustering stringencies, and rich orthogroup annotations from functional genomic databases (as described in section 2.1). Additionally, sequences belonging to single or low-copy gene families that are commonly used in species tree inference can be determined with a built-in command for this tool. Next, the classified input gene coding sequences can be integrated into their corresponding orthogroup's scaffold gene model files using the *GeneFamilyIntegrator* tool for downstream analyses.

2.2.3 Gene family alignment estimation

The *GeneFamilyAligner* tool estimates protein and codon multiple sequence alignments of integrated orthologous gene family fasta files produced by the *GeneFamilyIntegrator* tool or from an external source. Orthogroup alignments are estimated using either MAFFT's L-INS-i algorithm (Katoh and Standley, 2013) or the divide and conquer approach implemented in the PASTA (Mirarab et al., 2015) pipeline for large alignments. Optional post-alignment processing includes trimming out sites that are predominantly gaps (Capella-Gutiérrez et al., 2009), removing sequences with very low global orthogroup alignment coverage, and performing realignment of orthogroup sequences following site trimming and sequence removal. In the Galaxy framework, the MSAViewer (Yachdav et al., 2016) plugin allows orthogroup fasta multiple sequence alignments produced by the *GeneFamilyAligner* to be visualized and edited using the Jalview Java Web Start (Waterhouse et al., 2009) (Figure 2).



History

search datasets

PlantTribes test data

39 shown, 9 deleted, 2233 hidden

113.39 MB

1048: GeneFamilyAligner (trimmed orthogroup codon alignments) on data 1046, data 1045, and others

a list with 164 items

1047: GeneFamilyAligner (trimmed orthogroup protein alignments) on data 1046, data 1045, and others

a list with 164 items

collection of estimated orthogroup protein multiple sequence alignments

History

Back to PlantTribes test data

GeneFamilyAligner (trimmed orthogroup protein alignments) on data 1046, data 1045, and others

a list with 164 items

Add tags

10002.faa.aln

10350.faa.aln

21 sequences

format: **fasta**, database: ?

display with IGV local

```
>gnl_Glyma1.01_PACid_16288422
MAND-----------------------QDQQ-DSPQQ
YS------YDEAEDEAF---RRAFGQRDDGDDSSSNPK
LHEELDPLIAEKKNPYDFPRDPEEWTEQDLREFWADGP
EQIADGEEPPIAPFYLPYRKHYPPIPDNHYDIATPKDA
```

10742.faa.aln

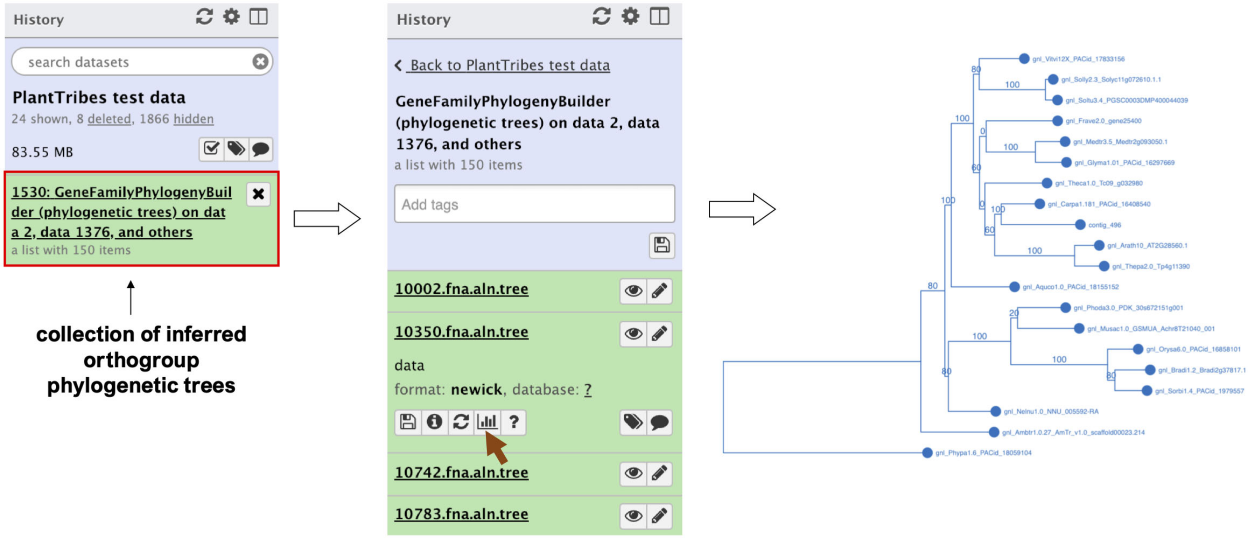
10783.faa.aln

Sequence 1 ID: gnl_Glyma1.01_PACid_16269735

FIGURE 2
An illustration of an orthogroup multiple sequence alignment produced by the Galaxy PlantTribes2 GeneFamilyAligner tool using the test dataset. Results can be visualized in Galaxy with the MSAViewer visualization plugin and manually edited with Jalview Java Web Start.

2.2.4 Gene family phylogenetic inference

The *GeneFamilyPhylogenyBuilder* tool performs a gene family phylogenetic inference of multiple sequence alignments produced by the *GeneFamilyAligner* tool or from an external source. PlantTribes2 estimates maximum likelihood (ML) phylogenetic trees using either RAxML (Stamatakis, 2014) or FastTree (Price et al., 2010) algorithms. Optional tree optimization includes setting the number of bootstrap replicates for RAxML to conduct a rapid bootstrap analysis, searching for the best-scoring ML tree, and rooting the inferred phylogenetic tree with the most distant taxon in the



History

search datasets

PlantTribes test data

24 shown, 8 deleted, 1866 hidden

83.55 MB

1530: GeneFamilyPhylogenyBuilder (phylogenetic trees) on data 2, data 1376, and others

a list with 150 items

collection of inferred orthogroup phylogenetic trees

History

Back to PlantTribes test data

GeneFamilyPhylogenyBuilder (phylogenetic trees) on data 2, data 1376, and others

a list with 150 items

Add tags

10002.fna.aln.tree

10350.fna.aln.tree

data

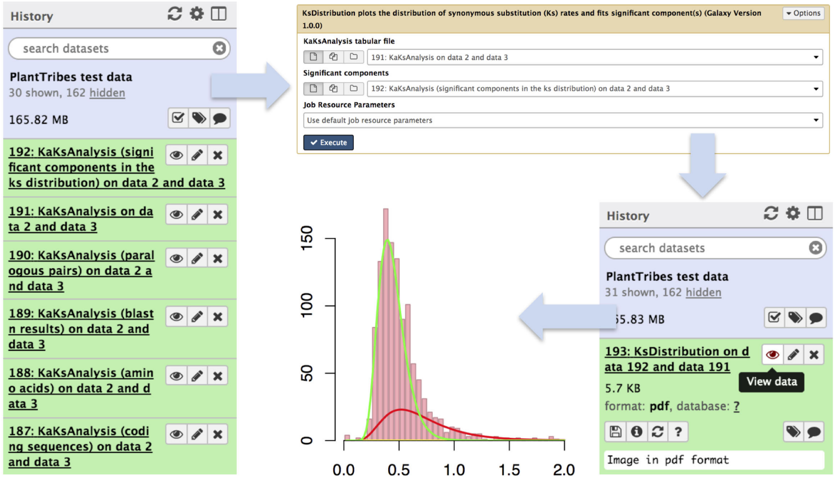
format: **newick**, database: ?

10742.fna.aln.tree

10783.fna.aln.tree

gnl_Vitvi12X_PACid_17833156
gnl_Solly2.3_Solyc11g072610.1.1
gnl_Soltu3.4_PGSC0003DMP400044039
gnl_Frave2.0_gene25400
gnl_Medtr3.5_Medtr2g093050.1
gnl_Glyma1.01_PACid_16297669
gnl_Theca1.0_Tc09_g032980
gnl_Carpa1.181_PACid_16408540
contig_496
gnl_Arath10_AT2G28560.1
gnl_Thepa2.0_Tp4g11390
gnl_Aquco1.0_PACid_18155152
gnl_Phoda3.0_PDK_30s672151g001
gnl_Musac1.0_GSMUA_Achr8T21040_001
gnl_Orysa6.0_PACid_16858101
gnl_Bradi1.2_Bradi2g37817.1
gnl_Sorbi1.4_PACid_1979557
gnl_Nelnu1.0_NNU_005592-RA
gnl_Ambtr1.0.27_AmTr_v1.0_scaffold00023.214
gnl_Phypa1.6_PACid_18059104
80
100
100
0
100
60
100
0
100
60
100
80
20
100
100
80
80

FIGURE 3
An illustration of an orthogroup phylogenetic tree produced by the Galaxy PlantTribes2 GeneFamilyPhylogenyBuilder using the test dataset. Results can be visualized in Galaxy using either the Phylocanvas (demonstrated here) or the PHYLOViZ plugin.



PlantTribes test data
30 shown, 162 hidden
165.82 MB

192: KaKsAnalysis (significant components in the ks distribution) on data 2 and data 3
191: KaKsAnalysis on data 2 and data 3
190: KaKsAnalysis (paralogous pairs) on data 2 and data 3
189: KaKsAnalysis (blastn results) on data 2 and data 3
188: KaKsAnalysis (amino acids) on data 2 and data 3
187: KaKsAnalysis (coding sequences) on data 2 and data 3

KsDistribution plots the distribution of synonymous substitution (Ks) rates and fits significant component(s) (Galaxy Version 1.0.0)

KaKsAnalysis tabular file: 191: KaKsAnalysis on data 2 and data 3

Significant components: 192: KaKsAnalysis (significant components in the ks distribution) on data 2 and data 3

Job Resource Parameters: Use default job resource parameters

PlantTribes test data
31 shown, 162 hidden
55.83 MB

193: KsDistribution on data 192 and data 191
5.7 KB
format: pdf, database: ?
Image in pdf format

FIGURE 4
An illustration of genome duplication events detected using the Galaxy PlantTribes2 KaKsAnalysis tool. The KaKs analysis tool produces a list of outputs including self blastn results (item 189), a list of paralogous pairs (item 190), Ka (non-synonymous) and Ks (synonymous) substitution rates (item 191), and the significant components in the Ks distribution (item 192). Then the distribution of estimated paralogous pair Ks values is clustered into components using a mixture of multivariate normal distributions to identify significant duplication event(s) (item 193) and is visualized using Galaxy built-in tools.

orthogroup or specified taxa. In the Galaxy framework, either the Phylocanvas plugin (https://phylocanvas.org/) or the PHYLOViZ 2.0 (Nascimento et al., 2016) plugin provides several options for visualizing and rendering the phylogenetic trees produced by the *GeneFamilyPhylogenyBuilder* (Figure 3).

2.2.5 Estimation of genome duplications

The *KaKsAnalysis* tool estimates paralogous and orthologous pairwise synonymous (*Ks*) and non-synonymous (*Ka*) substitution rates using PAML (Yang, 2007) for a set of protein coding genes (*i.e.*, produced by the *AssemblyPostProcessor*), with duplicates inferred from the phylogenomic analysis (using both the *GeneFamilyClassifier* and *GeneFamilyPhylogenyBuilder*) or from an external source. Optionally, the resulting set of estimated *Ks* values can be clustered into components using a mixture of multivariate normal distributions, implemented in the EMMIX (McLachlan and Peel, 1999) software, to identify significant duplication event(s) in a species or a pair of species. The *KsDistribution* tool then plots the *Ks* rates and fits the estimated significant component(s) onto the distribution (Figure 4).

3 Results

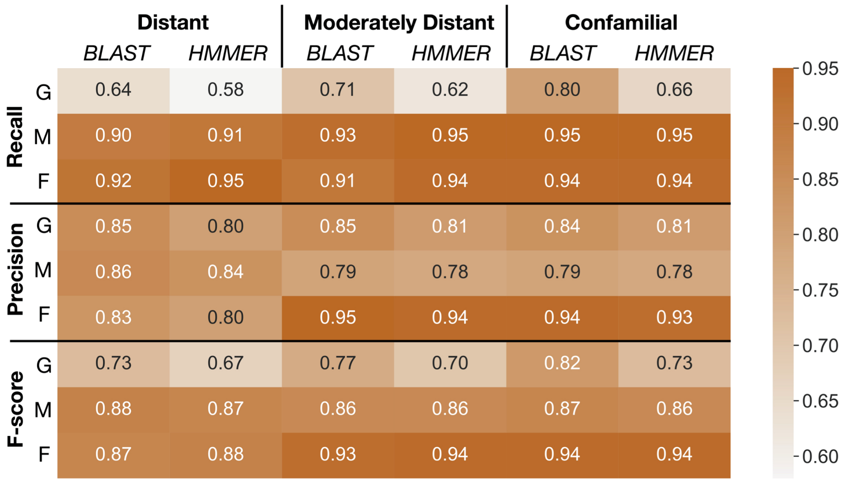
3.1 Performance evaluation of sequence classifiers

PlantTribes2 uses BLAST (blastp) and HMMER (hmmscan and hmmsearch) algorithms to classify inferred protein sequences into orthologous gene family clusters, a foundational step for many downstream analyses. To demonstrate the versatility of these two classifiers on gene family clusters, we present evaluations for classification algorithms using the pre-computed 22Gv1.1 gene family scaffold (Supplemental Table 2). This scaffold contains annotated protein coding sequences (CDSs) for 22 representative land plant genomes, including nine rosids, three asterids, two basal eudicots, five monocots, one basal angiosperm, one lycophyte, and one moss.

Three taxa with varying evolutionary distances in relationship to all the other taxa in the 22Gv1.1 gene family scaffold were selected: the only moss species, *Physcomitrella patens*, and two asterid sister species, *Solanum lycopersicum* and *Solanum tuberosum*. These three taxa were removed from the scaffold and then classified back to assess recall and precision of the BLAST and HMMER classifiers (Vihinen, 2012). Only protein sequences reassigned to their original orthologous clusters were considered true positives. In addition, F-score, a single metric that considers both recall and precision to measure the overall performance of the two classifiers, was calculated (Vihinen, 2012). The procedure is performed as described below:

(1) **Distant**: *Physcomitrella patens* was removed and sorted back into the scaffold to evaluate the performance of classifiers with distant species. No other moss or bryophyte species are present in this scaffold.

(2) **Moderately Distant**: Both *Solanum lycopersicum* and *Solanum tuberosum* were removed, and *S. lycopersicum* was sorted back into the scaffold to evaluate the



		Distant		Moderately Distant		Confamilial	
		BLAST	HMMER	BLAST	HMMER	BLAST	HMMER
Recall	G	0.64	0.58	0.71	0.62	0.80	0.66
	M	0.90	0.91	0.93	0.95	0.95	0.95
	F	0.92	0.95	0.91	0.94	0.94	0.94
Precision	G	0.85	0.80	0.85	0.81	0.84	0.81
	M	0.86	0.84	0.79	0.78	0.79	0.78
	F	0.83	0.80	0.95	0.94	0.94	0.93
F-score	G	0.73	0.67	0.77	0.70	0.82	0.73
	M	0.88	0.87	0.86	0.86	0.87	0.86
	F	0.87	0.88	0.93	0.94	0.94	0.94

FIGURE 5
Summaries of performance evaluation of classification rates for BLAST and HMMER classifiers. Recall, precision, and F-score (Vihinen, 2012) for the two classifiers are measured on GFam (G), OrthoMCL (M), and OrthoFinder (F) clustering methods to determine how well taxa at different distances are classified into the PlantTribes2 22Gv1.1 gene family scaffold. Larger values are better. Distant: remove and sort back *Physcomitrella patens*, a species distantly related to all other scaffolding species; Moderately distant: remove *Solanum lycopersicum* and *S. tuberosum*, then sort back *S. lycopersicum*. No other Solanaceae species are present in the scaffold, but moderately distant species, *i.e.*, other asterids, are used as scaffolding species; Confamilial: *S. lycopersicum* was removed and sorted back. A confamilial species, *S. tuberosum*, is present in the scaffold.

performance of classifiers with moderately distant species. After removing both *S. lycopersicum* and *S. tuberosum*, no other sister species in the same plant family are present in the scaffold. However, close lineages, including three asterids and nine rosids, are present in the scaffold.

(3) **Confamilial**: *Solanum lycopersicum* was removed and sorted back into the scaffold to evaluate the performance of classifiers with confamilial species. *Solanum tuberosum*, a sister species from the same plant family, is present in the scaffold.

As shown in Figure 5, the overall classification performance for BLAST and HMMER is similar based on the F-scores across different evolutionary distances (73%-94% for BLAST, 67%-94% for HMMER). In addition, both classifiers have a higher recall rate when classifying into OrthoMCL and OrthoFinder clusters (90% - 96%) compared to GFam clusters (58% - 80%). HMMER is slightly more sensitive than BLAST when the evolutionary distance is significant, while BLAST is much more sensitive when classifying into GFam clusters at any evolutionary distance. Precision for both classifiers is similar across the evolutionary distance of the scaffold (78% - 95%). Classifying into OrthoFinder clusters yields much higher precision (80%-95%) than classifying into OrthoMCL (78%-86%) and GFam (81%-85%) clusters. These findings suggest that, regardless of the sequence classifier algorithm used or evolutionary distance, clusters inferred by orthology methods (OrthoFinder and OrthoMCL) result in better clustering performance compared to clusters inferred by a consensus domain-based method (GFam). We recommend using the merged classification results from BLAST and HMMER, as implemented in the pipeline, because it leverages the strength of both classifiers.

3.2 Examples of application

Here we provide examples of how to use PlantTribes2 to answer specific questions regarding (1) alleviating fragmentation issues in a *de novo* transcriptome assembly, (2) evaluation and improvement of gene families and gene models, and (3) assessing the quality of genomes in closely related species.

3.2.1 Evaluation of targeted gene family assembly

De novo assembly of RNA-Seq data is commonly used to reconstruct expressed transcripts for non-model species that lack quality reference genomes. However, heterogeneous sequence coverage, sequencing errors, polymorphism, and sequence repeats, among other factors, cause algorithms to generate contigs that are fragmented (Zhang et al., 2014; Honaas et al., 2016). In order to demonstrate the utility of the targeted gene family assembly function in PlantTribes2, we obtained raw Illumina transcriptome datasets sequenced by the Parasitic Plant Genome Project (http://ppgp.huck.psu.edu) that represent key life stages of three parasitic species in the Orobanchaceae family (Westwood et al., 2012; Yang et al.,

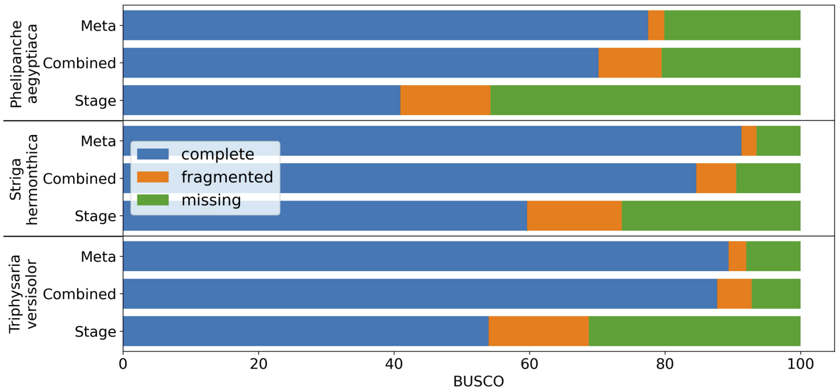


FIGURE 6
BUSCO completeness assessment of transcriptome assemblies to illustrate the results from targeted gene family assembly (meta-assembly) function in the PlantTribes2 *AssemblyPostProcessor* tool compared to Trinity approaches. Color bars indicate complete (blue), fragmented (orange), and missing (green) BUSCOs. Assemblies of parasitic plants, *Phelipanche*, *Striga*, and *Triphysaria*, examined include (1) developmental stage-specific assemblies (Stage, only the average of all the stages were shown in the plot), (2) assemblies combining all stage-specific raw data (Combined), and (3) meta-assembly of stage-specific assemblies and combined assembly (Meta) using *AssemblyPostProcessor*.

2015b). These species span the full spectrum of plant parasitism (Westwood et al., 2010; Westwood et al., 2012), and include *Triphysaria versicolor*, *Striga hermonthica*, and *Phelipanche aegyptiaca*. Species-specific transcriptome assemblies were performed with Trinity (Haas et al., 2013) using two approaches: (1) combining raw Illumina reads from all development stages of the plant in a single assembly, and (2) multiple assemblies of individual developmental stages of the plant. A BUSCO (benchmarked universal single-copy orthologs) (Manni et al., 2021) assembly quality assessment using 1,440 universally conserved land plants' single-copy orthologs suggests that the assembly combining all raw data recovers more conserved single-copy genes than any developmental stage-specific assembly (Combined *v.s.* Stage in Figure 6 and Supplemental Table 4). However, a meta-assembly of transcripts from both approaches with the targeted gene family function of the *AssemblyPostProcessor* tool using the 26Gv1.0 gene family scaffold recovers even more full-length conserved single-copy genes (Meta *v.s.* others in Figure 6 and Supplemental Table 4). Therefore, the meta-assembly implementation of the PlantTribes2 *AssemblyPostProcessor* tool can benefit many comparative transcriptome studies of

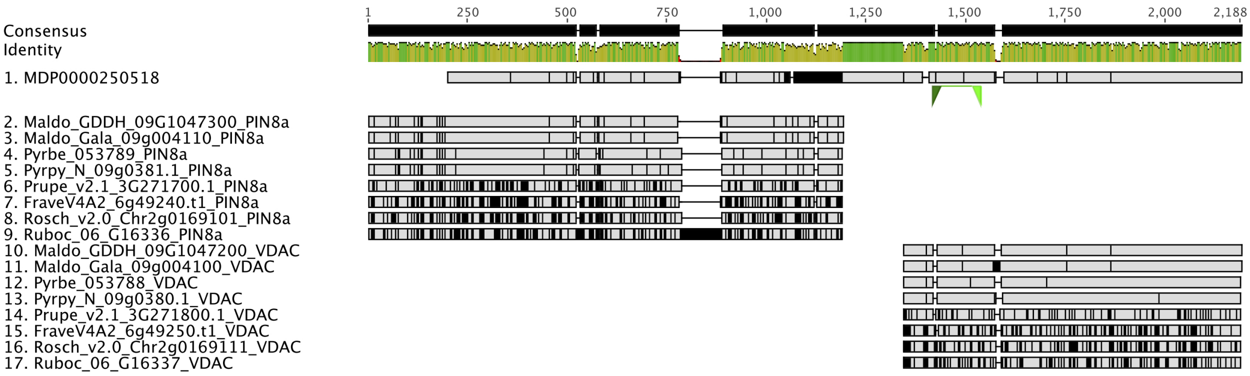

1
250
500
750
1,000
1,250
1,500
1,750
2,000
2,188
Consensus
Identity
1. MDP0000250518
2. Maldo_GDDH_09G1047300_PIN8a
3. Maldo_Gala_09g004110_PIN8a
4. Pyrbe_053789_PIN8a
5. Pyrpy_N_09g0381.1_PIN8a
6. Prupe_v2.1_3G271700.1_PIN8a
7. FraveV4A2_6g49240.t1_PIN8a
8. Rosch_v2.0_Chr2g0169101_PIN8a
9. Ruboc_06_G16336_PIN8a
10. Maldo_GDDH_09G1047200_VDAC
11. Maldo_Gala_09g004100_VDAC
12. Pyrbe_053788_VDAC
13. Pyrpy_N_09g0380.1_VDAC
14. Prupe_v2.1_3G271800.1_VDAC
15. FraveV4A2_6g49250.t1_VDAC
16. Rosch_v2.0_Chr2g0169111_VDAC
17. Ruboc_06_G16337_VDAC

FIGURE 7
Identification of an incorrect auxin transporter gene model, *MdPIN8a*, in *Malus domestica* genome annotation version 1. Nucleotide sequence alignment of putative *PIN8a* and *VDAC* genes from 9 Rosaceae genomes were shown here. *MDP0000250518* (sequence 1) gene model is a combination of two genes: The 5' end of *MDP0000250518* shares high sequence similarity with the *PIN8a* gene from other Rosaceae species (sequence 2 to 9), while its 3' end shows evidence of homology to a neighboring gene, *VDAC*, in the investigated genomes (sequence 10 to 17). Green triangles below *MDP0000250518* show the binding sites of the qRT-PCR primers used in the Song et al., 2016 research. Gray color indicates identical nucleotides compared to the consensus, while black color indicates different nucleotides. Genome abbreviations can be found in Supplemental Table 7.

non-model species to alleviate transcript fragmentation in gene families of interest.

3.2.2 Application in evaluating and improving gene families

Gene and gene family studies in non-model organisms are challenging due to the varying quality of genome assemblies and annotations, as well as the lack of closely related species as an annotation reference. Thousands of genes lack accurate gene models in draft and early version genomes (Darwish et al., 2015; Marx et al., 2016; Li et al., 2017; Pilkington et al., 2018; Li et al, 2019; Liu et al., 2021) creating pitfalls for global-scale analyses, but especially for researchers conducting reverse genetics studies. For example, in the first version of the apple (*Malus domestica*) genome annotation, we discovered that the gene model of *MDP0000250518*, annotated as *MdPIN8a* by Song et al. (2016), is problematic. A nucleotide sequence comparison of *MDP0000250518* and its orthologous genes in other Rosaceae genomes, identified using the PlantTribes2 orthogroup classification function, showed that this gene model is likely a combination of the putative *MdPIN8a* and a neighboring gene, which encodes a voltage dependent anion channel (VDAC) (Figure 7). These two genes are located about 3000bp apart on the same chromosome in most Rosaceae genomes (Supplemental Table 5). Analyses carried out using this incorrect gene model may confound or compromise the work. For example, in absence of the contextual gene family information we now have from analyses with PlantTribes2, the authors in Song et al. (2016) unknowingly designed primers for the *MDP0000250518* gene model that targeted the *VDAC* gene rather than the actual gene of interest, *MdPIN8a* (Figure 7). We identified the mis-annotated gene using contextual gene family information; a reliable way to avoid such pitfalls.

Better gene models can be obtained from re-annotating existing or new genome assemblies with additional transcriptome data. For instance, tens of thousands of gene models were improved or added in the subsequent annotations in several strawberry genomes (Darwish et al., 2015; Li et al., 2017; Li et al, 2019; Liu et al., 2021). In later versions of apple genome annotations, erroneous gene models such as *MdPIN8a* and the neighboring gene, *VDAC*, are corrected and are now concordant with other Rosaceae (Figure 7). This improved gene information provides a better starting point for studies like Song et al., 2016, however, full reannotation of complex plant genomes is a time-consuming and a resource-intensive undertaking.

A more efficient solution is targeted gene model improvement by evaluation of genes of interest (GOIs) from a gene family perspective. The comparative genomic and phylogenomic tools offered by PlantTribes2 allows researchers to efficiently compare orthologous genes across many closely related species and identify problematic genes in a high-throughput fashion. In a recent study with a goal to identify tree architecture genes in *Pyrus* (pear), functions from PlantTribes2 were used at the core of the workflow (Zhang et al., 2022, the companion paper in this issue). Using the

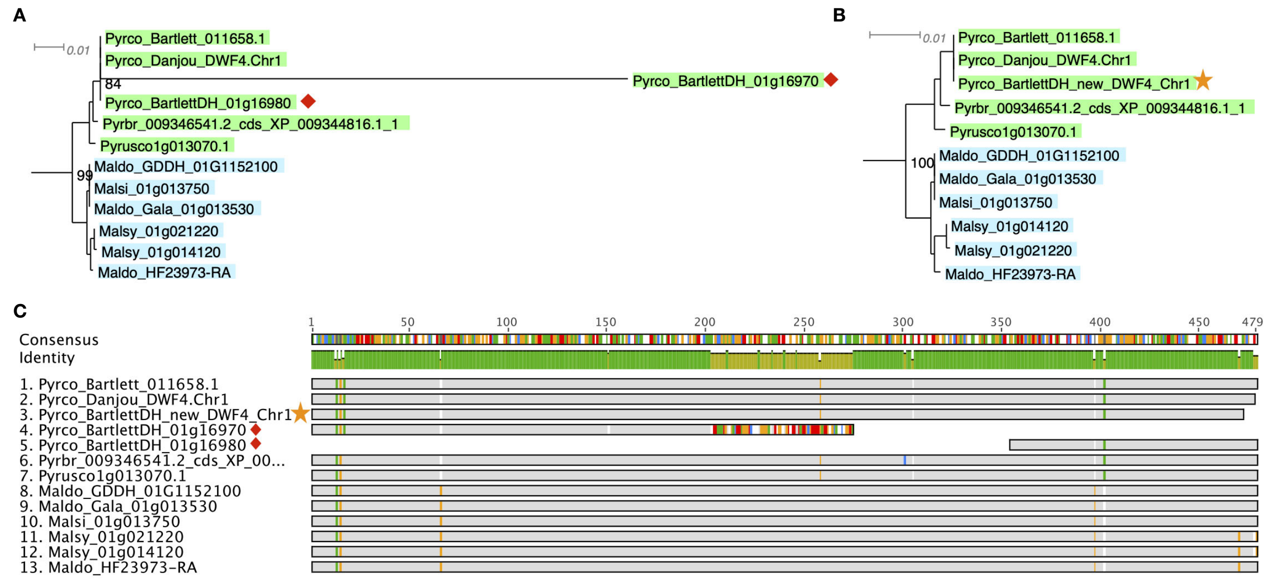


FIGURE 8
Example of a putative *DWF4* gene before (red diamond) and after (yellow star) improvement. **(A, B)** show a section of the *DWF4* gene family tree with Bartlett_DH gene models before and after improvement, respectively. **(C)** is the sequence alignment of the *DWF4* gene coding region in *Malus* and *Pyrus* genomes. Gray color indicates identical nucleotides compared to the consensus, while colors indicate different nucleotides. Genome abbreviations can be found in Supplemental Table 7.

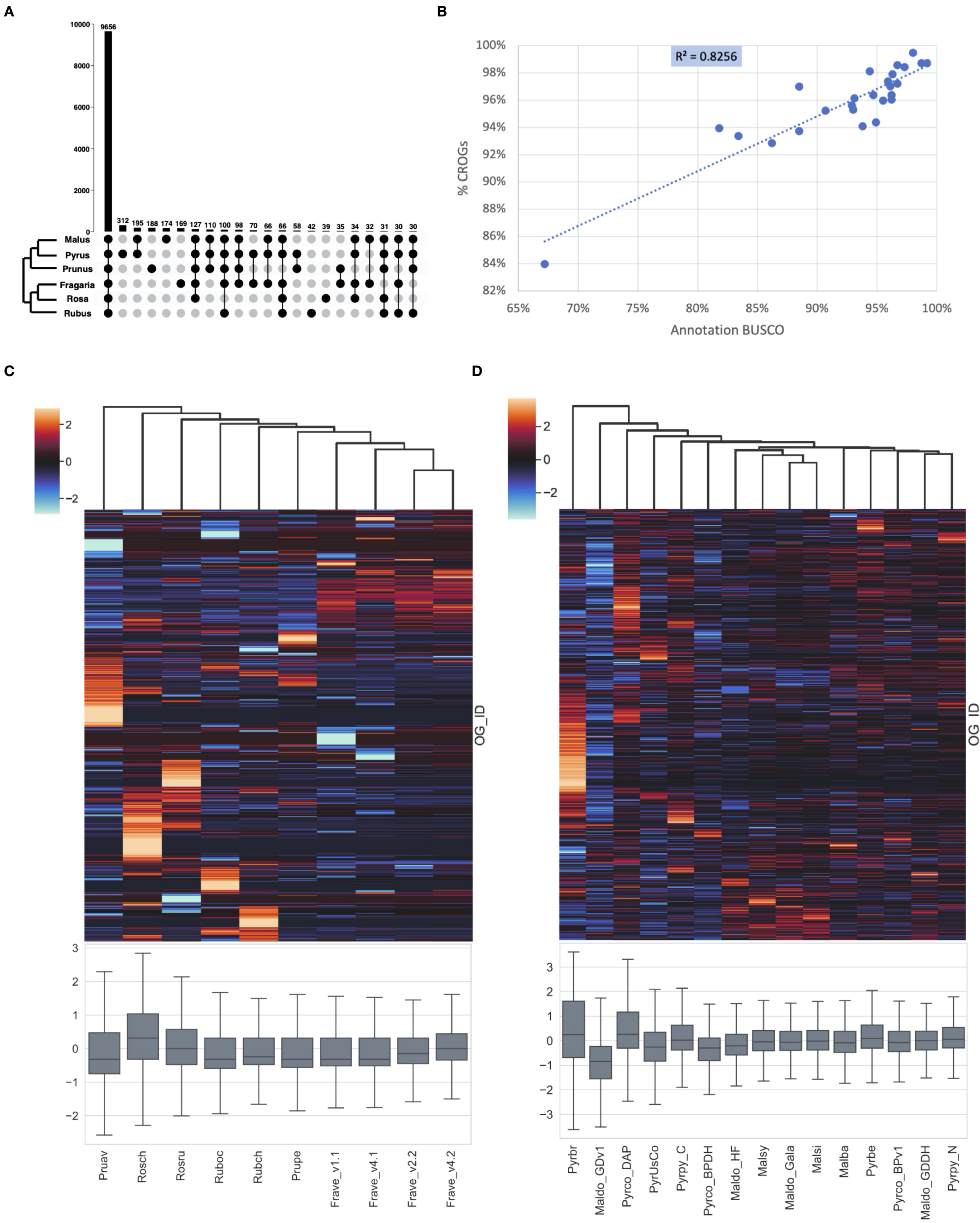


FIGURE 9
CoRe OrthoGroups - Rosaceae (CROGs). **(A)** Upset plot showing overlapping orthogroups between six Rosaceae genera, including 9656 orthogroups shared by all six genera (designated as "CROGs – Rosaceae"). **(B)** High correlation between Rosaceae genome annotation BUSCOs and % CROGs captured in the genomes (p<0.01). **(C)** Z-score distribution of gene counts in CROGs among selected Rosaceae genomes excluding *Malus* and *Pyrus*, shown as a clustermap (upper) and a box plot (lower). Each column represents a genome and each row in the clustermap represents a CROG. **(D)** Z-score distribution of gene counts in CROGs among selected *Malus* and *Pyrus* genomes, shown as a clustermap (upper) and a box plot (lower). Genome abbreviations can be found in Supplemental Table 7.

alignments and phylogenies generated by the *GeneFamilyAligner* and *GeneFamilyPhylogenyBuilder* tools from PlantTribes2, hundreds of problematic gene models were identified. For instance, two fragments of a putative pear *DWARF4* (*DWF4*) gene were found in the *Pyrus communis* 'Bartlett' Double Haploid (Bartlett.DH) genome annotation (Linsmith et al., 2019), one of which showed little evidence of homology at the 3' end of its coding sequence compared to other apple and pear *DWF4* genes. This problem was easily recognized in the nucleotide sequence alignment and phylogeny produced by PlantTribes2 (Figures 8A, C). Moreover, the homologous sequences from the PlantTribes2 orthogroups were readily used as resources for target-gene family annotation tools, such as TGFam-Finder (Kim et al., 2020) and Bitacora (Vizueta et al., 2020). In the case of *DWF4*, using the PlantTribes2 derived orthogroup information as reference, a more complete *DWF4* gene homologous to other Maleae sequences was annotated from the Bartlett.DH genome (Figures 8B, C). More examples like the *DWF4* gene are presented in Zhang et al., 2022.

3.2.3 Application in evaluating genome quality

A BUSCO analysis is a widely accepted benchmark to assess the completeness and accuracy of genomic resources (Manni et al., 2021). However, it only takes into consideration a very small fraction of the gene space. By definition, BUSCOs appear as highly conserved single copy genes in many organisms and return rapidly to single copy following gene and genome duplication. BUSCO genes may not reflect the quality of more challenging regions of the genome and the integrity of complex and divergent gene families. With more genomic resources being produced, especially in some agronomically important genera/species, lineage-specific BUSCO databases have been developed, bringing in larger numbers of markers. For instance, the poales_odb10 contains 3 times more markers than the generic embryophyta_odb10. However, this type of database has only been developed for 4 plant orders (Brassicales, Solanales, Poales, and Fabales), and like other BUSCO databases, only single copy genes are used. Following the same philosophy as the lineage-specific BUSCO databases, the natural next step is a gene-by-gene assessment on a genome scale, as proposed by Honaas et al (2016) regarding *de novo* transcriptome assembly evaluation. Here we present a case study of using the objective orthogroup classification offered by PlantTribes2 to evaluate the quality of genome annotations from a comparative perspective in Rosaceae, a step towards a gene-by-gene approach.

The number of publicly available Rosaceae genomes, generated by researchers all around the world using different technologies, has increased exponentially in the last decade (Jung et al., 2019). To better estimate the accuracy and sensitivity of genome annotation across a wide range of Rosaceae species, we created family-specific "CoRe OrthoGroups (CROGs) - Rosaceae". First, 26 representative genomes from six genera (*Malus*, *Pyrus*, *Prunus*, *Fragaria*, *Rosa*, and *Rubus*. Supplemental Tables 6, 7) in five major Rosaceae tribes were classified into the PlantTribes2 26Gv2.0 scaffold. Next, the union of orthogroups from each genus was generated, creating genus-level master orthogroups. Then the overlap of the six master orthogroups, consisting of 9656 orthogroups, were designated as the CROGs (Figure 9A, Supplemental Table 8), which is so far the most complete list of cores Rosaceae genes. Rich information from the CROGs, *i.e.*, the percentage of CROGs captured in each genome, gene counts in CROGs, and sequence similarity compared to the CROG consensus, can be used to assess annotation quality, pinpoint areas needing improvement, and find potentially interesting biology.

First, we calculated the percentage of CROGs captured in 26 Rosaceae genomes and correlated the %CROGs with the corresponding annotation BUSCO scores (Supplemental Table 6). The high positive correlation (R^2 = 0.82, Figure 9B) indicates that these two philosophically similar approaches draw the same conclusions for most genomes, however, CROGs provide additional information allowing more in-depth explorations of annotation quality.

Next, we calculated gene counts in each CROG. Due to the difference in chromosome numbers (17 chromosomes in Maleae and 9 in other genera) and a unique recent whole genome duplication event in the common ancestor of Maleae (Hodel et al, 2022), apple and pear genomes have more gene copies in most orthogroups than other Rosaceae. To make more appropriate comparisons, we generated two CROG gene count matrices, one for Maleae and one for other Rosaceae (Supplemental Tables 9, 10, respectively). Our hypothesis is that a high-quality genome will have a predictable and consistent number of genes in a large majority of CROGs. This is because issues that have predictable impacts on genome assembly and annotation are dependent on individual genome characteristics, the data used in assembly and annotation, and the various methodologies employed therein - thus creating a comparative framework with complementary error structure. Simply put, it is unlikely that a gene family will show a consistent yet erroneous shift in gene content due to methodological reasons alone. This perspective can reduce the false positive rate for evolutionary inference of lineage-specific shifts in gene family content by flagging changes in individual genomes that may be due to methodological bias.

As expected, in the non-Maleae matrix, nearly half of the CROGs (4,728) have the same number of genes or different gene counts in only 1 or 2 genomes. When we visualized the gene count matrix using the Seaborn z-score clustermap package (CROGs with standard deviation of 0 were removed prior to plotting), the four different versions of *Fragaria vesca* annotations clustered together (Figure 9C). They shared similar z-score patterns in most CROGs, but fewer low z-score regions (shown as cooler colors) were found in the later versions of annotation (v2.2 and v4.2). These two annotations also have a

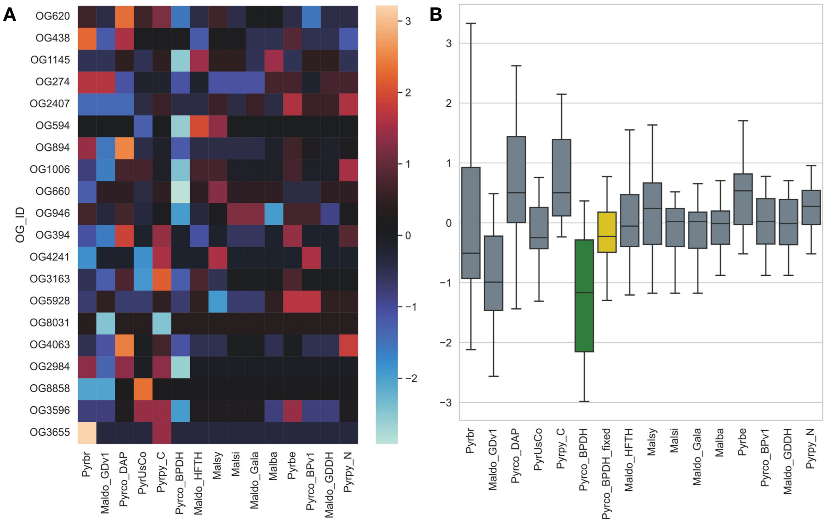


FIGURE 10
The gene count z-score of selected tree architecture gene families across *Pyrus* and *Malus* genomes. Pyrco_BPDH orthogroups have lower z-scores than most others, which is shown with a cooler color in the heatmap **(A)** and lower average z-score (green box in **B**), indicating fewer than expected gene counts. These missing genes were discovered after the targeted re-annotation process, which brought the average gene count z-score closer to 0 (yellow box in **B**) and comparable to other high-quality genomes. Genome abbreviations can be found in Supplemental Table 7.

mean z-score closer to 0 and relatively small variance compared to the earlier annotations. A similar pattern was seen while comparing the first version of the apple genome, Maldo_GDv1, to the more recent ones (Maldo_Gala and Maldo_GDDH in Figure 9D). Our results are consistent with previous reports (Daccord et al., 2017; Li et al., 2017; Li et al, 2019) and the CROG approach provided a fast and easy-to-visualize way to summarize these findings.

The clustermaps also allowed us to gain new insights from these genomes. For instance, there is not clear clustering of *Malus* or *Pyrus* at the genus level, however, the more recent genomes, which have less variable z-score distribution centered near 0, are clustered together (Figure 9D). We hypothesize that the current clustering is mainly driven by genome annotation strategy and quality, and therefore it is showing methodological similarities rather than biological patterns. The fact that *Malus domestica* Gala (Maldo_Gala) is clustered with *M. sieversii* (Malsi) and *M. sylvestris* (Malsy), genomes generated using the same method, rather than the other high-quality *M. domestica* genome, Maldo_GDDH, supports this hypothesis (Daccord et al., 2017; Sun et al., 2020).

Another unexpected observation is that the earlier version of the European pear (*Pyrus communis*) genome, Pyrco_BPv1 (Chagné et al., 2014), shared a more similar gene count pattern with some of the best Maleae genomes. On the contrary, the second version, Pyrco_BPDH (Linsmith et al., 2019), a double haploid genome, does not. Apple and pear are highly heterozygous, which is known to cause fragmented genome assembly and introduce multiple alleles to the annotation. Sequencing isogenic genotypes, such as a double haploid, is a common solution (Daccord et al., 2017; Linsmith et al., 2019; Zhang et al., 2019). This process will reduce the complexity in genome assembly and should have little to no influence on the number of genes in a genome, or even in individual gene families. When a smaller number of protein coding genes were annotated from the Pyrco_BPDH genome compared to version one, the authors hypothesized that the difference is resulted from removal of allelic sequences annotated as genes ("allelic genes") in the much more contiguous double haploid genome (Linsmith et al., 2019). However, our CROG gene count matrix indicates that the smaller gene number in Pyrco_BPDH is caused, at least in part, by CROG genes and gene families missing from the annotation - indeed the Pyrco_BPv1 genome captured a vast majority of the CROGs with the expected gene count, despite annotation of some allelic genes. This statement is supported by an investigation in putative tree architecture gene families by Zhang et al., 2022 (the companion paper). About half of the genes of interest were missing in the original Pyrco_BPDH annotation, but were recovered using a polished assembly and targeted annotation approaches (Figure 10).

The "hot" zones in the clustermaps also attract attention. To investigate the hot zones in the Maleae matrix, we examined the gene counts and annotation of 150 CROGs with the highest z-score from each genome. In most genomes, these CROG annotations lack a pattern, and the high z-score is caused by one or few extra copies, which may be caused by the introduction of alleles from fragmented assembly or could indicate genome-specific duplications. However, in some high z-score CROGs in *Pyrus betulifolia* (Pyrbe) (Dong et al., 2020), *Malus sieversii* (Malsi), *M. sylvestris* (Malsy), and *M. domestica* 'Gala' (Maldo_Gala) (Sun et al., 2020), the targeted genome has up to 10 times more genes than the others and the annotation of these CROGs are often related to transposons and repeat-containing genes (Supplemental Table 11). This finding suggests certain downstream analyses, such as repeat type comparison and gene family expansion estimation, can be bolstered against such pitfalls by a CROG analysis.

Using the PlantTribes2 orthogroup classification, we created a new method to evaluate genome quality in more depth, leveraging resources across an important plant family. The CROG gene count matrix does not only provide a highly effective way to visualize differences in gene numbers from a comparative genomic perspective, but also pinpoints where improvements could be made. As genomic resources are rapidly increasing, a CROG analysis can also help to inform the selection of the most appropriate genomes for comparative genomic studies, by avoiding specific issues related to assembly and annotation. Moreover, this approach can be applied to any groups of plants, creating custom CROGs for assessing the quality of genomes of interest.

4 Conclusions

PlantTribes2 uses pre-computed or expert gene family classifications for comparative and evolutionary analyses of gene families and transcriptomes for all types of organisms. The two main goals of PlantTribes2 are: (1) continual development of a scalable and modular set of analysis tools and methods that leverage gene family classifications for comparative genomics and phylogenomics to gain novel insight into the evolutionary history of genomes, gene families, and the tree of life; (2) to make these tools broadly available to the research community as a stand-alone package and also within the Galaxy Workbench. Many genomic studies, including inference of species relationships, the timing of gene duplication and polyploidy, reconstruction of ancestral gene content, the timing of new gene function evolution, detection of reticulate evolutionary events such as horizontal gene transfer, assessment of gene family and genome quality, and many others, can all be performed using PlantTribes2 tools. The modular structure, which allows component tools of the pipeline to be independent from each other, makes the PlantTribes2 tools easy to enhance over time.

Data availability statement

The datasets presented in this study can be found in online repositories: Project name: PlantTribes2 Archived version: 1.0.4 Project home page: https://github.com/dePamphilis/PlantTribes; Galaxy: https://usegalaxy.org Bioconda: https://bioconda.github.io/search.html?q=PlantTribes; Tutorials: https://github.com/dePamphilis/PlantTribes/blob/master/docs/Tutorial.md.; https://galaxyproject.org/tutorials/pt_gfam/; Operating system(s): Linux, Mac OS X; Programming language: Perl, Python; Other requirements: Web browser for Galaxy; 553 License: GNU.

Author contributions

EW, JL-M., and CD conceived and designed the research. EW, HZ, LH, and GK performed the analyses. All authors contributed to the article and approved the submitted version.

Funding

We acknowledge funding from the National Science Foundation (NSF) DBI-1238057, NSF IOS-0922742, the NSF Plant Cyberinfrastructure Program through iPlant (now CyVerse; DBI-0735191), the The iPlant Tree of Life Grand Challenge Project, USDA ARS, WTFRC grant AP-19-103.

Acknowledgments

The authors thank Heidi Hargarten for editing and revising the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1011199/full#supplementary-material

References

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., et al. (2019). OMA standalone: Orthology inference among public and custom genomes and transcriptomes. *Genome Res.* 29, 1152–1163. doi: 10.1101/gr.243212.118

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets - update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bel, M. V., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., et al. (2021). PLAZA 5.0: Extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Res.* 50, D1468–D1474. doi: 10.1093/nar/gkab1024

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* 53, 474–485. doi: 10.1002/dvg.22877

Blankenberg, D., Kuster, G. V., Bouvier, E., Baker, D., Afgan, E., Stoler, N., et al. (2014). Dissemination of scientific software with galaxy ToolShed. *Genome Biol.* 15, 403. doi: 10.1186/gb4161

Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C., et al. (2016). EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res.* 44, W22–W28. doi: 10.1093/nar/gkw255

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events *Nat.* 422, 433–438. doi: 10.1038/nature01521

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinf.* 10, 421. doi: 10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., et al. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

Carvalho, D. S., Schnable, J. C., and Almeida, A. M. R. (2018). Integrating phylogenetic and network approaches to study gene family evolution: The case of the *AGAMOUS* family of floral genes. *Evol. Bioinform. Online* 14, 1176934318764683. doi: 10.1177/1176934318764683

Chagné, D., Crowhurst, R. N., Pindo, M., Thrimawithana, A., Deng, C., Ireland, H., et al. (2014). The draft genome sequence of European pear (*Pyrus communis* l. 'Bartlett'). *PloS One* 9, e92644. doi: 10.1371/journal.pone.0092644

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, F., Mackey, A. J., Stoeckert, C. J., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34, D363–D368. doi: 10.1093/nar/gkj123

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., et al. (2017). High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* 49, 1099–1106. doi: 10.1038/ng.3886

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PloS One* 5, e11147. doi: 10.1371/journal.pone.0011147

Darwish, O., Shahan, R., Liu, Z., Slovin, J. P., and Alkharouf, N. W. (2015). Re-annotation of the woodland strawberry (*Fragaria vesca*) genome. *BMC Genomics* 16, 29. doi: 10.1186/s12864-015-1221-1

Deelman, E., Vahi, K., Juve, G., Rynge, M., Callaghan, S., Maechling, P. J., et al. (2015). Pegasus, A workflow management system for science automation. *Future Gener. Comp. Sy* 46, 17–35. doi: 10.1016/j.future.2014.10.008

Derelle, R., Philippe, H., and Colbourne, J. K. (2020). Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol. Biol. Evol.* 37, msaa159. doi: 10.1093/molbev/msaa159

Dong, X., Wang, Z., Tian, L., Zhang, Y., Qi, D., Huo, H., et al. (2020). *De novo* assembly of a wild pear (*Pyrus betuleafolia*) genome. *Plant Biotechnol. J.* 18, 581–595. doi: 10.1111/pbi.13226

Dunn, C. W., Howison, M., and Zapata, F. (2013). Agalma: an automated phylogenomics workflow. *BMC Bioinf.* 14, 330–330. doi: 10.1186/1471-2105-14-330

Ebmeyer, S., Coertze, R. D., Berglund, F., Kristiansson, E., and Larsson, D. G. J. (2021). GEnView: a gene-centric, phylogeny-based comparative genomics pipeline for bacterial genomes and plasmids. *Bioinformatics* 38, 1727–1728. doi: 10.1093/bioinformatics/btab855

Eddy, S. R. (2011). Accelerated profile HMM searches. *PloS Comput. Biol.* 7, e1002195. doi: 10.1371/journal.pcbi.1002195

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157. doi: 10.1186/s13059-015-0721-2

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Emms, D. M., and Kelly, S. (2022). SHOOT: phylogenetic gene search and ortholog inference. *Genome Biol.* 23, 85. doi: 10.1186/s13059-022-02652-8

Enright, A. J., Dongen, S. V., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Fuentes, D., Molina, M., Chorostecki, U., Capella-Gutiérrez, S., Marcet-Houben, M., and Gabaldón, T. (2021). PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res.* 50, D1062–D1068. doi: 10.1093/nar/gkab966

Gabaldón, T. (2008). Large-Scale assignment of orthology: back to phylogenetics? *Genome Biol.* 9, 235. doi: 10.1186/gb-2008-9-10-235

Gao, Y., Yang, Q., Yan, X., Wu, X., Yang, F., Li, J., et al. (2021). High-quality genome assembly of "Cuiguan" pear (*Pyrus pyrifolia*) as a reference genome for identifying regulatory genes and epigenetic modifications responsible for bud dormancy. *Hortic. Res.* 8, 197. doi: 10.1038/s41438-021-00632-w

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Gremme, G., Steinbiss, S., and Kurtz, S. (2013). GenomeTools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE ACM Trans. Comput. Biol. Bioinform.* 10, 645–656. doi: 10.1109/tcbb.2013.68

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hodel, R. G. J., Zimmer, E. A., Liu, B.-B., and Wen, J. (2022). Synthesis of nuclear and chloroplast data combined with network analyses supports the polyploid origin of the apple tribe and the hybrid origin of the maleae-gillenieae clade. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.820997

Honaas, L. A., Wafula, E. K., Wickett, N. J., Der, J. P., Zhang, Y., Edger, P. P., et al. (2016). Selecting superior *De novo* transcriptome assemblies: Lessons learned by leveraging the best plant genome. *PloS One* 11, e0146062. doi: 10.1371/journal.pone.0146062

Huang, X., and Madan, A. (1999).). CAP3: A DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., et al. (2016). Resolution of brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.* 33, 394–412. doi: 10.1093/molbev/msv226

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi: 10.1093/nar/gkv1248

Iseli, C., Jongeneel, C. V., and Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf Intelligent Syst. Mol. Biol. Ismb Int. Conf Intelligent Syst. Mol. Biol.*, 138–148.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13, R3–R3. doi: 10.1186/gb-2012-13-1-r3

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., et al. (2019). 15 years of GDR: New data and functionality in the genome database for rosaceae. *Nucleic Acids Res.* 47, D1137–D1145. doi: 10.1093/nar/gky1000

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kim, S., Cheong, K., Park, J., Kim, M., Kim, J., Seo, M., et al. (2020). TGFam-finder: a novel solution for target-gene family annotation in plants. *New Phytol.* 227, 1568–1581. doi: 10.1111/nph.16645

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., et al. (2018). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47, gky1053. doi: 10.1093/nar/gky1053

Lanza, V. F., Baquero, F., de la Cruz, F., and Coque, T. M. (2016). AcCNET (Accessory genome constellation network): Comparative genomics software for accessory genome analysis using bipartite networks. *Bioinformatics* 33, 283–285. doi: 10.1093/bioinformatics/btw601

Li, Z., Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., Rieseberg, L. H., et al. (2015). Early genome duplications in conifers and other seed plants. *Sci. Adv.* 1, e1501084. doi: 10.1126/sciadv.1501084

Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.-M., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* 4, 460–472. doi: 10.1038/s41477-018-0188-8

Linsmith, G., Rombauts, S., Montanari, S., Deng, C. H., Celton, J.-M., Guérif, P., et al. (2019). Pseudo-chromosome-length genome assembly of a double haploid "Bartlett" pear (*Pyrus communis* l.). *GigaScience* 8, 1-17. doi: 10.1093/gigascience/giz138

Li, Y., Pi, M., Gao, Q., Liu, Z., and Kang, C. (2019). Updated annotation of the wild strawberry *Fragaria vesca* V4 genome. *Hortic. Res.* 6, 61. doi: 10.1038/s41438-019-0142-6

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Liu, T., Li, M., Liu, Z., Ai, X., and Li, Y. (2021). Reannotation of the cultivated strawberry genome and establishment of a strawberry genome database. *Hortic. Res.* 8, 41. doi: 10.1038/s41438-021-00476-4

Li, Y., Wei, W., Feng, J., Luo, H., Pi, M., Liu, Z., et al. (2017). Genome re-annotation of the wild strawberry *Fragaria vesca* using extensive illumina- and SMRT-based RNA-seq datasets. *DNA Res.* 25, dsx038. doi: 10.1093/dnares/dsx038

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53, 661–673. doi: 10.1111/j.1365-313x.2007.03326.x

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38, 4647–4654. doi: 10.1093/molbev/msab199

Marks, R. A., Hotaling, S., Frandsen, P. B., and VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* 7, 1571–1578. doi: 10.1038/s41477-021-01031-8

Martinez, M. (2016). Computational tools for genomic studies in plants. *Curr. Genomics* 17, 509–514. doi: 10.2174/1389202917666160520103447

Marx, H., Minogue, C. E., Jayaraman, D., Richards, A. L., Kwiecien, N. W., Siahpirani, A. F., et al. (2016). A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. *Nat. Biotechnol.* 34, 1198–1205. doi: 10.1038/nbt.3681

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., et al. (2014). Data access for the 1,000 plants (1KP) project. *GigaScience* 3, 17. doi: 10.1186/2047-217x-3-17

McLachlan, G. J., and Peel, D. (1999). The EMMIX algorithm for the fitting of normal and t -components. *J. Stat. Softw* 4, 1-14. doi: 10.18637/jss.v004.i02

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., et al. (2020). PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* 49, gkaa1106. doi: 10.1093/nar/gkaa1106

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: Ultra-Large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.* 22, 377–386. doi: 10.1089/cmb.2014.0156

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, gkaa913. doi: 10.1093/nar/gkaa913

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., et al. (2021). Sustainable data analysis with snakemake. *F1000research* 10, 33. doi: 10.12688/f1000research.29032.1

Nagy, L. G., Merényi, Z., Hegedüs, B., and Bálint, B. (2020). Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic Acids Res.* 48, 2209–2219. doi: 10.1093/nar/gkz1241

Nakaya, A., Ichihara, H., Asamizu, E., Shirasawa, S., Nakamura, Y., Tabata, S., et al. (2017). Plant genome DataBase Japan (PGDBj). *Methods Mol. Biol. Clifton N J.* 1533, 45–77. doi: 10.1007/978-1-4939-6658-5_3

Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carriço, J. A., and Vaz, C. (2016). PHYLOViZ 2.0: Providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinform. Oxf Engl.* 33, 128–129. doi: 10.1093/bioinformatics/btw582

Oliveira, M. S., Alves, J. T. C., de Sá, P. H. C. G., and de Veras, A. A. O. (2021). PAN2HGENE–tool for comparative analysis and identifying new gene products. *PloS One* 16, e0252414. doi: 10.1371/journal.pone.0252414

One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Pabón-Mora, N., Wong, G. K.-S., and Ambrose, B. A. (2014). Evolution of fruit development genes in flowering plants. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00300

Perrin, A., and Rocha, E. P. C. (2021). PanACoTA: a modular tool for massive microbial comparative genomics. *NAR Genom Bioinform.* 3, lqaa106. doi: 10.1093/nargab/lqaa106

Pilkington, S. M., Crowhurst, R., Hilario, E., Nardozza, S., Fraser, L., Peng, Y., et al. (2018). A manually annotated *Actinidia chinensis* var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. *BMC Genomics* 19, 257. doi: 10.1186/s12864-018-4656-3

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - approximately maximum-likelihood trees for Large alignments. *PloS One* 5, e9490. doi: 10.1371/journal.pone.0009490

Pucker, B. (2022). Automatic identification and annotation of MYB gene family members in plants. *BMC Genomics* 23, 220. doi: 10.1186/s12864-022-08452-5

Pucker, B., Reiher, F., and Schilbert, H. M. (2020). Automatic identification of players in the flavonoid biosynthesis with application on the biomedicinal plant *Croton tiglium*. *Plants* 9, 1103. doi: 10.3390/plants9091103

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., et al. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol. Plant* 11, 414–428. doi: 10.1016/j.molp.2018.01.002

Rothfels, C. J., Larsson, A., Li, F.-W., Sigel, E. M., Huiet, L., Burge, D. O., et al. (2013). Transcriptome-mining for single-copy nuclear markers in ferns. *PloS One* 8, e76957. doi: 10.1371/journal.pone.0076957

Sasidharan, R., Nepusz, T., Swarbreck, D., Huala, E., and Paccanaro, A. (2012). GFam: a platform for automatic annotation of gene families. *Nucleic Acids Res.* 40, e152–e152. doi: 10.1093/nar/gks631

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2018). GenBank. *Nucleic Acids Res.* 47, D94–D99. doi: 10.1093/nar/gky989

Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M., and Bateman, A. (2014). TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.* 42, D922–D925. doi: 10.1093/nar/gkt1055

Shahid, S., Kim, G., Johnson, N. R., Wafula, E., Wang, F., Coruh, C., et al. (2018). MicroRNAs from the parasitic plant *Cuscuta campestris* target host messenger RNAs. *Nature* 553, 82. doi: 10.1038/nature25027

Song, C., Zhang, D., Zhang, J., Zheng, L., Zhao, C., Ma, J., et al. (2016). Expression analysis of key auxin synthesis, transport, and metabolism genes in different young dwarfing apple trees. *Acta Physiol. Plant* 38, 43. doi: 10.1007/s11738-016-2065-2

Sonnhammer, E. L. L., and Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18, 619–620. doi: 10.1016/s0168-9525(02)02793-2

Spannagl, M., Nussbaumer, T., Bader, K. C., Martis, M. M., Seidel, M., Kugler, K. G., et al. (2016). PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* 44, D1141–D1147. doi: 10.1093/nar/gkv1130

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.-C., Sjödin, A., et al. (2015). The plant genome integrative explorer resource: PlantGenIE.org. *New Phytol.* 208, 1149–1156. doi: 10.1111/nph.13557

Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., et al. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* 52, 1423–1432. doi: 10.1038/s41588-020-00723-9

Tello-Ruiz, M. K., Naithani, S., Gupta, P., Olson, A., Wei, S., Preece, J., et al. (2020). Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.* 49, gkaa979. doi: 10.1093/nar/gkaa979

Thanki, A. S., Soranzo, N., Haerty, W., and Davey, R. P. (2018). GeneSeqToFamily: a galaxy workflow to find gene families based on the ensembl compara GeneTrees pipeline. *Gigascience* 7, giy005. doi: 10.1093/gigascience/giy005

The Amborella Genome Project. (2013). The *Amborella* genome and the evolution of flowering plants. *Sci. (New York N.Y.)* 342, 1241089. doi: 10.1126/science.1241089

The Galaxy Community. (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research* 50 (W1), W345-W351. doi: 10.1093/nar/gkac247

The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi: 10.1093/nar/gkaa1100

Timilsena, P. R., Barrett, C. F., Nelson, A. P., Wafula, E. K., Ayyampalayam, S., McNeal, J. R., et al. (in press). Phylotranscriptomic analyses of mycoheterotrophic monocots show a continuum of convergent evolutionary changes in expressed nuclear genes from three independent nonphotosynthetic lineages. *Genome Biology and Evolution*.

Timilsena, P. R., Wafula, E. K., Barrett, C. F., Ayyampalayam, S., McNeal, J. R., Rentsch, J. D., et al (2022). Phylogenomic resolution of order- and family-level monocot relationships using 602 single-copy nuclear genes and 1375 BUSCO genes. *Front Plant Sci* 13, 876779. doi: 10.3389/fpls.2022.876779

Timme, R. E., Bachvaroff, T. R., and Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PloS One* 7, e29696. doi: 10.1371/journal.pone.0029696

Tomcal, M., Stiffler, N., and Barkan, A. (2013). POGs2: A web portal to facilitate cross-species inferences about protein architecture and function in plants. *PloS One* 8, e82569. doi: 10.1371/journal.pone.0082569

Tommaso, P. D., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Valentin, G., Abdel, T., Gaëtan, D., Jean-François, D., Matthieu, C., and Mathieu, R. (2020). GreenPhylDB v5: a comparative pangenomic database for plant genomes. *Nucleic Acids Res.* 49, D1464–D1471. doi: 10.1093/nar/gkaa1068

Vihinen, M. (2012). How to evaluate performance of prediction methods? measures and their interpretation in variation effect analysis. *BMC Genomics* 13, S2. doi: 10.1186/1471-2164-13-s4-s2

Viruel, J., Conejero, M., Hidalgo, O., Pokorny, L., Powell, R. F., Forest, F., et al. (2019). A target capture-based method to estimate ploidy from herbarium specimens. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00937

Vizueta, J., Sánchez-Gracia, A., and Rozas, J. (2020). Bitacora: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Mol. Ecol. Resour* 20, 1445–1452. doi: 10.1111/1755-0998.13202

Wall, P. K., Leebens-Mack, J., Müller, K. F., Field, D., Altman, N. S., and dePamphilis, C. W. (2008). PlantTribes: a gene and gene family resource for comparative genomics in plants. *Nucleic Acids Res.* 36, D970–D976. doi: 10.1093/nar/gkm972

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Westwood, J. H., dePamphilis, C. W., Das, M., Fernández-Aparicio, M., Honaas, L. A., Timko, M. P., et al. (2012). The parasitic plant genome project: New tools for understanding the biology of *Orobanche* and *Striga*. *Weed Sci.* 60, 295306. doi: 10.1614/ws-d-11-00113.1

Westwood, J. H., Yoder, J. I., Timko, M. P., and dePamphilis, C. W. (2010). The evolution of parasitism in plants. *Trends Plant Sci.* 15, 227–235. doi: 10.1016/j.tplants.2010.01.004

Whittle, C. A., Kulkarni, A., and Extavour, C. G. (2021). Evolutionary dynamics of sex-biased genes expressed in cricket brains and gonads. *J. Evol. Biol.* 34, 1188–1211. doi: 10.1111/jeb.13889

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.* 111, E4859–E4868. doi: 10.1073/pnas.1323926111

Williams, J. S., Der, J. P., dePamphilis, C. W., and Kao, T.-H. (2014). Transcriptome analysis reveals the same 17 s-locus f-box genes in two haplotypes of the self-incompatibility locus of *Petunia inflata*. *Plant Cell* 26, 2873–2888. doi: 10.1105/tpc.114.126920

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., et al. (2017). Evolution of rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* 34, 262–281. doi: 10.1093/molbev/msw242

Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., et al. (2016). MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinform. Oxf Engl.* 32, 3501–3503. doi: 10.1093/bioinformatics/btw474

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, Y., Moore, M. J., Brockington, S. F., Soltis, D. E., Wong, G. K.-S., Carpenter, E. J., et al. (2015a). Dissecting molecular evolution in the highly diverse plant clade caryophyllales using transcriptome sequencing. *Mol. Biol. Evol.* 32, 2001–2014. doi: 10.1093/molbev/msv081

Yang, Z., Wafula, E. K., Honaas, L. A., Zhang, H., Das, M., Fernandez-Aparicio, M., et al. (2015b). Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. *Mol. Biol. Evol.* 32, 767–790. doi: 10.1093/molbev/msu343

Yang, Z., Wafula, E. K., Kim, G., Shahid, S., McNeal, J. R., Ralph, P. E., et al. (2019). Convergent horizontal gene transfer and cross-talk of mobile nucleic acids in parasitic plants. *Nat. Plants* 5, 991–1001. doi: 10.1038/s41477-019-0458-0

Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N., and Ma, H. (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* 5, 4956. doi: 10.1038/ncomms5956

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., et al. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat. Commun.* 10, 1494. doi: 10.1038/s41467-019-09518-x

Zhang, Y., Sun, Y., and Cole, J. R. (2014). A scalable and accurate targeted gene assembly tool (SAT-assembler) for next-generation sequencing data. *PloS Comput. Biol.* 10, e1003737. doi: 10.1371/journal.pcbi.1003737

Zhang, H., Wafula, E. K., Eilers, J., Harkess, A. E., Ralph, P. E., Timilsena, P. R., et al. (2022). Building a foundation for gene family analysis in rosaceae genomes with a novel workflow: a case study in *Pyrus* architecture genes. *Front. Plant Sci* 13. doi: 10.3389/fpls.2022.975942

Zhang, N., Wen, J., and Zimmer, E. A. (2015). Expression patterns of *AP1*, *FUL*, *FT* and *LEAFY* orthologs in vitaceae support the homology of tendrils and inflorescences throughout the grape family. *J. Syst. Evol.* 53, 469–476. doi: 10.1111/jse.12138

Zwaenepoel, A., and de Peer, Y. V. (2019). Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol. Biol. Evol.* 36, 1384–1404. doi: 10.1093/molbev/msz088

TYPE Original Research
PUBLISHED 14 February 2023
DOI 10.3389/fpls.2023.1087023

Check for updates

OPEN ACCESS

EDITED BY
Concetta Licciardello,
CREA Research Centre for Olive, Fruit and Citrus Crops, Italy

REVIEWED BY
Pablo Aleza,
Instituto Valenciano de Investigaciones Agrarias, Spain
Angelo Ciacciulli,
Council for Agricultural Research and Economics, Italy
François Luro,
INRA Centre de Corse, France

*CORRESPONDENCE
Frederick G. Gmitter Jr.
✉ fgmitter@ufl.edu

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 01 November 2022
ACCEPTED 30 January 2023
PUBLISHED 14 February 2023

CITATION
Kumar K, Yu Q, Bhatia D, Honsho C and Gmitter Jr. FG (2023) Construction of a high density genetic linkage map to define the locus conferring seedlessness from Mukaku Kishu mandarin.
Front. Plant Sci. 14:1087023.
doi: 10.3389/fpls.2023.1087023

Construction of a high density genetic linkage map to define the locus conferring seedlessness from Mukaku Kishu mandarin

Krishan Kumar[1,2], Qibin Yu[1], Dharminder Bhatia[3], Chitose Honsho[1,4] and Frederick G. Gmitter Jr.[1*]

[1]Citrus Research and Education Center, Institute of Food and Agricultural Sciences, University of Florida, Lake Alfred, FL, United States, [2]Punjab Agricultural University, Dr. JC Bakhshi Regional Research Station, Abohar, India, [3]Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, India, [4]Laboratory of Pomology, Faculty of Agriculture, University of Miyazaki, Miyazaki, Japan

Mukaku Kishu ('MK'), a small sized mandarin, is an important source of seedlessness in citrus breeding. Identification and mapping the gene(s) governing 'MK' seedlessness will expedite seedless cultivar development. In this study, two 'MK'-derived mapping populations- LB8-9 Sugar Belle® ('SB') × 'MK' (N=97) and Daisy ('D') × 'MK' (N=68) were genotyped using an *Axiom_Citrus56 Array* encompassing 58,433 SNP probe sets, and population specific male and female parent linkage maps were constructed. The parental maps of each population were integrated to produce sub-composite maps, which were further merged to develop a consensus linkage map. All the parental maps (except 'MK_D') had nine major linkage groups, and contained 930 ('SB'), 810 ('MK_SB'), 776 ('D') and 707 ('MK_D') SNPs. The linkage maps displayed 96.9 ('MK_D') to 98.5% ('SB') chromosomal synteny with the reference Clementine genome. The consensus map was comprised of 2588 markers including a phenotypic seedless (*Fs*)-locus and spanned a genetic distance of 1406.84 cM, with an average marker distance of 0.54 cM, which is substantially lower than the reference Clementine map. For the phenotypic *Fs*-locus, the distribution of seedy and seedless progenies in both 'SB' × 'MK' (55:42, χ^2 = 1.74) and 'D' × 'MK' populations (33:35, χ^2 = 0.06) followed a test cross pattern. The *Fs*-locus mapped on chromosome 5 with SNP marker 'AX-160417325' at 7.4 cM in 'MK_SB' map and between two SNP markers 'AX-160536283' and 'AX-160906995' at a distance of 2.4 and 4.9 cM, respectively in 'MK_D' map. The SNPs 'AX-160417325' and 'AX-160536283' correctly predicted seedlessness of 25-91.9% progenies in this study. Based on the alignment of flanking SNP markers to the Clementine reference genome, the candidate gene for seedlessness hovered in a ~ 6.0 Mb region between 3.97 Mb (AX-160906995) to 10.00 Mb (AX-160536283). This region has 131 genes of which 13 genes (belonging to seven gene families) reportedly express in seed coat or developing embryo. The findings of the study will prove helpful in directing future research for fine mapping this region and eventually underpinning the exact causative gene governing seedlessness in 'MK'.

KEYWORDS

mandarin, consensus map, Mukaku Kishu, seedlessness, SNP array

Introduction

Edible citrus comprises a group of fruits mainly mandarins, sweet oranges, lime, lemons, pummelos, and grapefruit. These fruits are valued for their nutritive and health promoting abilities. Among the various citrus types, mandarins have primary utility as a fresh fruit. The major breeding objectives for mandarins are high eating quality, seedlessness, easy peelability, and round the year availability of fruit (Navarro et al., 2015). Conventional hybridization has been the most important method for genetic improvement of mandarins, but it is costly as well as challenging. For example, the release of LB8-9 Sugar Belle® ('SB') took 24 years from the year of its original cross. The integration of the molecular markers into the hybridization-based breeding program can expedite the pace of varietal development in citrus (Gmitter et al., 2007). The use of closely associated markers can allow selection of the desirable progenies many years before the evaluation for the targeted trait becomes possible, thus, help in compressing the breeding cycle (Dirlewanger et al., 2004).

Successful employment of marker assisted selection requires linkage maps with wide genomic coverage. The availability of genome sequences of different citrus cultivars and accessions has helped to improve the resolution of linkage maps through the informative expressed sequences (EST) derived EST-SSRs (Chen et al., 2008), gene-derived cleaved amplified polymorphic sequences (CAPS) (Shimada et al., 2014) and single nucleotide polymorphism based markers (SNPs) (Ollitrault et al., 2012; Chen and Gmitter, 2013; Yu et al., 2016). SNPs are the most abundant DNA markers which are evenly distributed on a whole genome and can tag almost any gene or locus of a genome (Brookes, 1999). With rapid developments in next generation sequencing technologies and availability of reference whole genome sequences, SNPs have become the marker of choice in genetics studies. SNP array-based genotyping platforms have been considered useful for developing high density linkage maps, gene/QTL mapping, and marker-assisted crop breeding. Their use over multiple populations also provides opportunity for development of integrated linkage maps for higher resolution of the loci conferring the target trait (Cui et al., 2017). High throughput markers such as DArTseq markers (Curtolo et al., 2017; Curtolo et al., 2018) and SNP markers through genotype by sequencing (GBS) technology (Huang et al., 2018) or by SNP array platforms (Yu et al., 2016) have been used to develop high density linkage maps in citrus. Previously, a medium density 1536 Illumina Golden Gate SNP-array was used to construct a mandarin linkage map (Yu et al., 2016), but a more dense *Axiom_Citrus56 Array* encompassing of 58,433 SNP probe sets became available for genotyping in citrus (Hiraoka, 2020).

Seedlessness is required for new mandarin cultivars by the marketplace. Citrus breeders have used Mukaku Kishu ('MK') as pollen parent in crosses with seedy monoembryonic diploid parents to produce seedless varieties (Nesumi et al., 2001). Seedlessness can be achieved through interploid hybridization in citrus (Grosser and Gmitter, 2005; Aleza et al., 2012a; Aleza et al., 2012b), or by exploiting cytoplasmic male sterility (CMS) of the Satsuma group of mandarins (Goto et al., 2018) or from self-incompatible pollen-pistil interaction (Montalt et al., 2021). However, triploid breeding requires the synthesis of tetraploids or exploitation of unreduced gametes, and recovery of triploid embryos generally is accomplished through embryo culture technique (Aleza et al., 2012a); further, some triploid hybrids can have occasionally very few seeds in their fruit, particularly with cross pollination. Similarly, induction of seedlessness using the CMS system is dependent on the cooperative action of the cytoplasmic and nuclear genes derived from Kishu and Satsuma mandarin (Yamamoto et al., 1997; Goto et al., 2018). On the other hand, self-incompatibility produces seedless fruits in self-pollination events. The response of self-incompatibility can also be influenced by environment (Aloisi et al., 2020). In contrast to these approaches, the seedlessness of 'MK' is governed by nuclear genes (Nesumi et al., 2001). The complete seedlessness in 'MK' is caused by developmental arrest of fertilized seeds. It is hypothesized that a two gene system possibly determines the seedlessness of 'MK': *Fs* a dominant gene that induces seedlessness, and *Is*, a repressor gene which in dominant state inhibits seedlessness. The allelic constitution of these two genes for 'MK' is reported to be '*Fsfs-isis*' and for seedy mandarins as '*fsfs-isis*' (Yamasaki et al., 2007). However, Nesumi et al. (2001) suggested that for mandarin crosses, this type of seedlessness is primarily determined by a single major locus. Several mandarin populations involving 'MK' as pollen parent have been developed, and high-resolution linkage maps can provide the exact location of genes controlling various traits as well as the markers closely associated with them. Previous attempts to identify and map genes governing seedlessness from 'MK' used RAPD markers or some limited number of gene-based markers (Shimada et al., 2014), but those identified were family-specific (Chavez and Chaparro, 2011). In this study, high density linkage maps for 'SB', Daisy ('D'), 'MK' were constructed by genotyping with *Axiom_Citrus56 SNP Array*, these maps were integrated to create a consensus mandarin map, which was used to identify and map the locus governing the seedlessness trait in 'MK'.

Materials and methods

Mapping populations

The 206 progeny individuals from five different F_1 mandarin populations growing at the University of Florida-IFAS Citrus Research and Education Center (Lake Alfred, FL) were used in this study (Figure 1). Mukaku Kishu (*Citrus reticulata* Blanco) ('MK'), a completely seedless mandarin cultivar, was the common male parent in all the populations. All five maternal mandarin parents, 'SB' (Clementine mandarin × Minneola tangelo), 'D' [(Clementine mandarin × Orlando tangelo) × (Clementine mandarin × Ponkan mandarin)], Temple ('T') (a natural mandarin × sweet orange hybrid), Lee ('L') (Clementine mandarin × Orlando tangelo), and Clementine × Valencia orange ('CVO') produce fruit containing monoembryonic seeds. All of these, except 'CVO' are released commercial cultivars. To preclude any inadvertent inclusion of off types/nucellars in the mapping population, the hybridity of the population individuals (for 'MK' paternity) was verified through few homozygous SNPs polymorphic between the maternal parents and 'MK'. The individuals with doubtful identity were not used in this study. All the populations were fruiting in the 2017-18 season.

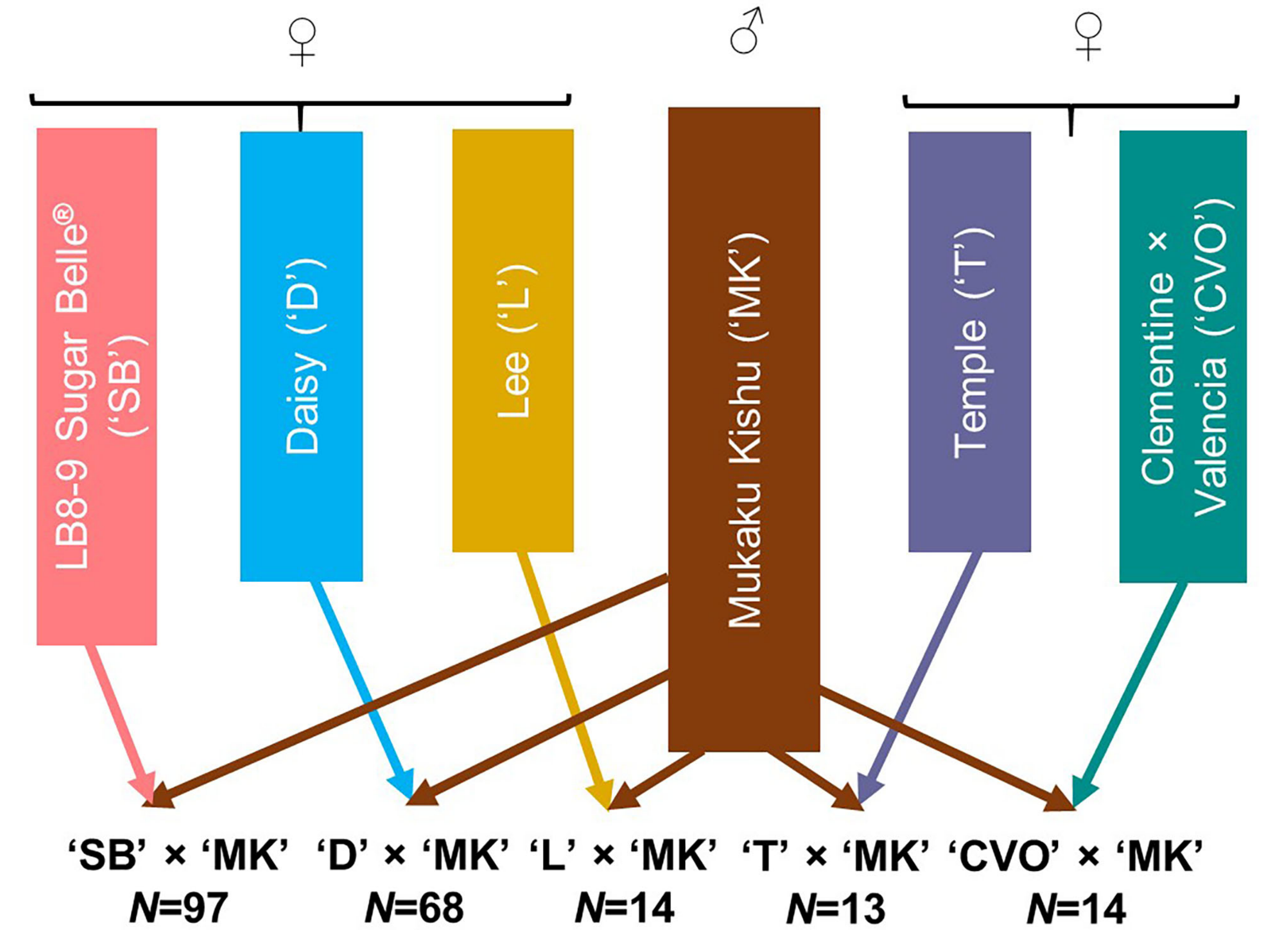


FIGURE 1
Description of the five F_1 mandarin populations used for genotyping with *Axiom_Citrus56AX array. N* represents the number of individuals in each population.

Phenotyping of progeny for seed content

The fruits of the 206 F_1 progenies were examined for presence or absence of seeds in December 2017. For each progeny, ten fruits were cut transversely into two halves, juice was squeezed, and the fruits were observed for the seeds. Progenies containing even a single seed in the fruit were scored as seedy, while those without seed as seedless.

Genotyping with *Axiom_Citrus56 Array* and data filtering

The genomic DNA of the parents and the progenies was isolated using *Qiagen DNeasy Plant Minikit* and quantified using NanoDrop^{TM}. The samples were prepared and outsourced to Thermo Fisher Scientific Inc. for genotyping with *Axiom_Citrus56 Array* encompassing 58,433 SNP probe sets. The genotyping data were filtered through Axiom Analysis Suite 5.1.1 software (Thermo Fisher Scientific Inc, 2020) by selecting *Axiom_Citrus56.r1* array and *Best Practices Workflow*. The sample QC (Dish quality control: ≥ 0.82, QC call rate: ≥ 0.97, Percent of passing samples: ≥ 95, Average call rate for passing samples: ≥ 98.5) and SNP QC were as per the default configurations of *Diploid.legacy.v5* under the *Threshold settings*. The data were processed together for the whole set of 210 samples [206 progenies and four parents namely 'SB', 'D', 'T' and 'MK')]; as well as separately for two major populations *i.e.* 'SB' × 'MK' (N=97) and 'D' × 'MK' (N=68). All three analyses were performed using *default* analysis configuration of *Axiom_Citrus56_96orMore.r1*, the only option available for *Axiom_Citrus56.r1* array in the software. The SNP probe sets were classified into six classes based on the properties of cluster formation: i. Poly High Resolution (PHR) - three highly resolved clusters (two homozygous and one heterozygous); ii. No Minor Homozygote (NMH) - two highly resolved clusters (one homozygous and one heterozygous); iii. Mono High Resolution (MHR) - only one homozygous cluster; iv. Off Target Variant (OTV) - three well resolved clusters with an additional off target cluster; v. Call Rate Below Threshold (CRBT) - Call rate was below threshold (0.97) but other cluster properties were above the threshold; and vi. Others- SNPs not grouped in any of the previous categories. The analyzed results were exported as text file.

Genotype coding and construction of parental linkage maps

The progenies of 'SB' × 'MK' and 'D' × 'MK' populations were used for linkage map construction using JoinMap version 4.1 (Van

Ooijen, 2006; Van Ooijen, 2011). For each population, the SNPs exhibiting 1:1 (heterozygous × homozygous for female parent, homozygous × heterozygous for male parent) and 1:2:1 segregation pattern (heterozygous for both parents) were used for genetic mapping. Depending upon the parental segregation direction, the original calls of the SNPs were substituted with the codes of *'nn'* and *'np'* (for SNPs segregating from male parent) and *'lm'* and *'ll'* (for SNPs segregating from female parent). The co-segregating (1:2:1) SNPs were coded as *'hh'*, *'hk'* or *'kk'*, depending upon their genotype. The linkage analyses and maps were constructed using cross pollination (CP) model in JoinMap 4.1 following two way pseudo test cross approach that allows generation of separate maps for male and female parents (Grattapaglia and Sederoff, 1994). Before linkage analysis, the following classes of SNPs were sequentially removed: SNPs with missing data for > 10% of the progenies; exhibiting 100% similarity in segregation pattern with another locus in the dataset; or having segregation pattern significantly skewed from 1:1 or 1:2:1 Mendelian ratio ($P < 0.005$). Further, the loci showing 99% similarity were also eliminated. In each population, initially, the *'nn × np'* and *'lm × ll'* datasets were used for developing male and female specific linkage maps. The selected '*hk × hk*' markers were then combined with male and female parent specific datasets to reconstruct male and female linkage maps. The linkage grouping was performed with grouping static independence LOD that permits use of markers with distorted segregation without inducing spurious linkage (Bernet et al., 2010; Ollitrault et al., 2012; Huang et al., 2018). The linkage groups were obtained at independence LOD threshold of 6.0 and recombination fraction of 0.4. The linkage groups were numbered according to the chromosomal ID of the SNPs in *Axiom_Citrus56 Array*. The SNP markers were ordered using regression mapping algorithm while map distance (cM) was calculated using Kosambi mapping function. All the linkage maps were drawn with MapChart 2.32 (Voorrips, 2002).

Construction of integrated 'MK' map and consensus linkage map

Since both the mapping populations were constructed using the common pollen parent 'MK', it enabled the development of an integrated 'MK' linkage map. The homologous linkage groups from individual 'MK' maps- 'MK_SB' (derived from 'SB' × 'MK' population) and 'MK_D' (derived from 'D' × 'MK' population) were selected and grouped using *'Combine Groups for Map Integration'* function in the JoinMap 4.1. The consensus linkage map was prepared in two steps. In the first step, the homologous male and female parent specific linkage groups were selected based on the sharing '*hk*' markers in each population. The homologues were combined to make sub-composite linkage maps using the function *'Combine Groups for Map Integration'* in the JoinMap. In the 2nd step, the homologous sub-composite linkage groups from two populations were combined to generate consensus linkage map using online Merge Map software (Wu et al., 2008; Peng et al., 2016) (http://www.mergemap.org/).

Evaluation of the selected SNPs to identify the seedless progenies

From the mapping analysis, seedless locus linked SNPs were identified. These SNPs were assessed for predicting the seedless progenies in different populations. For this purpose, the actual allelic calls of these SNPs were observed in the progenies and the allelic pattern associated with seedlessness was identified. The marker ability to identify the true seedless progenies was determined from its positive prediction value (PPV) for seedless progeny detection.

$$\text{PPV for seedless progeny detection } (\%) = \frac{\text{Actual seedless progenies among the predicted progenies}}{\text{Total seedless progenies predicted by the SNP}} \times 100$$

Prediction of candidate genes for seedlessness

The SNP markers flanking the seedless locus were identified on the individual 'MK' maps- 'MK_SB' and 'MK_D'. These SNPs were aligned to the annotated Clementine reference genome and the physical interval for the seedless locus was delineated. The total number of genes in these intervals was identified and their function was explored in the model plant *Arabidopsis* and other crops.

Results

Fruit phenotyping for seedlessness

For seeds in fruits, the progeny individuals could be classified into two categories: seedy (fully formed seeds), seedless (no seeds) (Figure 2). In the two major populations, the ratio of seeded to seedless was 55:42 (χ^2 = 1.74) in 'SB' × 'MK' population and 33:35 (χ^2 = 0.06) in 'D' × 'MK' populations, which fit the test cross distribution. These observations showed agreement with the hypothesis of Nesumi et al. (2001) that the seedlessness in crosses of seedy mandarins × 'MK' is governed by a major locus (*Fs*).

SNP genotyping and identification of polymorphic SNPs

Analysis of genotyping data of four parents ('SB', 'D', 'T' and 'MK') and 206 progeny individuals with Axiom Analysis Suite software revealed the presence of six types of SNP markers (described under Material and Methods), in variable proportions i.e. MHR: 52.13%, PHR: 22.49%, NMH: 17.71%, CRBT: 1.13%, OTV: 0.40% and Others: 6.14% (Table 1). The cumulative share of the polymorphic SNP markers (PHR and NMH) was 40.2% over the whole set of samples. The percentage of polymorphic SNP markers varied from 30.77% in 'D' × 'MK' population (10,352 NMH, and

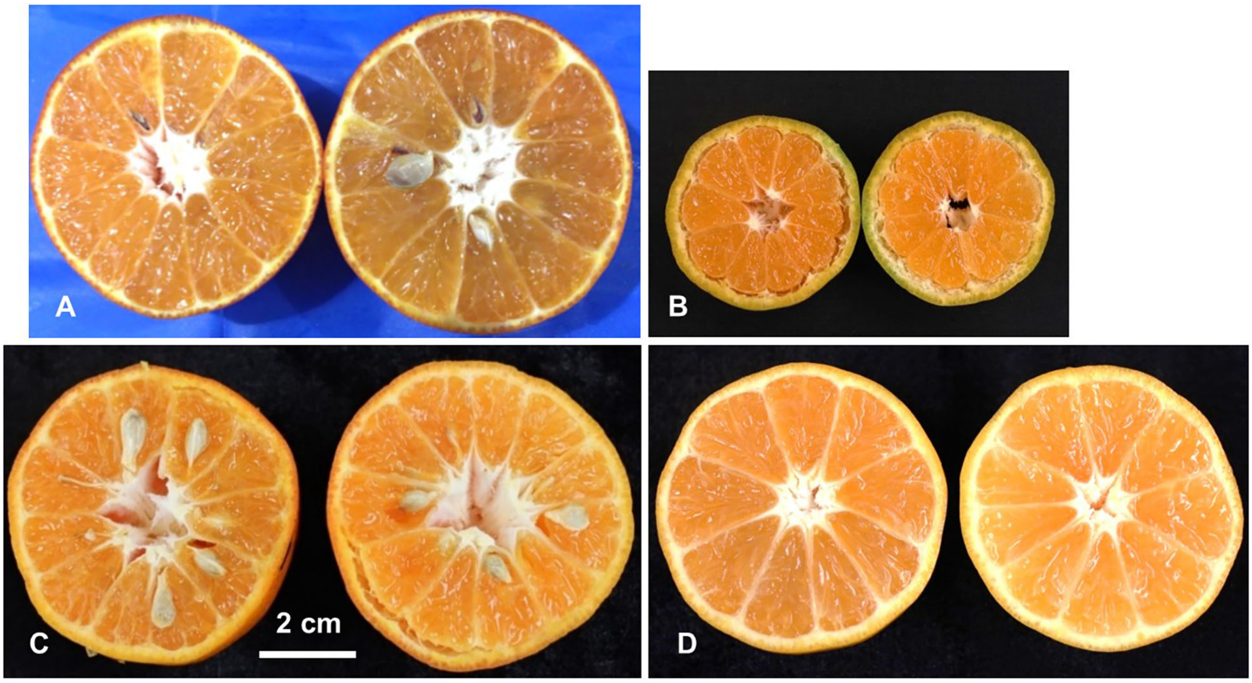


FIGURE 2
Phenotyping of the parents and population for presence/absence of seeds. Female parent LB8-9 Sugar Belle® ('SB') **(A)**, male parent Mukaku Kishu ('MK') **(B)**, and their derived F_1 progenies with seedy **(C)**, and seedless **(D)** phenotypes. Scale bar denotes 2 cm size.

7,623 PHR) to 40.1% in 'SB' × 'MK' population (12,669 NMH, and 10,761 PHR).

Processing of polymorphic SNPs

As the seedless locus (hereafter mentioned as *Fs*-locus) behaved as a test cross marker and segregated from 'MK', the seeded and seedless progenies were genotyped as *'nn'* and *'np'*, respectively. The *Fs*-locus was included in the male parent dataset (*'nn × np'*) of the two mapping populations for establishment of linkage groups and map generation. From the polymorphic SNP loci, the loci producing calls in both the parents of a population were selected. As a result, a total of 10,037 SNPs (5863 *'lm × ll'*, 2320 *'nn × np'* and 1854 *'hk × hk'*) for 'SB' × 'MK' and 11,744 SNPs (6241 *'lm × ll'*, 2964 *'nn × np'* and 2539 *'hk × hk'*) for 'D' × 'MK' populations were obtained (Table 2). After removal of SNPs with > 10% missing genotypes, across loci similarity of ≥ 0.99 and showing segregation distortion from 1:1 or 1:2:1 (P< 0.005), a total of 1521 parent specific SNP markers (817 *'lm × ll'*, 704 *'nn × np'*) in 'SB' × 'MK' and 1262 markers (664 *'lm × ll'*, 598 *'nn × np'*) in 'D' × 'MK' populations were left for mapping (Table 2). The proportion of loci showing ≥ 99% similarity was 76.9% in 'D' × 'MK' population and 57.1% in 'SB' × 'MK' population. The number of markers showing segregation distortion from 1:1 or 1:2:1 Mendelian ratio (P< 0.005) was higher in 'SB' × 'MK' population than that of 'D' × 'MK' population (Table 2). The processed parent specific SNP markers were combined with 114 *'hk × hk'* type markers (heterozygous in both parents) in 'SB' × 'MK' and with 112 *'hk ×*

TABLE 1 Genotyping results of *Axiom Citrus 56AX array* over all 210 samples and two major mapping populations.

Marker type	Genotyped samples					
	Total 210 samples		'SB' × 'MK' population		'D' × 'MK' population	
	Counts	Percentage	Counts	Percentage	Counts	Percentage
Poly High Resolution (PHR)	13138	22.49	10761	18.42	7623	13.05
No Minor Homozygote (NMH)	10350	17.71	12669	21.68	10352	17.72
Mono High Resolution (MHR)	30463	52.13	31682	54.22	36825	63.02
Call Rate Below Threshold (CRBT)	660	1.13	838	1.43	1164	1.99
Off Target Variant (OTV)	235	0.40	149	0.25	185	0.31
Others	3587	6.14	2334	3.99	2284	3.91
Total	58433		58433		58433	

TABLE 2 Processing of markers for constructing linkage maps in 'SB' × 'MK' and 'D' × 'MK' populations.

Particulars	Segregation pattern of markers							
	'SB' × 'MK' population				'D' × 'MK' population			
	lm × ll	*nn × np*[#]	*hk × hk*	Total	*lm × ll*	*nn × np*	*hk × hk*	Total
Total loci producing calls in two parents	5863	2320	1854	10037	6241	2964	2539	11744
Markers with > 10% missing genotypes	1189	424	–	1613	292	146	–	438
Markers with ≥ 99% similarity	3744	1130	858	5732	5279	2153	1606	9038
Markers with significantly skewed segregation (*P< 0.005*)	113	62	71	246	6	67	29	102
Total processed markers	817	704	114[$]	1635	664	598	112[$]	1374
After combining heterozygous markers	931	818	–		776	710	–	

[#] The nn × np dataset includes phenotypic Fs-locus for both the populations.

[$] The selection procedure of hk markers is described in Supplementary Information 1.

hk' type markers in 'D' × 'MK' populations to prepare male and female specific linkage maps. Of the '*hk'* markers, 91 were common in both populations (Supplementary Table 1). The step-by-step procedure used in selection of the '*hk*' markers is described in Supplementary Information 1.

Development of population specific parental linkage maps and integrated Mukaku Kishu map

In 'SB' × 'MK' population, a total of 931 (817 '*lm × ll*' and 114 '*hk × hk'* markers) and 818 SNP markers (704 '*nn × np'* and 114 '*hk × hk'* markers) were used for constructing female ('SB') and male parent ('MK_SB') specific linkage maps (Table 2). From the total used markers, 930 for 'SB' and 810 for 'MK_SB' grouped on to nine major linkage groups (LGs) at independence LOD score of 6.0. The remaining markers formed small sub-groups and were not considered for grouping. Except two LGs in 'SB' map (LGs 1 and 8) and three in 'MK_SB' map (LGs 5, 6 and 8), all other LGs conserved their integrity up to LOD score of 10.0. The 'SB' map had 922 distinct locations and spanned a total map length of 1012.87 cM with an average marker interval of 1.09 cM. The 'MK_SB' map of this population had 807 unique locations and covered a genetic map length of 1165.72 cM (Table 3). The distribution of the skewed markers ($P = 0.005$) was not uniform across the LGs in two maps. In 'SB' map, LG2 had maximum of 96 distorted markers while LG9 contained 16 distorted markers. Similarly, in 'MK_SB' map, LG2 had maximum of 38 distorted markers and the number of distorted markers in other groups ranged from 1 to 31 (Supplementary Table 2).

In 'D' × 'MK' population, 776 (664 *'lm × ll'* and 112 *'hk × hk'* markers) and 710 SNP markers (598 *'lm × ll'* and 112 *'hk × hk'* markers) were used for constructing female ('D') and male ('MK_D') parent specific maps, respectively. In 'D' map, the 776 SNPs grouped into nine LGs at independent LOD threshold of 6.0. A further increase of LOD score caused splitting of the majority of LGs into sub-groups. The map had a total genetic length of 879.37 cM with an average marker interval of 1.13 cM (Table 3). For 'MK_D' map, 707 of 710 markers, grouped into nine major linkage groups and a minor linkage group. The minor LG contained 9 markers, of which 4 were the part of LG3 in 'MK_SB' map. Hence, the minor LG was presumed to be a part of the LG3 in this map. Like 'D' map, most of the LGs tended to lose their integrity in 'MK_D' map at LOD score > 6.0. The map consisted of 707 SNPs with 674 unique mapping points and covered a genetic length of 1018.57 cM. Like 'SB' × 'MK' population, the presence of skewed markers was also variable across LGs on these two maps. In 'D' map, LG8 had 30 distorted markers while LG5 and LG9 contained 22 and 26 such markers (Supplementary Table 2). In 'MK_D' map, except LG1 and LG2, all other LGs had distorted markers, and their number varied from 7 (LG6) to 44 (LG7) (Supplementary Table 2).

In both populations, the female maps contained more SNPs, but their overall map length was shorter than the corresponding male parent map. The linkage maps of 'SB' and 'D' had 120 and 69 more markers than the corresponding 'MK' maps but were 152.85 and

TABLE 3 Summary of four parental maps and the integrated Mukaku Kishu ('MK') map.

Population	Parental Map	Total Mapped SNPs	Unique positions	Total syntenic SNPs	Number of distorted SNPs	Map Length (cM)	Average marker interval (cM)	Gaps (> 5 cM)
'SB' × 'MK'	'SB'	930 (9)[$]	922	916	129	1012.87	1.09	27
	'MK_SB'	810 (9)	807	790	184	1165.72	1.44	27
'D' × 'MK'	'D'	776 (9)	734	754	125	879.37	1.13	23
	'MK _D'	707 (9 + 1)	674	685	124	1018.57	1.44	18
Integrated	'MK'	1233 (9)	1222	1201	299	1140.16	0.92	06

[$] The values in parenthesis denote to the linkage groups for the mapped SNPs on each map.

139.20 cM shorter in length, respectively. Except two LGs in 'SB' × 'MK' population (LGs 1 and 5) and two in 'D' × 'MK' population (LGs 4 and 5), all other LGs recorded higher ratio of male to female map length (Supplementary Table 2). Among the four parental maps, 18 ('MK_D') to 27 ('SB' and 'MK_SB') gaps of > 5 cM length were noted (Table 3).

The integrated 'MK' map was based on 1233 SNPs of which 284 (193 '*nn* × *np*', 91 '*hk* ×*hk*') were common between 'MK_SB' and 'MK_D' maps. The nine markers of subgroup 2 of LG 3 of 'MK_D' map easily integrated with homologous 'MK_SB' group to form integrated LG3. Due to increase in the number of markers, the average marker interval in the integrated 'MK' map was reduced to 0.92 from 1.44 in the two population specific 'MK' maps (Table 3). The integrated map had only six gaps of > 5 cM length (Table 3). The information of the detailed maps is provided in Supplementary Table 2.

Development of sub-composite and consensus linkage maps

The use of heterozygous ('*hk* × *hk*') markers allowed building of sub-composite linkage maps between male and female parent specific maps in the two populations. In 'SB' × 'MK' population, the sub-composite linkage map ('SB'-'MK_SB') was based on 1626 SNPs (816 '*lm* × *ll*', 696 '*nn* × *np*' and 114 '*hk* × *hk*' markers) and spanned over a map length of 1105.14 cM with 1620 unique marker positions (Table 4). The average marker interval in the sub-composite 'SB'-'MK_SB' map was 0.68 while for the individual LGs, it ranged from 0.59 (LG6) to 0.83 cM (LG1) (Table 4). Five gaps of > 5 cM were noted on this sub-composite map (Table 4) while one of them present on LG1 was even longer than 10 cM (data not shown).

In 'D' × 'MK' population, the sub-composite map ('D'-'MK_D') contained 1363 SNPs (664 '*lm* × *ll*', 587 '*nn* × *np*' and 112 '*hk* × *hk*' markers). The nine markers of subgroup 2 of LG3 of 'MK_D' map (8 '*nn* × *np*' type and 1 '*hk* × *hk*' type) did not converge with homologous linkage group of female parent and were thus, not part of the sub-composite map. The mapped SNPs were present over 1303 distinct locations with a total genetic length of 938.52 cM. The average marker interval was 0.69 cM and for individual groups, it ranged from 0.49 (LG9) to 0.87 cM (LG1). Except two gaps of > 5 cM on LG5, the markers on all other LGs were evenly distributed (Table 4).

The consensus linkage map was composed of 2588 markers (2587 SNPs and a phenotypic *Fs*-locus) of which 401 markers were shared between the two sub-composite maps. Of the shared markers, 211 between the maps of 'SB' and 'D' (120 '*lm* × *ll*' and 91 '*hk* × *hk*' markers), and 280 between the 'MK_SB' and 'MK_D' (190 '*nn* × *np*' and 90 '*hk* × *hk*' markers) were common (Supplementary Table 1). The common markers between 'SB' and 'D' have also been shown on the skeleton consensus map (Figure 3, Supplementary Figure 1). The consensus map had 2495 distinct positions and spanned a map length of 1406.84 cM. The marker density in the consensus linkage map was very high as the average marker to marker distance was 0.54 cM. The map length ranged from 123.84 cM for LG1 to 226.59 cM for LG3 (Table 4, Figure 3). The LG3 was also the largest linkage group in the two sub-composite maps.

In the consensus map, there were four gaps of > 5 cM (three on LG5 and one on LG8) (Table 4). One of the three gaps on LG5 was of more than 10 cM length (Figure 2, Supplementary Figure 1).

Based on the common SNP markers, the marker collinearity was also examined between the two sub-composite maps and consensus map. Except for very few markers, the order of most of the markers was consistent across the three maps (Supplementary Figure 2).

TABLE 4 Attributes of two sub-composite and consensus mandarin maps.

Linkage group	Sub-composite maps										Consensus linkage map				
	'SB'-'MK_SB' map					'D'-'MK-D' map									
	Total markers	Unique positions	Map length (cM)	Average marker interval (cM)	Gaps (> 5 cM)	Total markers	Unique positions	Map length (cM)	Average marker interval (cM)	Gaps (> 5 cM)	Total markers[#]	Unique positions	Map length (cM)	Average marker interval (cM)	Gaps (> 5 cM)
1	117	117	97.11	0.83	1	104	103	90.77	0.87	–	169	165	123.84	0.73	–
2	184	183	137.30	0.75	1	213	206	137.41	0.65	–	353	343	175.71	0.50	–
3	270	269	160.85	0.60	–	205	193	150.48	0.73	–	408	396	226.59	0.56	–
4	205	205	132.72	0.65	–	119	116	76.39	0.64	–	276	269	139.09	0.50	–
5	165	165	112.87	0.68	1	138	136	117.51	0.85	2	273	269	159.73	0.59	3
6	164	164	96.27	0.59	–	142	139	94.60	0.67	–	272	264	133.15	0.49	–
7	168	166	128.01	0.76	1	159	146	111.80	0.70	–	288	272	157.26	0.55	–
8	146	144	106.69	0.73	1	109	98	74.89	0.69	–	224	206	145.18	0.65	1
9	207	207	133.31	0.64	–	174	166	84.67	0.49	–	325	311	146.29	0.45	–
Total	1626	1620	1105.14	0.68	5	1363	1303	938.52	0.69	2	2588	2495	1406.84	0.54	4

[#] In consensus map, 401 of the total 2588 markers were common to both subs-composite maps. Details are given in Supplementary Table 1.

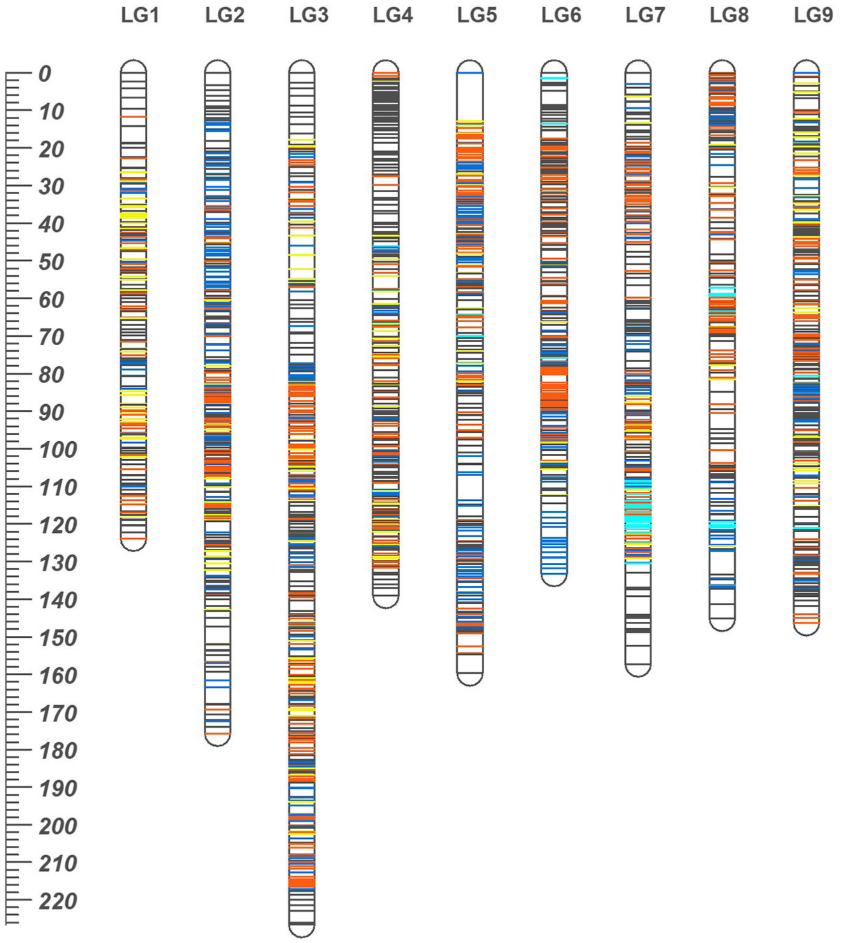


FIGURE 3
The high density consensus linkage map of mandarin. Nine linkage groups are numbered according to nine scaffolds of Clementine genome. On the map, *orange*, *blue* and *dark grey* lines represent marker sites specific to 'SB', Daisy ('D') and 'MK', respectively. The *yellow* lines represent the markers common to both 'SB' and 'D' while *cyan* lines are for markers translocated from other chromosomes with respect to their actual location on Clementine reference genome. The ruler at left indicates the map distances in cM.

Marker synteny

The grouping of the mapped SNPs showed high degree of synteny with their chromosomal location on the Clementine reference genome. Based on the number of syntenic markers, the parental maps of 'SB', 'D', 'MK_SB' and 'MK_D' shared 98.5, 97.2, 97.5 and 96.9% synteny with the Clementine reference genome. However, four LGs in both 'D' and 'SB' maps (LG6, LG7, LG8 and LG9), five LGs in 'MK_D' map (LG4, LG6, LG7, LG8 and LG9) and six LGs in 'MK_SB' map (LG4, LG5, LG6, LG7, LG8 and LG9) contained the markers translocated from other chromosomes relative to *C. clementina* reference genome (Figure 4). The LG7 showed maximum number of inter-chromosomal translocations in all of the parental maps and the consensus map (Figures 3, 4).

Mapping of seedlessness

The *Fs*-locus was mapped on the LG5 in both individual 'MK' maps ('MK_SB' and 'MK_D') (Figure 5). It was linked to the SNP marker AX-160417325 at 7.4 cM on 'MK_SB' map. In 'MK_D' map, it localized between two SNP markers AX-160536283 and AX-160906995 at a distance of 2.4 and 4.9 cM, respectively. On the integrated 'MK' map of LG5, the three SNPs AX-160417325, AX-160536283, AX-160906995 maintained their proximity to the *Fs*-locus and were at a relative distance of 4.0, 5.5 and 3.4 cM distance (Figure 5). The proximity of *Fs*-locus with two of the three SNP markers was also supported by the high LOD value and relative low recombination frequency values during the grouping of different markers in the male parent maps (Table 5). The SNP marker AX-160536283 showed strong linkage with *Fs*-locus both in 'MK_SB' (LOD- 16.58, RF-0.09) and 'MK_D' (LOD- 14.02, RF-0.06) maps. On the other hand, the SNP marker AX-160417325 showed strong linkage with *Fs*-locus (LOD-17.4, RF-0.08) only in 'MK_SB' map (Table 5). Notably, AX-160417325 was used for mapping only in 'SB' × 'MK' population as it was an intercross marker for 'D' × 'MK' population with its allelic composition of T/T, T/C and T/C for 'SB', 'D' and 'MK', respectively.

On the consensus map, the closely linked SNP markers AX-160536283 and AX-160906995 were located at 2.8 cM and 9.6 cM from *Fs*-locus while the marker AX-160417325 mapped at a remote distance. The physical location of SNP markers AX-160417325 and

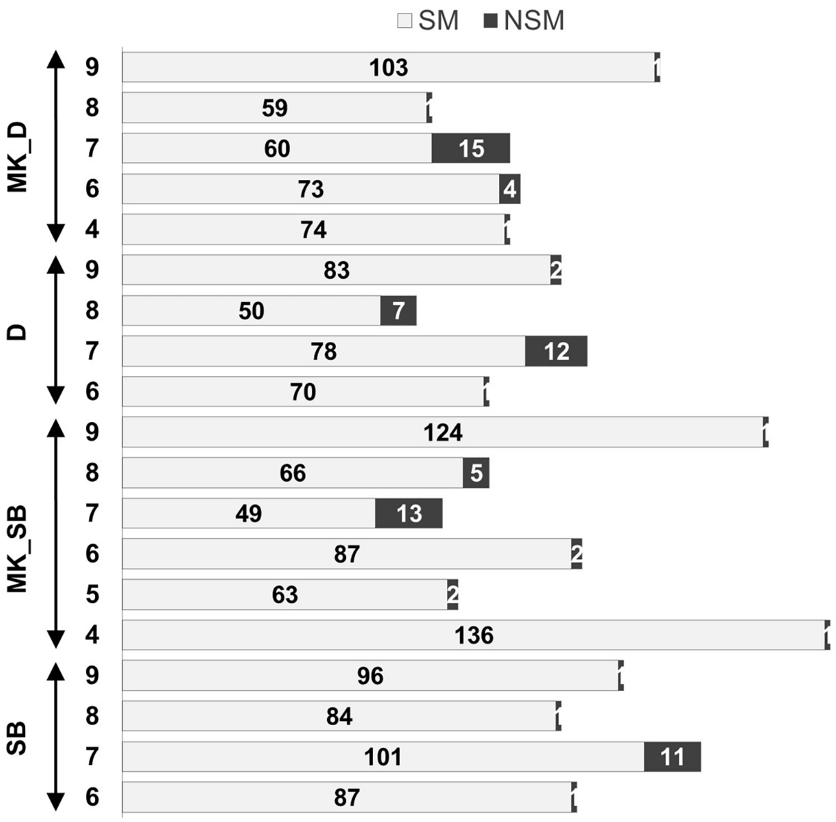


FIGURE 4
Non-syntenic linkage groups in the parental maps and proportion of syntenic (SM) and non-syntenic markers (NSM). 'SB', 'MK_SB' (Mukaku Kishu map derived from 'SB' × 'MK' population), 'D' and 'MK_D' (Mukaku Kishu map derived from 'D' × 'MK' population).

AX-160536283 was explored on the Clementine annotated genome. The SNP markers AX-160417325 and AX-160536283 were located at 12.2 Mb and 10.0 Mb on the physical map of chromosome 5 (Supplementary Information 2). Thus, the physical map supported that the *Fs*-locus was more proximate to SNP marker AX160536283 than to AX-160417325.

The association of the *Fs*-locus with SNP markers was also independently examined for grouping statistics in two other F_1 populations: 'L' × 'MK' (14 individuals) and 'CVO' × 'MK' (14 individuals). Here, none of the three closely linked markers showed proximity with the *Fs*-locus. The most closely associated SNP marker to the *Fs*-locus in these two populations was AX-159840260 (Table 5). This marker was at 29.9 and 20.4 cM in 'MK_SB' and 'MK_D' maps, respectively (Figure 5). Even the marker AX-160906995 located at 4.9 cM from *Fs*-locus on 'MK_D' map lacked direct proximity with it (Table 5).

Evaluation of the selected SNPs to identify the seedless progenies

The two *Fs*-locus associated SNPs, AX-160417325 and AX-160536283, were assessed for predicting the seedless progenies in different populations (Supplementary Table 3). AX-160417325 was found to be the test cross marker for all the populations, except for 'D' × 'MK', where it behaved as an intercross marker. Its allelic pattern was T:T in maternal parents and T:C in the male parent 'MK'. A selection based on the T:C allelic pattern, predicted 46 of the total 97 'SB' × 'MK' cross progenies as seedless. Among the predicted 46, 40 were actual seedless progenies (true positives) and six were the false positives (seedy recombinants). Thus, the marker showed 87% PPV for seedless progeny detection in this population. The remaining two seedless progenies of this population exhibited the alternate allelic pattern and hence, were considered as false negatives. In the progenies of two minor populations ('CVO' × 'MK' and 'T' × 'MK'), apart from the T:C allelic pattern, an improbable new C:C pattern was also found associated with seedlessness. Since the new variant carried allele 'C' from 'MK', we considered it linked with seedlessness. Based on T:C or C:C allelic patterns, all the seedless progenies could be identified in the three minor populations (Supplementary Table 3). However, abundant seedy individuals (false positives) also shared the seedless associated allelic patterns in these populations. Due to this reason, the AX-160417325 PPV for seedless progeny detection ranged from 25 to 50% in these populations. The AX-160536283 showed association with *Fs*-locus only in 'SB' × 'MK' and 'D' × 'MK' cross populations. It displayed 85.1 to 91.9% PPV for seedless progeny detection in these populations. The maternal ('SB' and 'D') and paternal ('MK') parents had the allelic patterns- A:A and A:G, respectively. Instead of 'MK' A:G allelic pattern, the majority of the seedless progenies had maternal A:A allelic pattern, indicating of its repulsive linkage with *Fs*-locus.

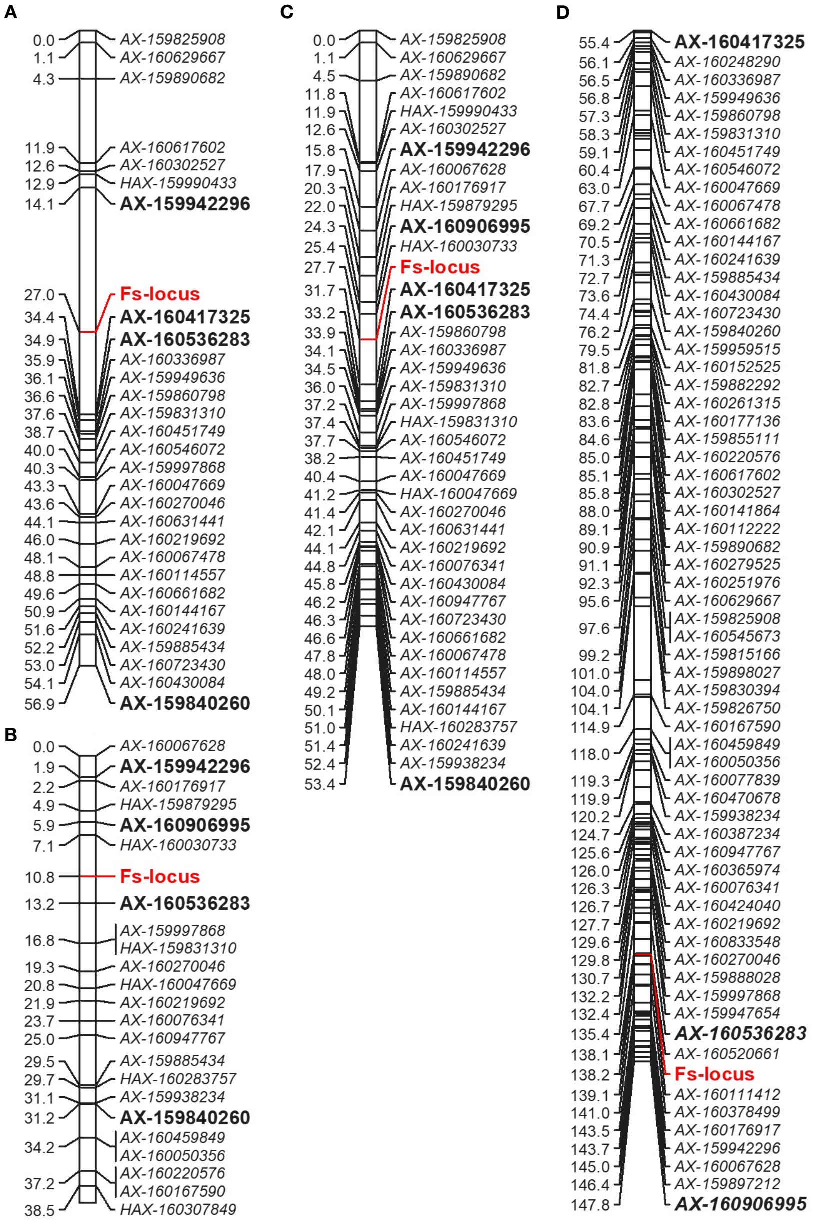


FIGURE 5
Comparative partial linkage map of LG5 of Mukaku Kishu derived from the 'SB' × 'MK' population ('MK_SB') **(A)**, 'D' × 'MK' population ('MK_D') **(B)**, integrated Mukaku Kishu from the two populations ('MK') **(C)**, and of consensus linkage map **(D)**. The markers showing linkage in the individual maps are bold and of large size. For easy comparison, the markers of female parents ('SB' and 'D') are not shown on consensus linkage map.

The presence of some seedy individuals (false positives) among the sorted progenies in every population indicated that both markers are not co-segregating and that recombination is occurring in their vicinity (Supplementary Table 3). It is worth mentioning that there were gaps of 12.9 and 7.4 cM in the vicinity of the *Fs*-locus on 'MK_SB' map and of 2.4 and 4.9 cM on 'MK_D' map (Figure 5). Thus, to find the markers co-segregating with the *Fs*-locus, there is a need to bridge these gaps with additional polymorphic markers. However, the use of *Fs*-locus associated SNPs can reduce the effective population size preselected for seedlessness to be finally evaluated for all other fruit quality characteristics. The physical location of these SNPs on the Clementine genome and their flanking sequences are provided in Supplementary Table 4.

Prediction of candidate genes for seedlessness

Based on the alignment of flanking SNP markers to the annotated Clementine genome, the *Fs*-locus on 'MK_SB' and 'MK_D' maps was delineated to a physical interval of 2.2 to 12.2 Mb and 3.97 to 10.0 Mb, respectively (Figure 5 and Supplementary Information 2). These physical intervals house a total of 171 and 131 genes, respectively. The region of 3.97 to 10.0 Mb was common between the two maps which encompasses 131 genes. These genes are mainly related to floral and reproductive development, seed development, seed dormancy, seed germination, and are also involved in ABA signaling, and biotic and abiotic stress tolerance. In 'MK', the seedlessness is reportedly due

TABLE 5 Proximity of SNP markers to *Fs*-locus in MK maps of different populations.

SNP markers	Attributes of linkage in different maps							
	'MK_SB'		'MK_D'		'MK_L'		'MK_CVO'	
	Recombination frequency	Linkage LOD	Recombination frequency	Linkage LOD	Recombination frequency	Linkage LOD	Recombination frequency	Linkage LOD
AX-160417325	0.0825	17.42	–	–	–	–		
AX-160536283	0.0928	16.58	0.0588	14.02	–	–		
AX-160906995	–	–	–	–	–	–		
AX-159840260	–	–	–	–	0.0000	3.96	0.0000	3.96
AX-159855111					0.0714	2.79	0.0714	2.79

to the developmental arrest of the embryos at an early stage. The functional exploration of the 131 genes across Arabidopsis and other crops helped in identification of 13 candidate genes which are reportedly expressed in developing embryos or seed coats (Table 6). These genes belong to seven gene families. One of these genes should be involved in imparting seedlessness in 'MK' and its derived seedless progenies.

Discussion

High density linkage maps are essential for genetic analysis and mapping of economic traits. Apart from the factors of mapping population and their size (Ferreira et al., 2006; Li et al., 2010), use of high throughput whole genome genotyping methodologies is essential to develop high density linkage maps. High density citrus linkage maps have been generated for pummelo (Guo et al., 2015), sweet orange (Huang et al., 2018), *C. sunki* (Curtolo et al., 2018), and trifoliate orange (Curtolo et al., 2018; Huang et al., 2018) through high throughput whole genome genotyping methodologies using DArTseq markers (Curtolo et al., 2018) and SNP markers (Guo et al., 2015; Huang et al., 2018). The above methodologies were based on the simultaneous discovery and genotyping of markers over a population. These methodologies sometimes miss substantial genotypic information because of low sequencing depth (Gouesnard et al., 2017). The availability of whole genome sequences for a vast number of citrus accessions (Wu et al., 2014; Wu et al., 2018) has enabled discovery of abundant SNPs (Yu et al., 2016). The use of SNP probe sets as a fixed array may allow consistent genotyping for these SNPs across individuals and populations (Pandey et al., 2017), and minimize the possibility of missing information, and can lead to higher density genetic linkage maps. In this study, we evaluated the polymorphism content of the *Axiom_Citrus56 Array* (Hiraoka, 2020) over 210 individuals [206 F_1 progenies derived from five different mandarin populations and four parent samples – LB8-9 Sugar Belle® ('SB'), Daisy ('D'), Temple ('T') and Mukaku Kishu ('MK') (Table 2)]. The genotypic information of progenies of two populations: 'SB' × 'MK' and 'D' × 'MK', was used for constructing high density linkage maps for three mandarins- 'SB', 'D', and 'MK' and an integrated 'MK' map. The maps were built using SNP markers showing segregation

TABLE 6 Prediction of putative candidate genes for seedlessness in 'MK'.

Sr. No.	Gene ID	Position (bp)	Predicted function
1	Ciclev10003724m.g	4985850[#]-	UDP-Glycosyltransferase superfamily protein
2	Ciclev10003145m.g	4990017-	Wall associated kinase-like 1
3	Ciclev10000464m.g	5001178-5004379	Wall associated kinase-like 1
4	Ciclev10003390m.g	5037558-5047193	Wall associated kinase-like 1
5	Ciclev10003989m.g	5816248-5818999	Wall associated kinase-like 1
6	Ciclev10001614m.g	5921681-	Cytochrome P450, family 82, subfamily C, polypeptide 4
7	Ciclev10003405m.g	5938251-5940692	Cytochrome P450, family 82, subfamily C, polypeptide 3
8	Ciclev10003581m.g	6455788-6455828	ATP phosphoribosyl transferase 2
9	Ciclev10004097m.g	6458000-6461223	ATP phosphoribosyl transferase 2
10	Ciclev10003937m.g	6672504-	Homeobox protein 33
11	Ciclev10000860m.g	7064394-7068371	Cytochrome P450, family 706, subfamily A, polypeptide 6
12	Ciclev10000069m.g	8533088-8537312	T-complex protein 11
13	Ciclev10001947m.g	9480020-	6-phosphogluconate dehydrogenase family protein

[#] The genes mentioned without end positions had undefined length in the annotated reference Clementine genome.

pattern of test cross (‘*lm* × *ll*’ and ‘*nn* × *np*’ markers). The selected SNP markers segregating in 1:2:1 ratio (‘*hk* × *hk*’ markers) (114 for ‘SB’ × ‘MK’ population and 112 for ‘D’ × ‘MK’ population) were also used in the dataset that allowed integration of male and female parent specific maps in the two maps. We also chose to use markers deviating from Mendelian pattern at significance level of P =0.005 with the assumption that these do not induce spurious linkage when coupled with independence LOD test statistic. The population specific sub-composite maps were further merged to construct a consensus mandarin map. The individual (‘MK_SB’, ‘MK_D’), integrated ‘MK’ map were used for deciphering the location of genes

Axiom_Citrus56 Array and its polymorphism evaluation

Previously, a comparatively low density, 1536 SNP derived Illumina Golden Gate assay has been used in citrus for mapping fruit related traits of mandarins (Yu et al., 2016). The present array was the assembly of 58,433 SNP probe sets representing the SNP polymorphism of 41 citrus accessions, representing cultivated (different varieties of mandarin, sweet orange, grapefruit, pummelo, lime and lemons) and wild citrus (Hiraoka, 2020). The 40.2% of the SNPs (PHR and NMH) showed polymorphism for the set of 210 samples. Between the two mapping populations used in this study, percentage of polymorphic SNPs was substantially higher in ‘SB’ × ‘MK’ population (40.1%) relative to ‘D’ × ‘MK’ population (30.77%). Both ‘SB’ and ‘D’ are complex hybrids. The variability in polymorphism index may be due to different levels of heterozygosity in their genomes, resulting from variable degrees of admixture of the two progenitor species, *C. reticulata* Blanco and *C. maxima* [(Burm.) Merr] (Wu et al., 2014; Wu et al., 2018).

The polymorphism index of this array seemed to be lower than the previously reported high density SNP arrays in other fruit crops like pear (> 93% PHR SNPs) (Montanari et al., 2019) and apple (74% PHR and 2% NMH SNPs) (Bianco et al., 2016). The SNP frequencies tend to vary among crops, mainly influenced by their domestication history and reproductive habits. The original diversity of the samples under assessment, and the type (coding or non-coding) of the analyzed regions also influence the frequency of SNPs (Leonforte et al., 2013).

Evaluation of map quality

The quality of the parental maps obtained in our study matched with other high quality linkage maps published for mandarins (Gulsen et al., 2010; Ollitrault et al., 2012). Like these maps, the markers were arranged on nine LGs in all the maps of our study (except for ‘MK_D’ map), corresponding to the haploid set of chromosomes of citrus. At LOD score of > 6.0, the affinity of the markers to remain grouped in different LGs was stronger in ‘SB’ × ‘MK’ population compared to ‘D’ × ‘MK’ population. This may be due to the more progeny individuals in ‘SB’ × ‘MK’ population, as grouping properties of markers are reported to improve significantly with increase of population size (Ferreira et al., 2006).

The average marker spacing on individual ‘SB’ (1.09 cM), ‘D’ (1.13 cM) and integrated ‘MK’ maps (0.92 cM) were comparable to the reference Clementine map reported by Ollitrault et al. (2012). The reference Clementine map was based on 961 co-dominant markers (677 SNPs, 258 SSRs and 26 indels) with total genetic length of 1084.1 cM and an average marker spacing of 1.13 cM. The markers number, map length and marker density was substantially higher over all other previously published maps for mandarins (Omura et al., 2000; Sankar and Moore, 2001; Oliveira et al., 2007; Gulsen et al., 2010; Curtolo et al., 2017).

Most of the mapped SNPs in the four parental maps showed a high degree of chromosomal synteny to the Clementine reference genome. The LG7 in all the four maps possessed the maximum number of translocations. This observation is consistent with the earlier findings of Huang et al. (2018) for molecular maps of trifoliate orange and sweet orange. The male ‘MK’ maps in both populations (‘MK_SB’ and ‘MK_D’) though, contained fewer segregating markers than the corresponding female maps, but had 15-16% greater map length. The variability of genetic distances among sex specific maps has also earlier been reported in citrus (Ollitrault et al., 2012; Huang et al., 2018). Ollitrault et al. (2012) reported lower recombination rates for the male compared to the female Clementine. Huang et al. (2018) observed greater map length for the pollen parent trifoliate orange compared to the seed parent sweet orange map. The mechanism of heterochiasmy *i.e.* the presence of differential cross over frequencies in male and female meiosis, may be responsible for this variability of genetic distances (Lenormand and Dutheil, 2005). According to this mechanism, gametic selection determines the heterochiasmy and the sex experiencing more stringent selection pressure during gametes production tend to have lower recombination frequencies. Bernet et al. (2010) in his experiments of cross pollination with compatible citrus parents found that the proportion of fertilized ovules is much greater than that of successful male gametes. In this study, the percentage of distorted markers was also greater in the male parent in both populations (Supplementary Table 2). Therefore, these results indicate that gametic selection is much more active in male than female gametes in citrus. The findings of the study are consistent with the recent reports of Garavello et al. (2020) in citrus. They studied the possible causes of segregation distortion by independently genotyping the pollen nuclei of the male parent [Clementine × sweet orange hybrid (‘CSO’)] and of the population progenies (‘RTSO’) resulting from the cross of female parent [(mandarin × sweet orange; ‘RTO’ tangor)] with ‘CSO’ pollen. The percentage segregation distortion was found to be lower in ‘CSO’ pollen than the ‘RTSO’ population nuclei. Hence, instead of the single sex based differential cross over frequency factor, the male gametes also experience selection pressure through other mechanisms like female-male gametic interactions or zygotic selection mechanisms (Garavello et al., 2020). From these other mechanisms, the S-RNase based gametophytic incompatibility system (GIS) causes male-female gamete interactions and contributes to segregation distortion in citrus (Garavello et al., 2020). GIS is the inability of a fertile hermaphrodite seed plant to produce zygotes after self-pollination. It has been found to be located on LG7 (Liang et al., 2019). In this study, ‘SB’ (Clementine mandarin × Minneola tangelo) and ‘D’ [(Clementine mandarin × Orlando tangelo) × (Clementine mandarin × Ponkan mandarin)] have Clementine as the common ancestor and therefore, could share an S-RNase haplotype, which could also be contributing towards this segregation distortion. Alternatively, the degree of heterozygosity of the parents has also been postulated to determine the rates of recombination and the genetic distance of maps. A higher degree of heterozygosity was found to correlate with lower recombination rates, as high heterozygosity suppresses recombination (Huang et al., 2018). The female parents used in the study are hybrids with relatively high degrees of heterozygosity, as a consequence of varying levels of admixture of two parental species, *C. reticulata* Blanco and *C. maxima* [(Burm.) Merr] (Wu et al., 2014; Wu et al., 2018). Mukaku Kishu is a somatic mutant of Kishu Mikan (*C. reticulata*) (Yamasaki et al., 2007), with less *C. maxima* introgression and therefore lower heterozygosity (Wu et al., 2018). Thus, multiple mechanisms could be the causes for the differences of genetic distance between male and female maps in this study.

Distribution of distorted markers

The markers which deviate from Mendelian segregation ratios are referred to as distorted markers. The factors like statistical bias, errors during genotyping or scoring, or biological mechanisms have been proposed as possible causes of segregation distortion (Bradshaw and Stettler, 1994). The distribution of the distorted markers was not uniform across different linkage groups or the different maps. For instance, on LG2 of 'SB' map, LG7 of 'MK_D' map and LG8 of 'D' map, the proportion of the distorted markers was 87.3, 58.7 and 52.6%, respectively (Supplementary Table 2), but the distorted markers did not affect the grouping in our study as revealed by high LOD score for different linkage groups, thus suggesting that distortion may be due to a biological mechanism (Fishman et al., 2001). Similar observations were also noted in citrus by Ruiz and Asins (2003), and exclusion of such markers may result in a loss of significant information (Cervera et al., 2001). The selection operating in male or female gametes, their interactions with the cytoplasm, or differential selection of zygotic individuals are the possible biological mechanisms responsible for segregation distortion (Reflinur et al., 2014; Garavello et al., 2020; Ollitrault et al., 2021). The variable distribution of skewed markers on different LGs suggests that all these mechanisms were involved in distortion in the two populations of this study. For instance, the respective female maps 'D' and 'SB' had 12.8% and 87.3% skewed markers in LG2. On 'MK_SB' and 'MK_D' maps, this LG had 44.2% and no skewed markers, respectively (Supplementary Table 2). This highlights the possible role of female gametic selection, nuclear-cytoplasm interactions, and zygotic selection in segregation distortion in this LG (Reflinur et al., 2014).

Consensus genetic map

Integration of maps across the parents (male and female) and populations is a useful approach to increase the marker density (Schlautman et al., 2017). The integration of maps is dependent upon the type of mapping population and the cross homology of the linkage groups. The use of intraspecific populations provides better opportunities to build integrated high density linkage maps compared to interspecific or intergeneric populations. Guo et al. (2015) successfully constructed an integrated map with 1543 SNP and 20 SSR markers using an intraspecific F_1 population of 124 individuals. In contrast, Curtolo et al. (2017) attempted to construct a genetic map in an interspecific full-sib F_1 population of 278 individuals derived from the cross of Murcott tangor and Pera sweet orange; only 661 SNP-based DArTseq markers were finally mapped on the integrated map.

In this study, we used two intraspecific mandarin populations generated using a common male parent Mukaku Kishu. This allowed us to prepare population specific, integrated male-female linkage maps (sub-composite maps) in JoinMap, followed by their merging into a consensus linkage map using MergeMap algorithm (Wu et al., 2008). MergeMap considers the marker order of individual maps to provide the consensus map order (Close et al., 2009). The consensus linkage map was based on 2588 markers (2587 distinct SNP markers and a phenotypic *Fs*-locus) positioned at 2495 locations and spanned over a genetic distance of 1406.84 cM. The reference Clementine map had an average marker density of 0.88 markers/cM (Ollitrault et al., 2012). The average marker density on the consensus map is 1.83 markers/cM with an average inter-marker distance of 0.54 cM. Thus, the present consensus map is a highly saturated map, and with an average genome size estimated to be 1500 to 1700 cM (Ollitrault et al., 1994), it covers 83 to 94% of the citrus genome. Due to high density of markers, high genome coverage and segregation information of two populations, the present map can serve as a future reference map for mandarins.

Inheritance of seedlessness and prediction of candidate genes

Seedlessness is an important trait for fresh citrus scion breeding, and 'MK' is an attractive genetic resource for breeding seedless citrus. Previous reports showed that seedlessness in 'MK' and its seedless progenies was due to the arrest of embryo development at zygotic or pre-cotyledonary stage or due to formation of small seeds (Yamasaki et al., 2007; Yamasaki et al., 2009). Yamasaki et al. (2009) characterized the expression of seedlessness in 'MK' and its descendants. They found two types of seeds in 'MK' and its seedless descendants: ovule-like seeds without seed coat (1 mm in size), type A seeds (2-3 mm size) with an immature soft seed coat. They found that 10 weeks after pollination was a defining point in embryo development of different types of seeds. In either of the 'MK' type seeds, an arrest of embryo development or slow development was a common feature while the perfect seeds from seedy individuals had faster embryo development after this time point. Two genes are hypothesized to govern the seedlessness in 'MK'; *Fs*, a dominant gene is proposed to support the seedlessness while *Is*, a repressor gene in dominant state inhibits seedless expression (Nesumi et al., 2001). The seedless 'MK' and seedy mandarins for these two genes were proposed to have the genetic constitution of *Fsfs-isis* and *fsfs-isis*, respectively (Nesumi et al., 2001; Yamasaki et al., 2007). Thus, in mandarin crosses with 'MK', a single gene could control the seedlessness. Based on this hypothesis, identification of RAPD markers linked to this *Fs*-locus was attempted earlier by Chavez and Chaparro (2011) using bulked segregant analysis (BSA) approach in the F_1 progeny of GS (an open pollinated seedling selection of Robinson tangerine) and 'MK'. They reported four RAPD markers, namely OPAI11-0.8, OPAJ19-1.0, OPM06r-0.85, and OPAJ04r-0.6, linked to the seedless locus at a distance of 4.3 to 8.7 cM. But, later these markers were found to be family specific. The results of RAPD markers are often found to be less repeatable (Li and Quiros, 2001) and are also sensitive to lab conditions. In this study, we mapped the seedless locus in two populations to identify and map the gene(s) governing the seedless traits. The putative candidate locus mapped on the LG5 of male parent 'MK'. This observation is consistent with the findings of Shimada et al. (2014) who found seedlessness to be located on chromosome 5 of 'Okitsu 46' × 'Kankitsu Chukanbohon Nou 5 gou' map. The second parent 'Kankitsu Chukanbohon Nou 5 gou' in their study was a hybrid between 'Lee' (Clementine mandarin × Orlando tangelo) and 'MK'. Their map was based on 708 EST-based CAPS markers, and the *Fs*-locus was mapped between markers Vs0015 and Edp005 at a distance of 7.2 and 3.1 cM, respectively. High-density maps can provide high resolution mapping of traits in one step and even preclude the need of fine mapping (Khan et al., 2012). In the present study, two markers AX-160536283 (at 2.4 cM on 'MK_D' map) and AX-160417325 (at 7.4 cM on 'MK_SB' map) showed close association with the *Fs*-locus. The first marker maintained its proximity in both populations, while the latter exhibited test cross segregation in the 'SB' × 'MK' population but appeared as an intercross marker for 'D' × 'MK' population. The consensus map also had AX-160536283 closer to the *Fs*-locus. The AX-160536283 showed 85.1 and 91.9% PPV for seedless progenies detection in 'SB' × 'MK' and 'D' × 'MK' populations, respectively. The AX-160417325 PPV for seedless progeny detection ranged from 25 to 87% in different observed populations. In two of the minor populations ('CVO' × 'MK' and 'T' × 'MK'), an unexpected allelic constitution (C:C) was also observed in the seedless progenies in addition to the expected 'MK' T:C allelic pattern. This type of genotypic variation is a deviation from the expected Mendelian segregation, and is classified under the category of offspring Mendelian error. For most of the populations, the female parent and male parents were of T:T and T:C allelic patterns, the appearance of additional homozygotes for C allele (other than T:T) points to a case of an allele drop in ($\mathrm{ADI_{hom}}$) (Arias et al., 2022). This type of call in SNP array based genotyping may result from the presence of partial null allele in one of the parents. Since we did not observe such unexpected homozygotes in 'SB' × 'MK', 'D' × 'MK' and 'L' × 'MK' populations, such partial null allele in the remaining two minor populations is probably being passed from their female parents- 'CVO' and 'T' (Arias et al., 2022).

Based on the 'MK_SB' and 'MK_D' maps, the genomic region for the *Fs*-locus corresponded to a physical interval of 3.97-10.0 Mb between SNPs AX-160906995 and AX-160536283 on chromosome 5. This region encompasses 131 genes of which many are multiple copies (Supplementary Information 2). Seedlessness in 'MK' and its seedless progenies is due to the arrest of zygote growth from globular to pre-cotyledonary stage (pro-embryo stages) (Yamasaki et al., 2007). Thus, a gene which causes zygotic arrest at pro-embryo stage could be the potential candidate gene for imparting seedlessness in 'MK' and its progenies. From 131 genes, the 13 candidate genes representing seven gene families namely homeobox protein 33, wall associated kinase-like 1, cytochrome P450, T-complex proteins, 6-phosphoglucanate dehydrogenase, ATP phosphoribosyl transferase 2, and UDP-Glycosyltransferase superfamily protein are expressed in developing embryos or seed coat. Of these seven family genes, the homeobox gene is earlier reported to influence seed number in grapes (Li et al., 2019). Contrary to this, the 'MK' fruits are completely seedless (Yamasaki et al., 2007). The genes encoding cytochrome P450 genes and wall associated proteins have a specific expression pattern in embryo sac units and embryos (Wang et al., 2012; Sotelo-Silveira et al., 2013). The ATP phosphoribosyl transferase 2 is involved in histidine biosynthesis, which is essential for normal embryo development; its mutation reportedly caused embryo abortion in Arabidopsis (DeFraia and Leustek, 2004). DeFraia and Leustek (2004) found that the wild types (homozygous dominant for this gene) could produce sufficient histidine for embryo growth while the heterozygous mutants had insufficient histidine to support embryo development, eventually leading to their abortion. The UDP-Glycosyltransferase superfamily protein is expressed in seed coats post-torpedo stage (Barvkar et al., 2012). Any mutation in the T-complex protein coding gene is known to cause impaired embryo development (Garcia et al., 2017). The 6-phosphoglucanate dehydrogenase genes code for an enzyme which is involved in oxidative pentose phosphate pathway. The plastidial pentose phosphate pathway is essential for post-globular stage development of embryos in Arabidopsis (Andriotis and Smith, 2019). Any defect or mutation in the above-described genes should be imparting complete seedlessness in 'MK' and its seedless progenies.

In this study, we found the location of the *Fs*-locus on LG5 of 'MK' (Mukaku Kishu) and identified two closely associated SNPs, AX-160417325 and AX-160536283. These SNPs reduced the effective population size and positively predicted seedlessness in 25.0-91.9% of the progenies in studied populations. These markers should prove useful for reducing the effective population size at seedling stage in crosses involving 'MK' paternity. Yet, the presence of seedless allelic pattern in some seedy individuals (false positives) and very few seedless individuals sharing the alternate allelic pattern indicate that these are not co-segregating markers. There are some gaps in immediate vicinity of *Fs*-locus in both 'MK_SB' and 'MK_D' maps. To underpin the exact causative gene and find a co-segregating marker for this trait, the region need to be delimited with use of additional cross populations, increase in size of the mapping population, and inclusion of additional polymorphic markers. Further, complementation of the study with expression analysis could assist in confirming the gene governing the seedlessness in Mukaku Kishu.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author/s.

Author contributions

FG and QY conceived and designed the study. KK and CH performed the experiment. KK, DB, and QY performed data analysis and prepared genetic maps. KK, DB, and QY wrote the rough draft of manuscript. FG and QY supervised the project. FG managed funds and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The research was partially supported by the New Varieties Development and Management Corporation, on behalf of the citrus growers in Florida, and Punjab Agricultural University, Ludhiana, India under the scheme – 'Training of PAU faculty members in new technologies, (MISC. I)' to KK.

Acknowledgments

The authors thank Fabieli Irizarry and Francis Gonzalez Rodriguez for their help during fruit phenotyping and DNA isolation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1087023/full#supplementary-material

References

Aleza, P., Juarez, J., Cuenca, J., Ollitrault, P., and Navarro, L. (2012a). Extensive citrus triploid hybrid production by 2x × 4x sexual hybridizations and parent-effect on the length of the juvenile phase. *Plant Cell Rep.* 31, 1723–1735. doi: 10.1007/s00299-012-1286-0

Aleza, P., Juarez, J., Hernandez, M., Ollitrault, P., and Navarro, L. (2012b). Implementation of extensive citrus triploid breeding programs based on 4x × 2x sexual hybridisations. *Tree Genet. Genomes* 8, 1293–1306. doi: 10.1007/s11295-012-0515-6

Aloisi, I., Distefano, G., Antognoni, F., Potente, G., Parrotta, L., Faleri, C., et al. (2020). Temperature-dependent compatible and incompatible pollen-style interactions in *Citrus clementina* hort. ex tan. show different transglutaminase features and polyamine pattern. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01018

Andriotis, V. M. E., and Smith, A. M. (2019). The plastidial pentose phosphate pathway is essential for postglobular embryo development in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15297–15306. doi: 10.1073/pnas.1908556116

Arias, K. D., Álvarez, I., Gutiérrez, J. P., Fernandez, I., Menéndez, J., Menéndez-Arias, N. A., et al. (2022). Understanding mendelian errors in SNP arrays data using a gochu asturcelta pig pedigree: Genomic alterations, family size and calling errors. *Sci. Rep.* 12, 19686. doi: 10.1038/s41598-022-24340-0

Barvkar, V. T., Pardeshi, V. C., Kale, S. M., Kadoo, N. Y., and Gupta, V. S. (2012). Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in *Linum usitatissimum* identified genes with varied expression patterns. *BMC Genomics* 13, 175. doi: 10.1186/1471-2164-13-175

Bernet, G. P., Fernandez-Ribacoba, J., Carbonell, E. A., and Asins, M. J. (2010). Comparative genome-wide segregation analysis and map construction using a reciprocal cross design to facilitate citrus germplasm utilization. *Mol. Breed.* 25, 659–673. doi: 10.1007/s11032-009-9363-y

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denance, C., Theron, A., et al. (2016). Development and validation of the axiom^{TM} Apple480K SNP genotyping array. *Plant J.* 86, 62–74. doi: 10.1111/tpj.13145

Bradshaw, H. D., and Stettler, R. F. (1994). Molecular genetics of growth and development in *Populus*. II. segregation distortion due to genetic load. *Theor. Appl. Genet.* 89, 551–558. doi: 10.1007/BF00222447

Brookes, A. J. (1999). The essence of SNPs. *Gene* 234, 177–186. doi: 10.1016/S0378-1119(99)00219-X

Cervera, M. T., Storme, V., Ivens, B., Gusmao, J., Liu, B. H., Hostyn, V., et al. (2001). Dense genetic linkage maps of three *Populus* species (*Populus deltoids*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. *Genetics* 158, 787–809. doi: 10.1093/genetics/158.2.787

Chavez, D. J., and Chaparro, J. X. (2011). Identification of markers linked to seedlessness in *Citrus kinokuni* hort. ex Tanaka and its progeny using bulked segregation analysis. *HortScience* 46, 693–697. doi: 10.21273/HORTSCI.46.5.693

Chen, C. X., Bowman, K. D., Choi, Y. A., Dang, P. M., Rao, M. N., Huang, S., et al. (2008). EST-SSR genetic maps for *Citrus sinensis* and *Poncirus trifoliata*. *Tree Genet. Genomes* 4, 1–10. doi: 10.1007/s11295-007-0083-3

Chen, C. X., and Gmitter, F. G. (2013). Mining of haplotype-based expressed sequence tag single nucleotide polymorphisms in citrus. *BMC Genomics* 14, 746. doi: 10.1186/1471-2164-14-746

Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., et al. (2009). Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* 10, 582. doi: 10.1186/1471-2164-10-582

Cui, F., Zhang, N., Fan, X. L., Zhang, W., Zhao, C. H., Yang, L. J., et al. (2017). Utilization of a Wheat660K SNP array-derived high-density genetic map for high-resolution mapping of a major QTL for kernel number. *Sci. Rep.* 7, 3788. doi: 10.1038/s41598-017-04028-6

Curtolo, M., Cristofani-Yaly, M., Gazaffi, R., Takita, M. A., and Figueira, A. (2017). QTL mapping for fruit quality in *Citrus* using DArTseq markers. *BMC Genomics* 18, 289. doi: 10.1186/s12864-017-3629-2

Curtolo, M., Soratto, T., Gazaffi, R., Takita, M., Machado, M. A., and Cristofani-Yaly, M. (2018). High-density linkage maps for *Citrus sunki* and *Poncirus trifoliata* using DArTseq markers. *Tree Genet. Genomes* 14, 5–10. doi: 10.1007/s11295-017-1218-9

DeFraia, C., and Leustek, T. (2004)Functional genomics study in arabidopsis thaliana of histidine biosynthesis. In: *The Rutgers scholar*. Available at: https://rutgersscholar.libraries.rutgers.edu/index.php/scholar/article/view/75 (Accessed 17 December 2022).

Dirlewanger, E., Graziano, E., Joobeur, T., Garriga-Caldere, F., Cosson, P., Howad, W., et al. (2004). Comparative mapping and marker-assisted selection in rosaceae fruit crops. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9891–9896. doi: 10.1073/pnas.0307937101

Ferreira, A., da Silva, M. F., Silva, L., and Cruz, C. D. (2006). Estimating the effects of population size and type on the accuracy of genetic maps. *Genet. Mol. Biol.* 29, 187–192. doi: 10.1590/S1415-47572006000100033

Fishman, L., Kelly, A. J., Morgan, E., and Willis, J. H. (2001). A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. *Genetics* 159, 1701–1716. doi: 10.1093/genetics/159.4.1701

Garavello, M., Cuenca, J., Dreissig, S., Fuchs, J., Navarro, L., Houben, A., et al. (2020). Analysis of crossover events and allele segregation distortion in interspecific citrus hybrids by single pollen genotyping. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00615

Garcia, N., Li, Y., Dooner, H. K., and Messing, J. (2017). Maize defective kernel mutant generated by insertion of a ds element in a gene encoding a highly conserved TTI2 cochaperone. *Proc. Natl. Acad. Sci. U.S.A.* 114, 5165–5170. doi: 10.1073/pnas.1703498114

Gmitter, F. G., Grosser, J. W., Castle, W. S., and Moore, G. A. (2007). "A comprehensive citrus genetic improvement program," in *Citrus genetics, breeding and biotechnology*. Ed. I. A. Kahn (Wallingford: CAB International), 9–18.

Goto, S., Yoshioka, T., Ohta, S., Kita, M., Hamada, H., and Shimizu, T. (2018). QTL mapping of male sterility and transmission pattern in progeny of Satsuma mandarin. *PLoS One* 13, e0200844. doi: 10.1371/journal.pone.0200844

Gouesnard, B., Negro, S., Laffray, A., Glaubitz, J., Melchinger, A., Revilla, P., et al. (2017). Genotyping by sequencing highlights original diversity patterns within a

European collection of 1191 maize flint lines, as compared to the maize USDA genebank. *Theor. Appl. Genet.* 130, 2165–2189. doi: 10.1007/s00122-017-2949-6

Grattapaglia, D., and Sederoff, R. (1994). Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: Mapping strategy and RAPD markers. *Genetics* 137, 1121–1137. doi: 10.1093/genetics/137.4.1121

Grosser, J. W., and Gmitter, F. G. J. (2005). 2004 SIVB congress symposium proceedings "Thinking outside the cell": Applications of somatic hybridization and cybridization in crop improvement, with citrus as a model. *In Vitro Cell. Dev. Biol. Plant* 41, 220–225. doi: 10.1079/IVP2004634

Gulsen, O., Uzun, A., Canan, I., Seday, U., and Canihos, E. (2010). A new citrus linkage map based on SRAP, SSR, ISSR, POGP, RGA and RAPD markers. *Euphytica* 173, 265–277. doi: 10.1007/s10681-010-0146-7

Guo, F., Yu, H. W., Tang, Z., Jiang, X. L., Wang, L., Wang, X., et al. (2015). Construction of a SNP-based high-density genetic map for pummelo using RAD sequencing. *Tree Genet. Genomes* 11, 1–11. doi: 10.1007/S11295-014-0831-0

Hiraoka, Y. (2020). Application of high-density SNP genotyping array in citrus germplasm characterization and genetic dissection of traits. Ph.D. Dissertation, University of California, Riverside, USA.

Huang, M., Roose, M. L., Yu, Q., Du, D., Yu, Y., Zhang, Y., et al. (2018). Construction of high density genetic maps and detection of QTLs associated with huanglongbing tolerance in *Citrus*. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01694

Khan, M. A., Han, Y., Zhao, Y. F., and Korban, S. S. (2012). A high-throughput apple SNP genotyping platform using the GoldenGateTM assay. *Gene* 494, 196–201. doi: 10.1016/j.gene.2011.12.001

Lenormand, T., and Dutheil, J. (2005). Recombination difference between sexes: a role for haploid selection. *PLoS Biol.* 3, e63. doi: 10.1371/journal.pbio.0030063

Leonforte, A., Sudheesh, S., Cogan, N. O., Salisbury, P. A., Nicolas, M. E., Materneet, M., et al. (2013). SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity tolerance in field pea (*Pisum sativum*L.). *BMC Plant Biol.* 13, 161. doi: 10.1186/1471-2229-13-161

Li, H., Hearne, S., Banziger, M., Li, Z., and Wang, J. (2010). Statistical properties of QTL linkage mapping in biparental genetic populations. *Heredity* 105, 257–267. doi: 10.1038/hdy.2010.56

Li, G., and Quiros, C. F. (2001). Sequence-related amplified polymorphism (SRAP) a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in *Brassica*. *Theor. Appl. Genet.* 103, 455–461. doi: 10.1007/s001220100570

Li, Y. D., Zhang, S. L., Dong, R. Z., Wang, L., Yao, J., Nocker, S. V., et al. (2019). The grapevine homeobox gene VvHB58 influences seed and fruit development through multiple hormonal signaling pathways. *BMC Plant Biol.* 19, 523. doi: 10.1186/s12870-019-2144-9

Liang, M., Cao, Z., Zhu, A., Liu, Y., Tao, M., Yang, H., et al. (2019). Evolution of self-compatibility by a mutant Sm-RNase in citrus. *Nat. Plants* 6, 131–142. doi: 10.1038/s41477-020-0597-3

Montalt, R., Vives, M. C., Navarro, L., Ollitrault, P., and Aleza, P. (2021). Parthenocarpy and self-incompatibility in mandarins. *Agronomy* 11, 2023. doi: 10.3390/agronomy11102023

Montanari, S., Bianco, L., Allen, B. J., Martínez-García, P. J., Bassil, N. V., Postman, J., et al. (2019). Development of a highly efficient axiom™ 70 K SNP array for *Pyrus* and evaluation for high-density mapping and germplasm characterization. *BMC Genomics* 20, 331. doi: 10.1186/s12864-019-5712-3

Navarro, L., Aleza, P., Cuenca, J., Juarez, J., Pina, J. A., Ortega, C., et al. (2015). The mandarin triploid breeding program in Spain. *Acta Hortic.* 1065, 389–395. doi: 10.17660/ActaHortic.2015.1065.48

Nesumi, H., Nakano, M., and Yoshida, T. (2001). Mode of inheritance on the abnormal development of impregnated ovules derived from mukaku-kishu. *J. Jpn. Soc. Hortic. Sci.* 70 (Suppl. 2), 403.

Oliveira, A. C., Bastianel, M., Cristofani-Yaly, M., do Amara, A. M., and Machado, M. A. (2007). Development of genetic maps of the citrus varieties 'Murcott' tangor and 'Pera' sweet orange by using fluorescent AFLP markers. *J. Appl. Genet.* 48, 219–231. doi: 10.1007/BF03195216

Ollitrault, P., Ahmed, D., Costantino, G., Evrard, J. C., Cardi, C., Mournet, P., et al. (2021). Segregation distortion for male parents in high density genetic maps from reciprocal crosses between two self-incompatible cultivars confirms a gametophytic system for self-incompatibility in citrus. *Agriculture* 11, 379. doi: 10.3390/agriculture11050379

Ollitrault, P., Dambier, D., Luro, F., and Duperray, C. (1994). Nuclear genome size variation in *Citrus*. *Fruits* 49, 390–393.

Ollitrault, P., Terol, J., Chen, C. X., Federici, C. T., Lotfy, S., Hippolyte, I., et al. (2012). A reference genetic map of *C. clementina* hort. ex tan.: Citrus evolution inferences from comparative mapping. *BMC Genomics* 13, 593. doi: 10.1186/1471-2164-13-593

Omura, M., Ueda, T., Kita, M., Komatsu, A., Takanokura, Y., Shimada, T., et al. (2000). "EST mapping of *Citrus*," in Proceedings of the International Society of Citriculture: IX Citrus Congress (Orlando, Florida, USA). 71–74.

Pandey, M. K., Agarwal, G., Kale, S. M., Clevenger, J., Nayak, S. N., Sriswathi, M., et al. (2017). Development and evaluation of a high density genotyping 'Axiom_*Arachis*' array with 58 K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* 7, 40577. doi: 10.1038/srep40577

Peng, W., Xu, J., Zhang, Y., Feng, J., Dong, C., Jiang, L., et al. (2016). An ultra-high density linkage map and QTL mapping for sex and growth-related traits of common carp (*Cyprinus carpio*). *Sci. Rep.* 6, 26693. doi: 10.1038/srep26693

Reflinur, K. B., Jang, S. M., Chu, S. H., Bordiya, Y., Akter, M. B., Lee, J., et al. (2014). Analysis of segregation distortion and its relationship to hybrid barriers in rice. *Rice* 7, 3. doi: 10.1186/s12284-014-0003-8

Ruiz, C., and Asins, M. J. (2003). Comparison between *Poncirus* and *Citrus* genetic linkage maps. *Theor. Appl. Genet.* 106, 826–836. doi: 10.1007/s00122-002-1095-x

Sankar, A. A., and Moore, G. A. (2001). Evaluation of inter-simple sequence repeat analysis for mapping in citrus and extension of the genetic linkage map. *Theor. Appl. Genet.* 102, 206–214. doi: 10.1007/s001220051637

Schlautman, B., Covarrubias-Pazaran, G., Diaz-Garcia, L., Iorizzo, M., Polashock, J., Grygleski, E., et al. (2017). Construction of a high-density American cranberry (*Vaccinium macrocarpon* ait.) composite map using genotyping-by-sequencing for multi-pedigree linkage mapping. *G3: Genes- Genomes-Genetics* 7, 1177–1189. doi: 10.1534/g3.116.037556

Shimada, T., Fuji, H., Endo, T., Ueda, T., Sugiyama, A., Nakano, M., et al. (2014). Construction of a *Citrus* framework genetic map anchored by 708 gene-based markers. *Tree Genet. Genomes* 10, 1001–1013. doi: 10.1007/s11295-014-0738-9

Sotelo-Silveira, M., Cucinotta, M., Chauvin, A. L., Chávez Montes, R. A., Colombo, L., Marsch-Martínez, N., et al. (2013). Cytochrome P450 CYP78A9 is involved in *Arabidopsis* reproductive development. *Plant Physiol.* 162, 2779–2799. doi: 10.1104/pp.113.218214

Thermo Fisher Scientific Inc (2020) *AxiomTMAnalysis suite (AxAS) v5.1 USER GUIDE*. Available at: https://downloads.thermofisher.com/Axiom_Analysis/Axiom_Analysis_Suite_v5.1.

Van Ooijen, J. W. (2006). *JoinMap 4: Software for the calculation of genetic linkage maps in experimental populations* (Wageningen, Netherlands: Kyazma).

Van Ooijen, J. W. (2011). Multipoint maximum likelihood mapping in a fullsib family of an outbreeding species. *Genet. Res.* 93, 343–349. doi: 10.1017/S0016672311000279

Voorrips, R. E. (2002). Map chart: Software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wang, N., Huang, H. J., Ren, S. T., Li, J. J., Sun, Y., Sun, D. Y., et al. (2012). The rice wall-associated receptor-like kinase gene OsDEES1 plays a role in female gametophyte development. *Plant Physiol.* 160, 696–707. doi: 10.1104/pp.112.203943

Wu, Y., Close, T. J., and Lonardi, S. (2008). "On the accurate construction of consensus genetic maps," in Proceedings of LSS Computational Systems Bioinformatics Conference (USA: Stanford). 285–296.

Wu, G. A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., et al. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* 32, 656–662. doi: 10.1038/nbt.2906

Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C., et al. (2018). Genomics of the origin and evolution of *Citrus*. *Nature* 554, 311–316. doi: 10.1038/nature25447

Yamamoto, M., Matsumoto, R., Okudai, N., and Yamada, Y. (1997). Aborted anthers of *Citrus* result from gene-cytoplasmic male sterility. *Sci. Hortic.* 70, 9–14. doi: 10.1016/S0304-4238(97)00017-4

Yamasaki, A., Kitajima, A., Ohara, N., Tanaka, M., and Hasegawa, K. (2007). Histological study of expression of seedlessness in *Citrus kinokuni* 'Mukaku kishu' and its progenies. *J. Amer. Soc Hortic. Sci.* 132, 869–875. doi: 10.21273/JASHS.132.6.869

Yamasaki, A., Kitajima, A., Ohara, N., Tanaka, M., and Hasegawa, K. (2009). Characteristics of arrested seeds in mukaku kishu-type seedless citrus. *J. Jpn. Soc. Hortic. Sci.* 78, 61–67. doi: 10.2503/jjshs1.78.61

Yu, Y., Chen, C., and Gmitter, F. G. (2016). QTL mapping of mandarin (*Citrus reticulata*) fruit characters using high-throughput SNP markers. *Tree Genet. Genomes* 12, 77. doi: 10.1007/s11295-016-1034-7

TYPE Original Research
PUBLISHED 20 February 2023
DOI 10.3389/fpls.2023.1090711

Role of long non-coding RNA in regulatory network response to *Candidatus* Liberibacter asiaticus in citrus

Xiaokang Zhuo[†], Qibin Yu[†], Riccardo Russo, Yi Zhang, Xu Wei, Yuanzhi Zimmy Wang, Paige Marie Holden and Fred G. Gmitter Jr.*

Citrus Research and Education Center, Institute of Food and Agricultural Sciences, University of Florida, Lake Alfred, FL, United States

Check for updates

OPEN ACCESS

EDITED BY
Concetta Licciardello,
CREA Research Centre for Olive, Fruit and Citrus Crops, Italy

REVIEWED BY
Fabienne Micheli,
Institut National de la Recherche Agronomique (INRA), France
Hongliang Zhu,
China Agricultural University, China

*CORRESPONDENCE
Fred G. Gmitter Jr.
✉ fgmitter@ufl.edu

[†]These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 05 November 2022
ACCEPTED 16 January 2023
PUBLISHED 20 February 2023

CITATION
Zhuo X, Yu Q, Russo R, Zhang Y, Wei X, Wang YZ, Holden PM and Gmitter FG Jr (2023) Role of long non-coding RNA in regulatory network response to *Candidatus* Liberibacter asiaticus in citrus. *Front. Plant Sci.* 14:1090711. doi: 10.3389/fpls.2023.1090711

Long non-coding RNAs (lncRNAs) serve as crucial regulators in plant response to various diseases, while none have been systematically identified and characterized in response to citrus Huanglongbing (HLB) caused by *Candidatus* Liberibacter asiaticus (*C*Las) bacteria. Here, we comprehensively investigated the transcriptional and regulatory dynamics of the lncRNAs in response to *C*Las. Samples were collected from leaf midribs of *C*Las- and mock-inoculated HLB-tolerant rough lemon (*Citrus jambhiri*) and HLB-sensitive sweet orange (*C. sinensis*) at week 0, 7, 17, and 34 following inoculation using *C*Las+ budwood of three biological replicates in the greenhouse. A total of 8,742 lncRNAs, including 2,529 novel lncRNAs, were identified from RNA-seq data with rRNA-removed from strand-specific libraries. Genomic variation analyses of conserved lncRNAs from 38 citrus accessions showed that 26 single nucleotide polymorphisms (SNPs) were significantly correlated with HLB. In addition, lncRNA-mRNA weighted gene co-expression network analysis (WGCNA) showed a significant module correlated with *C*Las-inoculation in rough lemon. Notably, the most significant *LNC_28805* and multiple co-expressed genes related to plant defense in the module were targeted by *miRNA5021*, suggesting that *LNC28805* might compete with endogenous *miR5021* to maintain the homeostasis of immune gene expression levels. Candidate *WRKY33* and *SYP121* genes targeted by *miRNA5021* were identified as two key hub genes interacting with bacteria pathogen response genes based on the prediction of protein-protein interaction (PPI) network. These two genes were also found within HLB-associated QTL in linkage group 6. Overall, our findings provide a reference for a better understanding of the role of lncRNAs involved in citrus HLB regulation.

KEYWORDS

long non-coding RNA, citrus Huanglongbing, weighted gene co-expression network analysis, miRNA5021, plant immunity

1 Introduction

Transcripts with a length of more than 200 nt and lower protein-coding potential are operationally termed long non-coding RNAs (lncRNAs), which are widespread non-coding RNAs (ncRNAs) in eukaryotes. In animals and plants, lncRNA can function as important and versatile regulators in a variety of cellular and biological processes (Kim and Sung, 2012). Biochemical experiments and genetics studies have demonstrated that lncRNAs are associated with chromatin modification (Rinn and Chang, 2012), mRNA splicing (Bardou et al., 2014), transcriptional gene silencing (Wierzbicki, 2012), and posttranscriptional gene regulation (Yoon et al., 2013). Studies indicate that lncRNAs can help the host to prevent pathogen replication or be used by pathogens to promote pathogen proliferation (Li et al., 2016; Zaynab et al., 2018; Shirahama et al., 2020). In tomato, yellow leaf curl virus intergenic siRNAs target a host long noncoding RNA to modulate disease symptoms (Yang et al., 2019); also, tomato lncRNA23468 functions as a competing endogenous miR482b to enhance the defense against *Phytophthora infestans* (Jiang et al., 2019). It was also found that lncRNAs function as precursors of miRNAs having stable hairpin structures in wheat (Xin et al., 2011). In a word, lncRNAs are involved in pathogenic infection by acting as miRNA targets, miRNA precursors, or endogenous target mimics (eTMs) to regulate the expression of their target genes (Shirahama et al., 2020; Song et al., 2021).

Recent studies have revealed a set of the important regulatory functions of lncRNAs in response to pathogen infection. Transcriptome analyses revealed that a large number of lncRNAs were differentially expressed in response to pathogen infection in plants. For instance, *lncRNA16397* was involved in resistance to *P. infestans* infection by co-expressing glutaredoxin in tomato (Cui et al., 2017). In melon, lncRNAs function as miRNA precursors and are involved in the response of powdery mildew fungi (Gao et al., 2020). Also, lncRNAs are involved in the response of *Arabidopsis thaliana* to *Fusarium oxysporum* infection (Zhu et al., 2014) and cotton to *Verticillium dahlia* infection (Zhang et al., 2018). In addition, a study also found that interaction between Sl-lncRNA15492 and Sl-miR482a can affect *Solanum lycopersicum* immunity against *P. infestans* (Jiang et al., 2020). Thus, lncRNAs are important components of regulated networks in response to pathogen infection. Although many lncRNAs have been identified from transcriptome data in diverse plant species, most of them are not well characterized.

Huanglongbing (HLB), a disease caused by the phloem-limited bacterium *Candidatus* Liberibacter asiaticus (*C*Las), is the most prevalent and destructive citrus epidemic. It has devastated the citrus industry in Florida and is threatening the global citrus industries (Bové, 2006; Graham et al., 2020). Thus far, citrus HLB has not been controlled effectively, and some research directions are precluded because of the challenge of *C*Las' unculturable and phloem-limited nature. HLB and its vector, Asian citrus psyllid (ACP, Diaphorina citri) is still rapidly spreading in citrus-producing areas, which leads to billions of dollars in annual economic loss (Alvarez et al., 2016; Wang, 2019; Monzó and Stansly, 2020). Current strategies for insecticide and antibiotics application are limited and unsustainable. One of the most effective and eco-friendly strategies is strengthening host plant defense and immunity. Usually, the plant innate immune response can be triggered when they are infected by the pathogen (Nobori and Tsuda, 2019). In citrus, *C*Las-triggered plant immune responses are delayed 5–9 weeks after inoculation (Albrecht and Bowman, 2008; Yu et al., 2017). Traditional molecular biology, genetic, and multi-omics analyses also incompletely revealed the nature of pathogenesis of citrus HLB (da Graça et al., 2016; Wang, 2019). A study indicated that *Citrus tristeza virus* (CTV) can produce LMT1 lncRNA to suppress salicylic acid (SA) accumulation and mitigate reactive oxygen species (ROS) accumulation (Kang et al., 2019). These cases related to lncRNA involved in plant disease regulation bring us a promising direction to explore the role of lncRNA against citrus HLB disease (Jiang et al., 2020; Hong et al., 2022; Sharma et al., 2022). Thus far, no studies have explored such roles and characteristics in citrus.

Several citrus relatives such as *Poncirus trifoliata* and *Microcitrus* were considered as tolerant or resistant (Ramadugu et al., 2016; Godfrey et al., 2017), however, they are genetically distant from commercial citrus varieties which have originated mainly from three common ancestors, wild mandarin (*C. reticulata*), pummelo (*C. maxima*), and citron (*C. medica*) through a long domestication evolution (Wu et al., 2018). Rough lemon shares wild mandarin as a common ancestor with sweet orange. Our previous study showed that rough lemon is HLB tolerant compared with sweet orange (Fan et al., 2012). Once rough lemon trees are infected and symptomatic, they can be rejuvenated by the continued growth of new shoots with few or no foliar symptoms of the disease, and they repeat this cycle for many growing seasons; in contrast, sweet orange exhibits continuous growth inhibition and eventual dieback (Fan et al., 2012). By comparative transcriptional and anatomical analysis of rough lemon and sweet orange in response to *C*Las, phloem transport activity and the expression of defense-related genes are much greater in rough lemon than in sweet orange (Fan et al., 2012; Yu et al., 2017), suggesting the ability to maintain good phloem transport with extensive *C*Las titer is likely critical to good HLB tolerance. To further explore the contributions of lncRNA in response to HLB, we systematically identified lncRNAs from rough lemon and sweet orange at four different time points of a greenhouse experiment and characterized their genomic transcriptional and regulatory dynamics. We predicted their potential regulatory genes and functions and constructed a co-expression network. Our study provides valuable information and expands the knowledge of the role of citrus lncRNA in HLB disease expression.

2 Materials and methods

2.1 Plant materials

The plant inoculation was performed using the method as previously described (Fan et al., 2012). Briefly, two-year-old HLB-sensitive sweet orange (*C. sinensis* L Osb.) and HLB-tolerant rough lemon (*C. jambhiri* Lush.) plants were graft-inoculated with *C*Las positive bud wood collected from Carrizo citrange (*C. sinensis* × *P. trifoliata* L. Raf.) grown in a protected greenhouse, and mock-inoculated controls used bud wood from pathogen tested and healthy Carrizo trees. Each treatment had three biological replicates. All these plants were kept in a state-certified disease-free greenhouse (a United States Department of Agriculture Animal and Plant Health Inspection Service and Center for Disease Control-approved and secured greenhouse) at the University of Florida, Citrus Research and Education Center, Lake Alfred. Midribs of mature leaves were sampled from *C*Las-inoculated and mock-inoculated trees every two weeks after inoculation (WAI) at early stages (before ten weeks) and every one week at later stages; quantitative real-time PCR (qRT-PCR) was performed to test for the presence of *C*Las as previously described (Li et al., 2006). Plants were considered HLB positive when PCR cycle threshold (CT) values were below 30 (Yu et al., 2017). Positive plants were not detected until 17 WAI. The typical blotchy mottled HLB symptom was observed around 34 WAI in rough lemon and sweet orange. Based on the presence of positive samples and HLB symptoms, midribs of mature leaf from *C*Las-inoculated and mock-inoculated rough lemon and sweet orange trees at 0, 7, 17, and 34 WAI were collected. A total of 48 samples from four different time-points were used for RNA-seq in this study. The information about plant materials and the HLB test is shown in Table S1. Midribs from the mature leaves of rough lemon and sweet orange, which were under Huanglongbing (HLB) disease stress for more than ten years in the field, were also used for qPCR validation to explore whether the regulatory relationship between candidate lncRNAs and genes also existed in the plants from different growth condition.

2.2 Strand-specific RNA sequencing

Total RNA was extracted using TRIzol® Reagent and purified using a RNeasy Mini Kit (Valencia, CA, United States) following the manufacturer's protocol. Ribosomal RNA was removed from the total RNA using a Ribo-Zero rRNA removal kit (Epicenter-Illumina, Madison, WI, United States) following the manufacturer's protocol. High-quality RNA was used to construct strand-specific RNA (ssRNA) libraries at the Interdisciplinary Center for Biotechnology Research (ICBR) Gene Expression Core, University of Florida (UF) described by Yu et al. (Yu et al., 2017). The prepared libraries were sequenced on an Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA) producing paired-end 100 bp reads.

2.3 Transcript assembly and lncRNA identification

Low-quality reads and the adaptor sequences were removed using the fastp tool (Chen et al., 2018). After filtering, the clean reads were mapped to the *Citrus clementina* genome v1.0 (JGI) (Wu et al., 2014) using HISAT2 software (Kim et al., 2019). Next, StringTie was used to assemble transcripts of each sample, merge transcripts to get a consensus transcriptome assembly, and compute the abundance of these transcripts (Pertea et al., 2016). Subsequently, the newly assembled transcripts were compared with the *C. clementina* reference genome annotations using the GffCompare program (Pertea and Pertea, 2020). Transcripts overlapped with the known genes were discarded. The resulting transcripts with length ≥ 200 bp and fragments per kilobase of transcript per million fragments (FPKM) ≥ 1 in more than three samples were extracted, and then the tRNAs and rRNA were removed from the extracted transcripts using tRNAccan-SE (Lowe and Eddy, 1997) and RNAmmer (Lagesen et al., 2007), respectively. To reduce the noise of transcripts encoding proteins, TBtools software (Chen et al., 2020) was used to identify the open reading frames (ORFs) of these transcripts. Transcripts with significant ORFs were aligned to the Swiss-Prot, Nr, and Pfam databases, and the transcripts with E-value ≤ 10^{-5} were excluded. Finally, we further evaluated the coding ability of the remaining transcripts using the Coding Potential Calculator version2 (CPC2) (Kang et al., 2017). We searched the lncRNAs in the PLncDB V2.0 database (Jin et al., 2021) and CANTATAdb (Szcześniak et al., 2016) to identify novel lncRNAs. The information on lncRNAs is listed in Table S2. Based on the genomic coordinates of protein-coding genes, the lncRNAs were divided into five groups (Scheuermann and Boyer, 2013; Ransohoff et al., 2018): Intergenic lncRNAs (LINC), intronic lncRNA (INTRONIC), natural antisense transcripts (NAT), genic lncRNA (GENIC), and exonic lncRNA (EXON).

2.4 Prediction of lncRNAs targets, precursors, and eTMs

Mature miRNA and miRNA precursors were downloaded from the miRbase database (Release 22.1) (Kozomara et al., 2019). The psRNATarget (Dai et al., 2018) was used to predict lncRNAs acting as putative miRNA targets with the default settings. All 8742 identified lncRNAs were used to predict eTMs using the psMimic software with the default settings (Wu et al., 2013). The lncRNAs transcripts were aligned to the miRNA precursors sequences from the miRbase database to predict the precursors using BLASTN software based on the best hits with E-value < 1e-10 and query identify > 80%. The miRNA target of candidate genes was predicted using TAPIR (http://bioinformatics.psb.ugent.be/webtools/tapir/) online tool.

2.5 Differential expression analysis of lncRNAs and mRNAs

The expression level of lncRNAs and mRNAs were quantified based on the position of mapped reads using StringTie software (Pertea et al., 2016) and evaluated by FPKM. The sample biological replicates were examined using principal component analysis (PCA) and correlation analysis, and the differentially expressed (DE) analysis was performed using the DEseq2 package in R software (Love et al., 2014). DE mRNAs and lncRNAs were determined with false discovery rate (FDR) < 0.01 and fold change (FC) |log2FC| ≥ 1. Protein coding (PC) genes were annotated using Swiss-Prot (Boeckmann et al., 2003) and non-redundant (NR) database (Pruitt et al., 2005). Gene functional enrichment was analyzed using Metascape (Zhou et al., 2019) and MapMan (Thimm et al., 2004).

2.6 Orthologous identification and phylogenetic analysis of lncRNAs

The orthologous lncRNAs were identified based on reciprocal best blast hits (RHB) using the Basic Local Alignment Search Tool

(BLAST) with query coverage > 80% and E-value < 1e-10 (Moreno-Hagelsieb and Latimer, 2008). The orthologous lncRNA sequences of *Arabidopsis thaliana*, *Oryza sativa*, and *Populus trichocarpa* were downloaded from the PLncDB V2.0 database (Jin et al., 2021). A total of 36 orthologues were identified. For phylogenetic analysis, sequences of 38 accessions (Table S3) were from previous publications except for *Citrus latipes* (Wu et al., 2014; Wu et al., 2018). Variant calling and filtering were according to the method previously described (Peng et al., 2020). Briefly, sequences were mapped to the *Citrus clementina* genome v1.0 (JGI) (Wu et al., 2014) *via* BWA-MEM (Li, 2013). Raw aligned reads were sorted and duplicate reads removed *via* samtools V1.7 (Li et al., 2009) and sambamba V0.6.7 (Tarasov et al., 2015), respectively. Variant calling and filtering were performed using The Genome Analysis Toolkit (GATK) v4.1.2 (McKenna et al., 2010) and VCFtools (Danecek et al., 2011), respectively. Finally, 1,658 bi-allelic variants derived from 36 conserved orthologous lncRNAs genomic loci were extracted and used to construct a maximum-likelihood (ML) phylogenetic tree (Data S1). The best substitution model general time-reversible (GTR) for the ML tree was inferred using Smart Model Selection (SMS) web server (Lefort et al., 2017).

2.7 Construction of co-expression network for lncRNAs and mRNAs

First, we excluded the transcripts that had similar expression patterns between mock-inoculated and *C*Las-inoculated plants to reduce the noise caused by gene spatiotemporal-specific expression using the 'Mfuzz' package in R software with the k-nearest neighbor method (Peterson, 2009). Co-expression analysis was performed using the weighted co-expression network analysis (WGCNA) package in R software with threshold power = 6, minimum module size = 3, and a branch merge cut height = 0.25 (Langfelder and Horvath, 2008). The co-expression network was plotted using Gephi software (Bastian et al., 2009). Proteins of *Arabidopsis thaliana* were used as model to infer the protein-protein interaction (PPI) of co-expressed genes in citrus, based on STRING database (Szklarczyk et al., 2015).

2.8 cDNA synthesis and qRT-PCR

Twelve genes with high amplification efficiency and primer specificity were selected to validate the RNA-seq data of rough lemon and sweet orange using qRT-PCR. In addition, we also validate *LNC28805*, *WRKY33*, and *SYP121* and their targeting *miRNA5021* in rough lemon and sweet orange, which were under Huanglongbing (HLB) disease stress for more than ten years in the field in Lake Alfred. The STEM-LOOP RT-qPCR method was carried out for *miRNA5021* based on the method of Kramer (2011) (Kramer, 2011). The qPCR method for mRNA and lncRNA was described by Yu et al. (2017). Briefly, the first strand cDNA was synthesized using an Affinityscript qPCR cDNA Synthesis Kit (Agilent Technologies, Santa Clara, CA, USA), and RT-qPCR was performed using SYBR Green qPCR Master Mix (Agilent Technologies) in a 20-µl volume. 18S rRNA gene was used for an internal reference according to previous studies (Yan et al., 2012; Wei et al., 2021). The primers are listed in Table S4.

3 Results

3.1 Identification and characterization of lncRNAs in citrus

Approximately 2,523 million paired-end reads from 48 sample libraries were produced and mapped to the *C. clementina* v1.0 reference genome (Table S5). After a comprehensive pipeline of filtering, a total of 8,742 lncRNAs with FPKM > 1 in at least three samples were identified in both sweet orange and rough lemon, including 2,529 novel lncRNAs (Figure 1A and Table S2). Most of lncRNAs belong to LINC and EXON groups (Figure 1C and Table S2). The EXON group has a greater exon number and transcript length in comparison to other groups (Figure S1A, B).

To investigate the characteristics of citrus lncRNAs response to HLB, we analyzed the correlation coefficients between different samples using the expression profiles of lncRNAs. A relatively high correlation between biological replicates was observed (Figure S2). Analysis of lncRNAs distribution in the citrus genome showed that lncRNAs were widely expressed across all citrus chromosomes, and the highest number of expressed lncRNA was around the 40 Mb region on scaffold_3 (Figure 1B). We then investigated the transcript length and exon number distribution of lncRNAs and mRNAs. The exon number of lncRNA and mRNA had similar distribution patterns in rough lemon and sweet orange (Figure S3), while mRNA had a greater exon number and sequence length than lncRNA (Figures 1D, E; Figure S3). Meanwhile, we also investigated the difference in the overall gene expression levels of lncRNA and mRNA in four different periods. The results show that expression levels of mRNA were always higher than lncRNAs in both mock- and *C*Las-inoculated plants (Figure S1C-D).

3.2 Differential expression dynamics of lncRNAs and mRNA after CLas-inoculation

The repeatability of samples was evaluated using PCA analysis. The VH1_W0 sample was eliminated due to poor replication (Figure S4). After that, twelve genes were selected to further validate the RNA-seq data, indicating good reliability of the data (Figure S5). Based on the analysis, we systematically compared the expression levels of lncRNAs and mRNAs in rough lemon and sweet orange plants at different time points of mock- and *C*Las-inoculated conditions. A total of 1,943 and 25,118 differentially expressed (DE) lncRNA and mRNAs were identified from 14 pairwise comparisons of rough lemon and sweet oranges (Data S2-S3). The percentage of DE lncRNAs and mRNAs was similar in both mock-inoculated sweet orange and rough lemon plants, but there was a prominent difference between them in comparisons of the *C*Las-inoculated groups (Figure 2A). A larger percentage of DE lncRNAs and mRNAs was found at week 7 and week 34 in *C*Las compared to groups of sweet orange, whereas there was no prominent difference in rough lemon (Figure 2A). These results indicate *C*Las altered the spatiotemporal

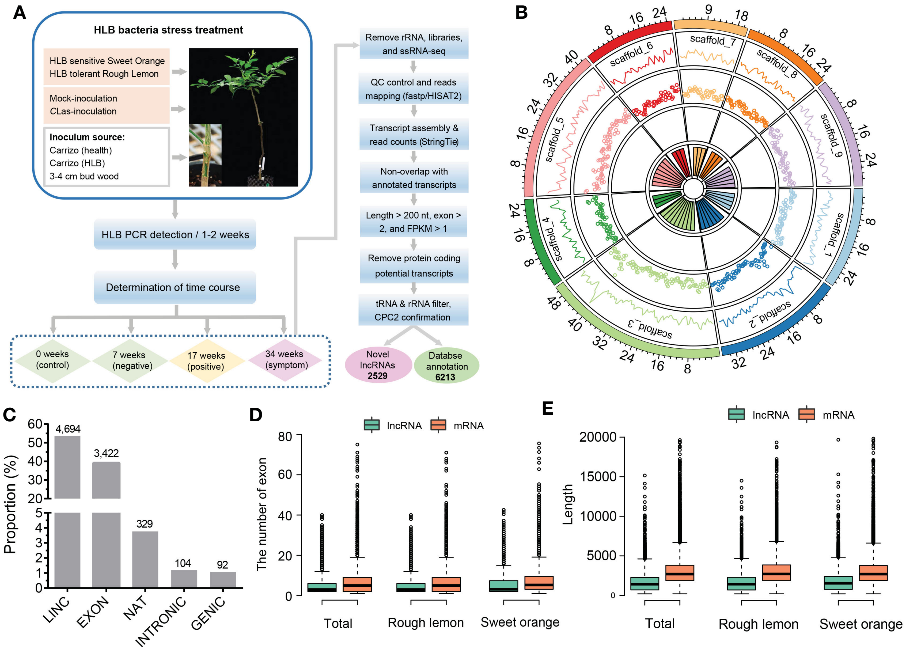


FIGURE 1
Identification and characterization of lncRNAs in HLB-tolerant rough lemon and HLB-sensitive sweet orange. **(A)** Pipeline of lncRNA identification. **(B)** Genomic distribution of lncRNAs on the *Citrus clementina* genome v1.0. From outer to inner rings indicates chromosome-level scaffolds, line plots, and dot plots of the lncRNA number distributed in each scaffold. **(C)** The numerical distribution of different types of lncRNAs. LINC, intergenic lncRNAs; INTRONIC, intronic lncRNA; NAT, natural antisense transcripts; genic lncRNA, GENIC; EXON, exonic lncRNAs. **(D)** Exon number and **(E)** length comparison of lncRNAs and mRNAs.

expression pattern of plants and reflect the significant difference of genes in response to Clas infection between sweet orange and rough lemon. We further investigated the number of up- and down-regulated DE lncRNA in sweet orange and rough lemon after Clas-inoculation (Figures 2B, C). Comparing W0 and different time points after CLas inoculation, we found that the number of DE lncRNAs significantly increased at week 7 (W7) and week 34 (W34) in sweet orange (Figure 2B), and the largest number of DE lncRNA was at W7 (down/up, 103/176). However, the largest number of DE lncRNA was found at W34 in rough lemon, and only included 75 down-regulated and 77 up-regulated DE lncRNAs (Figure 2B). Comparing healthy plants with CLas infected plants at the same time point, we found that the number of upregulated DE lncRNAs significantly increased at W34 in rough lemon (down/up, 16/76), while it was the opposite in sweet orange at the same points (down/up, 36/9) (Figure 2C).

Heatmaps showed completely different expression patterns of lncRNAs between rough lemon and sweet orange species but a similar expression pattern in different time points of mock- and CLas-inoculated plants of the same species (Figure 2D), indicating lncRNAs have higher species specificity in response to Clas infection. In addition, we further investigated the dynamic changes of specific DE lncRNAs and mRNAs in rough lemon and sweet orange (Figure S6). The number of specific DE lncRNAs in the pairwise groups of mock-inoculation vs. CLas-inoculation at the

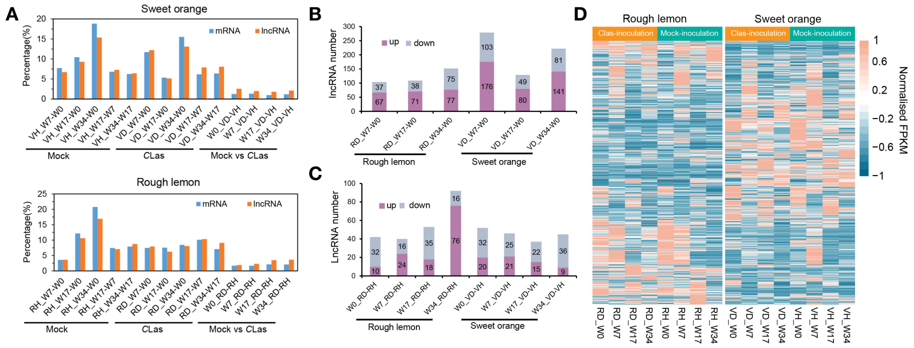


FIGURE 2
Global comparative analysis of lncRNAs and mRNA expressional dynamics. **(A)** Distribution of the differentially expressed (DE) mRNA and lncRNAs of 14 pairwise comparison groups in HLB-sensitive sweet orange and HLB-tolerant rough lemon. **(B)** Comparison of up- and down-regulated DE lncRNAs from W7-W0, W17-W0, and W34-W0 pairwise groups in *C*Las- inoculated rough lemon and sweet orange. **(C)** Comparison of up- and down-regulated DE lncRNAs between mock-inoculation and *C*Las-inoculation pairwise groups of rough lemon and sweet orange at the same stages. **(D)** Heatmap representing the expression patterns of DE lncRNAs in rough lemon and sweet orange. Data for lncRNAs expression were normalized to the Z-score.

same time point was higher in rough lemon than in sweet orange at W17 and W34 groups (Figure S6E). However, in the pairwise groups of W0 -W7, W0-W17, and W0-W34 of *C*Las-inoculation, we found the number of specific DE lncRNAs of sweet orange was near two-fold higher than rough lemon (Figure S6A), and a mass of specific DE lncRNAs were found at week 7 and week 34 after *C*Las-inoculation, and most of them were upregulated in both rough lemon and sweet orange. However, a larger number of DE lncRNAs were downregulated in the same pairwise mock inoculation groups (Figure S6B). In addition, we also found the specific DE lncRNAs were overall upregulated in rough lemon or sweet orange in the other compared groups (Figure S6C, D).

3.3 Evolution and function of specific DE lncRNAs in citrus

To further explore the characteristics of specific DE lncRNAs, we analyzed their genomic origins from rough lemon and sweet orange. The result shows that a total of 35 conserved lncRNAs, 166 specific lncRNAs in rough lemon, and 170 specific lncRNAs in sweet orange, were identified, and the distribution landscape of these specific DE lncRNAs was different between rough lemon and sweet orange on chromosome-level scaffolds (Figure 3A). Compared with the W0 time point (W7-W0, W17-W0, and W34-W0 pairwise groups of *C*Las-inoculated plants), most of DE lncRNAs were distributed in scaffold_1, 2, 3, 5, and their numbers exhibited significant differences between rough lemon and sweet orange (Figure 3B). The largest number was located on scaffold_5 and scaffold_3 in rough lemon and sweet orange, respectively (Figure 3B). In the pairwise groups between mock- and *C*Las-inoculated plants at the same time point, the largest percentage of DE lncRNAs (31.6%) was found in unanchored scaffolds in rough lemon, while the largest percentage was found on scaffold_3 in sweet orange (Figure 3C). We further investigated the number of common and unique genomic origin DE lncRNAs in rough lemon and sweet orange (Figure 3D). Only 43 specific DE lncRNAs derived from the same genomic loci of these two citrus species were identified, indicating that lncRNAs have high species-specific expression profiles in response to *C*Las.

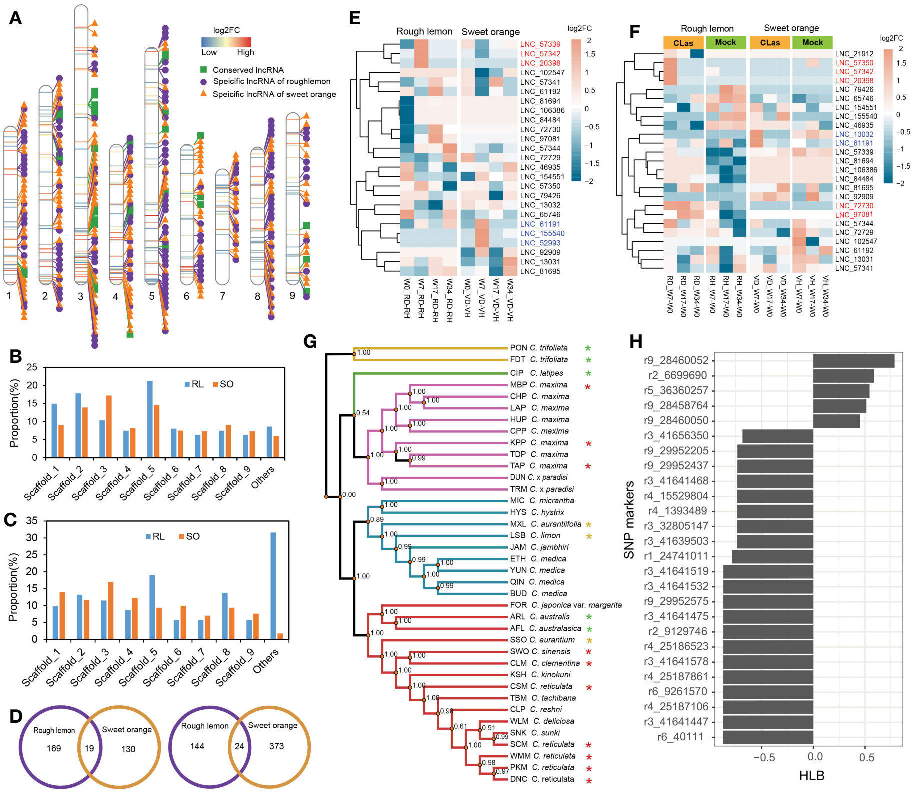


FIGURE 3

Identification and characterization of conserved lncRNAs in HLB-tolerant rough lemon and HLB-sensitive sweet orange. **(A)** Chromosome distribution of specific differentially expressed **(D, E)** lncRNAs and conserved DE lncRNAs. **(B)** Comparison of DE lncRNAs number of rough lemon and sweet orange distributed on the chromosome-level among W7-W0, W17-W0, and W34-W0 pairwise groups of *C*Las-inoculated plants, and **(C)** among comparison groups between mock- and *C*Las-inoculation at same time-point. RL, rough lemon; SO, sweet orange. **(D)** The number of DE lncRNAs specifically originated in the genome of rough lemon or sweet orange. DE lncRNAs of mock- and *C*Las-inoculated pairwise at same time-point (left); DE lncRNAs of W7-W0, W17-W0, and W34-W0 pairwise groups of *C*Las-inoculated plants (right). **(E, F)** Expression patterns of conserved lncRNAs across ten different pairwise comparison groups in rough lemon and sweet orange. Red and blue color font indicate lncRNAs significantly up-regulated at the early stage in rough lemon and in sweet orange, respectively. **(G)** A maximum likelihood (ML) phylogenetic tree of 38 citrus accessions. The tree was constructed based on the genomic variations derived from conserved lncRNAs. Asterisks indicate HLB tolerant levels (green, yellow, and red color indicate HLB tolerance, moderate tolerance, and sensitivity, respectively). The data of HLB evaluation were from Ramadugu et al. (2016) and Godfrey et al. (2017). **(H)** SNP markers derived from conserved lncRNAs significantly correlated with HLB response traits.

Analysis of the evolutionary conservation of lncRNAs showed that thirty-one sequences of conserved lncRNAs located in seven different scaffolds were identified with e-value > 1×10^{-10} (Figure 3A and Table S6), and the expression levels of them represented specifically spatiotemporal different expression patterns between mock- and *C*Las-inoculated plants (Figures 3E, F). For instance, three lncRNAs (*LNC_57342*, *LNC_20398*, and *LNC_61191*) were significantly upregulated at W7 stage in rough lemon, while they were downregulated in sweet orange; the other three lncRNAs (*LNC_61191*, *LNC_155540*, and *LNC_52993*) exhibited reverse expression patterns in sweet orange (Figure 3E). Compared with W0, we found that seven lncRNAs exhibited higher differentially expressed levels in W7 and W17 time points after *C*Las inoculation in rough lemon and sweet orange. which might be involved in response to HLB in early phases after *C*Las inoculation. We further identified 1,658 bi-allelic variants derived from these homologous lncRNAs based on resequencing data of 38 citrus accessions, and a phylogenetic tree was constructed (Table S3 and Data S1). The result showed that citrus species with close relatives were clustered together with higher support values (Figure 3G). Based on the HLB symptom evaluation of citrus relatives (Folimonova et al., 2009; Ramadugu et al., 2016; Godfrey et al., 2017), correlation analysis between bi-allelic variants and HLB symptom evaluation showed that 26 variants were significantly correlated with the HLB traits (Figure 3H). Most of the significant SNPs were located in 41.64-41.65 Mb region on scaffold_3 and 28.45-29.95 Mb region on scaffold_9. Six conserved DE lncRNAs with these correlated SNPs derived from four genomic loci (XLOC_007316, XLOC_030383, XLOC_072976, and XLOC_037329) were identified (Table S7).

In this study, we also predicted the miRNA targets, precursors, and eTMs of lncRNAs (Data S4). A total of 133 lncRNAs were identified as precursors for 33 miRNA families, 116 lncRNAs were identified to be targeted by 35 miRNA families, and 40 lncRNAs were predicted as eTMs for 14 miRNA families. Notably, 16 lncRNAs predicted as miRNA targets were simultaneously acting as miRNA precursors or eTMs, and nine lncRNAs were differentially expressed in sweet orange (Table S8), suggesting that these DE lncRNAs may be involved in response to HLB by interacting with miRNAs in sweet orange.

3.4 Identification of lncRNA-mRNA co-expression modules related to HLB response

To identify the DE lncRNAs and mRNAs potential response to HLB and to reduce the noise caused by gene spatiotemporal-specific expression, we first excluded the DE lncRNAs or mRNAs that had similar expression patterns in different time points of mock- and *C*Las-inoculation plants by using the k-nearest neighbor method (Figure 4A and Figure S7, 8). For instance, 53 and 49 similar expression pattern lncRNAs were excluded in the comparison of cluster1 (*C*Las) vs. cluster7 (Mock) and cluster3 (*C*Las) vs. cluster3

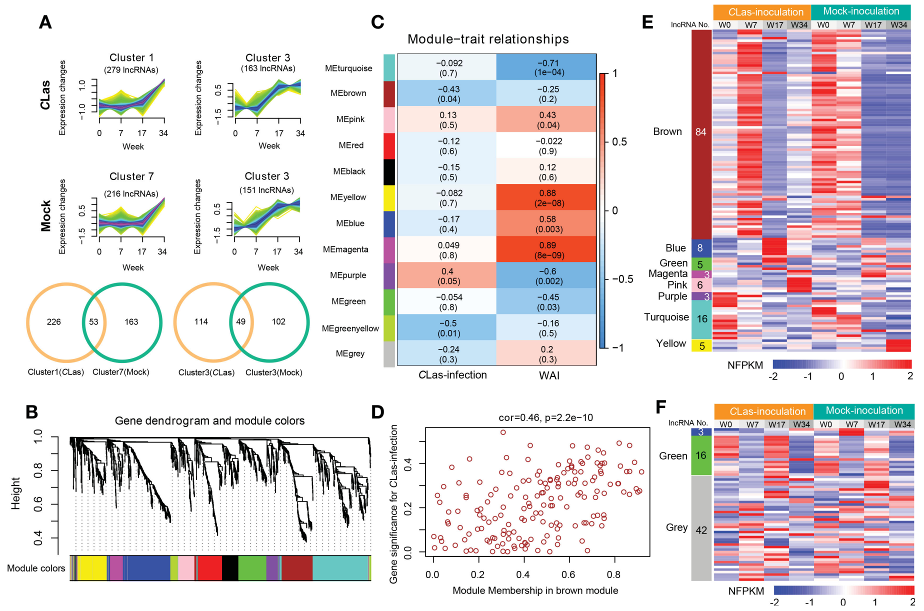


FIGURE 4
Hierarchical clustering and lncRNA-mRNA co-expression modules in rough lemon. **(A)** Similar expression tendency of lncRNAs in *C*Las-inoculation and mock-inoculation plants. Venn diagrams showing the specifically expressed lncRNAs. **(B)** Hierarchical cluster tree showing lncRNAs and mRNAs co-expression modules identified by Weighted Gene Co-expression Network Analysis (WGCNA). Twelve different modules were constructed and labeled by different colors. **(C)** Module-*C*Las-infection and module-WAI (weeks after inoculation) relationships. Each row corresponds to a module; Left column corresponds to inoculation approach and right column corresponds to the time point of after inoculation. Each cell is color-coded by correlation coefficient and contains corresponding *P-value*. **(D)** A scatterplot showing the relationship between gene significance for inoculation and module membership in brown module of rough lemon. **(E, F)** Heatmap showing the normalized FPKM (NFPKM) of lncRNAs in each significant module in **(E)** rough lemon and **(F)** sweet orange. FPKM were normalized to the Z-score.

(Mock) in rough lemon (Figure 4A), respectively. Finally, a total of 2133 (including 246 lncRNAs and 1887 mRNAs) and 2863 (including 295 lncRNAs and 2568 mRNAs) transcripts from rough lemon and sweet orange, respectively, were potentially responsible for HLB response and were used for co-expression analysis. This analysis resulted in 12 and 7 distinct modules, which are clusters of highly interconnected genes (Langfelder and Horvath, 2008) in rough lemon (Figure 4B) and sweet orange (Figure S9A), respectively. Module eigengene is considered a representative gene expression profile in a module and correlated with the corresponding tissue type or trait (Langfelder and Horvath, 2008). We found that 3 out of 12 co-expression modules are comprised of genes that are significantly correlated with Mock- or *C*Las-inoculation, and 7 out of 12 modules are significantly correlated with time points (0, 7, 17, 34 WAI) ($P \leq 0.05$; Figure 4C) in rough lemon. However, no significant modules were correlated with *C*Las inoculation in sweet orange except for three significant modules correlated with the WAI (Figure S9B).

Modules with high trait significance may represent potential pathways associated with the trait, and genes with high module membership in modules significantly correlated with traits are hub candidate genes (Langfelder and Horvath, 2008). In this study, scatterplots showed that genes with high module membership in 8 modules in rough lemon (Figure 4D and Figure S10) and three modules in sweet orange (Figure S9C) were identified as potential hub genes with high significance ($P < 0.01$), including a total of 130 lncRNAs and 61 lncRNAs in rough lemon and sweet orange, respectively (Table S9). Among these lncRNAs, 23 lncRNAs were predicted to function as miRNA targets, miRNA precursors, or eTM (Table S10).

A heatmap showing the relative normalized FPKM (NFPKM) of lncRNAs revealed that most of the co-expression lncRNAs were from brown module correlated with *C*Las-inoculation and significantly upregulated at 7 WAI in rough lemon, and lncRNAs in different modules exhibited highly specific expression (Figure 4E). However, few co-expression lncRNAs showed spatiotemporal-specific expression in sweet orange (Figure 4F). Heatmap of co-expressed mRNAs also showed prominently different expression patterns between rough lemon and sweet orange (Figure S11). Genes in the pink and blue modules were specifically upregulated at 17 WAI and 34 WAI after CLas-inoculation in rough lemon, respectively (Figure S11A), while genes in the blue module of sweet orange were significantly downregulated at 7 WAI and 34 WAI (Figure S11B). These specific DE lncRNAs and mRNAs might specifically contribute to the response to *C*Las infection.

3.5 Functional annotation and enrichment of genes in significant modules

A total of 1146 and 879 were included in the eight significant modules in rough lemon and three significant modules in sweet orange, respectively. GO enrichment indicated that stress response-related biological process terms (such as response to reactive oxygen species, plant-pathogen interaction, and response to stimulus processes) were significantly over-represented in rough lemon (Figure 5A and Figure S12A). However, the development and growth of biological processes, negatively regulated cell proliferation, and response to stimulus terms were significantly over-represented in sweet orange (Figure 5B and Figure S12B). MapMan functional categories related to biotic stress show that redox state, glutathione-S-transferase, and secondary metabolites were significantly enriched in rough lemon (Figure S12C). Notably, genes enriched in the Glutathione-S-transferase category were mainly involved in pathogenic effector-triggered immunity, systemic acquired resistance, and WRKY33-dependent plant immunity in rough lemon, and most of these genes were significantly upregulated (Data S5 and Figure S12C). In sweet orange, most genes were mainly enriched in the redox state and secondary metabolites categories related to cell wall organization, but few genes were enriched in the Glutathione-S-transferase category related to pathogen response (Data S5 and Figure S12D). This result indicates pathogenic response genes in the Glutathione-S-transferase category play an important role in HLB tolerance of rough lemon. In addition, we also found that the functional roles of genes in the redox state and secondary metabolite categories were quite different between rough lemon and sweet orange. Compared with rough lemon, many cell wall related genes were enriched in sweet orange, including genes related to arabinogalactan, callose, and lipid biosynthesis (Data S5). According to the previous study, callose deposition can play a role as a defensive fortification in response to *C*Las bacteria in citrus (Achor et al., 2010). Based on the results of gene function enrichment, we suggest that the

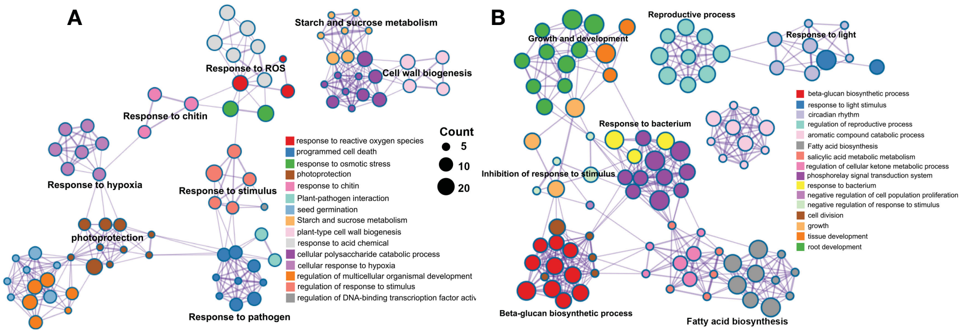

A
Starch and sucrose metabolism
Response to ROS
Cell wall biogenesis
Response to chitin
Response to hypoxia
Response to stimulus
photoprotection
Response to pathogen
Count
5
10
20
🟥 response to reactive oxygen species
🟦 programmed cell death
🟩 response to osmotic stress
■ photoprotection
■ response to chitin
■ Plant-pathogen interaction
■ seed germination
🟧 Starch and sucrose metabolism
■ plant-type cell wall biogenesis
■ response to acid chemical
🟪 cellular polysaccharide catabolic process
■ cellular response to hypoxia
■ regulation of multicellular organismal development
■ regulation of response to stimulus
■ regulation of DNA-binding transcrioption factor activ
B
Growth and development
Reproductive process
Response to light
Response to bacterium
Inhibition of response to stimulus
Beta-glucan biosynthetic process
Fatty acid biosynthesis
🟥 beta-glucan biosynthetic process
🟦 response to light stimulus
■ circadian rhythm
■ regulation of reproductive process
■ aromatic compound catabolic process
■ Fatty acid biosynthesis
■ salicylic acid metabolic metabolism
■ regulation of cellular ketone metabolic process
🟪 phosphorelay signal transduction system
🟨 response to bacterium
■ negative regulation of cell population proliferation
■ negative regulation of response to stimulus
🟫 cell division
■ growth
🟧 tissue development
🟩 root development

FIGURE 5
Pathway and process enrichment analysis of genes in the significant WGCNA modules in **(A)** rough lemon and **(B)** sweet orange. Network of enriched terms colored by cluster ID, where nodes that share the same cluster ID are typically close to each other.

mechanisms of HLB response are quite different between rough lemon and sweet orange.

3.6 Co-expression regulatory network of lncRNA-mRNA in response to *C*Las infection

Based on the results of GO and MapMan functional enrichment, co-expression genes associated with pathogen stimuli response in rough lemon and genes associated with callose synthase in sweet orange were of particular interest (Table S11). Multiple genes (such as NLR, RIN4, and CBP60/SARD) were suggested to play an important role in plant innate immunity (Abramovitch et al., 2006; da Graça et al., 2016). Genes showing the most connections with those immunities might be mutually involved in the pathway of pathogen response.

Therefore, we further identified genes first connected to stimuli response and callose synthase related genes and constructed their potential co-expression regulatory network. A total of 205 and 73 first neighboring genes with weight > 0.2 were identified in rough lemon and sweet orange, respectively (Figure S13A, B and Data S6). Multiple defense-related genes were identified in the network, such as *BOTRYTIS-INDUCED KINASE 1* (*BIK1*)/MSTRG.37943.1 and *NECROTIC SPOTTED LESIONS 1* (*NLS1*)/MSTRG.17571.7 in rough lemon; peroxidase superfamily protein (*PRX52*)/*MSTRG.57748.1* and subtilisin-like protease (*SBT1.5*)/*MSTRG.58400.1* in sweet orange. GO enrichment showed that the neighboring genes in rough lemon and sweet orange were significantly enriched in plant-type hypersensitive response and cutin biosynthetic process, respectively (Figure S13C-D). In addition, seven pathogen-related genes derived from MapMan pathogen stimuli response category showed a strong association with each other (Figure S13A). Among these seven genes, effector-triggered immunity related *RPM1-interacting factor 4* (*RIN4*)/*MSTRG.29058.1* and WRKY33-dependent plant immunity related SIGMA FACTOR BINGD PROTEIN (SIB)/Ciclev10002803m.v1.0 with the highest edge weight are suggested to be two key hub genes in the co-expression network.

To know more about the protein-protein interaction (PPI) relationships between these connected genes, proteins of *A. thaliana* were used as model to infer the potential interaction network (Szklarczyk et al., 2015). A total of 43 and 13 PPI genes (edge confidence > 0.4) were identified in rough lemon and sweet oranges, respectively (Figure S14 and Data S7). Most of them in the PPI network were related to stress response (Figure 6A; Figure S14). For instance, WRKY33 (Ciclev10011386m.v1.0) is important for plant pathogen immunity and interacted with multiple disease related proteins in rough lemon. ACS6 is involved in bacterial flagellin-induced ethylene production (Gravino et al., 2015), SYP121 is involved in ABA signaling (Eisenach et al., 2012), NECROTIC SPOTTED LESIONS (NSL1) negatively regulated cell death programs, and salicylic acid-related defense responses; *nsl1* mutants exhibited higher levels of salicylic acid (SA) and callose deposition (Noutoshi et al., 2006). As to the first neighboring genes of callose synthase in sweet orange, LONG-CHAIN ACYL-COA SYNTHETASE 2 (LACS2) is required for the accumulation of cuticular wax to enhance plant stress resistance (Figure S14B), which strongly interacts with CYP86A2 and is involved in the processes of cuticle development and repression of bacterial type III gene (Xiao et al., 2004). Notably, WRKY33 and its interacted SIB (Ciclev10002803m.v1.0) had the highest confidence in the PPI and co-expression networks (Figure 6A and Figure S13A), implying an extent regulatory relationship between them in response to *C*Las infection. Based on the PPI network, genes with high confidence (edge confidence > 0.7) were identified, and a set of them related to pathogenic response, such as effector receptor, systemic acquired resistance protein, callose synthase, and calcium-dependent protein et al., are identified as key genes in response to *C*Las infection (Table 1). The analysis of gene expression showed that pathogen response genes were distinctly upregulated in rough lemon at 17 weeks after *C*Las inoculation, while callose synthase related *CALS1* was not expressed in rough lemon (Figure 6E). Interestingly, effector-triggered immunity related leucine-rich repeat receptor-like protein kinase family protein (*AT4G08850*), pathogen-associated molecular pattern (PAMP)-triggered immunity (PTI) signaling related *CPK7*, and ethylene signaling response gene *ERF-1* were not expressed in sweet orange, which partially explained greater HLB tolerance.

Next, we further investigated the co-expression relationship between pathogen stimuli response genes and lncRNAs. We found three lncRNAs (*LNC_28805*, *LNC_24384*, and *LNC_128248*) that were co-expressed with pathogen response genes in rough lemon (Figure 6B). In this network, *LNC_28805* was a hub lncRNA first connecting to all five pathogen stimuli response genes, suggesting it might be a key regulator involved in response to *C*Las. LncRNAs could regulate gene transcription levels and cellular processes by acting as miRNA precursors or miRNA mimics (Chekanova, 2015). Among the three lncRNAs, *LNC28805* was solely found to act as *miR167* and *miR5021* targets in both rough lemon and sweet orange and exhibited similar expression patterns after *C*Las inoculation (Table S10 and Figure 6C). *LNC28805* was significantly upregulated at 17 WAI in rough lemon. To realize whether other pathogenic response genes could be co-expressed with *LNC28805*, we further identified the first neighboring genes of *LNC28805* with edge weight > 0.2 from rough lemon and sweet orange. A total of 145 differentially expressed transcripts first connect to *LNC28805* in rough lemon (Data S8). However, none of the genes co-expressed with *LNC28805* were found in the co-expression network of sweet orange. Functional enrichment analysis of the 145 transcripts showed that plant defense-related genes were the most significantly enriched (Figure 6D).

3.7 A hypothetical regulatory pathway of *LNC28805* involved in HLB regulation

LNC28805 was not predicted to act as an eTM or miRNA precursor, but it did appear to act as a target for *miR167* and *miR5021*, thus suggesting a role in the regulation of HLB response by interacting with miRNAs. To validate this point, we analyzed the *miR5021* and *miR167* potential targets among the co-expression genes of *LNC28805*, as well as pathogen stimuli response genes and callose synthase genes. A total of 25 transcripts were identified to be targeted by *miR5021* in rough lemon and sweet orange, respectively (Table S12 and Data S9). However, none of the transcripts were identified to be targeted by *miR167*. Most of these genes were involved in disease regulation, such as *WRKY33*, *SYP121*, and *NB-ARC* domain-

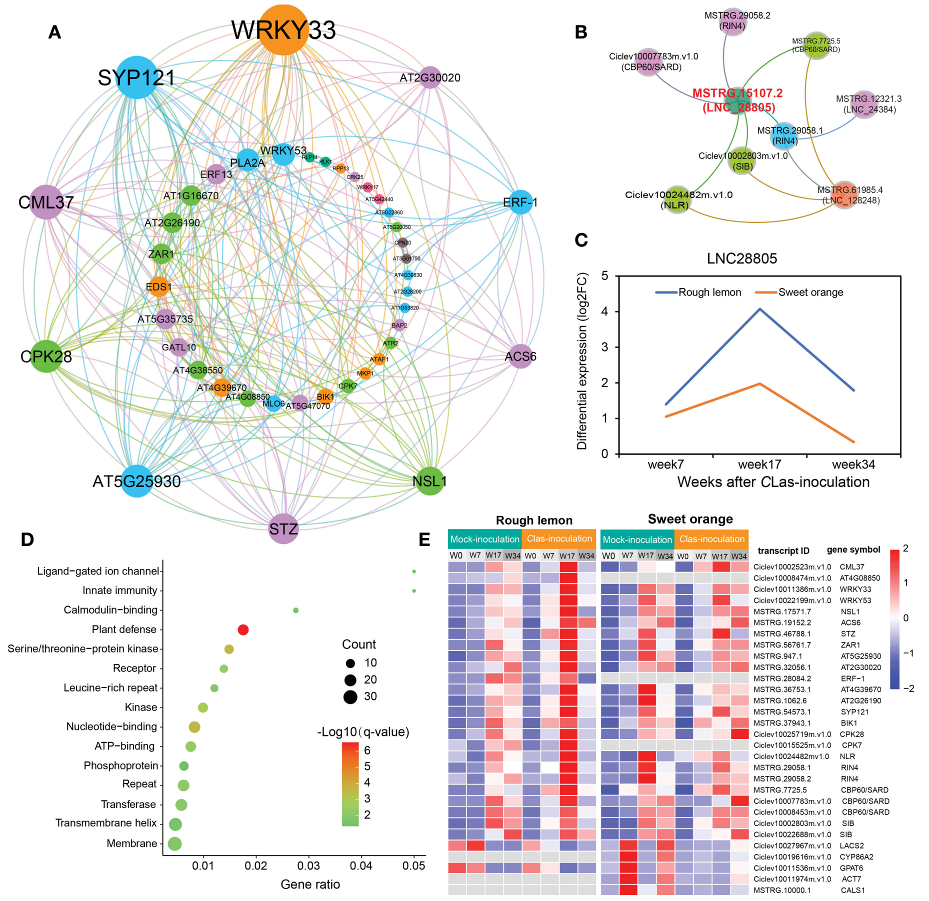


FIGURE 6
Construction of pathogen-related protein-protein interaction (PPI) network and lncRNA-mRNA co-expression network in response to CLas infection. **(A)** PPI network of co-expressed genes based on the prediction of the STRING database with edge confidence > 0.4. *A thaliana* was used as model to infer the PPI network of the co-expressed genes in citrus species. The circle size indicates the edge degree of the adjacent genes in the PPI network; nodes and edges were colored by modularity class based on Gephi software analysis. **(B)** lncRNAs co-expressed with pathogen stimuli response genes. Red font indicates the most significant hub lncRNA. **(C)** Differential expression of *LNC28805* across three different stages after CLas inoculation. FC indicates fold-change. **(D)** Functional enrichment of 145 co-expressed differentially expressed transcripts first connected to *LNC28805*. **(E)** Heatmap showing the expression patterns of co-expressed transcripts related to pathogen response in rough lemon and sweet orange. FPKM were normalized to the Z-score.

TABLE 1 Function and annotation of key genes related to disease response in the co-expression networks.

Transcript ID	Arabidopsis homolog	Description	Function	miRNA target
Ciclev10002523m.v1.0	CML37	Calcium-binding protein CML37	Potential calcium sensor	
Ciclev10008474m.v1.0	AT4G08850	Leucine-rich repeat receptor-like protein kinase family protein	Plant immunity and defense response	
Ciclev10011386m.v1.0	WRKY33	Probable WRKY transcription factor 33	Mediating responses to the bacterial pathogen and the necrotrophic pathogen.	ath-miR5021
Ciclev10022199m.v1.0	WRKY53	Probable WRKY transcription factor 53	Regulate the early events of leaf senescence	
MSTRG.17571.7	NSL1	MAC/Perforin domain-containing protein	Negatively regulating salicylic acid-related defense responses and cell death programs	–
MSTRG.19152.2	ACS6	1-aminocyclopropane-1-carboxylic acid (acc) synthase 6	Involved in bacterial flagellin-induced ethylene production	ath-miR5021
MSTRG.46788.1	STZ	Related to Cys2/His2-type zinc-finger proteins	Acts as a transcriptional repressor and is responsive to chitin oligomers	–

(Continued)

TABLE 1 Continued

Transcript ID	Arabidopsis homolog	Description	Function	miRNA target
MSTRG.56761.7	ZAR1	Disease resistance RPP13-like protein 4	CC-NB-LRR receptor-like protein required for recognition of the Pseudomonas syringae type III effector HopZ1a	–
MSTRG.947.1	AT5G25930	Protein kinase family protein with leucine-rich repeat domain	Involved in protein amino acid phosphorylation; a crucial component of early immune responses	–
MSTRG.32056.1	AT2G30020	Protein phosphatase 2C family protein	Negatively regulates defense response; inactivates MPK4 and MPK6	–
MSTRG.28084.2	ERF-1	Ethylene responsive element binding factor 1	Ethylene signaling response	–
MSTRG.36753.1	AT4G39670	Glycolipid transfer protein (GLTP) family protein	Involved in glycolipid transport	ath-miR5021
MSTRG.1062.6	AT2G26190	IQ domain-containing protein IQM4	Involved in biotic and abiotic stress responses	–
MSTRG.54573.1	SYP121	Syntaxin of plants 121	A component of a complex of SNARE proteins that plays a role in ABA signaling and against fungal invaders	ath-miR5021
MSTRG.37943.1	BIK1	Serine/threonine-protein kinase BIK1	Required to activate the resistance responses to necrotrophic pathogens	
Ciclev10025719m.v1.0	CPK28	Calcium-dependent protein kinase 28	Involved in pathogen-associated molecular pattern (PAMP)-triggered immunity (PTI) signaling	–
Ciclev10015525m.v1.0	CPK7	Calcium-dependent protein kinase 7	Involved in pathogen-associated molecular pattern (PAMP)-triggered immunity (PTI) signaling	–
Ciclev10024482m.v1.0	NLR	Disease resistance protein (TIR-NBS-LRR class) family	Effector receptor; involved in signal transduction, defense response, apoptosis, innate immune response;	
MSTRG.29058.1 MSTRG.29058.2	RIN4	RPM1-interacting factor	The plant immune regulator; required for activation of RPM1-dependent inhibition of bacterial growth.	
MSTRG.7725.5 Ciclev10007783m.v1.0 Ciclev10008453m.v1.0	CBP60/SARD	calmodulin binding protein 60 (CBP60) family transcription factors	Systemic acquired resistance (SAR) positively regulate immunity	
Ciclev10002803m.v1.0 Ciclev10022688m.v1.0	SIB	Sigma factor binding protein (SIB)	Stimulate the DNA binding activity of WRKY33	
Ciclev10027967m.v1.0	LACS2	Long chain acyl-CoA synthetase 2	Activation of long-chain fatty acids for both synthesis of cellular lipids, and degradation *via* beta-oxidation	–
Ciclev10019616m.v1.0	CYP86A2	Cytochrome P450, family 86, subfamily A, polypeptide 2	Involved in the biosynthesis of hydroxylated fatty acids r, cuticle development and repression of bacterial type III gene expression	–
Ciclev10011536m.v1.0	GPAT6	Glycerol-3-phosphate 2-O-acyltransferase 6	Esterifies acyl-group from acyl-ACP to the sn-2 position of glycerol-3-phosphate, a step in cutin biosynthesis	–
Ciclev10011974m.v1.0	ACT7	Actin-7	Involved in the regulation of hormone-induced plant cell proliferation and callus formation	–
MSTRG.10000.1	CALS1	Callose synthase 1	Involved in callose synthesis at the forming cell plate during cytokinesis	–

containing disease resistance protein (Table S12). Interestingly, we found eight genes targeted by *miR5021* were found with the QTLs identified in our previous study (Huang et al., 2018) (Table S13). Intriguingly, QTLs on scaffold_6 were simultaneously detected for foliar symptoms (FS) and canopy damage (CD) in two different years, and there two key hub genes *(WRKY33* and *SYP121*) in the PPI network, strongly linked to the QTL peak markers, were found (Figure 7A). This result further supports a putatively important regulatory relationship between *LNC28805* and these disease response genes. It is tempting to speculate that *LNC28805* might act as competing endogenous *miR5021* to regulate HLB response genes by attenuating their cleavage or translated inhibition caused by *miR5021*. The mechanism of miRNA-LncRNA interactions regulating host immunity-related genes has been reported in tomato (Jiang et al., 2020). Overall, a hypothetical model of *LNC28805* and its potential role in regulatory processes was proposed (Figure 7B). This model showed that WRKY33 might regulate cross-talk between jasmonate-, abscisic acid (ABA)-, and salicylic acid (SA)-regulated disease response pathways (Zheng et al., 2006; Birkenbihl et al., 2012; Liu et al., 2015).

To explore if the regulatory relationship also exists in the plants under HLB stress for a long term in the field, we further tested the

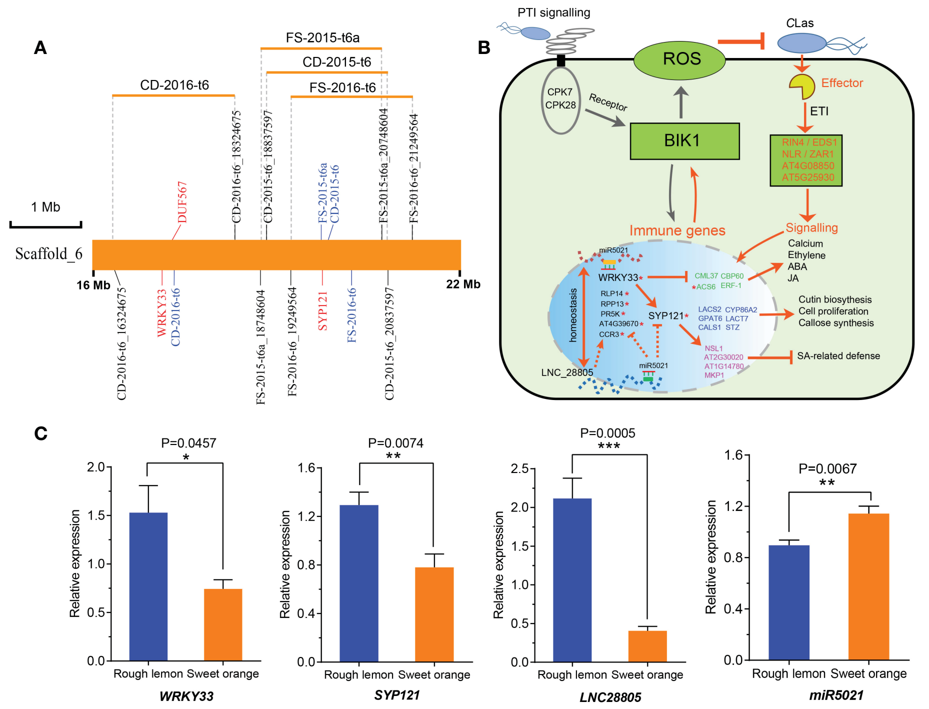


FIGURE 7

A hypothetical regulatory pathway of *LNC28805* involved in citrus HLB regulation. **(A)** Most significant hub genes targeted by *miRNA5021* located in an overlapping QTL identified by Huang et al., 2018. Word colored by blue indicates the QTL peak marker; CD, canopy damage; FS, foliar symptom. **(B)** A hypothetical regulatory model depicting the *LNC28805* and co-expression genes related to disease response based on protein-protein interaction (PPI) networks (shown in Figure 6A and Figure S14) and reported studies. This network indicated that *WRKY33* and defense related genes targeted by *miR5021* putatively regulate cross-talk between jasmonate-, abscisic acid (ABA)-, and salicylic acid (SA)-regulated pathways (Zheng et al., 2006; Birkenbihl et al., 2012; Liu et al., 2015). *LNC28805* might importantly compete with endogenous *miR5021* to maintain the homeostasis of these immune gene expression levels. Red asterisks indicate *miR5021* targets. **(C)** Relative expression of *miRNA5021* and its targets in rough lemon and sweet orange under Huanglongbing (HLB) stress more than ten years in the field (shown in Figure S15). Asterisks indicate significant differences (Students' t-test, *p < 0.05, **p < 0.01, ***p < 0.001).

expression levels of *miR5021* and its targets (*WRKY33*, *SYP121*, and *LNC28805*) using midribs of rough lemon and sweet orange under HLB stress more than ten years in the field (Figure S15). We found *LNC28805* was significantly upregulated, and *miR5021* was significantly downregulated in rough lemon, which supports the hypothesis that *LNC28805* is involved in regulating the expression level of pathogenic response genes by competing for endogenous *miR5021*.

4 Discussion

It has been known that lncRNAs act as versatile regulators involved in diverse biological processes in plants. In citrus, thousands of lncRNAs from different species have been identified in flowers, leaves, seeds, and fruits based on RNA-seq transcriptome data (Xu et al., 2018). However, lncRNAs have not been systematically identified in response to citrus HLB, and little is known about their biological function during the response to the disease. Here, we present a comprehensive picture of lncRNAs and mRNAs transcription patterns in different time points after CLas-inoculation in HLB-tolerant rough lemon and HLB-sensitive sweet orange. A set of DE lncRNAs exhibited stage-specific and species-specific expression, implying specific roles for lncRNAs in response to infection of citrus with CLas, leading to HLB.

4.1 Identification, conservation, and phylogenetic relationship of lncRNAs

In this study, we have identified 8742 high confident lncRNAs from mock- and CLas-inoculated rough lemon and sweet orange. There are more than 50% lncRNAs from intergenic regions, which is consistent with a study of lncRNAs and flowering in trifoliate orange (Wang et al., 2017). We also identified 2529 novel lncRNAs by aligning lncRNA sequences to the annotated lncRNA database (Szcześniak et al., 2016; Jin et al., 2021). These lncRNAs might specifically respond to HLB disease and need more investigation. Therefore, our study has expanded the information on disease-related lncRNAs in citrus. Compared to protein-coding genes, lncRNAs identified from rough lemon and sweet orange shared most of the common features of lncRNAs reported in other citrus species and model plants (Di et al., 2014; Yuan et al., 2018; Ke et al., 2019), such as low conservation, fewer exons, shorter transcript length, and lower expression levels. Similar features in different species might be explained by the high rates of origin sequence evolution (Mercer et al., 2009). These results are consistent with previous studies in other species (Ke et al., 2019; Tang et al., 2021), also supporting the reliability of lncRNAs identified in this study. Analysis of the primary sequence evolutionary conservation by comparing with *A. thaliana*, *O. sativa*, and *P. trichocarpa* lncRNAs showed that only 33 lncRNAs derived from 22 genomic loci were identified. Compared with *Arabidopsis* and rice, a relatively high proportion of conserved lncRNAs was found in *P. trichocarpa*, indicating high sequence similarity in a more closely related taxon. The phylogenetic tree of 38 citrus relatives also displayed high conservation of these lncRNAs genomic sequence at the intra-species level (Figure 3G), suggesting that these conserved lncRNAs might share a common ancestor and ancient evolutionary origin at genomic levels (Ulitsky, 2016). Similar results were also observed in animals and other plants (Pang et al., 2006; Di et al., 2014). Intriguingly, correlation analysis between SNPs derived from the conserved lncRNAs indicated 26 SNPs were significantly associated with HLB tolerance, and six conserved lncRNAs contained these SNPs were significantly upregulated at W7 time point after *C*Las-inoculation (Figure 3E and Table S7). Some studies reported that conserved lncRNAs might play an important role in conserved functions, such as interacting with RNA-binding proteins by conserving secondary structures (Wang et al., 2014; Ulitsky, 2016). Although experimental evidence about the biogenesis and functions of conserved lncRNAs is limited, most current studies suggest specific secondary structures of conserved lncRNAs might be important for a conserved function (Johnsson et al., 2014; Ulitsky, 2016). More extensive experiments are needed to validate whether these conserved lncRNAs have important contributions to HLB response at an early stage in our future studies.

4.2 Tissue-specific and species-specific expression pattern of lncRNAs

It is difficult to deduce and validate the lncRNA function due to its low conservation and non-coding property. The tissue-specific expression may help to understand the potential functions underlying lncRNAs. In the current study, lncRNAs exhibited more prominent species-specific expression patterns compared with spatiotemporal-specific expression patterns (Figure 2D). Pairwise comparisons between mock- or *C*Las-inoculated tissues or different time points revealed that the number of DE lncRNAs and mRNAs distinctly increased at week 7 and decreased at week 17 after CLas inoculation in sweet orange (Figure 2A). However, it did not change significantly in rough lemon. The greater numbers of DE lncRNAs and mRNAs in sweet orange may imply greater sensitivity and activity in response to external stimuli at transcriptional regulation levels. Changes in the environment and rhythmic plant growth also can affect plant gene expression levels (López-Maury et al., 2008; Covington et al., 2008). The pairwise comparison of lncRNAs between mock- and *C*Las-inoculated plants was further performed to reduce the environmental noise, presenting obvious differences of lncRNA number at 34 WAI between rough lemon and sweet orange (79 upregulated lncRNAs in rough lemon and nine upregulated lncRNAs in sweet orange). Moreover, most of these lncRNAs presented high species-specificity. These results indicated that species-specific DE lncRNA might play an essential role in the regulation of HLB tolerance in rough lemon. High tissue- and species-specific expression of lncRNAs, such as *LNC_57342* and *LNC_61191* specifically expressed in rough lemon and sweet orange at W7 time point, respectively, also make them potentially useful as biomarkers for HLB detection or screening HLB tolerant species in early stages. Based on the aforementioned results, we suggest these specific DE lncRNAs might contribute to citrus HLB tolerance or sensitivity and imply important dynamics of lncRNAs response to HLB in citrus.

4.3 Putative interaction between DE lncRNAs and miRNAs

The lncRNA-miRNA interaction is a vital regulatory mechanism of lncRNAs. Several studies have shown that lncRNAs can serve as miRNA and siRNA precursors to assist in target gene cleavage or translation inhibition, function as traps for miRNA binding, or are directly targeted by miRNA to attenuate miRNA presence in plant immunity (Song et al., 2021). For instance, lncRNA *MuLnc1* was identified to be cleaved by *mul-miR3954* to produce *si16157* and negatively regulated *Botrytis cinerea* and *Pseudomonas syringae* resistance by inhibiting the functions of *calmodulin-like protein 27* (*CML27*) in mulberry (Gai et al., 2018). In tomato, *lncRNA42705* and *lncRNA08711* increase the expression levels of MYB genes by acting as decoys for *miR159* and enhance resistance against *P. infestans* (Cui et al., 2019); and, *Sl-lncRNA15492* was targeted by *SI-miR482a* to maintain *Sl-NBS-LRR1* at an appropriate expression level during the immune response to *P. infestans* (Jiang et al., 2020). In this study of citrus and HLB, we identified a set of high confidence lncRNAs that potentially serve as precursors, eTMs, and targets of miRNAs. Among them, it is noteworthy that five lncRNAs simultaneously serve as eTMs and targets of miRNA, and eleven lncRNAs simultaneously serve as precursors and miRNA targets (Table S8). Notably, nine DE lncRNAs were specifically identified in sweet orange. Several miRNAs that interacted with these lncRNAs have been reported to be involved in disease resistance, such as *miR858* negatively regulated *Arabidopsis* immunity (Camargo-Ramírez et al., 2018), *miR477* enhanced the susceptibility of the tea plants to *Pseudopestalotiopsis* species infection (Wang et al., 2020), and *miRNA482* suppressed the expression of NBS-LRR defense genes in cotton (Zhu et al., 2014). Interestingly, *LNC_40405* and *LNC_69103* were predicted to act as eTMs of *miR477* and *miR482*, respectively, and were significantly downregulated in *C*Las-inoculated sweet orange at 7 WAI and 17 WAI. In addition, multiple DE lncRNAs were also targeted by *miR2111* and *miR5021*. A recent study indicated that *miR2111* positively regulates shoot-to-root systemic effectors of rhizobia and promotes nodule formation (Moreau et al., 2021). Surprisingly, *miR5021* was also predicted to be one having maximal matches against 19910 ESTs from periwinkle (*Catharanthus roseus*) (Pani and Mahapatra, 2013), which is an alternate host of the HLB bacteria (Zhang et al., 2010). Most of the predicted *miR5021* targets were the genes involved in cell growth and development, signaling, and metabolism in periwinkle. In citrus, we also found multiple genes associated with disease response were targeted by *miR5021* (Table S12). *LNC_69103* was not only acting as an eTM of *miR482* but also acting as a target of *miR5021*. Compared to healthy plants, *LNC_69103* was extremely downregulated (log2FC < -9) in *C*Las-infected sweet orange plants. These findings indicated the expression level of these lncRNAs might be tightly related to the sensitivity of citrus plants to HLB.

4.4 A potential regulatory model of *LNC_28805* in response to CLas

One of the main objectives of this study was to understand the expression dynamics and co-expression networks responding to citrus HLB in tolerant rough lemon and sensitive sweet orange. The WGCNA co-expression network revealed inoculation and temporal specific modules and important hub genes involved in citrus HLB response. Brown module was a unique significant module associated with *C*Las-inoculation in rough lemon and included a large group of co-expressed lncRNAs. Cluster analysis of expression pattern showed that lncRNAs displayed more temporal specific expression patterns than mRNAs and were significantly upregulated at 7 WAI (Figure 4E), suggesting that the functions of these lncRNAs might be closely related to HLB response at early stages after *C*Las-inoculation. We speculate that HLB-tolerant genes in rough lemon might be mediated by these early response lncRNAs. Evidence from most disease response genes exhibiting strong protein-protein interactions was identified in the blue module, but not the brown module, and these genes were significantly upregulated at 17 WAI, though not at 7 WAI (Figure S11 and Figure 6E). Functional enrichment analysis indicated that co-expressed genes in rough lemon were mainly responsible for the response to reactive oxygen species, programmed cell death, and plant pathogen, which are highly related to disease defense (Figure 6E). Although genes in response to bacteria were also enriched in sweet orange, most of them were responsible for growth and development, fatty acids biosynthesis, and reproductive processes (Figure 5B). A fraction of genes was even associated with negative regulation of response to stimulus in sweet orange. These results also reflect the greater sensitivity of sweet orange to HLB than rough lemon.

Additionally, MapMan analysis identified multiple genes involved in pathogenic effector-triggered immunity, systemic acquired resistance, WRKY33-dependent plant immunity, and callose synthase. Their first neighboring genes in the co-expression network were highly correlated with pathogen response and exhibited strong interactions with each other. Interestingly, *WRKY33* and *SYP121* were two key hub genes with the highest edge weight in the PPI network (Figure 6A), and both of them are located in significant QTLs identified by the previous study (Huang et al., 2018), suggesting that these two genes might play important roles in HLB regulation. In addition, we found a hub *LNC_28805* was the most probable lncRNA involved in HLB regulation, which was co-expressed with multiple genes associated with the effector-triggered immunity (ETI) network, including *RIN4*, *CBP60*, *SIB*, and *NLR*. Moreover, *LNC_28805* has 145 first neighboring DE mRNAs (including *WRKY33* and *SYP121*) in the co-expression network. These neighboring genes were the most significantly enriched in the plant defense component (Figure 6D). Meanwhile, we also found that *LNC_28805*, belonging to intergenic lncRNAs type, was predicted to be targeted by the homologous *miR5021* of *Arabidopsis* with one mismatch, indicating that *LNC_28805* may be an evolutionarily conserved and functionally maintained lncRNA. We notably found that multiple disease resistance genes were targeted by *miR5021*, such as *WRKY33*, *SYP121*, and *NB-ARC* domain-containing disease resistance genes (Table S12), suggesting *LNC28805* might compete with endogenous *miR5021* to maintain the homeostasis of expression levels between immune-related genes and growth genes. All these results further indicate that *LNC28805* might be an important regulator involved in the HLB tolerance of rough lemon.

Plant WRKY transcription factors (TF) play key roles in plant responses to microbial infection. The PPI network showed WRKY33 was involved in multiple disease response pathways (Figure 6A and Figure S14). WRKY33 can regulate the expression of defense-related genes toward the necrotrophic fungus *B. cinerea*, but WRKY33 has also been shown to regulate cross-talk between jasmonate-, abscisic acid (ABA)-, and salicylic acid (SA)-regulated disease response pathways (Zheng et al., 2006; Birkenbihl et al., 2012; Liu et al., 2015). In *Arabidopsis*, the ectopic expression of *WRKY33* results in enhanced susceptibility to the bacterial pathogen *P. syringae* caused by the reduced expression of the salicylate-regulated PR-1 gene (Zheng et al., 2006). Loss of WRKY33 function results in activation of the ABA- and salicylic acid (SA)-related host response (Birkenbihl et al., 2012; Liu et al., 2015). The evidence indicated that high expression *WRKY33* might suppress the expression of ABA- or SA-regulated genes response to HLB bacteria (Figure 7B). WRKY33 also can activate the expression of RING-type ubiquitin ligase *ATL31* involved in vesicle trafficking with PEN1/SYP121 SNARE protein (Reyes et al., 2015), which function to guard cell membrane transport and stomatal control (Eisenach et al., 2012). Some others candidate genes in the PPI networks, such as *ACS6*, *ERF-1*, *CML37*, *CALS1*, *NSL1*, and *MPK1* involved in bacteria-induced ethylene, ABA, and JA signaling, callose synthesis, and SA-related defense (Tsuda and Katagiri, 2010; Luna et al., 2010; Fukunaga et al., 2017), were also putatively involved the defense response of *C*Las invasion (Figures 6A and 7B). Based on the regulatory relationship of the genes in the networks, we suggested that *LNC_28805* may play an important role in maintaining the homeostasis of antagonistic relationship between defense pathways mediating WRKY33 associated with ABA- or SA-regulated genes involved *C*Las response (Zheng et al., 2006; Liu et al., 2015).

In addition, WRKY33 also can activate the expression of *Arabidopsis* RING-type ubiquitin ligase *ATL31* involved in vesicle trafficking with PEN1/SYP121 SNARE protein (Reyes et al., 2015), which functions to guard cell membrane transport and stomatal control (Eisenach et al., 2012). Because *LNC28805* and some of its co-expressed genes were targeted by *miRNA5021*, we suggest that *LNC28805* is probably involved in regulating the expression level of pathogenic response genes by competing for endogenous *miR5021*. If this hypothetical mechanism for HLB tolerance is correct, it should also be observed in plants grown in the field. The expression levels of *miR5021* targets (*WRKY33*, *SYP121*, *LNC28805*) presented significantly higher in rough lemon than in sweet orange under HLB stress for more than ten years in the field (Figure 7C). However, the expression levels of *miRNA5021* were reversed between them. It indicates that this relationship also exists in the naturally infected plants grown in the field. Thus, we suggest that *miR5021* targeting *WRKY33* and *SYP121* might promote the expression of genes responding to *C*Las. Dynamic expression of *WRKY33* might be required to balance the expression levels between immune-related genes and growth genes. *LNC28805* probably plays an important role in regulating the expression level of these pathogenic response genes by competing for endogenous *miR5021*.

Though callose deposition plays a role in defense against the pathogen, overaccumulation of callose inhibits phloem transport activities in Clas-infected citrus (Achor et al., 2010). According to the previous study, callose-plugged phloem sieve elements were less serious in HLB-diseased rough lemon than in HLB-diseased sweet orange (Fan et al., 2012). A potential possibility of HLB tolerance mechanisms of rough lemon might be that *LNC28805* is involved in competing for endogenous *miR5021* to promote the expression of *WRKY33* and *SYP121*, which might function to suppress the immune-related genes overresponse to *C*Las infection and enhance the activity of phloem transport by reducing callose deposition (Fan et al., 2012; Deng et al., 2019). Taken together, our results not only represent the gene modules of lncRNAs and mRNAs related to pathogenic response but also bring new insights into the roles of lncRNAs acting as potential regulatory factors for citrus HLB tolerance.

5 Conclusion

To conclude, we systematically identified and characterized 8,742 lncRNAs among HLB-tolerant rough lemon and HLB-sensitive sweet orange from different time points after *C*Las-inoculation. Based on the integrated analysis of sequence conservation and variation, spatiotemporal-specific expression, functional enrichment, and lncRNA-mRNA co-expression networks with WGCNA, we identified a fraction of lncRNAs and mRNAs that were potentially responsive to *C*Las bacterium infection in citrus. *LNC_28805* was identified as one of the most important candidate lncRNAs involved in citrus HLB regulation. Two key candidates (*WRKY33* and *SYP121*) in the PPI network are known to negatively regulate bacteria pathogen responses and were found within overlapping QTLs identified in our previous study. Based on the reported studies and PPI network and gene co-expression networks in this study, a putative hypothesis for the regulatory pathway of *LNC_28805* is proposed (Figure 7B). This study will be useful in understanding the role of lncRNAs involved in citrus HLB regulation and provide a foundation for further investigation of their regulatory functions.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material. The raw data of RNA sequencing presented in the study are deposited in the NCBI repository, GEO accession number is GSE215306.

Author contributions

Conceptualization, XZ, QY, and FG; Data curation, XZ, QY, and RR; Methodology, XZ, QY, RR, YZ, XW, PH, and YW; Formal analysis, XZ, QY, and RR; Funding acquisition, FG; Investigation, XZ, QY, YZ, RR, and XW; Software and visualization, XZ and RR; Writing—original draft, XZ; and Writing—review and editing, XZ, QY, PH, RR, YW, and FG. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by grants from the Citrus Research and Development Foundation (CRDF 15-010 and 18-011) and the USDA-NIFA-SCRI-HLB Research Program (2015–70016–23027).

Acknowledgments

We thank Misty Holt from the Gmitter lab for providing technical support in the greenhouse and field.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1090711/full#supplementary-material

References

Abramovitch, R. B., Anderson, J. C., and Martin, G. B. (2006). Bacterial elicitation and evasion of plant innate immunity. *Nat. Rev. Mol. Cell Biol.* 7 (8), 601–611. doi: 10.1038/nrm1984

Achor, D. S., Etxeberria, E., Wang, N., Folimonova, S. Y., Chung, K. R., and Albrigo, L. G. (2010). Sequence of anatomical symptom observations in citrus affected with huanglongbing disease. *Plant Pathol. J.* 9 (2), 56–64. doi: 10.3923/ppj.2010.56.64

Albrecht, U., and Bowman, K. D. (2008). Gene expression in citrus sinensis (L.) osbeck following infection with the bacterial pathogen candidatus liberibacter asiaticus causing huanglongbing in Florida. *Plant Sci.* 175 (3), 291–306. doi: 10.1016/j.plantsci.2008.05.001

Alvarez, S., Rohrig, E., Solís, D., and Thomas, M. H. (2016). Citrus greening disease (Huanglongbing) in Florida: Economic impact, management and the potential for biological control. *Agric. Res.* 5 (2), 109–118. doi: 10.1007/s40003-016-0204-z

Bardou, F., Ariel, F., Simpson Craig, G., Romero-Barrios, N., Laporte, P., Balzergue, S., et al. (2014). Long noncoding RNA modulates alternative splicing regulators in arabidopsis. *Dev. Cell* 30 (2), 166–176. doi: 10.1016/j.devcel.2014.06.017

Bastian, M., Heymann, S., and Jacomy, M. (Eds.) (2009). Gephi: an open source software for exploring and manipulating networks. *Proceedings of the international AAAI conference on web and social media*. 3(1), 361–36. doi: 10.1609/icwsm.v3i1.13937

Birkenbihl, R. P., Diezel, C., and Somssich, I. E. (2012). Arabidopsis WRKY33 is a key transcriptional regulator of hormonal and metabolic responses toward *Botrytis cinerea* infection. *Plant Physiol.* 159 (1), 266–285. doi: 10.1104/pp.111.192641

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31 (1), 365–370. doi: 10.1093/nar/gkg095

Bové, J. M. (2006). Huanglongbing: A destructive, newly-emerging, century-old disease of citrus. *J. Plant Pathology.* 88 (1), 7–37. Available at: http://www.jstor.org/stable/41998278

Camargo-Ramírez, R., Val-Torregrosa, B., and San Segundo, B. (2018). MiR858-mediated regulation of flavonoid-specific MYB transcription factor genes controls resistance to pathogen infection in arabidopsis. *Plant Cell Physiol.* 59 (1), 190–204. doi: 10.1093/pcp/pcx175

Chekanova, J. A. (2015). Long non-coding RNAs and their functions in plants. *Curr. Opin. Plant Biol.* 27, 207–216. doi: 10.1016/j.pbi.2015.08.003

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13 (8), 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17), i884–ii90. doi: 10.1093/bioinformatics/bty560

Covington, M. F., Maloof, J. N., Straume, M., Kay, S. A., and Harmer, S. L. (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol.* 9 (8), R130. doi: 10.1186/gb-2008-9-8-r130

Cui, J., Jiang, N., Hou, X., Wu, S., Zhang, Q., Meng, J., et al. (2019). Genome-wide identification of lncRNAs and znalysis of ceRNA networks during tomato resistance to phytophthora infestans. *Phytopathology®* 110 (2), 456–464. doi: 10.1094/PHYTO-04-19-0137-R

Cui, J., Luan, Y., Jiang, N., Bao, H., and Meng, J. (2017). Comparative transcriptome analysis between resistant and susceptible tomato allows the identification of lncRNA16397 conferring resistance to *Phytophthora infestans* by co-expressing glutaredoxin. *Plant J.* 89 (3), 577–589. doi: 10.1111/tpj.13408

da Graça, J. V., Douhan, G. W., Halbert, S. E., Keremane, M. L., Lee, R. F., Vidalakis, G., et al. (2016). Huanglongbing: An overview of a complex pathosystem ravaging the world's citrus. *J. Integr. Plant Biol.* 58 (4), 373–387. doi: 10.1111/jipb.12437

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psRNATarget: A plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* 46 (W1), W49–W54. doi: 10.1093/nar/gky316

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics.* 27 (15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Deng, H., Achor, D., Exteberria, E., Yu, Q., Du, D. , Stanton, D., et al. (2019). Phloem regeneration is a mechanism for huanglongbing-tolerance of "Bearss" lemon and "LB8-9" sugar belle mandarin. *Front. Plant Science.* 10, 277. doi: 10.3389/fpls.2019.00277

Di, C., Yuan, J., Wu, Y., Li, J., Lin, H., Hu, L., et al. (2014). Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. *Plant J.* 80 (5), 848–861. doi: 10.1111/tpj.12679

Eisenach, C., Chen, Z.-H., Grefen, C., and Blatt, M. R. (2012). The trafficking protein SYP121 of *Arabidopsis* connects programmed stomatal closure and k+ channel activity with vegetative growth. *Plant J.* 69 (2), 241–251. doi: 10.1111/j.1365-313X.2011.04786.x

Fan, J., Chen, C., Yu, Q., Khalaf, A., Achor, D. S., Brlansky, R. H., et al. (2012). Comparative transcriptional and anatomical analyses of tolerant rough lemon and susceptible sweet orange in response to '*Candidatus* liberibacter asiaticus' infection. *Mol. Plant-Microbe Interactions®.* 25 (11), 1396–1407. doi: 10.1094/MPMI-06-12-0150-R

Folimonova, S. Y., Robertson, C. J., Garnsey, S. M., Gowda, S., and Dawson, W. O. (2009). Examination of the responses of different genotypes of citrus to huanglongbing (citrus greening) under different conditions. *Phytopathology®.* 99 (12), 1346–1354. doi: 10.1094/PHYTO-99-12-1346

Fukunaga, S., Sogame, M., Hata, M., Singkaravanit-Ogawa, S., Piślewska-Bednarek, M., Onozawa-Komori, M., et al. (2017). Dysfunction of arabidopsis MACPF domain protein activates programmed cell death *via* tryptophan metabolism in MAMP-triggered immunity. *Plant J.* 89 (2), 381–393. doi: 10.1111/tpj.13391

Gai, Y.-P., Yuan, S.-S., Zhao, Y.-N., Zhao, H.-N., Zhang, H.-L. , and Ji, X.-L. (2018). A novel lncRNA, *muLnc1*, associated with environmental stress in mulberry (*Morus multicaulis*). *Front. Plant Science.* 9, 669. doi: 10.3389/fpls.2018.00669

Gao, C., Sun, J., Dong, Y., Wang, C., Xiao, S., Mo, L., et al. (2020). Comparative transcriptome analysis uncovers regulatory roles of long non-coding RNAs involved in resistance to powdery mildew in melon. *BMC Genomics* 21 (1), 125. doi: 10.1186/s12864-020-6546-8

Godfrey, P. M., Ed, S., Chandrika, R., Manjunath, L. K., and Richard, F. L. (2017). Apparent tolerance to huanglongbing in citrus and citrus-related germplasm. *HortScience horts.* 52 (1), 31–39. doi: 10.1111/jipb.12437

Graham, J., Gottwald, T., and Setamou, M. (2020). Status of huanglongbing (HLB) outbreaks in Florida, California and Texas. *Trop. Plant Pathology.* 45 (3), 265–278. doi: 10.1007/s40858-020-00335-y

Gravino, M., Savatin, D. V., Macone, A., and De Lorenzo, G. (2015). Ethylene production in *Botrytis cinerea-* and oligogalacturonide-induced immunity requires calcium-dependent protein kinases. *Plant J.* 84 (6), 1073–1086. doi: 10.1111/tpj.13057

Hong, Y., Zhang, Y., Cui, J., Meng, J., Chen, Y., Zhang, C., et al. (2022). The lncRNA39896–miR166b–HDZs module affects tomato resistance to phytophthora infestans. *J. Integr. Plant Biol.* 64 (10), 1979–1993. doi: 10.1111/jipb.13339

Huang, M., Roose, M. L., Yu, Q., Du, D., Yu, Y., Zhang, Y., et al. (2018). Construction of high-density genetic maps and detection of QTLs associated with huanglongbing tolerance in citrus. *Front. Plant Science.* 9, 1694. doi: 10.3389/fpls.2018.01694

Jiang, N., Cui, J., Hou, X., Yang, G., Xiao, Y., Han, L., et al. (2020). Sl-lncRNA15492 interacts with sl-miR482a and affects solanum lycopersicum immunity against phytophthora infestans. *Plant J.* 103 (4), 1561–1574. doi: 10.1111/tpj.14847

Jiang, N., Cui, J., Shi, Y., Yang, G., Zhou, X., Hou, X., et al. (2019). Tomato lncRNA23468 functions as a competing endogenous RNA to modulate NBS-LRR genes by decoying miR482b in the tomato-phytophthora infestans interaction. *Horticulture Res.* 6, 28. doi: 10.1038/s41438-018-0096-0

Jin, J., Lu, P., Xu, Y., Li, Z., Yu, S., Liu, J., et al. (2021). PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. *Nucleic Acids Res.* 49 (D1), D1489–D1D95. doi: 10.1093/nar/gkaa910

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K. V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta (BBA) - Gen. Subjects.* 1840 (3), 1063–1071. doi: 10.1016/j.bbagen.2013.10.035

Kang, S.-H., Sun, Y.-D., Atallah, O. O., Huguet-Tapia, J. C., Noble, J. D., and Folimonova, S. Y. (2019). A long non-coding RNA of *Citrus tristeza virus*: Role in the virus interplay with the host immunity. *Viruses.* 11 (5), 436. doi: 10.3390/v11050436

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., et al. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45 (W1), W12–WW6. doi: 10.1093/nar/gkx428

Ke, L., Zhou, Z., Xu, X.-W., Wang, X., Liu, Y., Xu, Y., et al. (2019). Evolutionary dynamics of lincRNA transcription in nine citrus species. *Plant J.* 98 (5), 912–927. doi: 10.1111/tpj.14279

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37 (8), 907–915. doi: 10.1038/s41587-019-0201-4

Kim, E.-D., and Sung, S. (2012). Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Science.* 17 (1), 16–21. doi: 10.1016/j.tplants.2011.10.008

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: From microRNA sequences to function. *Nucleic Acids Res.* 47 (D1), D155–DD62. doi: 10.1093/nar/gky1141

Kramer, M. F. (2011). Stem-loop RT-qPCR for miRNAs. *Curr. Protoc. Mol. Biol.* 95 (1), 15.0.1–15.0.0. doi: 10.1002/0471142727.mb1510s95

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35 (9), 3100–3108. doi: 10.1093/nar/gkm160

Langfelder, P., and Horvath, S. (2008). WGCNA: An r package for weighted correlation network analysis. *BMC Bioinf.* 9 (1), 559. doi: 10.1186/1471-2105-9-559

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: Smartmodel selection in PhyML. *Mol. Biol. Evolution.* 34 (9), 2422–2424. doi: 10.1093/molbev/msx149

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv.* arXiv:1303.3997v2 [q-bio.GN]. doi: 10.48550/arXiv.1303.3997

Li, J., Chen, C., Ma, X., Geng, G., Liu, B., Zhang, Y., et al. (2016). Long noncoding RNA NRON contributes to HIV-1 latency by specifically inducing tat protein degradation. *Nat. Commun.* 7 (1), 11730. doi: 10.1038/ncomms11730

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinf. (Oxford England).* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, W., Hartung, J. S., and Levy, L. (2006). Quantitative real-time PCR for detection and identification of *Candidatus* liberibacter species associated with citrus huanglongbing. *J. Microbiological Methods* 66 (1), 104–115. doi: 10.1016/j.mimet.2005.10.018

Liu, S., Kracher, B., Ziegler, J., Birkenbihl, R. P., and Somssich, I. E. (2015). Negative regulation of ABA signaling by WRKY33 is critical for *Arabidopsis* immunity towards *Botrytis cinerea* 2100. *eLife.* 4, e07295. doi: 10.7554/eLife.07295

López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* 9 (8), 583–593. doi: 10.1038/nrg2398

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi: 10.1186/s13059-014-0550-8

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25 (5), 955–964. doi: 10.1093/nar/25.5.955

Luna, E., Pastor, V., Robert, J., Flors, V., Mauch-Mani, B., and Ton, J. (2010). Callose deposition: A multifaceted plant defense response. *Mol. Plant-Microbe Interactions®.* 24 (2), 183–193. doi: 10.1094/MPMI-07-10-0149

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9), 1297–1303. doi: 10.1101/gr.107524.110

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: Insights into functions. *Nat. Rev. Genet.* 10 (3), 155–159. doi: 10.1038/nrg2521

Monzó, C., and Stansly, P. A. (2020). Economic value of conservation biological control for management of the Asian citrus psyllid, vector of citrus huanglongbing disease. *Pest Manage. Science.* 76 (5), 1691–1698. doi: 10.1002/ps.5691

Moreau, C., Gautrat, P., and Frugier, F. (2021). Nitrate-induced CLE35 signaling peptides inhibit nodulation through the SUNN receptor and miR2111 repression. *Plant Physiol.* 185 (3), 1216–1228. doi: 10.1093/plphys/kiaa094

Moreno-Hagelsieb, G., and Latimer, K. (2008). Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics.* 24 (3), 319–324. doi: 10.1093/bioinformatics/btm585

Nobori, T., and Tsuda, K. (2019). The plant immune system in heterogeneous environments. *Curr. Opin. Plant Biol.* 50, 58–66. doi: 10.1016/j.pbi.2019.02.003

Noutoshi, Y., Kuromori, T., Wada, T., Hirayama, T., Kamiya, A., Imura, Y., et al. (2006). Loss of *NECROTIC SPOTTED LESIONS 1* associates with cell death and defense responses in arabidopsis thaliana. *Plant Mol. Biol.* 62 (1), 29–42. doi: 10.1007/s11103-006-9001-6

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22 (1), 1–5. doi: 10.1016/j.tig.2005.10.003

Pani, A., and Mahapatra, R. K. (2013). Computational identification of microRNAs and their targets in *Catharanthus roseus* expressed sequence tags. *Genomics Data.* 1, 2–6. doi: 10.1016/j.gdata.2013.06.001

Peng, Z., Bredeson, J. V., Wu, G. A., Shu, S., Rawat, N., Du, D., et al. (2020). A chromosome-scale reference genome of trifoliate orange (*Poncirus trifoliata*) provides insights into disease resistance, cold tolerance and genome evolution in citrus. *Plant J.* 104 (5), 1215–1232. doi: 10.1111/tpj.14993

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. *Nat. Protoc.* 11 (9), 1650–1667. doi: 10.1038/nprot.2016.095

Pertea, G., and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Research* 9 (304), 304. doi: 10.12688/f1000research.23297.1

Peterson, L. E. (2009). K-Nearest neighbor. *Scholarpedia* 4 (2), 1883. doi: 10.4249/scholarpedia.1883

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33 (suppl_1), D501–D5D4. doi: 10.1093/nar/gki025

Ramadugu, C., Keremane, M. L., Halbert, S. E., Duan, Y. P., Roose, M. L., Stover, E., et al. (2016). Long-term field evaluation reveals huanglongbing resistance in citrus relatives. *Plant Disease.* 100 (9), 1858–1869. doi: 10.1094/PDIS-03-16-0271-RE

Ransohoff, J. D., Wei, Y., and Khavari, P. A. (2018). The functions and unique features of long intergenic non-coding RNA. *Nat. Rev. Mol. Cell Biol.* 19 (3), 143–157. doi: 10.1038/nrm.2017.104

Reyes, T. H., Maekawa, S., Sato, T., and Yamaguchi, J. (2015). The arabidopsis ubiquitin ligase *ATL31* is transcriptionally controlled by WRKY33 transcription factor in response to pathogen attack. *Plant Biotechnol.* 32 (1), 11–19. doi: 10.5511/plantbiotechnology.14.1201b

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81 (1), 145–166. doi: 10.1146/annurev-biochem-051410-092902

Scheuermann, J. C., and Boyer, L. A. (2013). Getting to the heart of the matter: Long non-coding RNAs in cardiac development and disease. *EMBO J.* 32 (13), 1805–1816. doi: 10.1038/emboj.2013.134

Sharma, Y., Sharma, A., Singh, K., and Upadhyay, S. K. (2022). Long non-coding RNAs as emerging regulators of pathogen response in plants. *Non-coding RNA.* 8 (1), 4. doi: 10.3390/ncrna8010004

Shirahama, S., Miki, A., Kaburaki, T., and Akimitsu, N. (2020). Long non-coding RNAs involved in pathogenic infection. *Front. Genet.* 11 (454). doi: 10.3389/fgene.2020.00454

Song, L., Fang, Y., Chen, L., Wang, J., and Chen, X. (2021). Role of non-coding RNAs in plant immunity. *Plant Commun.* 2 (3), 100180. doi: 10.1016/j.xplc.2021.100180

Szcześniak, M. W., Rosikiewicz, W., and Makałowska, I. (2016). CANTATAdb: A collection of plant long non-coding RNAs. *Plant Cell Physiol.* 57 (1), e8–ee. doi: 10.1093/pcp/pcv201

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43 (D1), D447–DD52. doi: 10.1093/nar/gku1003

Tang, Y., Qu, Z., Lei, J., He, R., Adelson, D. L., Zhu, Y., et al. (2021). The long noncoding RNA FRILAIR regulates strawberry fruit ripening by functioning as a noncanonical target mimic. *PloS Genet.* 17 (3), e1009461. doi: 10.1371/journal.pgen.1009461

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics.* 31 (12), 2032–2034. doi: 10.1093/bioinformatics/btv098

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). Mapman: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37 (6), 914–939. doi: 10.1111/j.1365-313X.2004.02016.x

Tsuda, K., and Katagiri, F. (2010). Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. *Curr. Opin. Plant Biol.* 13 (4), 459–465. doi: 10.1016/j.pbi.2010.04.006

Ulitsky, I. (2016). Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* 17 (10), 601–614. doi: 10.1038/nrg.2016.85

Wang, N. (2019). The citrus huanglongbing crisis and potential solutions. *Mol. Plant* 12 (5), 607–609. doi: 10.1016/j.molp.2019.03.008

Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., et al. (2014). *Arabidopsis* noncoding RNA mediates control of photomorphogenesis by red light. *Proc. Natl. Acad. Sci.* 111 (28), 10359. doi: 10.1073/pnas.1409457111

Wang, S., Liu, S., Liu, L., Li, R., Guo, R., Xia, X., et al. (2020). *miR477* targets the *phenylalanine ammonia-lyase* gene and enhances the susceptibility of the tea plant (*Camellia sinensis*) to disease during *Pseudopestalotiopsis* species infection. *Planta.* 251 (3), 59. doi: 10.1007/s00425-020-03353-x

Wang, C.-Y., Liu, S.-R., Zhang, X.-Y., Ma, Y.-J., Hu, C.-G., and Zhang, J.-Z. (2017). Genome-wide screening and characterization of long non-coding RNAs involved in flowering development of trifoliate orange (*Poncirus trifoliata* l. raf.). *Sci. Rep.* 7 (1), 43226. doi: 10.1038/srep43226

Wei, X., Mira, A., Yu, Q., and Gmitter, F. G. (2021). The mechanism of citrus host defense response repression at early stages of infection by feeding of diaphorina citri transmitting candidatus liberibacter asiaticus. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.635153

Wierzbicki, A. T. (2012). The role of long non-coding RNA in transcriptional gene silencing. *Curr. Opin. Plant Biol.* 15 (5), 517–522. doi: 10.1016/j.pbi.2012.08.008

Wu, G. A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., et al. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* 32 (7), 656–662. doi: 10.1038/nbt.2906

Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C., et al. (2018). Genomics of the origin and evolution of citrus. *Nature.* 554 (7692), 311–316. doi: 10.1038/nature25447

Wu, H.-J., Wang, Z.-M., Wang, M., and Wang, X.-J. (2013). Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants. *Plant Physiol.* 161 (4), 1875–1884. doi: 10.1104/pp.113.215962

Xiao, F., Mark Goodwin, S., Xiao, Y., Sun, Z., Baker, D., Tang, X., et al. (2004). *Arabidopsis CYP86A2* represses *Pseudomonas syringae* type III genes and is required for cuticle development. *EMBO J.* 23 (14), 2903–2913. doi: 10.1038/sj.emboj.7600290

Xin, M., Wang, Y., Yao, Y., Song, N., Hu, Z., Qin, D., et al. (2011). Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. *BMC Plant Biol.* 11 (1), 61. doi: 10.1186/1471-2229-11-61

Xu, W., Yang, T., Wang, B., Han, B., Zhou, H., Wang, Y., et al. (2018). Differential expression networks and inheritance patterns of long non-coding RNAs in castor bean seeds. *Plant J.* 95 (2), 324–340. doi: 10.1111/tpj.13953

Yang, Y., Liu, T., Shen, D., Wang, J., Ling, X., Hu, Z., et al. (2019). Tomato yellow leaf curl virus intergenic siRNAs target a host long noncoding RNA to modulate disease symptoms. *PloS Pathogens.* 15 (1), e1007534. doi: 10.1371/journal.ppat.1007534

Yan, J., Yuan, F., Long, G., Qin, L., and Deng, Z. (2012). Selection of reference genes for quantitative real-time RT-PCR analysis in citrus. *Mol. Biol. Rep.* 39 (2), 1831–1838. doi: 10.1007/s11033-011-0925-9

Yoon, J.-H., Abdelmohsen, K., and Gorospe, M. (2013). Posttranscriptional gene regulation by long noncoding RNA. *J. Mol. Biol.* 425 (19), 3723–3730. doi: 10.1016/j.jmb.2012.11.024

Yuan, J., Li, J., Yang, Y., Tan, C., Zhu, Y., Hu, L., et al. (2018). Stress-responsive regulation of long non-coding RNA polyadenylation in oryza sativa. *Plant J.* 93 (5), 814–827. doi: 10.1111/tpj.13804

Yu, Q., Chen, C., Du, D., Huang, M., Yao, J., Yu, F., et al. (2017). Reprogramming of a defense signaling pathway in rough lemon and sweet orange is a critical element of the early response to 'Candidatus liberibacter asiaticus'. *Horticulture Res.* 4 (1), 17063. doi: 10.1038/hortres.2017.63

Zaynab, M., Fatima, M., Abbas, S., Umair, M., Sharif, Y., and Raza, M. A. (2018). Long non-coding RNAs as molecular players in plant defense against pathogens. *Microbial Pathogenesis.* 121, 277–282. doi: 10.1016/j.micpath.2018.05.050

Zhang, M., Duan, Y., Zhou, L., Turechek, W. W., Stover, E., and Powell, C. A. (2010). Screening molecules for control of citrus huanglongbing using an optimized regeneration system for '*Candidatus* liberibacter asiaticus'-infected periwinkle (*Catharanthus roseus*) cuttings. *Phytopathology.* 100 (3), 239–245. doi: 10.1094/PHYTO-100-3-0239

Zhang, L., Wang, M., Li, N., Wang, H., Qiu, P., Pei, L., et al. (2018). Long noncoding RNAs involve in resistance to *Verticillium dahliae*, a fungal disease in cotton. *Plant Biotechnol. J.* 16 (6), 1172–1185. doi: 10.1111/pbi.12861

Zheng, Z., Qamar, S. A., Chen, Z., and Mengiste, T. (2006). *Arabidopsis* WRKY33 transcription factor is required for resistance to necrotrophic fungal pathogens. *Plant J.* 48 (4), 592–605. doi: 10.1111/j.1365-313X.2006.02901.x

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10 (1), 1523–. doi: 10.1038/s41467-019-09234-6

Zhu, Q.-H., Fan, L., Liu, Y., Xu, H., Llewellyn, D., and Wilson, I. (2014). miR482 regulation of *NBS-LRR* defense genes during fungal pathogen infection in cotton. *PloS One* 8 (12), e84390. doi: 10.1371/journal.pone.0084390

Zhu, Q.-H., Stephen, S., Taylor, J., Helliwell, C. A., and Wang, M.-B. (2014). Long noncoding RNAs responsive to *Fusarium oxysporum* infection in arabidopsis thaliana. *New Phytol.* 201 (2), 574–584. doi: 10.1111/nph.12537

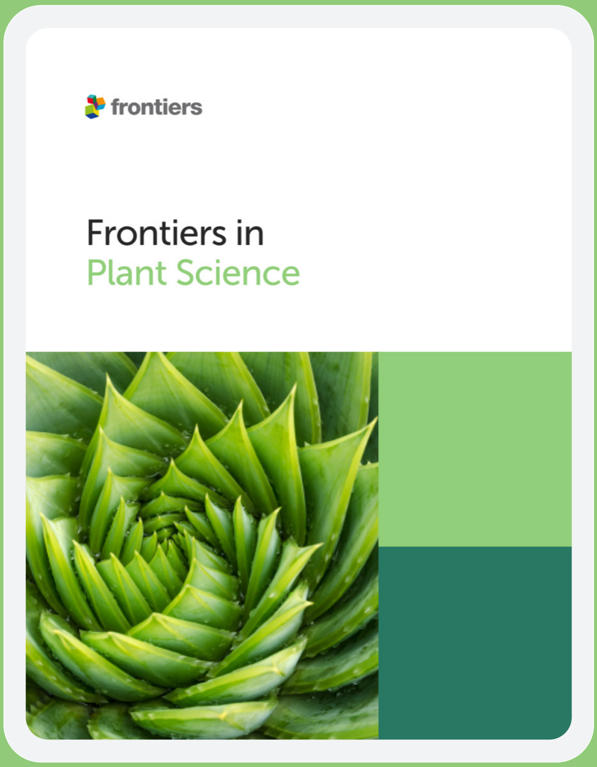
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Plant Science

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact