UNIVERSITAT POLITÈCNICA DE VALÈNCIA

PhD THESIS

A genomic approach to the evolution, diversification and domestication of the genus *Citrus*

Presented by: Carles Borredá Fernández

Supervised by: Manuel Talón Cubillo

Javier Terol Alcayde

Valencia, June 2021

PhD THESIS

A genomic approach to the evolution, diversification and domestication of the genus *Citrus*

For the degree of

Doctor in Biotechnology

Presented by: Carles Borredá Fernández

Supervised by: Manuel Talón Cubillo

Javier Terol Alcayde

This doctoral thesis was carried out at the Instituto Valenciano Investigaciones Agrarias (IVIA).

Valencia, June 2021

TABLE OF CONTENTS

Acknowledgements

ACKNOWLEDGEMENTS

Quisiera agradecer a todas aquellas personas que han hecho posible la realización de la presente Tesis Doctoral. En primer lugar, quiero agradecer a mis directores de tesis Manolo y Javier el interés y apoyo recibido durante estos años en el Centro de Genómica, que han permitido que llevara a cabo este proyecto. También quiero agradecer al resto de compañeros que han colaborado de una manera u otra en esta tesis doctoral: a Mati y a Tony por su ayuda en el laboratorio; a Vicky, a Dani, y a Toni por su experiencia; y a Quico y Concha por los buenos ratos a la hora de comer. Y por supuesto quiero agradecer al resto de doctorandos y becarios: a Juanlu y a Dani por su ayuda y por enseñarme "cómo funcionan las cosas en el IVIA"; a Julia, Mireia y Guillermo por su apoyo e interés, y en especial a mi coetánea Estela, por las largas conversaciones entre comando y comando. Al final ha resultado que al trabajo también se va a hacer amigos.

I would like to thank Albert G. Wu for his help during my short stay at the Joint Genome Institute, and for receiving me in the group he is part of. Our conversations, during my stay and since I left the US, have been always very enlightening.

También quiero agradecer a mis amigos el apoyo recibido durante estos años. A Bea y Jose por las rutas y las comilonas, que parece que no tengamos punto medio. A Vero y a Sara, por los viajes, los escapes y los buenos ratos que hemos pasado juntos. A Alberto, a Jose y a Sara, por todas las tardes y noches de juegos y sushi, y algún que otro gintonic, siempre tendréis un hogar allá donde esté yo. Al sector astur, muchas gracias por estar siempre ahí, aunque sea en la distancia, y por tener siempre abierta la puerta para mí; sabéis que podéis venir a secar cuando queráis. Y muchas gracias también a mis amigos, a los de toda la vida y a los que he hecho en los últimos años que, sin saber lo que supone hacer un doctorado, han estado ahí y me han brindado tantos buenos ratos. Gracias.

Quiero agradecer también a mi familia el apoyo constante a lo largo de todos estos años. A mis padres, muchas gracias por el apoyo incondicional en este proyecto. Gracias por aguantar horas y horas hablando de genes y transposones pese a que os resulte un tema totalmente marciano, y por brindarme todo vuestro apoyo y toda la ayuda que me habéis podido dar. Gracias de verdad.

Muchas gracias también a Víctor y a Merche, por ayudarme a desconectar de los problemas. Cada rato que hemos pasado en Pedralvilla ha sido en el ambiente más despreocupado y alegre que me puedo imaginar. Gracias a los dos.

Muchas gracias a Ximo, a Fina, a Paqui y a Pere por la paciencia y comprensión brindada durante todos estos años, y por acogerme como a uno más entre vosotros. Gracias también a Pau y a Shara, y desde hace apenas un año a Alicia, por esos viernes noche de kebab o chinos, o brócoli si se tercia, para cerrar la semana laboral con alegría.

Por último, quiero agradecer a mi pareja. Neus, saps que no existeixen paraules per agrair-te el esforç i la dedicació que has posat en aquest projecte. Aquesta tesi doctoral no haguera sigut possible sense la teua paciència i, més encara, sense la teua ajuda. Moltíssimes gràcies per ajudar-me a seguir amb el mateix coratge.

Abstract

ABSTRACT

Citrus is a highly diverse genus within the Aurantioideae subfamily that comprises a still undetermined number of pure species, natively found in a vast territory that extends from India to Japan and Australia. Indeed, a pivotal unsolved issue concerning the genus *Citrus* is related to the taxonomy and evolution of these species, obscured by the frequency of the admixed *Citrus* germplasm and its huge phenotypical diversity. Besides pure species, countless citrus cultivars of commercial interest, such as mandarins, oranges, grapefruits and lemons, have been traditionally included in this genus. Commercial citrus are the product of several interspecific crosses between these pure species, that occurred during the first events of *Citrus* domestication. In addition, a genome-wide analysis has recently provided the backbone of the *Citrus* phylogeny. This study suggests that the native current species diverged from an ancestral citrus in a relatively rapid radiation triggered by a global climate change, about 8 million years ago during the Late Miocene. Understanding the processes that shaped the evolution and subsequent domestication of the genus will prove useful for citrus breeders while providing novel insights in the field of plant genome evolution.

To this end, the genus *Citrus* has been rooted into the Aurantioideae subfamily tree to generate the most complete and detailed citrus phylogeny presented so far, including several members of all citrus types and clades currently known. An alignment-free method was used to generate a genome-wide Aurantioideae phylogeny, revealing that their native distribution is compatible with several independent dispersal events in the last 10 million years, spanning vast distances from Asia to Africa and Australia. The *Citrus* phylogeny has been studied under a novel evolutionary model that takes into consideration the process of incomplete lineage sorting and is better suited to capture the variability generated during fast radiations. The data suggests that the citrus original radiation occurred so fast that most of the extant citrus species emerged and diversified simultaneously, migrating in several directions and colonizing practically the whole South East Asian region. The dating of these events has also allowed to advance new and original proposals on the paleogeographic and climatological environments leading to these migrations.

The consequences of the *Citrus* radiation can be appreciated in the great genetic and phenotypic diversity found among the pure species of this genus. In order to investigate the effects of the Late Miocene climate change on the genomic structure of the *Citrus* pure species, the activity and evolution of retrotransposons was analyzed, as they represent a major force generating genomic variability. Most of the retrotransposon families found in *Citrus* species were also present in *Severinia*, a member of the subfamily Aurantioideae that diverged from *Citrus* more than 10 million years ago. This implies that only few families were specifically acquired after the divergence of these two genera. However, estimations of the retrotransposon insertion rate in the last 15 million years suggest that, shortly after the radiation, the transposon activity profiles displayed profound differences even among closely related species. Hence, it seems plausible that the retrotransposon insertion dynamics are linked to the stress caused by the Late Miocene climate change, although specific responses seem to be largely governed by the particular evolutionary history of each individual species. Overall, the data indicates that retrotransposon activity is in a substantial way associated with the process of citrus speciation.

The differences of gene expression in fruits of domesticated varieties and wild species have been also studied in an attempt to elucidate how the interspecific hybridizations that produced commercial citrus altered the expression of key genes during *Citrus* domestication. Indeed, the data suggest that interspecific hybridizations were key for this process, very possibly aided with the asexual propagation of the admixed individuals. Different mechanisms explaining some commercially relevant *Citrus* traits are also proposed. For example, pulp acidity in citrons and lemons appears to be linked to the increased proton influx to the vacuolar lumen. In parallel, the data also suggest that the peel pigmentation is not controlled by a single gene or mechanism, as the additive effect of several minor genes altogether appears to determine the final carotenoid accumulation. Finally, an allele-dependent expression pattern of the *chalcone synthase* gene, which codes for a rate limiting enzyme in the flavonoid biosynthetic pathway, was found. This observation might advocate for the existence of stepwise evolution in the mandarin flavonoid accumulation profile. All in all, the transcriptomic approach used in this work allowed to generate broader hypotheses that stand from a genus-wide perspective.

In this doctoral thesis, multiple genomic approaches have been used in order to expand the existing knowledge on major determinants driving the processes of evolution, diversification and domestication in *Citrus*. Overall, the results provide a comprehensive framework of the genus *Citrus* and its phylogenetic and genealogic relationships. These analyses are completed with the finding that transposons are deeply involved in the processes of citrus speciation and with the study of the relevance of gene expression in wild and commercial citrus and its association with their phenotypical traits. The insights exposed in the following sections reveal the inherent complexity of the evolutionary history of this fascinating genus.

RESUMEN

El género *Citrus*, perteneciente a la subfamilia Aurantioideae, abarca un número aún desconocido de especies puras, extremadamente diversas, que crecen salvajes en un amplio territorio que se extiende desde la India hasta Japón y Australia. La taxonomía y evolución de este género son cuestiones que han permanecido sin resolver durante décadas, en parte debido al origen mestizo de las variedades comerciales de cítricos y a la enorme diversidad fenotípica que existe entre ellas. Además de las especies puras, un gran número de variedades comerciales de cítricos, como mandarinas, naranjas, pomelos o limones, han sido tradicionalmente incluidas en el género *Citrus*. Hoy sabemos que los cítricos comerciales son el producto de múltiples cruces interespecíficos entre las especies puras del género que ocurrieron al inicio del proceso de domesticación del mismo. Además, la estructura básica de la filogenia del género *Citrus* ha sido publicada recientemente. Este estudio propone que las especies actuales de cítricos surgieron desde un ancestro común en un proceso de radiación rápida, desencadenado por un cambio climático global que tuvo lugar en el Mioceno tardío, hace aproximadamente 8 millones de años. Una mejor comprensión de los procesos involucrados en la evolución y posterior domesticación del género *Citrus* podría ser de utilidad para los mejoradores, además de proporcionar nuevas perspectivas dentro del ámbito de la evolución del genoma de plantas.

Para ello, se ha anclado el género *Citrus* dentro la subfamilia Aurantioideae, generando una filogenia de cítricos que incluye distintas especies pertenecientes a todos los clados de cítricos conocidos, siendo así la filogenia más completa presentada hasta la fecha. Se ha empleado un método de inferencia filogenética libre de alineamiento para generar una filogenia de las aurantioideas empleando datos de todo el genoma. Esta filogenia ha revelado que la distribución geográfica de estas especies es compatible con la existencia de varios eventos de dispersión de largo alcance, desde Asia hacia África u Oceanía. La filogenia del género *Citrus* ha sido estudiada bajo un modelo evolutivo novedoso, considerando el proceso de coalescencia profunda para que la filogenia obtenida refleje la variabilidad inherente a los procesos de radiación rápida, como es el caso del género *Citrus*. Los resultados aquí presentados sugieren que la radiación original del género *Citrus* ocurrió de una forma tan súbita que la mayoría de las especies de cítricos que

existen hoy en día aparecieron de manera simultánea, migrando en varias direcciones y colonizando prácticamente la totalidad del sudeste asiático. La datación de estos eventos ha permitido hacer nuevas propuestas sobre los eventos paleogeográficos y climatológicos que dieron lugar a estas migraciones.

Las consecuencias de la radiación de los cítricos se ven reflejadas en la enorme diversidad genética y fenotípica que existe entre las especies puras del género. Para investigar los efectos del enfriamiento global durante el Mioceno tardío en la estructura genómica de los cítricos, se ha analizado la actividad y la evolución de los retrotransposones en distintas especies de cítricos, dado que estos elementos representan una enorme fuente de variabilidad genética. La mayoría de los retrotransposones de los cítricos también se encuentran en *Severinia*¸ un género de las aurantioideas cuya divergencia con el ancestro de los cítricos data de hace 10 millones de años, lo que sugiere que tan sólo unas pocas familias de retrotransposones fueron adquiridas desde entonces. Sin embargo, la estimación de las tasas de inserción de los retrotransposones en las distintas especies de cítricos durante los últimos 15 millones de años sugiere que, poco después de la radiación de los cítricos, la actividad de estos elementos sufrió cambios drásticos incluso entre especies próximas. Por tanto, es posible que dicha actividad esté ligada al estrés causado por el enfriamiento global a finales del Mioceno, aunque también parece verse afectada por las condiciones evolutivas particulares de cada una de las especies estudiadas. De todo esto se deduce que la actividad de los retrotransposones podría estar sustancialmente asociada al proceso de la especiación de los cítricos.

Por último, también se ha estudiado la expresión génica diferencial en cítricos de variedades domesticadas y especies puras, para así elucidar cómo las hibridaciones interespecíficas que generaron las variedades comerciales de cítricos alteraron la expresión de genes clave en la domesticación de este género. Los datos obtenidos sugieren que estas hibridaciones jugaron un papel esencial en este proceso, posiblemente en conjunción con la propagación clonal de los individuos híbridos o mestizos. Los resultados también han permitido proponer un mecanismo que explica la acidez de la pulpa de cidros y limones basado en el flujo de protones al lumen vacuolar. Por otra parte, el color de la piel de los cítricos no parece estar controlado por un único gen o mecanismo, sino que el efecto aditivo de varios genes en conjunto parece determinar la concentración final de carotenoides. Finalmente, se ha encontrado una copia del gen de la chalcona

sintasa, enzima limitante en la ruta de biosíntesis de flavonoides, que tan solo se expresa en mandarinas y variedades derivadas. Esto permite sugerir la existencia de un proceso evolutivo escalonado para el perfil de acumulación de flavonoides de las mandarinas. En resumen, la estrategia de análisis transcriptómico empleada en este trabajo ha permitido generar hipótesis más amplias que se sostienen para todo el género *Citrus*.

A lo largo de esta Tesis Doctoral se han empleado diversas estrategias genómicas para ampliar el conocimiento existente sobre los procesos que dirigieron la evolución, diversificación y domesticación de los cítricos. Los resultados presentados aportan un marco de trabajo global para las relaciones filogenéticas del género *Citrus*. Estos análisis se completan con el descubrimiento de la asociación entre los transposones y la especiación de los cítricos, y con el estudio de la expresión génica durante el proceso de maduración en cítricos salvajes y domesticados, y cómo esto se asocia con sus rasgos fenotípicos. Los datos presentados en este trabajo revelan la complejidad inherente a la historia evolutiva de este género tan fascinante.

RESUM

El gènere *Citrus*, pertanyent a la subfamília Aurantioideae, comprèn un nombre encara desconegut d'espècies pures, extremadament diverses, que creixen salvatges en un ampli territori que s'estén des de l'Índia fins a Japó i Austràlia. La taxonomia i evolució d'aquest gènere són qüestions que han estat sense resoldre durant dècades, en part a causa de l'origen mestís de les varietats comercials de cítrics i a l'enorme diversitat fenotípica que existeix entre aquestes. A més de les espècies pures, un gran nombre de varietats comercials de cítrics, com a mandarines, taronges, pomelos o llimes, han sigut tradicionalment incloses dins del gènere *Citrus*. Hui sabem que els cítrics comercials són el producte de múltiples encreuaments interespecífics entre les espècies pures del gènere que van ocórrer a l'inici del procés de domesticació d'esta planta. A més, l'estructura bàsica de la filogènia del gènere *Citrus* ha sigut publicada recentment. Aquest estudi proposa que les espècies actuals de cítrics van sorgir des d'un avantpassat comú en un procés de radiació ràpida, desencadenat per un canvi climàtic global que va tindre lloc en el Miocè superior, fa aproximadament 8 milions d'anys. Una millor comprensió dels processos involucrats en l'evolució i posterior domesticació del gènere *Citrus* podria ser d'utilitat per als milloradors, a més de proporcionar noves perspectives dins de l'àmbit de l'evolució del genoma de plantes.

Per això, s'ha ancorat el gènere *Citrus* dins de la subfamília Aurantioideae, generant una filogènia de cítrics que inclou distintes espècies pertanyents a tots els clades de cítrics coneguts, sent així la filogènia més completa presentada fins a la data. S'ha empleat un mètode d'inferència filogenètica lliure d'alineament per a generar una filogènia de les aurantioideas utilitzant dades de tot el genoma. Esta filogènia ha revelat que la distribució geogràfica d'estes espècies és compatible amb l'existència de diversos esdeveniments de dispersió de llarg abast, des d'Àsia cap a Àfrica o Oceania. També s'ha aplicat un nou model evolutiu per a estudiar la filogènia dels cítrics, considerant el procés de coalescència profunda de manera que la filogènia obtinguda reflectisca la variabilitat inherent als processos de radiació ràpida, com és el cas del gènere *Citrus*. Els resultats ací presentats suggereixen que la radiació original del gènere *Citrus* va ocórrer d'una forma tan sobtada que la majoria de les espècies de cítrics que existeixen actualment van aparèixer de manera simultània, migrant en diverses direccions i colonitzant quasi la

totalitat del sud-est asiàtic. La datació d'aquests esdeveniments ha permès fer noves propostes sobre els esdeveniments paleogeogràfics i climatològics que van donar lloc a aquestes migracions.

Les conseqüències de la radiació dels cítrics es veuen reflectides en l'enorme diversitat genètica i fenotípica que existeix entre les espècies pures del gènere. Per a investigar els efectes del refredament global durant el Miocè superior en l'estructura genòmica dels cítrics, s'ha analitzat l'activitat i l'evolució dels retrotransposons en distintes espècies de cítrics, ja que aquests elements representen una enorme font de variabilitat genètica. La majoria dels retrotransposons dels cítrics també es troben en *Severinia*¸ un gènere de les aurantioideas la divergència del qual amb l'avantpassat dels cítrics data de fa 10 milions d'anys, la qual cosa suggereix que tan sols unes poques famílies de retrotransposons van ser adquirides des d'aleshores. No obstant això, l'estimació de les taxes d'inserció dels retrotransposons en les distintes espècies de cítrics durant els últims 15 milions d'anys suggereix que, poc després de la radiació dels cítrics, l'activitat d'aquests elements va patir canvis dràstics inclús entre espècies pròximes. Per tant, és possible que l'esmenada activitat estiga lligada a l'estrès causat pel refredament global a finals del Miocé, encara que també sembla veure's afectada per les condicions evolutives particulars de cada una de les espècies estudiades. De tot açò es dedueix que l'activitat dels retrotransposons podria estar substancialment associada al procés de l'especiació dels cítrics.

Finalment, també s'ha estudiat l'expressió gènica diferencial en cítrics de varietats domesticades i espècies pures, per a elucidar com les hibridacions interespecífiques que van generar les varietats comercials de cítrics van alterar l'expressió de gens clau en la domesticació d'este gènere. Les dades obtingudes suggereixen que aquestes hibridacions van jugar un paper essencial en aquest procés, possiblement en conjunció amb la propagació clonal dels individus híbrids o mestissos. Els resultats també han permès proposar un mecanisme que explica l'acidesa de la polpa de poncems i llimes basat en el flux de protons al lumen vacuolar. D'altra banda, el color de la pell dels cítrics no pareix estar controlat per un únic gen o mecanisme, sinó que sembla que l'efecte additiu de diversos gens en conjunt determina la concentració final de carotenoides. Finalment, s'ha trobat una còpia del gen de la chalcona sintasa, enzim limitant en la ruta de biosíntesi de flavonoides, que tan sols s'expressa en mandarines i varietats derivades. Açò permet suggerir l'existència d'un procés evolutiu escalonat per al perfil d'acumulació de

flavonoides en mandarines. En resum, l'estratègia d'anàlisi transcriptòmic empleada en este treball ha permès generar hipòtesis més àmplies que se sostenen per a tot el gènere *Citrus*.

Al llarg d'aquesta Tesi Doctoral s'han emprat diverses estratègies genòmiques per a ampliar el coneixement existent sobre els processos que van dirigir l'evolució, diversificació i domesticació dels cítrics. Els resultats presentats aporten un marc de treball global per a les relacions filogenètiques del gènere *Citrus*. Aquestes anàlisis es completen amb el descobriment de l'associació entre els transposons i l'especiació dels cítrics, i amb l'estudi de l'expressió gènica durant el procés de maduració en cítrics salvatges i domesticats, i com s'associa amb les seues característiques fenotípiques. Les dades presentades en aquest treball revelen la complexitat inherent en la història evolutiva d'aquest gènere tan fascinant.

Abbreviations

ABBREVIATIONS

Admx: admixed

ASE: allele-specific expression

ATP: adenosine triphosphate

ATPase: Adenylpyrophosphatase

BAM: Binary Alignment Map

BCE: before current era

bHLH: basic Helix-Loop-Helix

BLAST: Basic Local Alignment Search Tool

Bp: base pairs

C/M: citron/mandarin

C/P: citron/pummelo

CCD: carotenoid cleavage dioxygenase

cDNA: complementary deoxyribonucleic acid

Chr: chromosome

CHS: chalcone synthase

CNAG: Centre Nacional de Análisis Genómico

CRM: Centromere-specific Retrotransposon of Maize

DEGs: differentially expressed gene

Dels: Deletions

DNA: deoxyribonucleic acid

ERE: ethylene responsive element

ERF: ethylene responsive factor

ESS: effective sample size

FAO: Food and Agricultural Organization

FDR: False discovery ratio

FISH: fluorescence *in situ* hybridization

FOMT: phenylpropanoid O-methyltransferase

GABA: γ-aminobutyric acid

Gb: giga base pairs

GO: Gene Ontology

GQ: Genotype quality

GTR: General Time Reversible

GyDB: Gypsy Database

HK: hexokinase

HKY: Hasegawa-Kishino-Yano

HLB: Huanglongbing

i.e.: id est (this is)

IGV: Integrative Genome Viewer

ILS: incomplete lineage sorting

IN: integrase

indel: insertion-deletion

IR: illegitimate recombination

IUPAC: International Union of Pure and Applied Chemistry

IVIA: Instituto Valenciano de Investigaciones Agrarias

kb: kilo base pairs

KEGG: Kyoto Encyclopedia of Genes and Genomes

LCYb: Lycopene β-cyclase

LTR: long terminal repeat

LTR-TE: long terminal repeat transposable element

M/M: mandarin/mandarin

MAPA: Ministerio de Agricultura Pesca y Alimentación

Mb: mega base pairs

MCMC: Monte Carlo Markov chain

MITE: miniature inverted-repeat transposable element

MSC: multispecies coalescent model

MULE: mutator-like element

Mya: million years ago

NCBI: National Center for Biotechnology Information

P/P: pummelo/pummelo

PCA: principal component analysis

PCR: Polymerase Chain Reaction

PP: posterior probabilities

PSRF: potential scale reduction factor

PSY: phytoene synthase

qPCR: quantitative Polymerase Chain Reaction

RAD-seq: Restriction site associated DNA sequencing

RH: RNase H

RNA: ribonucleic acid

RNA-seq: ribonucleic acid sequencing

RT: reverse transcriptase

RT-qPCR: reverse transcription quantitative Polymerase Chain Reaction

SIRE: Soybean Interspersed Repetitive Element

SNP: single nucleotide polymorphism

SPP: Sucrose pyrophosphatase

SPS: Sucrose phosphate synthase

SRA: Sequence Read Archive

SuSy: Sucrose synthase

TA: titratable acid

tb1: teosinte branching 1

TCA: tricarboxylic acid

TSD: target site duplication

UBC: Ubiquitin C

UR: unequal recombination

UV: ultraviolet

V-ATPase: Vacoular adenylpyrophosphatase

WGD: whole genome duplication

ZDS: zeta-carotene desaturase

ZEP: zeaxanthin epoxidase

Introduction

INTRODUCTION

1 Relevance and history of citrus fruits

Citrus are amongst the most relevant fruit crops, both in terms of production and area cultivated. Sweet oranges and mandarins represent the fourth and fifth most cultivated crops in the world, with over 75 and 34 million tons produced in 2018, respectively (FAO, 2021). Citrus cultivation is concentrated in several warm areas of the world: the Mediterranean Basin, around the Caribbean Sea, in California, South of Africa, South East Asia and several regions of South America. The main producers are, in descending order, China, Brazil, India, Mexico, the United States and Spain (FAO, 2021). Despite occupying the sixth position in terms of total citrus production, Spain is the world major exporter, with most of its exported fruits being destined to fresh consumption. In 2018, more than half of the total Spanish citrus fruits were exported, accounting for over 3.5 million tons of fresh fruit. Within Spain, the Valencian region represents over 50% of the total fruit production, mostly as mandarins and oranges (MAPA, 2021).

The flagship of commercial citrus are mandarins, for instance clementines and satsumas, sweet "blonde and blood" oranges, grapefruits, limes and lemons (Figure 1). However, many other citrus are used throughout the world for a variety of reasons. For example, the juice of the Yuzu fruit, a small citrus, is a common ingredient in Korean and Japanese cuisine (Nile and Park, 2014), while the bergamot orange is an essential component of several teas (Orth *et al.*, 2013). Other less-known citrus, such as calamondins, kumquats and kaffir limes, are used in local cuisines of several other regions of South East Asia (Budiarto *et al.*, 2019). Some of these are also utilized as ornamental trees and for the production of essential oils and aromas (Chávez-González *et al.*, 2016; González-Mas *et al.*, 2019). Some inedible citrus, such as poncirus and citranges, are thoroughly employed as rootstocks due to their cold hardiness, abiotic stress tolerance or disease resistance (Castle, 2010).

That most of the local uses of *Citrus* are concentrated around East Asia is not coincidental. In fact, the genus *Citrus* originated in South East Asia and only came to Europe by human propagation. The first written records of a sweet orange date from the year 314 BCE, in

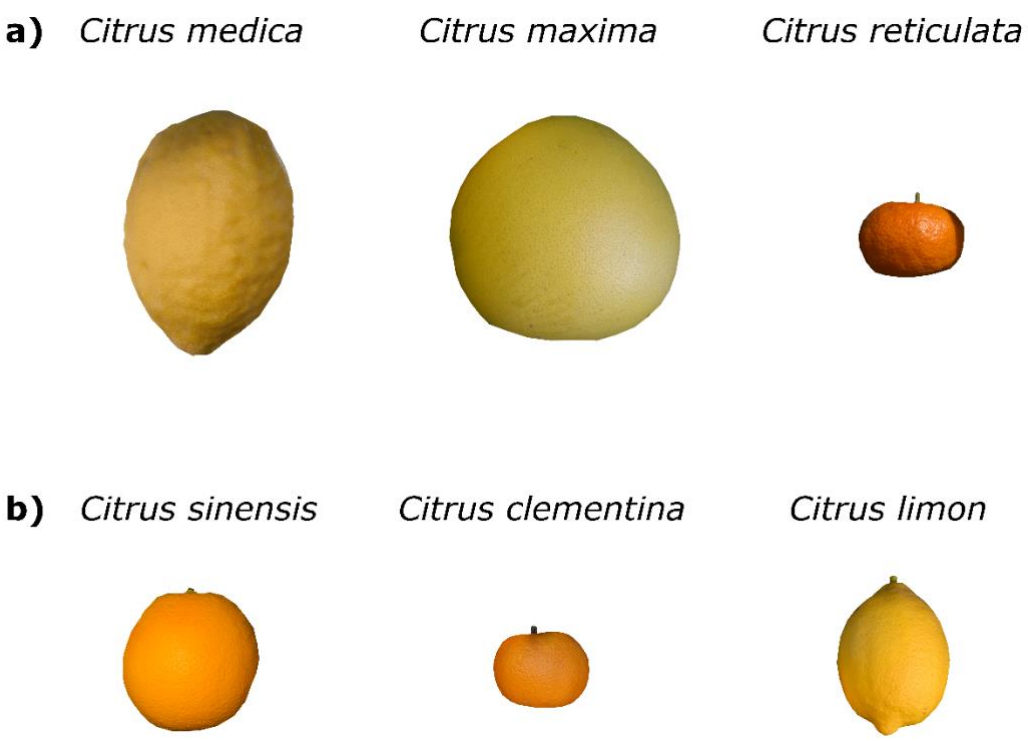


Figure 1: Morphological differences across wild and cultivated *Citrus*. a) Citron (*Citrus medica*), pummelo (*Citrus maxima*) and mandarin (*Citrus reticulata*), the progenitor species of most of the existing commercial *Citrus*, are shown. b) Sweet orange (*Citrus sinensis*), clementine (*Citrus clementina*) and lemon (*Citrus limon*), three commercially relevant citrus cultivars, are shown. Pictures were retrieved from the *Citrus* Variety Collection maintained at the University of California Riverside.

China, but if the whole genus *Citrus* is considered, earlier references exist, dated in 2200 BCE in the same country (Xu *et al.*, 2013). Based on written history, some authors suggested that the introduction of *Citrus* species in the Mediterranean took place much later in several independent events (see Deng *et al*. (2020) and the references therein). Citrons were the first citrus fruits that reached the Mediterranean, probably through Persia and Jerusalem, between the 4th and 5th century BCE. Lemons and sour oranges were found in roman mosaics dated from the first century BCE and are reported to be present in Andalusia in the 9th -11th centuries. Sweet oranges reached Europe near the 15th century through trade and spread shortly thereafter through the Mediterranean. Mandarins, despite their great relevance in current citriculture, did not arrive to Europe until 1805; they were imported to England, then sent to Italy and finally also spread through the Mediterranean. After that, other mandarin varieties were brought from South East Asia to Europe.

The Mediterranean basin offered a favorable climate for citrus cultivation, becoming a major producer region worldwide, and giving birth to many new varieties. For example, in 1902, a French priest named Clément Rodier found in Algeria a chance seedling of commercial relevance, which was named "Clementine" (known nowadays as "Fina clementine") after him (Trabut, 1902). Clementines rapidly expanded across the Mediterranean countries due to their high-standard agronomical traits. In 1953, near Nules (Castellón), it was found a bud sport mutation of Fina clementine, named Clemenules, that shortly became a gold standard in citriculture, since it retained the exceptional flavor of the original clementine while also exhibiting a considerably increase in size.

The emergence of Clemenules and many other clementine varieties boosted the Valencian citriculture, which soon became a major player in the citrus world market. However, in the last years Valencian citriculture has gone downhill and it is currently going through an unpreceded crisis. The increasing exports of competing countries such as South Africa or Turkey (FAO, 2016) have forced a price drop of the Valencian citrus, discouraging investors. It is predicted that global warming will pose a major challenge in a short term for citrus cultivation, as it will induce abiotic stresses associated with heat and drought. On the other hand, diseases like the citrus greening (Huanglongbing or HLB) have decimated the fruit production in regions such as Florida or Brazil and, while this disease has not been detected in Spain yet, it already constitutes a major threat to our citrus (Gottwald *et al.*, 2007; Siverio *et al.*, 2017). Any long-term solution for these problems requires the generation of new varieties capable of facing the upcoming challenges, which would favor the Valencian citriculture and allow it to keep its privileged position as the main exporter region of the world.

Unfortunately, the generation of new citrus varieties is a long-term investment. The extended juvenile period of citrus trees, up to 15 years, impedes the evaluation of key traits such as productivity or fruit quality shortly after the breeding process. Breeding novel traits is a problem by itself, since most commercial varieties come from a reduced number of genetically similar individuals, and their hybridizations do not generally produce as much variability as desired. Furthermore, many of them are apomictic (polyembryonic), and therefore, in addition to the sexual embryo, their fertilized seeds

develop numerous nucellar embryos genetically identical to the maternal parental, severely hindering the efficiency of the crosses.

Since traditional breeding in citrus is strongly limited because of these and another biological impediments, it is not surprising that most commercial varieties are somatic mutants or bud sport mutations (Luro *et al.*, 2018), a situation that today appears clearly insufficient to face current global challenges. In the last twenty years, genetics and genomics have provided a set of useful tools and knowledge related to the processes of crop evolution, selection and domestication for breeders and researchers. With a proper understanding of such processes, the generation and selection of new cultivars could be accelerated, and the associated costs reduced. These benefits are clearly desirable for any existing crop, but they are of vital importance in the case of tree crops, and especially in the case of members of the genus *Citrus* due to the biological particularities that surround this genus.

2 *Citrus* genomics

The application of genomic tools in breeding requires a proper understanding of the genome and the mechanisms that have shaped it. Prior to the genomics era, knowledge of the *Citrus* genomic structure was based on cytogenetic and microarray data. For example, cytogenetic data from different species already suggested a relatively compact diploid genome, of about 300-400 Mb (Arumuganathan and Earle, 1991; Seker *et al.*, 2003), organized in nine chromosome pairs (2n = 18), although few triploid and tetraploid cultivars have been reported (Reuther *et al.*, 1967). Some progress had also been achieved in *Citrus* applied genetics, mostly with the generation of early genetic markers (reviewed in Talon and Gmitter (2008)). However, many commercial varieties such as oranges and clementines are bud-sport mutations (i.e., somatic clones), which could not be discerned by most of these markers. Microarray-based transcriptome studies shed some light into the ripening process of citrus fruits (Cercós *et al.*, 2006; Martinez-Godoy *et al.*, 2008; Aprile *et al.*, 2011), a process directly related with fruit quality and therefore of great interest for breeders.

However, with the rise of genomics in the landscape of plant breeding, the international citrus research community focused in the generation of a reference genome for *Citrus*. In 2013, a draft genome sequence of sweet orange was published (Xu *et al.*, 2013), that was

soon followed by a high-quality genome from a haploid clementine, that rapidly became the reference genome for *Citrus* species (Wu *et al.*, 2014). These milestones defined the start of the genomics era for the genus *Citrus*. Both genome sequences and those obtained thereafter confirmed and extended some of the results previously determined by cytological analysis. Currently, ten complete genome sequences are available for *Citrus* species and close relatives. Among commercial varieties, the genome sequences of *Citrus clementina* (Wu *et al.*, 2014), *Citrus unshiu* (Shimizu *et al.*, 2017) and *Citrus sinensis* (Xu *et al.*, 2013) have been released. Assembled genomes for another five citrus species are available as well, including *Citrus medica*, *Citrus grandis*, *Citrus ichangensis* (Wang *et al.*, 2017*b*), *Citrus reticulata* (Wang *et al.*, 2018*a*) and *Fortunella hindsii* (Zhu *et al.*, 2019*b*). The genomes of *Poncirus trifoliata* and *Severinia buxifolia*, two outgroups related to genus *Citrus*, have also been obtained (Wang *et al.*, 2017*b*; Peng *et al.*, 2020).

High-quality reference genomes are required for standard high-throughput analysis, and. in recent years, the versatility and affordable cost of short-read sequencing projects allowed the re-sequencing of hundreds of citrus varieties, becoming a fundamental tool in genomic analysis. A direct application of short-read sequencing is the retrieval of single nucleotide polymorphisms (SNPs) from different individuals, which can be used in further analysis. This way, analyzing the heterozygosity distribution across different genomic regions Wu *et al*. (2014) reported the existence of islands of high heterozygosity in the genomes of most of the commercial citrus. These windows correspond to introgressions of one genome into another; the varieties carrying these introgressions are therefore admixtures in contrast with pure *Citrus* species, that by definition contain a single genome. These genomics insights have identified at least 10 *Citrus* pure species (Wu *et al.*, 2018). Citrons, pummelos and mandarins are considered the three fundamental species due to their role in generating most of the admixed commercial varieties (Figure 1). It is worth to mention that mandarin is a popular term that is not supported by neither botanical nor genetic data. In this doctoral thesis, a distinction between pure and admixture mandarins will be held: pure mandarins are wild species bearing inedible fruits, mostly due to their extreme acidity; in contrast, palatable commercial mandarins are admixtures between pure mandarins and pummelos. Not only commercial mandarins, but sweet oranges, grapefruits, tangors and many other hybrids are also mandarin-pummelo admixtures (Wu *et al.*, 2014; Oueslati *et al.*, 2017; Wang *et al.*, 2018*a*). On the other hand, lemons are hybrids between a citron and a mandarin/pummelo admixture, harboring

fragments coming from the three species, while limes are generally a direct cross between a citron and the Philippine papeda *Citrus micrantha* (Curk *et al.*, 2016).

Re-sequencing data has also proved useful for detection of small insertions and deletions (indels), that have been used among other goals to identify *Citrus* chloroplast haplotypes and their phylogenetic relationships (Carbonell-Caballero *et al.*, 2015; Maddi *et al.*, 2018), as well as molecular markers for cultivar identification in mandarin hybrids (Noda *et al.*, 2020). Currently, numerous analyses in *Citrus* have also been performed to detect mobile element insertions (Caruso *et al.*, 2019; Liu *et al.*, 2019). The relevance of these variants has been proved in other crops, where they have been linked to major agronomical traits (Zhang *et al.*, 2015; Zhou *et al.*, 2019; Alonge *et al.*, 2020).

The analysis of the reference genome sequences themselves confirmed previously reported data, like the *Citrus* chromosome count (2n = 18) or the genome size, which ranges from 250 to 400 Mb. Between 23000 and 30000 genes were annotated in these genomes, and the proportion of the genome covered by repetitive elements ranged from 20% to 40%. Synteny analyses based on these reference genomes highlighted a generally well-conserved genomic structure in citrus and their relatives, with a small number of inversions or translocations affecting large portions of the genome. Similar results were obtained recently by chromosome-specific FISH in six different *Citrus* species, suggesting that synteny in this genus is extremely conserved (He *et al.*, 2020).

It also became evident that *Citrus* genomes did not experience a specific whole genome duplication (WGD) event (Xu *et al.*, 2013; Peng *et al.*, 2020). In contrast with other plant species (Qiao *et al.*, 2019), the last WGD event taking place in the *Citrus* evolution was the gamma event, a genome triplication shared by all the core eudicots (Jiao *et al.*, 2012). Nevertheless, specific gene families have been expanded and contracted in both the *Citrus* clade and in particular species. For example, two gene families have been found to be specifically expanded in cold-resistant citrus such as *P. trifoliata*, *C. ichangensis* and *Fortunella spp.* (Peng *et al.*, 2020). Some gene families appear to be largely expanded across the whole genus *Citrus* when compared with unrelated plants, as for example the ones involved in terpenoid biosynthesis and flavonoid modifications, which might be in line with the huge diversity of carotenoids and flavonoids that citrus fruit accumulate to a large extent (Gonzalez-Ibeas *et al.*, 2021, in press). Other families involved in pathogen defense are also extremely expanded in *Citrus* compared with other angiosperms

(Magalhães *et al.*, 2016). All in all, these results suggest that, despite a conserved genome size, synteny and gene number, the genomes of the different *Citrus* species harbor key differences that produce the wide range of phenotypes observed within the genus.

3 Origin and phylogeny of the genus *Citrus*

The taxonomy of the genus *Citrus* has been a longstanding problem for scientists. Before the genomics era, the main works addressing this issue are those from citrus botanists Swingle and Tanaka (Tanaka, 1954; Swingle and Reece, 1967). While Tanaka identified more than 166 different species, Swingle reduced that number down to 16, arguing that most of Tanaka's species were actually different cultivars of other species not deserving the consideration of species by themselves (Luro *et al.*, 2018; Ollitrault *et al.*, 2020). Both took into account only what they called "true citrus", excluding genera such as *Clymenia*, *Microcitrus* or *Fortunella*, which were considered close relatives to the genus *Citrus*. Being solely based on botanical traits, the two classification systems generally failed to identify admixtures, which were either classified as independent species or clustered with other pure species.

Another issue subject to debate was the localization of a center of diversification or origin of the *Citrus* species (Talon *et al.*, 2020). Tanaka postulated that the Eastern Himalaya was a major center of dispersion and origin of *Citrus* considering that many *Citrus* were native to the surrounding areas of India, China and Indochina (Tanaka, 1959). Swingle hypothesized that the predecessors of *Citrus* inhabited Melanesia and close islands, and that a few of these ancestor species arrived to mainland Asia following a predominant East to West direction, and only then evolved into the major true citrus groups, i.e., pummelos, citrons and mandarins (Swingle and Reece, 1967). Despite the discrepancies between both botanists, Swingle and Tanaka performed an exhaustive work in describing many different *Citrus* and their native habitats.

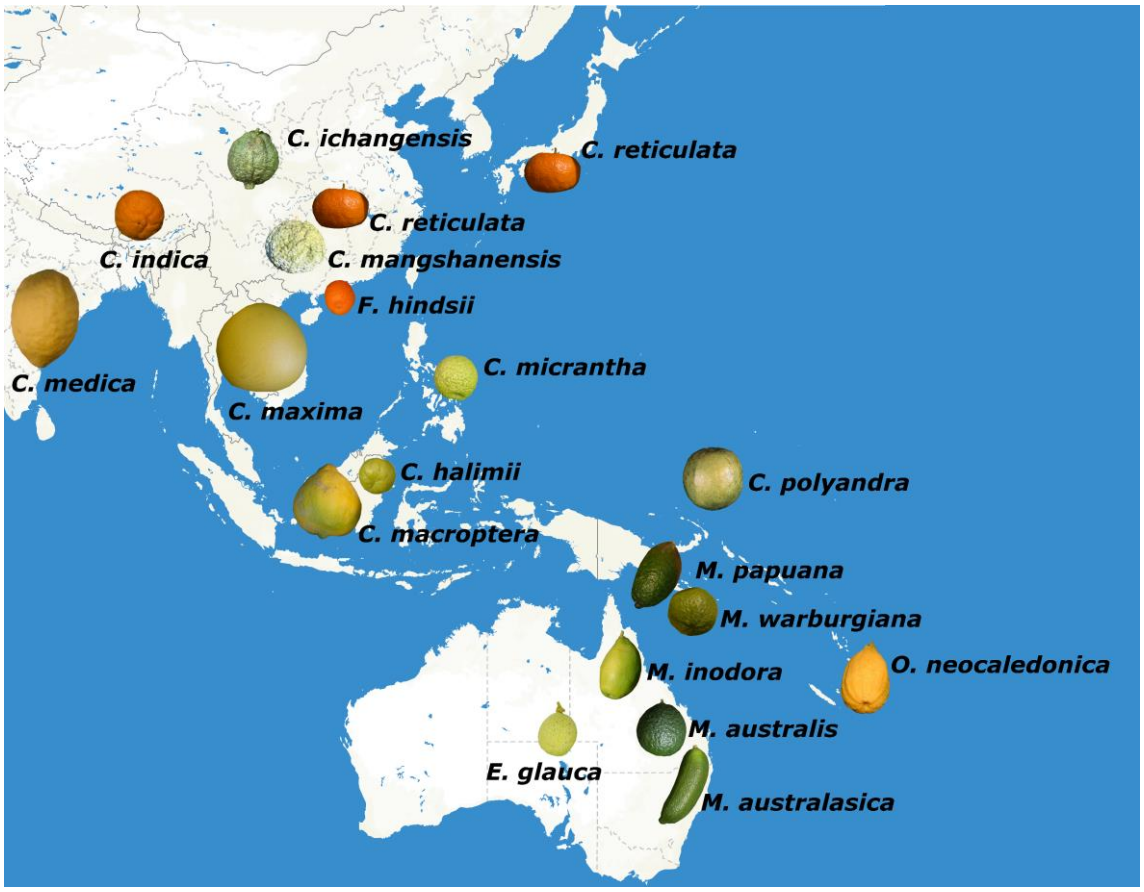


Figure 2: Geographical distribution of *Citrus* pure species. The current native area for the different existing *Citrus* pure species is shown for the species analyzed in this doctoral thesis. Members of *Citrus*, *Fortunella*, *Microcitrus*, *Eremocitrus*, *Clymenia* and *Oxanthera* are shown. The distribution of each species was retrieved from published data (Tanaka, 1954; Swingle and Reece, 1967; Deng *et al.*, 2020; Talon *et al.*, 2020). Pictures of the different *Citrus* species were retrieved from the Citrus Variety Collection maintained at the University of California Riverside and from Talon *et al.*, 2020. The map data was retrieved from MapBox and OpenStreetMap.

Citrus trees are found in the wild in a wide territory extending from India to New Guinea and from Korea to South Australia (Figure 2). In brief, India is the native habitat of citrons and *Citrus indica*, a mandarin-type *Citrus*. Pummelos are naturally found in Indochina and the West part of the Malay Archipelago. Mandarins are originally from the Nanling region in Southwest China (Wang *et al.*, 2018*a*), although some of them, most notably the satsumas and the Tachibana mandarin, are native from Japan. Other important *Citrus* are the papedas, which appear to form a paraphyletic group. Most papedas are found only in specific islands of the Malay Archipelago, such as the Philippine *C. micrantha*, while the Ichang papeda *Citrus ichangensis* grows in Central and Southwest China, where it resists the harsh winters of the area. *P. trifoliata*, another cold-resistant species closely

related to *Citrus*, is native to Central and Northern China. Many other *Citrus* relatives have been described in New Guinea and other Melanesian islands, while at least three different limes are native to Mainland Australia. Among the Australian limes, the round lime and the finger lime (*Microcitrus australis* and *Microcitrus australasica*) grow in the rainforests of the East Coast, but the desert lime *Eremocitrus glauca* is found in the desert and present multiple adaptations to cope with such an extreme environment.

3.1 *Citrus* taxonomy and early phylogenies

Since the studies of Swingle and Tanaka were published, many authors have tried to elucidate the taxonomical and phylogenetic relationships among different citrus cultivars. In 1976, based on phenotype data, an exhaustive characterization was performed on 146 traits from 43 different citrus including members of *Citrus*, *Microcitrus*, *Eremocitrus*, *Poncirus* and *Fortunella*. The authors identified citron, pummelo and mandarin as the three main species from which most commercial varieties were derived (Barrett and Rhodes, 1976). Subsequent works built different phylogenies, based on molecular markers such as microsatellites or random markers (Fang *et al.*, 1998; Nicolosi *et al.*, 2000), and more recently genic sequences (Ramadugu *et al.*, 2013). Unfortunately, these works usually disagreed one with each other and were generally inconclusive, hindering the establishment of a consensus phylogeny.

Several studies that used chloroplastic (Bayer *et al.*, 2009) or mitochondrial sequences (Froelicher *et al.*, 2011) were more coincident, although the placement of some clades, such as that of the papedas, for instance, was still elusive. Later, Carbonell-Caballero *et al.* (2015) generated a *Citrus* phylogenetic tree based on the whole chloroplast sequence of the major citrus groups. This tree clustered citrons and Australian limes together, as observed with previous phylogenies built on partial chloroplastic sequences (Bayer *et al.*, 2009), despite their clearly different phenotypes and traits. It is worth to note that chloroplast and mitochondria are maternally inherited organelles, and therefore, the phylogenies based on their DNA sequence do not necessarily agree with the nuclear phylogeny (Wang *et al.*, 2014). All in all, the diverging results obtained by all these works highlighted the inherent difficulties to clarify the taxonomy and phylogeny of the genus *Citrus*, an issue that remained unsolved until very recently.

In parallel with the debate about *Citrus* phylogeny, their centers of origin and dispersal were also under discussion. Although many researchers agreed with the East Himalayan origin, other works suggested Australia as the original habitat for the *Citrus* ancestor (Beattie *et al.*, 2006, 2009). Later, Xie *et al*. (2013) reported the occurrence of a leaf fossil that gathered several traits common to different *Citrus* clades. The fossil was assigned as to a new *Citrus* species, *Citrus linczangensis*, and was considered an ancestor of the whole *Citrus* genus. It was found in the Chinese province of Yunnan, in a geological layer dated from the Late Miocene, 5 to 11 million years ago (Mya), in an age range that agreed with the dating proposed with the chloroplastic phylogeny (Carbonell-Caballero *et al.*, 2015).

3.2 The phylogenomic approach to the *Citrus* taxonomy

In 2018, a comprehensive phylogeny of *Citrus* was released, using whole genome sequencing data of pure species while leaving aside the admixed varieties. The age estimation of *C. linczangensis* was used to date the speciation events disclosed in the tree (Wu *et al.*, 2018). The tree revealed that both Asian and Australian citrus diverged from a common ancestor that probably existed in a region limited by Northern India, South West China and north of current Myanmar, approximately 8 Mya (Figure 3). The speciation process occurred in a relatively short period of time of around 2 million years, generating many of the current *Citrus* species (Wu *et al.*, 2018). During this time, two main clades arose, one comprising true papedas (*C. micrantha*), citrons (*C. medica*) and pummelos (*Citrus maxima*), while the other included mandarins (*C. reticulata*), kumquats (*Fortunella spp*.) and Australian limes (*M. australis*, *M. australasica* and *E. glauca*). According to the authors, two other species, *Citrus mangshanensis* and the Ichang papeda *C. ichangensis* diverged earlier, at the beginning of the original *Citrus* radiation. This radiation occurred during the Late Miocene, when global temperatures progressively decreased all around the Earth and the current equator-poles temperature gradient were established (Herbert *et al.*, 2016). In South East Asia, the global cooling caused an aridification process, characterized by weaker summer monsoons and the establishment of a strong seasonal regime (Clift *et al.*, 2014; Holbourn *et al.*, 2018). This evidence led to propose that the original *Citrus* radiation was triggered by the sudden change in the environmental conditions that characterized the Late Miocene, and particularly those taking place in South East Asia (Wu *et al.*, 2018). Indeed, this radiation is not an isolated

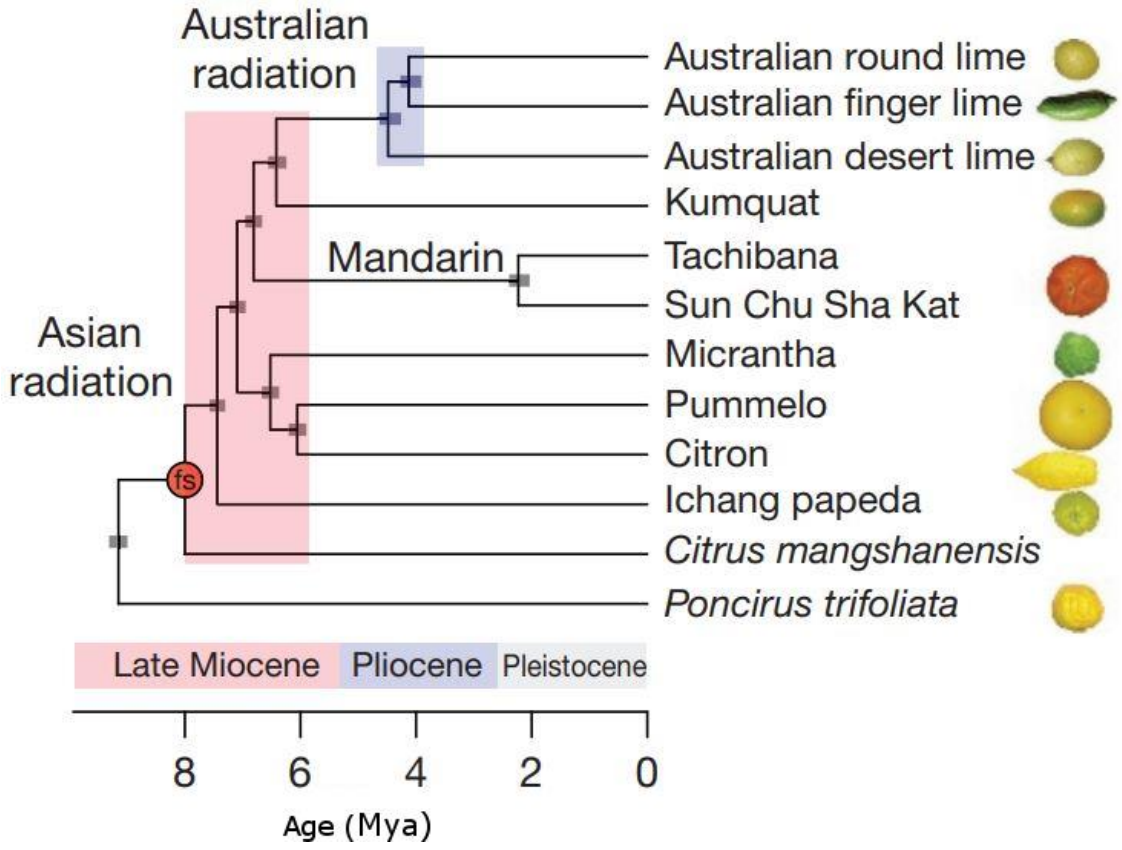


Figure 3: Dated phylogeny of the genus *Citrus*. Chronogram of *Citrus* speciation based on a concatenated genome-wide SNP set. Extracted from Wu *et al*., 2018.

rare event, as many other plant and animal species have gone through similar events in the Himalayan foothills (Wen *et al.*, 2014; Favre *et al.*, 2015; Xing and Ree, 2017).

That most of the *Citrus* species arose in a rapid speciation has some implications that may be related to previous incongruences among the several phylogenies presented. During the citrus radiation, most species diverged in around 2 million years, a period of time that may appear insufficient to fix the ancestral copies of specific locus that existed prior to a given speciation. This phenomenon, called incomplete lineage sorting (ILS; (Maddison, 1997), might have a major effect in distorting the phylogenetic inference.

In summary, several questions about the origin and evolution of *Citrus* now have plausible explanations, although a solid and complete phylogeny of the genus *Citrus* is still far from being resolved. For example, the work by Wu *et al*. (2018) includes representatives of the major *Citrus* clades (pummelos, citrons, mandarins, papedas, kumquats and Australian citrus) but leaves aside interesting species, such as the mandarin-like *C. indica,* also known as the Indian wild orange, as well as other Pacific

species such as the related genera *Clymenia* and *Oxanthera* or the Australian and New Guinean *Microcitrus* (Wu *et al.*, 2018).

The genus *Citrus*, *sensu stricto*, comprises a still elusive number of species that have been naturally evolving during few millions of years in a vast subtropical area of the world. The phylogeny of this genus was clearly marked by the radiation event, but the genomic determinants driving their evolution have not been yet explored. Indeed, *Citrus* genomes display a well-conserved synteny (He *et al.*, 2020), despite some specific gene families being expanded in the genome of some species of this genus (Gonzalez-Ibeas *et al.*, 2021*a*,*b*, in press). However, other factors may have also played a role during the speciation and diversification of the *Citrus* species. Mobile elements, for instance, represent an important source of variability, but their effects during the evolution of the genus *Citrus* remain largely unexplored.

4 Mobile elements and genome evolution

Mobile elements or transposons, initially described in maize (McClintock, 1950), are ubiquitous constituents of the genome that have been found in all eukaryotic genomes sequenced so far (Wicker *et al.*, 2007). Although they account for over 80% of the total genome size in maize (Baucom *et al.*, 2009), in other plants they generally represent at least 20% of the genome (Wicker *et al.*, 2018; Xie *et al.*, 2019; Quesneville, 2020), and 20 - 40% in the case of *Citrus* (Wu *et al.*, 2014; Shimizu *et al.*, 2017; Peng *et al.*, 2020). Their ability to propagate themselves within a genome and towards other species explains their high prevalence (Aziz *et al.*, 2010).

4.1 Transposon structure and classification

Mobile elements have been intuitively classified in two major classes: class I elements rely on an RNA intermediate for their transposition and are called retrotransposons, while class II elements lack such an intermediate and are called DNA transposons (Finnegan, 1989). In general, retrotransposons are transcribed by the host genome machinery, the RNA product is retrotranscribed back to DNA, and are finally inserted elsewhere into the genome, generating a new copy in each transposition cycle (Boeke *et al.*, 1985). In contrast, DNA transposons do not copy themselves on each transposition and rather jump across the genome, moving in a cut-and-paste manner (Greenblatt and Alexander Brink,

1963). For multiplication, they rely on mechanisms exploiting the DNA replication machinery of the host genome. These general rules have remarkable exceptions, as the transposition mechanisms differ enormously even between members belonging to the same class.

In principle, mobile elements carry the genes encoding for the machinery they require to transpose, but as every other genomic region, they are subject to mutation and selection. Transposon genes accumulate mutations, eventually losing their functionality and producing a non-autonomous transposon. This can occur by frameshift mutations or deletions that renders unfunctional genes (Wicker *et al.*, 2003; Fujino *et al.*, 2005; Jo and Kim, 2020). Nested insertions of mobile elements can also disrupt the transposition machinery (Gao *et al.*, 2012; Zhao and Jiang, 2014). Despite not carrying functional genes for their transposition, non-autonomous elements can make use of the machinery of similar autonomous elements, allowing them to replicate and move within the genome (Tsugane *et al.*, 2006; Tanskanen *et al.*, 2007). The most prominent class of non-autonomous family are the miniature inverted repeat elements or MITEs (Bureau and Wessler, 1992), the non-autonomous version of many subclasses of DNA transposons. In spite of their small size and lack of coding sequences, these short sequences appear in large copy numbers and are thought to play key roles in several plant genomes (Guo *et al.*, 2017; Keidar *et al.*, 2018).

4.2 Effect of mobile elements in plant genome evolution

Mobile elements are considered relevant sources of genome variation and evolution. One of the most straightforward effects of transposon activity in any genome is the increase of the genomic space; as new transposon copies are inserted in the genome, its size forcibly increases, as does the proportion of mobile elements in the genome. It is generally accepted that transposable elements represent a major force driving genome expansion in many plant species, including rice (Piegu *et al.*, 2006), sunflower (Mascagni *et al.*, 2017) or maize (Tenaillon *et al.*, 2011). However, genome expansion can be counteracted: in some species, non-homologous recombinations comprising one or more transposons can partially or completely delete a mobile element (Devos *et al.*, 2002). In their recent history, some species have undergone a net genomic shrinkage despite transposon insertions (Vitte *et al.*, 2007; Hu *et al.*, 2011), suggesting that the genome size effect of

the transposon activity depends on both the insertion rate and the retention of the novel insertions into the genome (Legrand *et al.*, 2019).

Besides the changes at the genomic scale, mobile elements can produce mutations on specific genes. For example, novel transposon insertions in a coding sequence or its surroundings might affect the genic structure or provide novel regulatory sequences that alter the expression pattern. A MITE insertion in maize, for instance, induces the expression of nearby genes only under drought stress, adding a new regulatory element to these genes (Mao *et al.*, 2015). Another study in maize found a transposon insertion enhancing the expression of the *tb1* gene, involved in apical dominance, which could explain the apical dominance of maize compared with teosinte (Studer *et al.*, 2011). Other examples are found in rice, where transposons from the *mPing* family appear as major players in generating novel regulatory networks, since their insertions render nearby genes stress-inducible (Naito *et al.*, 2009).

To sum up, mobile elements are a great source of variability, and as such, they have been relevant factors in the evolution and domestication of several crop. Comparative studies between wild and domesticated carrots found specific MITE families expanded only in domesticated individuals and preferentially inserted near transcription factors, suggesting that they might be linked with the domestication process (Macko-Podgórni *et al.*, 2019). In grapevine, where most new varieties are somatic mutants, transposon insertions represent by far the most common mutation event, being several times more frequent than SNPs or indels (Carrier *et al.*, 2012). Very recently, transposon insertion polymorphisms have been used as markers for genome-wide association analyses in tomato, and they have been linked to several domestication traits that were overlooked by SNP-based approaches (Domínguez *et al.*, 2020).

4.3 The importance of mobile elements in the genus *Citrus*

Mobile elements have definitely played an undeniable role in altering the genomic structure of *Citrus*, as revealed by the large amount of specific fruit quality traits linked to transposon insertions in *Citrus*. For example, a single retrotransposon insertion has been linked to the anthocyanin accumulation that characterizes blood oranges, granting them their darker pulp color that gave them their name (Butelli *et al.*, 2012). This is a classic example of the effect of mobile element insertions in the promoter region, but

since its publication, many other *Citrus* key traits have been also linked with such kind of insertions. Pulp acidity (Butelli *et al.*, 2019), carotenoid content (Zheng *et al.*, 2019) and even the mutation determining apomixis (Wang *et al.*, 2017*b*) are all at least partially related with transposon insertions in specific genes involved in these traits. Hence, it appears evident that mobile elements represent a recurrent force capable of altering the genomic structure of *Citrus*.

Besides these examples, focused on single genes, genome-wide transposon surveys have been also carried out in *Citrus*. MITE insertion patterns have been studied across six different *Citrus* genomes and, despite the existence of old insertions well-conserved in all the species analyzed, polymorphic insertions were detected among them (Liu *et al.*, 2019), suggesting that at least some of these elements are still active. Recent MITE insertions were preferentially found in promoter regions, possibly generating novel regulatory networks. Notably, most of these insertions occurred after the radiation of the genus *Citrus* and are therefore specific to species or lineages (Liu *et al.*, 2019). Retrotransposon abundance and distribution in the clementine genome has also been described (Du *et al.*, 2018). In addition, large genomic rearrangements have been related to mobile elements in *Citrus*. In a clementine somatic mutant, a 2 Mb deletion was reported, representing almost 1% of the total genome. This deletion spans over 250 genes, halving the gene dosage and providing an explanation for the earliness phenotype of the mutated cultivar. A detailed study of the deletion flanking regions revealed that the deletion might have been cause by the transposition of the DNA transposon CitMULE (Terol *et al.*, 2015).

In conclusion, the roles of mobile elements in genome evolution and domestication processes have been largely proven in several plants. In *Citrus*, these elements have undoubtedly played a key role in the domestication process, as suggested by the different studies based on their effects in specific genes. Several evidences highlight the potential of transposons as drivers for evolution, but their role in the evolutionary history of *Citrus* prior to domestication has not been fully explored yet.

5 *Citrus* domestication

Domestication in crops is defined as the evolutionary process that alters the phenotype of the plant favoring specific traits associated with human cultivation and use, generating

new varieties or species considerably different from their wild counterparts. In human history more than 1500 plant crops have been domesticated (Meyer and Purugganan, 2013). A general agreement exists around the idea that the first instances of crop domestication started with the discovery of agriculture, between 10000 and 13000 years ago, which occurred independently in several regions of the world (Larson *et al.*, 2014). Notably, many of the first domesticated plants were grains and short-lived crops, whose reduced generation times allow for several selection rounds in a human lifetime (Purugganan and Fuller, 2009). Indeed, grain crops display common traits that remind of an evolutionary convergence process, also known as "domestication syndrome" (Fuller *et al.*, 2014). These recurrent traits include the homogeneity of flowering and ripening times, reduced toxicity and the development of non-shattering seeds, i.e., seeds that are retained in the shoot after ripening, which is considered a turning point in the domestication of every grain crop as it establishes a direct dependence on humans for propagation (Purugganan and Fuller, 2009; Olsen and Wendel, 2013; Dong and Wang, 2015).

Tree crop domestication started later in history, progressed slower and generally produced a less severe domestication syndrome in the domesticated species (Meyer *et al.*, 2012). In fact, there are key differences in the domestication processes in annual and tree crops. In long-lived perennials, breeding by sexual reproduction is generally not feasible because of several biological and reproductive limitations, such as the extended juvenile period or the occurrence of incompatibilities and sterility. Instead, many domesticated perennials appear to be the product of recurrent selections of somatic mutants vegetatively propagated (Gaut *et al.*, 2015). Accordingly, sexual reproduction in perennial crops is a relatively rare event, involving a moderately low number of interspecific crosses between cultivated and wild individuals, as observed in grapevine, olive, apple or cocoa (Myles *et al.*, 2011; Duan *et al.*, 2017; Cornejo *et al.*, 2018; Julca *et al.*, 2020). As a consequence of the vegetative propagation of perennial trees, the genomic structure of hybrids and admixtures is conserved across generations, frozen in time, which allows for interspecific hybrids to be maintained without allele segregation.

Citrus domestication appears to have followed the same dynamics observed in other tree crops, with infrequent cross-breeding that took place early in the domestication process, or even in the wild before the domestication begun (Wu *et al.*, 2014). Indeed, most of the

current commercial varieties are admixtures, including lemons or sour oranges, which were cultivated by the romans more than 2000 years ago. This necessarily means that these admixtures appeared earlier in history, and hence represent ancient events that possibly predated the beginning of the domestication process. In the case of mandarins, a tightly packed relatedness network has been reported, implying that most of the commercial mandarins are related (Wu *et al.*, 2018). This complex network could have been set during *Citrus* evolution, but it is exclusive to mandarins, as it has not been observed in neither citrons nor pummelos. Instead, the authors suggest that such an intricate network is consistent with early human-assisted breeding and selection of mandarins, underpinning the importance of crosses and backcrosses at the beginning of the mandarin domestication (Wu *et al.*, 2018).

It is interesting to mention that crossbreeding commercial citrus in modern citriculture can still partially recover the ancestral phenotypes of the introgressed genomes. For example, a segregant population of two commercial mandarins with multiple pummelo introgressions generated an offspring displaying extreme phenotypes away from the traits of the parents, since pummelo and mandarin alleles segregated in the progeny (Terol, 2020). However, most of the elite *Citrus* cultivars are bud sports which were selected for their improved traits, as it occurs in many other tree crops (Caruso *et al.*, 2020).

In short, it appears that in the early domestication of *Citrus*, the generation of diversity was mostly based on interspecific hybridizations. The development of early grafting techniques (Mudge *et al.*, 2009) and, in particular, the apomictic nature of some citrus permitted a rapid fixation of elite genotypes via the generation of clonal individuals from seeds, which could in turn explain why many commercial *Citrus* today are apomictic (Wang *et al.*, 2017*b*). These admixed individuals, despite not belonging to any of the pure *Citrus* species, are undeniably a core part of the genus, and represent the most recent portion of the genus complex history.

5.1 Molecular basis of *Citrus* agronomical traits

Domestication implies the selection of specific traits of interest, that in fruit-bearing crops are generally linked to fruit production and quality. In *Citrus*, key traits that determine the overall fruit quality, such as acidity, sugar accumulation or fruit color, were very likely early domestication targets.

The characteristic red pigmentation of mandarins, oranges and kumquats was possibly selected because it produces visually appealing fruits that might have been of interest for early growers and breeders. This red color is the result of the accumulation of carotenoids such as violoxanthin, although the major contributors to this trait appear to be C30 apocarotenoids, such as β-citraurin and their derivatives. C30 apocarotenoids are produced by cleaving the carotenoid ring, a reaction carried out by a specific carotenoid cleavage dioxygenase encoded by the *CCD4b* gene (Alquézar *et al.*, 2009). According to these authors, however, this sole gene cannot explain the wide color range displayed by commercial cultivars. Indeed, many somatic mutants of oranges and mandarins with different colors have been described, generally harboring one or more mutations in genes involved in the carotenoid biosynthetic pathway (Liu *et al.*, 2007; Alquézar *et al.*, 2008; Rodrigo *et al.*, 2019). Another level of regulation of *Citrus* carotenoid biosynthesis appears to be linked with substrate availability, as the expression of two β-lycopene cyclases in ripening citrus correlates with the accumulation of downstream pigments (Zhang *et al.*, 2012). Despite the complexity of carotenoid accumulation in *Citrus*, a putative origin of the red peel trait in mandarins and derivatives has been recently suggested (Zheng *et al.*, 2019). In this vision, original mandarins were yellow, and successive mutations on *CCD4b* generated the current allele that confers red color. Duplication of *CCD4* gene produced *CCD4a* and *CCD4b*, allowing the neofunctionalization of the duplicated allele, which is a recurrent feature observed in plants (Flagel and Wendel, 2009). Then in some mandarins, an insertion of a MITE element in the *CCD4b* promoter, followed by a mutation on its sequence, allowed for an overall increased expression of *CCD4b*, ultimately producing the accumulation of C30 apocarotenoids. The authors of that work suggest that red mandarins were possibly selected during their early domestication in South China, and then spread by human action due to their appealing color (Zheng *et al.*, 2019).

Apocarotenoid accumulation is responsible of the red pigmentation of most of the commercial *Citrus* fruits, but other compounds also contribute to the coloration of specific varieties. For instance, anthocyanins are a family of flavonoids commonly found in young flowers and leaves of several *Citrus* species such as citrons, lemons or limes, although fruits generally lack these pigments (Butelli *et al.*, 2017). However, blood oranges are a prominent example of anthocyanin-accumulating *Citrus*. They are somatic mutants of blonde (i.e., regular) oranges, and their peel color is caused by the

accumulation of anthocyanins in their pulp and flavedo. Remarkably, the great majority of blood oranges share a common origin, and their abnormal anthocyanin accumulation is produced by the same mutation that affects the promoter of a MYB transcription factor named *Ruby* (Butelli *et al.*, 2012). *Ruby* does not express in *Citrus* fruits except in blood oranges, provided ripening takes place in cold conditions (Butelli *et al.*, 2012). As stated above, this is produced by the insertion of a mobile element in the *Ruby* promoter. Notably, the vast majority of blood oranges share the same mutation, implying that a single mutagenic event occurred but, as it produced a desirable phenotype, it was selected and propagated following a domestication pattern.

While color makes fruits visually appealing, citrus are commercialized for human consumption. Wild mandarins are in general extremely acidic and unpalatable, suggesting that the reduction of fruit acidity was an early target in *Citrus* domestication. The acidic flavor in commercial varieties happens as a consequence of the vacuolar pH of the vesicle cells of the citrus pulp, which can reach extremely low values in some varieties such as lemons, limes and several wild species (Müller *et al.*, 1996; Hussain *et al.*, 2017). The pH gradient is possible thanks to the buffering action of citric acid (Shimada *et al.*, 2006), whose concentration contributes the most to the final acidity of the pulp (Hussain *et al.*, 2017; Strazzer *et al.*, 2019). Several key genes have been suggested to be responsible of the decreased acidity of domesticated mandarins. For example, by comparing wild mandarins with local varieties, a consistent reduction in the titratable acid content in the pulp of the cultivated mandarins was found. This was associated with the reduced expression of the i*socitrate dehydrogenase* and *aconitase* genes, both implicated in citric acid degradation (Wang *et al.*, 2018*a*). In a parallel study, the *isocitrate dehydrogenase* was already proposed as a major domestication target, since most of the commercial mandarins invariably harbor a pummelo introgression spanning over a *isocitrate dehydrogenase* gene, regardless of the admixture pattern in the rest of the genome (Wu *et al.*, 2018). Changes in specific transcription factors have also been reported, as in the case of *Noemi*, a bHLH transcription factor that regulates both acidity accumulation and the expression of other genes such as *Ruby* (Butelli *et al.*, 2019). Indeed, many of the acidless varieties are somatic mutants of *Noemi*, some of them linked to transposon insertions, although the acidless phenotype of other accessions relies on mutations in other genes involved in citrate accumulation and vacuolar proton pumping (Lu *et al.*, 2016; Guo *et al.*, 2016; Strazzer *et al.*, 2019). The occurrence of that many acidless

varieties, of presumably independent origin, further supports the idea that acidity reduction is a desirable trait which has been long under selection.

Citrus domestication has profoundly shaped fruit appearance and taste, but other traits not linked with fruit quality have been possibly selected as well. The rapid fixation of desired genotypes, for instance, was possible thanks to the natural asexual propagation of citrus trees through apomictic seed dispersal. However, apomixis is not universal in all citrus types, as wild species and several commercial varieties do not present this trait. The apomixis origin and mechanism of action in *Citrus* was recently elucidated (Wang *et al.*, 2017*b*). Apomixis in *Citrus* is a dominant trait caused by a transposon insertion in the promoter region of *CitRWP*, a gene that was already linked with polyembryony in *Citrus* and other species (Waki *et al.*, 2011; Shimada *et al.*, 2018). When a polyembryonic plant acts as the female parent of a cross, its clonal progeny reproduces the maternal genotype, since several nucellar embryo develop in addition to the zygotic one. This does not occur when the polyembryonic plant acts as the male parent. Hence, prior to the development of grafting techniques, the only available method to maintain the genotype of admixed *Citrus* was through apomictic seed dispersal, allowing apomictic *Citrus* to become widely spread. This is supported by the fact that most commercial citrus classes are apomictic in spite of the drastic reduction of genetic diversity inherent to asexual propagation. Apomixis, therefore represents another example of domestication of a trait of very high relevance in the genus *Citrus* (Wang *et al.*, 2017*b*).

Overall, the domestication of the genus *Citrus* appears to have affected to several key genes linked with peel pigmentation and pulp palatability, aided by the asexual propagation of germplasm carrying desirable traits and by the initial interspecific crosses giving birth to the admixed germplasm.

Objectives

OBJECTIVES

To decipher the structure and composition of the genus *Citrus* and the processes that drove its evolution, diversification and domestication, the following objectives were proposed:

1. To elucidate the phylogenetic relationships of the species comprising the genus *Citrus* and to update the existing knowledge. The citrus crown will be rooted on the Aurantioideae subfamily, the phylogenetic placement of the traditional "citrus related genera" will be studied and the effect of rapid radiations on the topology of the *Citrus* phylogeny will be assessed.

2. To study the retrotransposon landscape across several *Citrus* species and close relatives, and to determine its role on the process of citrus speciation. The mechanisms driving their insertion and deletion rates in the different genomic regions will be analyzed. The insertion rates over time will be assessed to understand the effects of the *Citrus* radiation in shaping the retrotransposon activity and vice versa.

3. To characterize the transcriptome of ripening citrus fruits in non-edible pure species and admixed commercial varieties, as an approach to identify major determinants of citrus domestication. Commercially relevant traits such as peel pigmentation, acidity or sugar accumulation will be linked to the differential gene expression levels among species attending the genealogy of the studied varieties

Chapter 1

A reviewed phylogeny of the genus *Citrus* based on whole genome sequencing

ABSTRACT

The phylogeny of the genus *Citrus* has been a continuous matter of debate, hindered by the extensive phenotypic diversity and the prevalence of interspecific admixtures across the commercial species of this genus. Recently, the first published genome-wide phylogeny of the genus revealed that most of the major *Citrus* clades diverged from a common ancestor in a rapid radiation during the Late Miocene, about 8 Mya. Based on this pioneer work, we performed a comprehensive phylogenetic study of the genus *Citrus* and related genera to provide new insights and elucidate major determinants of the processes driving *Citrus* evolution. To this end, an alignment-free method was first used to establish a genome-wide phylogeny of the Aurantioideae subfamily and anchor the genus *Citrus* within, calibrating the speciation times with two independent fossils. Our results suggest that the Aurantioideae subfamily emerged during the early Oligocene, some 32 Mya, and then diversified during this geological epoch generating some of the major clades. During the Oligocene-Miocene boundary 25 Mya, a rapid radiation occurred in the Citreae tribe, followed by multiple long distance migrations from Asia to either Africa or Oceania during the last 10 million years. The phylogeny of the species of the genus *Citrus*, inferred under the multispecies coalescent model, revealed that the initial radiation of this genus 8 Mya cannot be significantly differentiated from a true polytomy. This indicates that the *Citrus* phylogeny adjusts more precisely to a multifurcating tree rather than to a dichotomic model, a proposal that resolve the incongruences presented in previous works and the associated debate. This work also expands the boundaries and the concept of the genus *Citrus* by including the genera *Oxanthera* and *Clymenia* within the *Citrus* clade. This implies the occurrence of at least one long range dispersal event to New Caledonia within *Citrus*. Multiple dispersals between Australia and New Guinea are also deduced from our results, and we hypothesize about plausible dispersal routes for other *Citrus* species.

Keywords: *Citrus*, phylogeny, Aurantioideae, radiation, incomplete lineage sorting, paleogeography

INTRODUCTION

The advancements in genome sequencing and computational biology have opened the door to genome-wide analyses. These analyses were initially costly and restricted to model species and annual crops (Goff *et al.*, 2002; Tuskan *et al.*, 2006; Schnable *et al.*, 2009), but the availability of high-quality reference genomes and the reduced cost of *de novo* sequencing allowed genome-wide studies to be performed in other organisms. For instance, the first *Citrus* reference genomes were published almost ten years ago (Xu *et al.*, 2013; Wu *et al.*, 2014), and since then the field of *Citrus* genomics has progressively gained relevance, given the importance of citrus fruits in the global market.

Citrus belongs to the Aurantioideae subfamily, within the Rutaceae family. Aurantioideae comprises around 30 genera, most of which are generally found in the wild in South East Asia, India, China and Oceania, although some are indigenous from tropical Africa. The wide distribution range of Aurantioideae and their economic importance makes this subfamily an interesting object of study. Several attempts have been made to establish a consistent Aurantioideae phylogeny, starting with the botanical classification of this subfamily in two tribes: Clauseneae and Citreae (Swingle and Reece, 1967). However, molecular phylogenies based on non-coding chloroplast regions reported later that this botanical classification failed to reflect the true phylogeny of Aurantioideae, especially in the case of Clauseneae, which was found to be polyphyletic (Samuel *et al.*, 2001). Subsequent studies based on other chloroplast regions agreed with the polyphyletic nature of the Clauseneae tribe (Bayer *et al.*, 2009; Morton, 2009; Oueslati *et al.*, 2016). The divergence times of different Aurantioideae genera were inferred by Pfeil and Crisp (2008), that using deep fossil calibrations established in 30 million years ago (Mya) the upper limit of the crown age. Based on these results, the authors rejected the vicariance of Aurantioideae ancestors in the Gondwana supercontinent, more than 100 Mya, as a valid hypothesis to explain their distribution across three continents. The authors suggested that these species probably relied in long range transoceanic dispersals for spreading to multiple continents. More recent analyses combining two nuclear genes and a chloroplastic non-coding region agreed with this timeframe and further supported the relevance of transoceanic dispersals (Schwartz *et al.*, 2016). Unfortunately, most of the work in the Aurantioideae phylogeny is based on a restricted number of loci, generally

from chloroplast sequences, which does not necessarily reflect the true phylogeny of these species.

Despite *Citrus* being amongst the most studied genus of Aurantioideae, its taxonomy and that of close genera has been questioned for a long time, and even *a priori* basic concepts such as the number of pure species or the boundaries of the genus have long remained unanswered (Talon *et al.*, 2020). For example, based on botanical traits, the "true citrus fruits" were defined as a group of six related genera: *Eremocitrus*, *Microcitrus*, *Clymenia*, *Poncirus*, *Fortunella* and *Citrus* (Swingle and Reece, 1967). *Eremocitrus* and *Microcitrus*, also known as Australian limes, include the desert lime *Eremocitrus glauca*, the finger lime *Microcitrus australasica* and the round lime *Microcitrus australis*. *Clymenia* was described as a monotypic genus native to New Ireland, an island located to the East of New Guinea. *Poncirus* was also assigned to the "true citrus" group, despite its marked physiological differences. *Fortunella*, which comprises multiple kumquat species, was also considered as a "true citrus". Finally, the genus *Citrus* included all citrus species popularly recognized as citrus, such as oranges, lemons, mandarins, pummelos, grapefruits, limes and citrons. Papedas, a type of wild citrus distributed in Melanesia and Central China, were classified in the subgenus *Papeda*, within the genus *Citrus* (Swingle and Reece, 1967).

Posterior analyses using genetic data supported a broader extension of the genus *Citrus*, combining *Eremocitrus*, *Microcitrus*, *Fortunella* and *Citrus* into a single genus, while *Poncirus* was identified as an outgroup by some authors (Ramadugu *et al.*, 2013; Wu *et al.*, 2018), but not others (Nesom, 2014). Moreover, within the genus *Citrus*, the phylogenetic placement of some botanical groups has been controversial. For example, chloroplast-based phylogenies recurrently cluster Australian limes and citrons (Pfeil and Crisp, 2008; Carbonell-Caballero *et al.*, 2015; Oueslati *et al.*, 2016), despite their marked phenotypical differences and their geographical distribution, which represent the extremes of the *Citrus* native area. Another longstanding problem of the early botanical classification is the term "Papeda" and the associated subgenus *Papeda*. As suggested by chloroplastic and nuclear phylogenies, the Chinese Ichang papeda *Citrus ichangensis* and the Melanesian and Philippine papedas including *Citrus micrantha* do not represent a monophyletic clade. Hence, the term "Papeda" lacks phylogenetic support (Ramadugu *et al.*, 2013; Wu *et al.*, 2018).

Most of these incongruences were resolved with the recent establishment of a genome-wide *Citrus* phylogeny (Wu *et al.*, 2018). In comparison with previous approaches, which were focused on few loci or markers, here the authors included more than 300000 single nucleotide polymorphisms (SNPs) distributed across noncoding regions of the genome. Furthermore, the *Citrus linczangensis* fossil, which is considered the ancestor of all *Citrus* (Xie *et al.*, 2013), was used to calibrate the age of the *Citrus* crown. All in all, this phylogeny and the associated chronogram provided the backbone of the *Citrus* phylogenetic tree and dated, for the first time, the citrus speciation process. For example, the polyphyly of the subgenus Papeda was confirmed, as well as the inclusion of *Eremocitrus*, *Microcitrus* and *Fortunella* in the genus *Citrus*. *Poncirus*, a more distantly related genus, was unambiguously located outside of the *Citrus* cladogram and therefore should be considered an outgroup of this genus. In brief, they reported the existence of two major clades, one including citrons (*Citrus medica*), pummelos (*Citrus maxima*), and the Philippine papeda (*C. micrantha*), and another clade comprising mandarins (*Citrus reticulata*), kumquats (*Fortunella spp.*), and the Australian limes (*E. glauca*, *M. australis* and *M. australasica*). Two species were placed out of the rest of *Citrus* crown: the Ichang papeda (*C. ichangensis*) and the Mangshanyegan (*Citrus mangshanensis*), which was initially described as a primitive mandarin (Liu *et al.*, 1990; Wang *et al.*, 2018*a*). In total, ten different citrus species were defined, based on the nucleotide diversity among them.

Remarkably, the citrus speciation processes giving rise to most of these species took place in a relatively short period of time (Carbonell-Caballero *et al.*, 2015; Wu *et al.*, 2018). Assuming that the *C. linczangensis* fossil represented the last common ancestor of all *Citrus*, the main *Citrus* speciation events were estimated to occur in a period no longer than 2 million years, starting 8 Mya (Wu *et al.*, 2018), during the Late Miocene. It is widely accepted that the global climate went through rapid changes in this epoch, producing a worldwide cooling, likely triggered by the reduction of the planetary CO_2 levels (Holbourn *et al.*, 2018; Tanner *et al.*, 2020). In South East Asia, this period of cooling was also accompanied with the aridification of the region (Herbert *et al.*, 2016), generating a series of climatic conditions that might have forced *Citrus* to rapidly diversify and radiate. In fact, many other species inhabiting the same area also underwent rapid radiations at that time (Hodkinson *et al.*, 2010; Favre *et al.*, 2015; Valcárcel *et al.*, 2017), suggesting that the trigger for these events was not intrinsic for each species but rather an external factor affecting all of them, i.e., very possibly the climate change.

While Wu *et al*. (2018) performed a comprehensive analysis of the *Citrus* phylogeny, some species fell out of the scope of their work, leaving their phylogenetic placement unsolved. This is the case of the so-called Indian wild orange *Citrus indica*, which is botanically similar to mandarins but has been clustered near citrons based on molecular data (Gulsen and Roose, 2001; Pfeil and Crisp, 2008). A relatively new *Citrus* species, the Mountain citron *Citrus halimii*, which solely grows over 1000 m above the sea level in Malaysia and Borneo (Stone *et al.*, 1973), is also yet to be studied. Furthermore, two Oceanic genera, *Clymenia* and *Oxanthera*, have been classified as near-citrus fruits possibly linked with the Australian limes, although the connection with the core *Citrus* clade remains unclear (Pfeil and Crisp, 2008; Oueslati *et al.*, 2016). The addition of these species into the current *Citrus* phylogeny would certainly provide a richer perspective evolutionary and biogeographic dynamics of the genus.

In this work, we have explored the taxonomy of Aurantioideae and in particular of the genus *Citrus*. An alignment-free method was used to infer the Aurantioideae phylogeny, reporting the first genome-wide phylogeny of this subfamily. Then, the *Citrus* phylogeny was studied, including the species mentioned above since they had not been classified so far. The addition of these new species provides a more complete framework of the *Citrus* phylogeny, expanding the concept and boundaries of the genus as defined so far. Rapid radiations, as in the case of *Citrus*, pose a major challenge for phylogenetic analysis for many reasons (Whitfield and Lockhart, 2007), generally linked to the succession of short branches followed by long branches and to the process of incomplete linage sorting (ILS). In here we have used a new approach, the multispecies coalescent model (MSC), to face the challenge of disentangling the radiation that gave birth to the genus *Citrus*. The results extend our understanding of the evolutionary, climatological and geographical factors determining the emergence and dispersal of the genus, offering a new perspective of this phenomenon from a phylogenomic perspective.

MATERIALS AND METHODS

Plant material

Most of the plant material used in this work was retrieved from germplasm collections in the Instituto Valenciano de Investigaciones Agrarias in Valencia (Spain). The three

Oxanthera samples were kindly provided by Stéphane Lebegin and Carole Martin from the Institut Agronomique néo-Calédonien, New Caledonia (France). Several samples including many Australian and New Guinean limes were kindly provided by Malcolm W. Smith from the Bundaberg Research Station in Queensland (Australia). Two more samples were retrieved from the Valencian Botanical Garden (Spain). The origin of each sample is shown in Supplementary Table 1. For the sequences retrieved from the NCBI Sequence Read Archive (SRA), the accession number is displayed.

DNA extraction and sequencing

In the de novo sequenced samples, DNA was extracted using the CTAB protocol with minor modifications (Webb and Knapp, 1990). In some specific samples, however, a nuclear DNA extraction method was used as described in Terol *et al*. (Terol *et al.*, 2015). The extraction protocols used for each sample are specified in Supplementary Table 1.

Libraries were prepared using the Illumina TruSeq DNA Sample Prep standard protocol following the manufacturer instructions. Then, fragments of 500 bp were selected and sequenced on a HiSeq2000 instrument using 100 bp paired-end read sequencing as in previous studies (Terol *et al.*, 2015; Wu *et al.*, 2018).

Alignment-free phylogeny

Raw sequencing files were sketched using mash 2.2 (Ondov *et al.*, 2016), with a k-mer size of 21 and a sketch size of 10000, selecting only k-mers with a frequency above 5 to avoid unique k-mers produced that might result from sequencing errors. This operation was repeated 100 times using different random seeds to generate 100 replicates of the analysis. A phylogenetic tree was estimated for each replicate with mashtree 1.1.2 (Katz *et al.*, 2019) and the maximum clade credibility tree was calculated in R with the packages ape 5.4 and phangorn 2.5.5 (Schliep, 2011; R Core Team, 2018; Paradis and Schliep, 2019). To estimate the speciation times, two independent calibration points were used: a *Clausena* fossil dated from 27 million years ago (Pan, 2010) and the *C. linczangensis* fossil (Xie *et al.*, 2013), which we considered to be 8 million years old as in previous studies (Wu *et al.*, 2018). Calibration was performed using the chronos function implemented in the R package ape 5.4 (Paradis and Schliep, 2019).

Read mapping and variant calling

Before mapping, reads with a base quality below 30 for at least 30% of their sequence length were discarded, as well as the read pair. The remaining reads were mapped against the *Citrus clementina* reference genome (Wu *et al.*, 2014) using bwa mem 0.7.17 (Li, 2013). SNPs were called independently in each sample using GATK 4.1.1 in GVCF mode (Van der Auwera *et al.*, 2013) and later merged using the GATK pipeline. A hard filter on SNP quality was applied following the GATK Best Practices. Then, SNPs with a Genotype Quality GQ < 20 in any species and singleton SNPs were discarded from the analysis using bcftools 1.10 (Danecek *et al.*, 2021).

Per sample heterozygosity profiles were generated by counting the proportion of heterozygous SNPs in non-overlapping 50 kb windows along the nine main scaffolds of the *Citrus clementina* reference genome. This process was performed in single-sample VCFs generated as above, requiring a Genotype Quality GQ over 20 on every SNP to be considered.

A total of 200 non-coding unique regions were selected for further analysis. Unique regions of the *C. clementina* reference genome (Wu *et al.*, 2014) were determined using genmap 1.3.0 (Pockrandt *et al.*, 2020). A region was considered unique only if all of the 50-mers of the region are unique in the reference genome. Annotated genes were discarded from the analysis to reduce the effects of natural selection in sequence variation, excluding also the 500 bp flanking each gene using bedtools 2.27.1 and bedmap 2.4.35 (Quinlan and Hall, 2010; Neph *et al.*, 2012). Then, the SNP coverage in each region was assessed and those with the highest nucleotide diversity were selected if they were at least 500 kb away from other selected sequences. In total, 200 regions of 5 kb each were selected. The sequence of each region was reconstructed *in silico* using the called SNPs and bcftools 1.10 (Danecek *et al.*, 2021), coding heterozygous SNPs according to the IUPAC nomenclature.

Concatenation and species tree phylogeny

For the concatenation-based phylogenetic estimation, the 200 reconstructed sequences were joined together to form a single alignment. Phylogenetic inference was performed using RAxML 8.2.12 with the GTR+GAMMA model and 500 bootstrap replicates

(Stamatakis, 2014). The best scoring tree was selected and branch support was estimated based on bootstrap values. Parallely, the concatenated dataset was used to generate a species network using the NeighborNet algorithm implemented in SplitsTree 4.17.0 (Bryant and Moulton, 2004; Huson and Bryant, 2006).

For the species tree summary-based phylogeny, the maximum likelihood tree was estimated for each of the 200 reconstructed sequences independently using RAxML with the GTR+GAMMA model and 500 bootstrap replicates. Prior to summarizing the gene trees into a single species trees, nodes with bootstrap support below 10% were collapsed into polytomies using the Newick Utilities toolset 1.6 (Junier and Zdobnov, 2010), as it can improve the performance of summary methods (Zhang *et al.*, 2017). Then, collapsed trees were used to reconstruct the species tree using ASTRAL 5.7.5 (Zhang *et al.*, 2018; Rabiee *et al.*, 2019). Species were defined based on the tree topology of the k-mer and the concatenation topologies as described in Table 1. The species assignment was provided to ASTRAL and *Poncirus* was used as the outgroup.

Table 1. Species assignment of the *Citrus* analyzed samples.

Assigned species	**Accession Number**
Citrus medica	ivia_217
	ivia_112
	ivia_320
	ivia_1051
	ivia_317
	ivia_322
Citrus indica	ivia_1091
	ivia_1163
	ivia_1018
Citrus maxima	ivia_326
	ivia_1209
	ivia_011
	ivia_328
	ivia_327
Citrus micrantha	ivia_135
Citrus macroptera	ivia_1176
Oxanthera neocaledonica	ivia_1085
	ivia_1084
	ivia_1086
Clymenia polyandra	ivia_1025
	ivia_1081
	ivia_1164

Table 1 (continued)

Assigned species	Accession Number
Eremocitrus glauca	ivia_1174
	ivia_1089
	ivia_1083
Microcitrus australis	ivia_324
	ivia_1160
	ivia_1179
Microcitrus australasica	ivia_107
	ivia_1172
	ivia_1159
Microcitrus garrawayi	ivia_1168
Citrus gracilis	ivia_1173
Microcitrus inodora	ivia_1177
Citrus wakonai	ivia_1175
Microcitrus papuana	ivia_1166
Microcitrus warburgiana	ivia_1178
	ivia_1024
Citrus mangshanensis	ivia_329
Citrus ichangensis	ivia_1053
	ivia_319
Citrus reticulata	ivia_5137
	ivia_5124
	ivia_5121
	ivia_5123
	ivia_5132
Citrus reticulata tachibana	ivia_5135
Citrus halimii	ivia_1068
Fortunella spp	ivia_1219
	ivia_1060
	ivia_1074
	ivia_1073
Poncirus trifoliata	ivia_114
	ivia_020

Finally, the presence of polytomies in each quartet tree of the species tree was studied using the polytomy test implemented in ASTRAL (Sayyari and Mirarab, 2018). In short, this approach tests in every possible the quartet trees the probability to reject a null hypothesis in which the branch length equals zero based on the quartet frequency of each possible quartet topology. Nodes where the null hypothesis cannot be rejected are not statistically different from a polytomy. In all cases, the significance threshold was corrected using the Bonferroni-Hochberg correction.

StarBeast2 phylogeny

The species tree was also inferred using the full Bayesian multispecies coalescent model implemented in StarBeast2 0.15.11 (Ogilvie *et al.*, 2017). This allows for the co-estimation of gene trees and species trees, as well as relevant parameters such as mutation rates, branch lengths or population sizes. Despite the computational improvements of StarBeast2, a direct analysis of a dataset of 200 loci across 23 species is still unfeasible, as the Monte Carlo Markov Chain (MCMC) can fail to converge in a reasonable amount of time. Hence, the dataset was divided into 10 subsets of 20 loci each.

Species were manually assigned as displayed in Table 1, and their effective population size (Ne) was estimated as a constant value per branch. For each species and ancestral branch, the molecular clock was allowed to vary following the uncorrelated log-normal relaxed clock implemented in the StarBeast2 package (Drummond *et al.*, 2006; Ogilvie *et al.*, 2017). For each locus, an independent HKY substitution model (Hasegawa *et al.*, 1985) was used with four different substitution rates, using empirical nucleotide frequencies. More complex models, such as the General Time Reversible model (Tavaré, 1986), were discarded as they failed to reach convergence in a reasonable amount of time.

Two independent calibration points were used to infer the speciation times. The *Citrus* most recent common ancestor was calibrated using a broad gamma distribution with 95% of the prior density found between 5 and 13 Mya and the mean at 8 Mya, that roughly matches the age of the fossil *C. linczangensis* (Xie *et al.*, 2013). Independently, the species tree root, that is, the *Citrus* and *Poncirus* most recent common ancestor, was dated using a gamma distribution with 95% of the prior density assigned of a wide interval between 4.2 and 17.4 Mya and a mean of 9 Mya, matching the early estimate of the *Poncirus-Citrus* divergence made based on deep fossils calibrations (Pfeil and Crisp, 2008). The species tree was estimated using the calibrated Yule model implemented in BEAST 2.6.3 (Bouckaert *et al.*, 2019).

For each set of 20 loci, two independent runs were performed in parallel to assess convergence of the MCMC chains. Each of these replicates ran a Metropolis-Coupled MCMC as implemented in the CoupledMCMC BEAST 2 package (Altekar *et al.*, 2004; Müller and Bouckaert, 2020). Specifically, three different chains ran in parallel: two hot chains that perform larger jumps in the parameter space and a cold chain that better

explores the parameter space in small regions. The CoupledMCMC adaptively adjusts the hot chain temperature to reach an optimal swap acceptance ratio of 0.234 (Atchadé *et al.*, 2011). Only the cold chain was logged in each run.

Each of the runs consisted of 500 million generations. The convergence of the different replicates of each set was assessed by two estimators: the Effective Sample Size (ESS) and the Potential Scale Reduction Factor (PSRF). Parameter convergence was further assessed by visually inspecting the trace plot in Tracer 1.7.1 (Rambaut *et al.*, 2018). Tree topology convergence was assessed using the split frequencies and topological ESS analyses implemented in the R package RWTY (Warren *et al.*, 2017).

RESULTS

Heterozygosity distribution across wild *Citrus* species

The distribution of heterozygosity across different genomic segments was used to discriminate and discard interspecific hybrids, as these are characterized by the presence of high heterozygosity regions in their genome (Wu *et al.*, 2014) and can negatively influence the phylogenetic inference (Wu *et al.*, 2018). Many known pure species, such as *C. medica, C. maxima, C. reticulata, C. micrantha* or *C. mangshanensis,* presented a single peak of low heterozygosity (Figure 1a). We also found a prominent single peak of around 0.1-0.3% heterozygosity in the sequenced members of *Clymenia, C. indica* and *C. halimii* samples, indicating that they are not a product of interspecific hybridization, and therefore are very likely pure species.

Members of *Fortunella* and *Oxanthera,* and most of the Australian and New Guinea limes, also presented a single peak in their heterozygosity distribution (Figure 1a). Even though in these cases it was slightly above 0.5% heterozygosity, it is still below the 1% heterozygosity threshold previously established for *Citrus* (Wu *et al.*, 2018). These species may well contain pure genomes. The case of *Citrus macroptera* deserves a specific comment. *C. macroptera,* as showed below, nested with *C. micrantha* and also contains a chloroplast genetically very close to that of this species (Wu, personal communication). Since its heterozygosity does not exceed 1%, we have included it in the

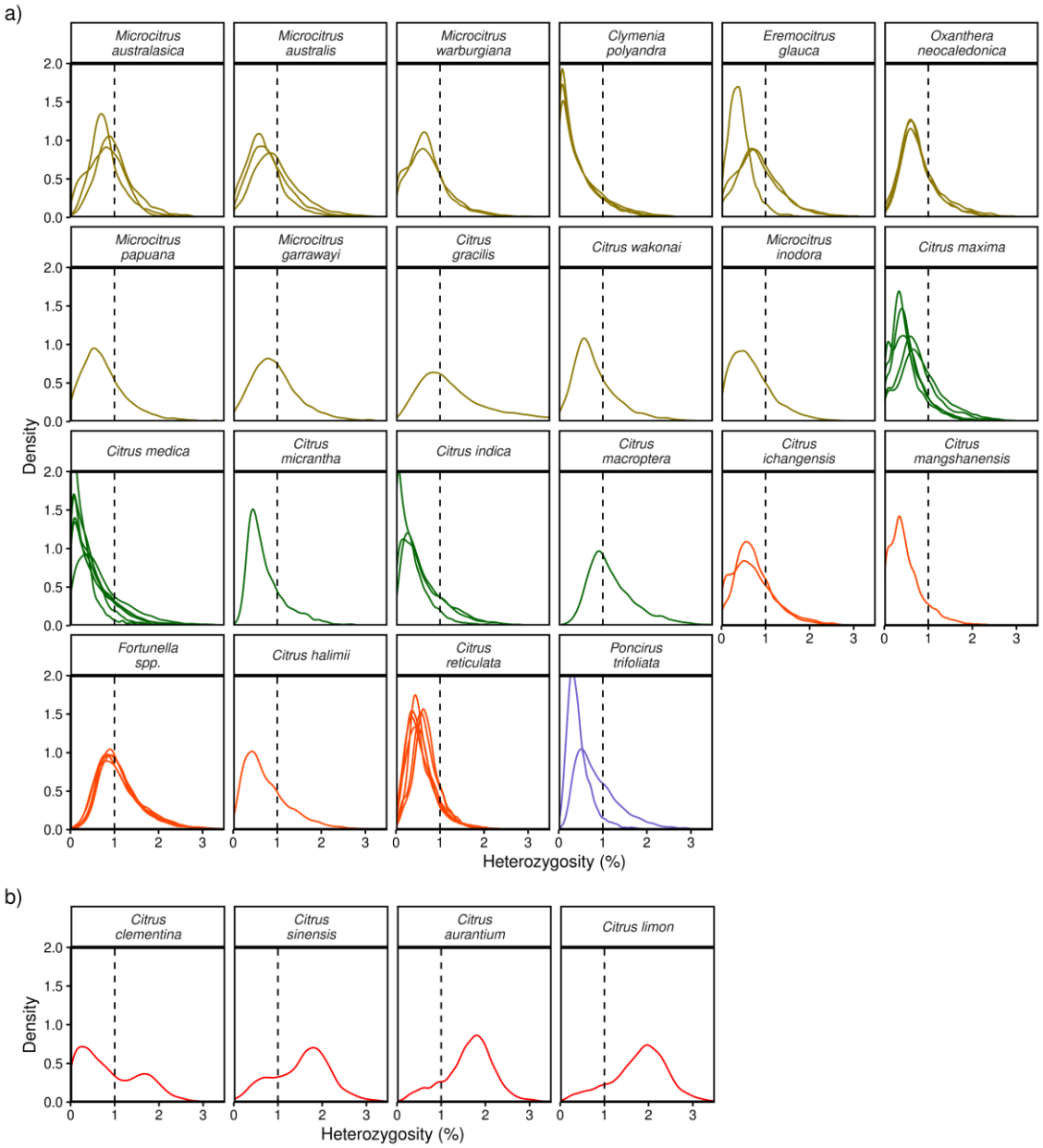


Figure 1: Heterozygosity profiles for pure and admixed species. Segmental heterozygosity was calculated across 500 kb windows of the *Citrus clementina* reference genome for each sample. a) Samples included in the *Citrus* phylogeny inference, belonging to any of the following genera: *Citrus*, *Microcitrus*, *Eremocitrus*, *Fortunella*, *Clymenia* and *Oxanthera*. b) Four admixed *Citrus* for comparison with pure *Citrus* species. Line colors represent the assignment of each species into one of the three major *Citrus* clades: yellow for Oceanic *Citrus*, orange for Chinese *Citrus* and green for South East Asian *Citrus*. The outgroup species *Poncirus trifoliata* is shown in blue and the admixed varieties are shown in red. The dashed line represents the 1% heterozygosity threshold used to define *Citrus* pure species.

set of pure species, but the possibility that *C. macroptera* is a hybrid between *C. micrantha* and a closely related but unknown papeda, although unlikely, may still persist.

Four well-known *Citrus* admixtures were included in the analysis: *C. clementina, Citrus sinensis, Citrus aurantium* and *Citrus limon*, which correspond to the clementine mandarin, sweet orange, sour orange and lemon, respectively (Wu *et al.*, 2018). These samples were included solely to compare their heterozygosity profile with those of the other analyzed samples, as their admixed nature would very likely hinder the phylogenetic inference. As shown in Figure 1b, the heterozygosity distribution of *Citrus* admixtures typically displays a peak at around 1.5-2% heterozygosity. This implies that, at least in some regions of the genome of these samples, both haplotypes are more distant than the 1% threshold, and hence belong to different species.

Alignment-free phylogeny of the Aurantioideae subfamily

Pure *Citrus* and non-citrus genomes were reduced into k-mer sketches, the genetic distances between each sample pair were estimated in 100 independent replicates, and the average distance between replicates was calculated (Supplementary Figure 1). Divergence estimates between the outgroup *Ruta chalepensis* and members of the Aurantioideae subfamily ranged between 15% and 19%, implying a great amount of sequence divergence between them. Within the Aurantioideae subfamily, the genetic distance among species was smaller (up to 13.6%) while within the "true citrus fruits", including *Citrus* and *Poncirus*, the maximum genetic distance was 5.7%, as these species diverged very recently.

The Aurantioideae phylogeny was estimated based on the genetic distances calculated in each replicate. In the maximum clade credibility tree, most of the clades were well supported, with confidence values over 95% (Figure 2). In general, when several samples from the same genus were analyzed, they formed a monophyletic clade (*Murraya, Citropsis, Atalantia, Poncirus* and *Citrus*), with one exception, the genus *Clausena*, that forms a paraphyletic clade with *Glycosmis*. Some of the clusters included genera with native ranges located in different continents. For example, a monophyletic clade clustered the Indian bael (*Aegle*) and three African species (*Aeglopsis*, *Balsamocitrus* and *Afraegle*). Similarly, the South East Asian species *Hesperethusa crenulata* clustered with several species of *Citropsis*, found in tropical Africa.

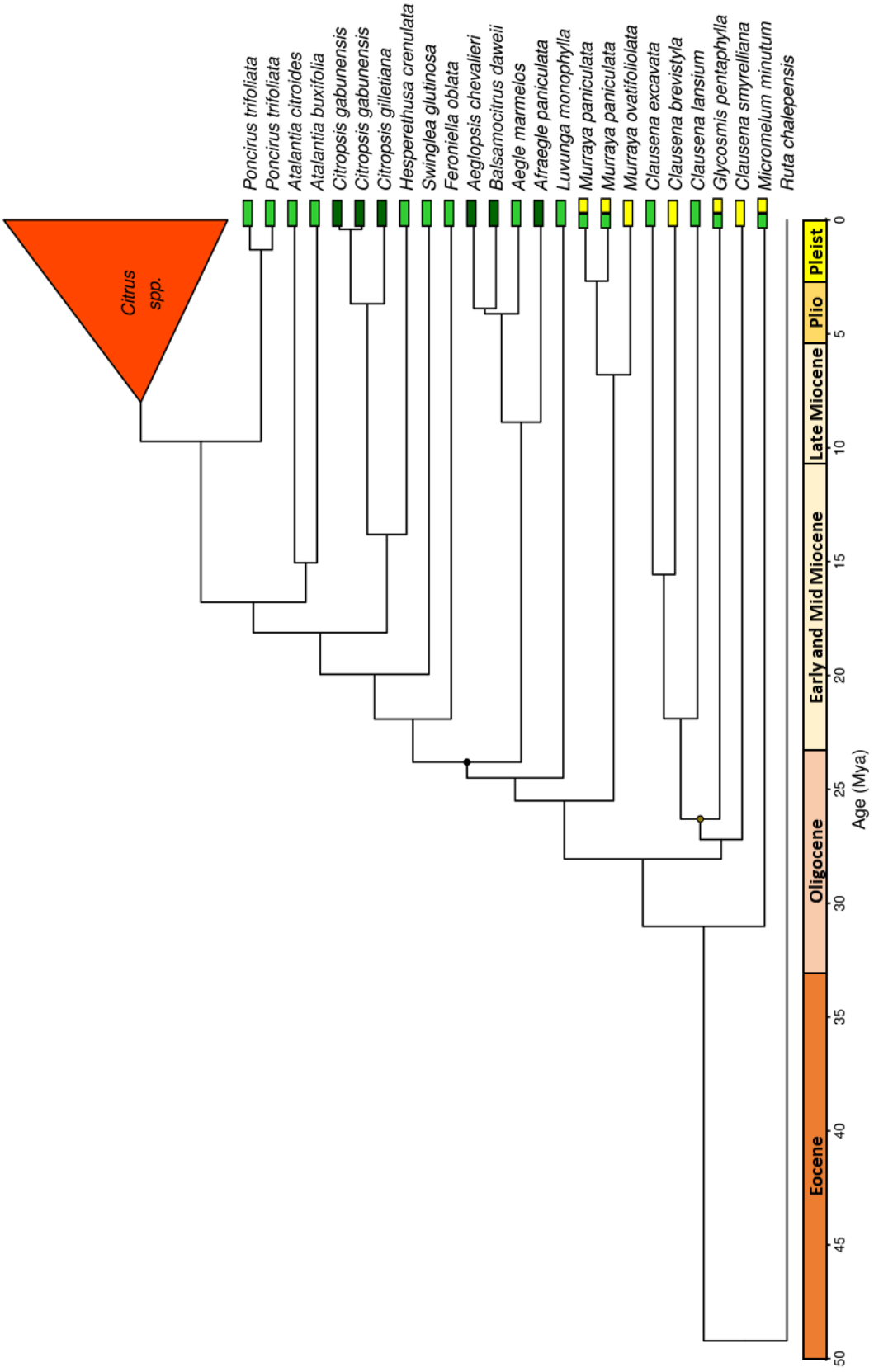

Citrus spp.
Poncirus trifoliata
Poncirus trifoliata
Atalantia citroides
Atalantia buxifolia
Citropsis gabunensis
Citropsis gabunensis
Citropsis gilletiana
Hesperethusa crenulata
Swinglea glutinosa
Feroniella oblata
Aeglopsis chevalieri
Balsamocitrus daweii
Aegle marmelos
Afraegle paniculata
Luvunga monophylla
Murraya paniculata
Murraya paniculata
Murraya ovatifoliolata
Clausena excavata
Clausena brevistyla
Clausena lansium
Glycosmis pentaphylla
Clausena smyrelliana
Micromelum minutum
Ruta chalepensis
Eocene
Oligocene
Early and Mid Miocene
Late Miocene
Plio
Pleist
50
45
40
35
30
25
20
15
10
5
0
Age (Mya)

Figure 2: Phylogenetic tree of Aurantioideae inferred by an alignment-free method. Majority clade credibility tree of from 100 trees calculated in each. The genus *Citrus* is collapsed here but it is displayed in Figure 3. Most nodes displayed bootstrap support values above 95% with minor exceptions. Node colors represent bootstrap support values: black for nodes with 80% - 95% support, yellow for nodes with 50% - 80% support. The native area of the different Aurantioideae species is marked next to their names: yellow for those native to Oceania, bright green for Asian clades and dark green for African clades. The time scale is in million years and was calculated using two calibration points (see Materials and Methods). Gelogical epochs are shown below the time scale.

It is worth to mention that the *Citrus* tree was generally worse resolved, with low support values especially concentrated around short branches, although most samples from each species clustered into monophyletic clades with high support, over 95 % (Figure 3). *Clymenia* and *Oxanthera* consistently clustered with the Australian limes, which formed a monophyletic clade within *Citrus*. The Indian wild orange *C. indica* clustered with citrons, while the mountain citron *C. halimii* clustered with *Fortunella* samples. The Ichang papeda (*C. ichangensis*) and the Melanesian and Philippine papedas (*C. macroptera* and *C. micrantha*, respectively) clustered into two unrelated clades. Three main clades can be distinguished: an Oceanic clade including all samples from Oceania (members of *Microcitrus*, *Eremocitrus*, *Clymenia* and *Oxanthera*, as well as *Citrus wakonai* and *Citrus gracilis*), a South East Asian clade including species found from India to maritime Indonesia (*C. maxima*, *C. medica*, *C. indica*, *C. macroptera* and *C. micrantha*), and a Chinese clade including *C. reticulata*, *C. mangshanensis*, *C. ichangensis*, *C. halimii* and *Fortunella spp*. These three clades generally match the native distribution of the different *Citrus*, although some exceptions exist. For example, *Citrus halimii* is native to Malaysia and Borneo, and *C. macroptera* is widely naturalized from North East India to New Guinea.

Figure 3: Phylogenetic tree of *Citrus* inferred by an alignment-free method. Close up of the genus *Citrus* collapsed in Figure 2, with *Poncirus* as an outgroup. Node colors represent bootstrap support values: black for nodes with 80% - 95% support, yellow for nodes with 50% - 80% support and red for nodes with < 50% support. The three shaded areas correspond to the Oceanic, Chinese and South East Asian clades in yellow, orange and green, respectively.

Citrus phylogeny using concatenation and summarized species tree

In order to explore the reliability and consistency of a specific citrus phylogeny, a total of 200 regions were selected as a representative sample of the *C. clementina* reference genome, to be used in two further approaches (Supplementary Figure 2). First, the representative set was analyzed using a concatenation approach. In the obtained phylogenetic tree, the vast majority of the clades were well-supported, with bootstrap values above 95% (Figure 4). Both the Oceanic and the South East Asian clades recovered in the above alignment-free phylogeny were present also in the concatenation tree with high support, but the Chinese clade was not found. Instead, *Fortunella* and *C. halimii* clustered with the Australian limes. The Mangshanyegan appeared as a sister clade to the rest of the *Citrus* species and the Ichang papeda formed a sister clade with the South East Asian clade. Mandarins were anchored near the base of the *Citrus* crown, where a series of short branches represent the ancestral radiation process that gave rise to the genus. In the Australian clade, a secondary radiation event was also evident.

The *Citrus* phylogeny was also assessed using a summary species tree approach, implemented in ASTRAL, which accommodates incomplete lineage sorting as a source of gene-tree discordance. An independent phylogenetic tree was computed for each locus and the summary species tree was inferred (Figure 5). About 79% of the quartets found in the locus trees were satisfied by the species tree, implying that the proportion of ILS is not negligible. Again, in the species tree the Oceanic and South East Asian clade appeared as highly supported monophyletic clades, while the Chinese clade was divided. In this case the Mangshanyegan clustered with *Fortunella* and *C. halimii*, while the Ichang papeda was close to the mandarins. Short branches characterized the *Citrus* and the Australian radiations, which again concentrated most of the less supported clades.

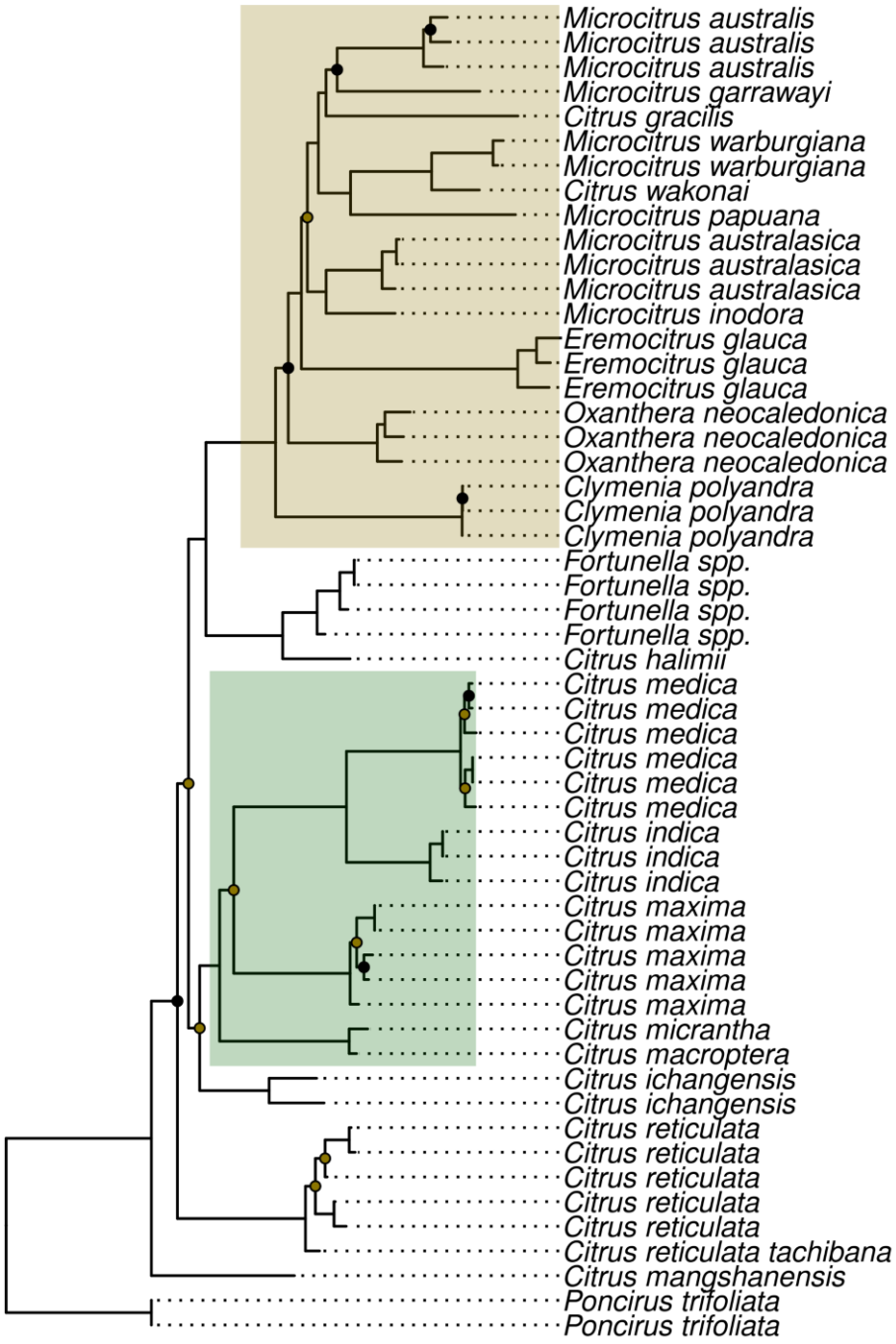


Figure 4: Concatenation-based phylogenetic tree of *Citrus*. Phylogenetic tree inferred by maximum likelihood from the concatenation of the 200 sets into a single alignment. Node colors represent bootstrap support values: black for nodes with 80% - 95% support and yellow for nodes with 50% - 80% support. The Oceanic and South East Asian clades are shown in yellow and green, respectively. The Chinese clade is not shown as it is polyphyletic according to this inference method.

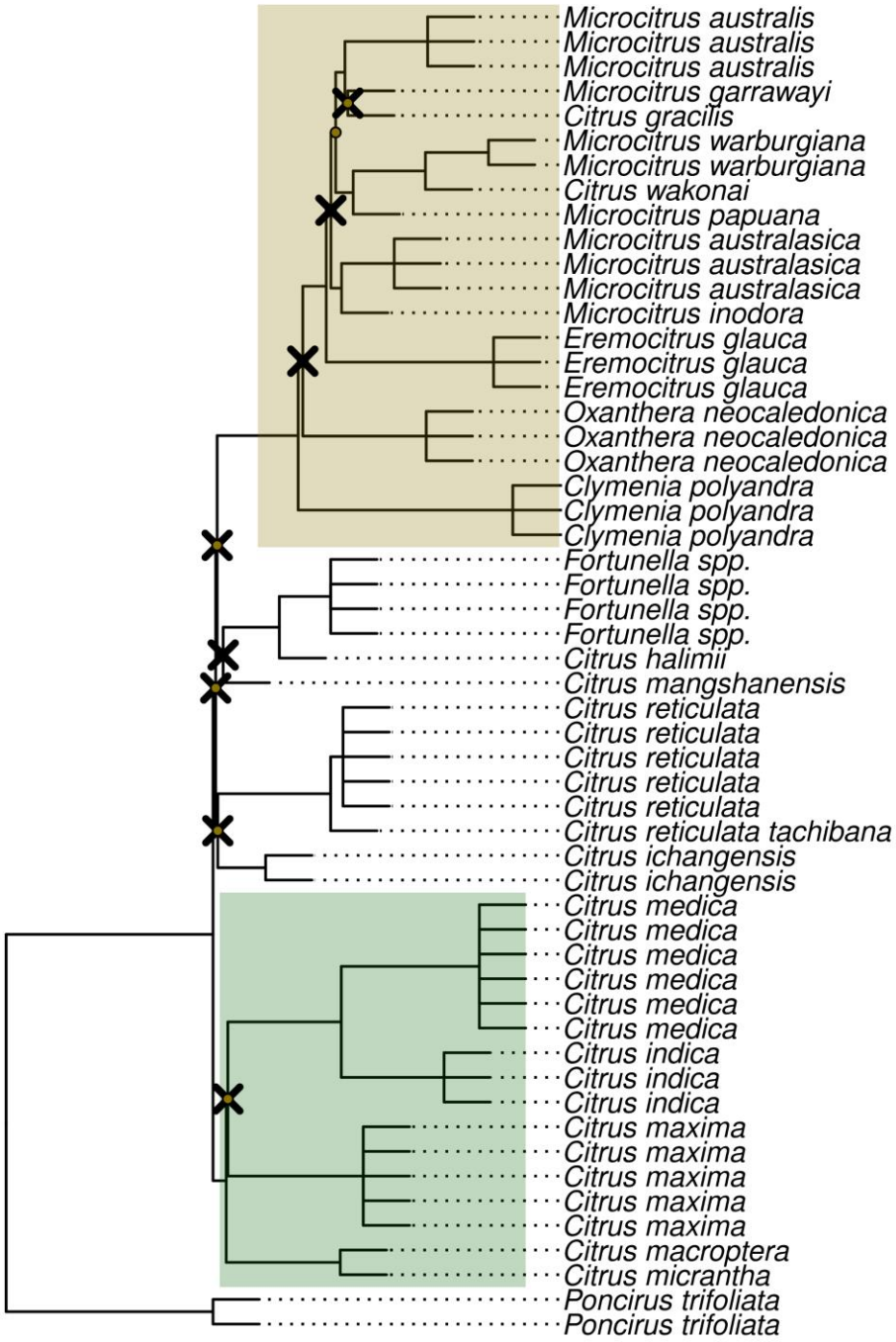


Figure 5: Summary species tree of *Citrus*. Phylogenetic tree inferred by ASTRAL from 200 independent gene trees. Node colors represent bootstrap support values: black for nodes with 80% - 95% support and yellow for nodes with 50% - 80% support. The Oceanic and South East Asian clades are shown in yellow and green, respectively. The Chinese clade is not shown. All the clades statistically indistinguishable from a polytomy according to the ASTRAL-implemented polytomy test are marked with an X.

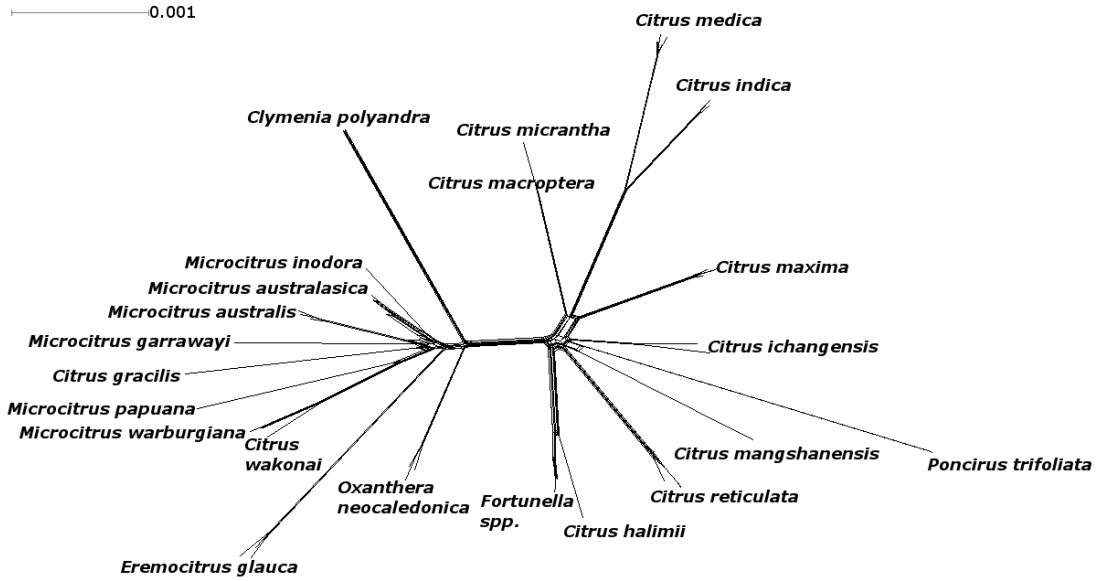


Figure 6: Phylogenetic network of the genus *Citrus*. Phylogenetic network inferred from the concatenated dataset using the NeighborNet algorithm implemented in SplitsTree.

The polytomy test included in ASTRAL was then used to infer which nodes could not be definitively resolved. After correcting the p-values for multiple hypothesis testing, the crown of the South East Asian clade, the *Citrus* crown and some internal nodes of the Chinese and Australian clades could not be considered statistically different from a polytomy, since the null hypothesis stating that the branch length is not zero could not be rejected.

We finally calculated a species network using SplitsTree. The Australian radiation and the *Citrus* radiation are displayed as intricate parts of the network while the rest of the speciation events appear better resolved as they are displayed by either a single edge or a sets of edges without connections to others, as in the case of the *C. indica – C. medica* clade, the *Fortunella spp. – C. halimii* clade or even the whole Australian clade.

Bayesian inference of the *Citrus* phylogeny

In a complementary approach, ten independent sets including 20 loci each were analyzed using StarBeast2. However, after 500 million generations, five sets did not reach convergence, denoted by the lack of topological convergence between replicates as reported by RWTY split frequencies. Only the five sets that reached convergence were further analyzed.

Each of these sets converged into a consistent tree topology: after 250 million generations of burn-in, the two different runs of each set explored the same area of the tree space, and an almost perfect linear correlation between their split frequencies was found. Each replicate of the set converged into similar values for most of the parameters of the model, and the ESS values of both the model parameters and the tree topology in the combined logs were high, as displayed by Tracer and RWTY (Warren *et al.*, 2017; Rambaut *et al.*, 2018). Some parameters displayed ESS values below 100 but visual inspection revealed that they had indeed reached a stationary state. This was not the case for the speciation rate parameter, which failed to converge in the different replicates, although all replicates from all sets produced low values, with speciation rates below 0.2. PRSF values revealed a similar pattern, with well-converged estimates except for the speciation rate parameter.

Despite the topological convergence within sets, when the species trees among sets were compared, different topologies were observed in some cases (Supplementary Figure 3). Notably, specific clades consistently appeared in the different species trees with high support (Supplementary Figure 4). This was the case for the South East Asian clade, including *C. indica* and the citrons as closely related species, with posterior probabilities (PP) above 90% in all but one cases. The *C. indica* and citrons closeness was particularly well supported, with a PP of 1 in every set. Similar results were observed for *C. halimii* and *Fortunella*, which consistently clustered together, as well as the Oceanic clade. Within the latter, *Clymenia* and *Oxanthera* were generally positioned in a basal position, and the Oceanic limes (Australian and New Guinean) were mostly monophyletic. *Eremocitrus* was generally at a basal position compared with other limes, although not always. Similarly, the New Guinean species, conformed by *Citrus wakonai*, *Microcitrus papuana* and *Microcitrus warburgiana* always clustered together with PP over 90%, but their position with respect with other Australian limes was less clear. Overall, the most variable clade is that of the Chinese species, as mandarins, the Ichang papeda, the Mangshanyegan and the *Fortunella – C. halimii* clade were shuffled among sets. Given that each independent set had reached convergence, this might arise from the usage of different loci in each of them.

Given the lack of convergence among sets, the combination of the different trees revealed the less supported clades (Figure 7). In short, the major clade credibility tree obtained from the combination of the five sets roughly matched that of the concatenation approach,

with the notable exception of *C. mangshanensis*, which clustered with the mandarins. Overall, the support values were low for many of the clades, given that each set converged into a slightly different topology.

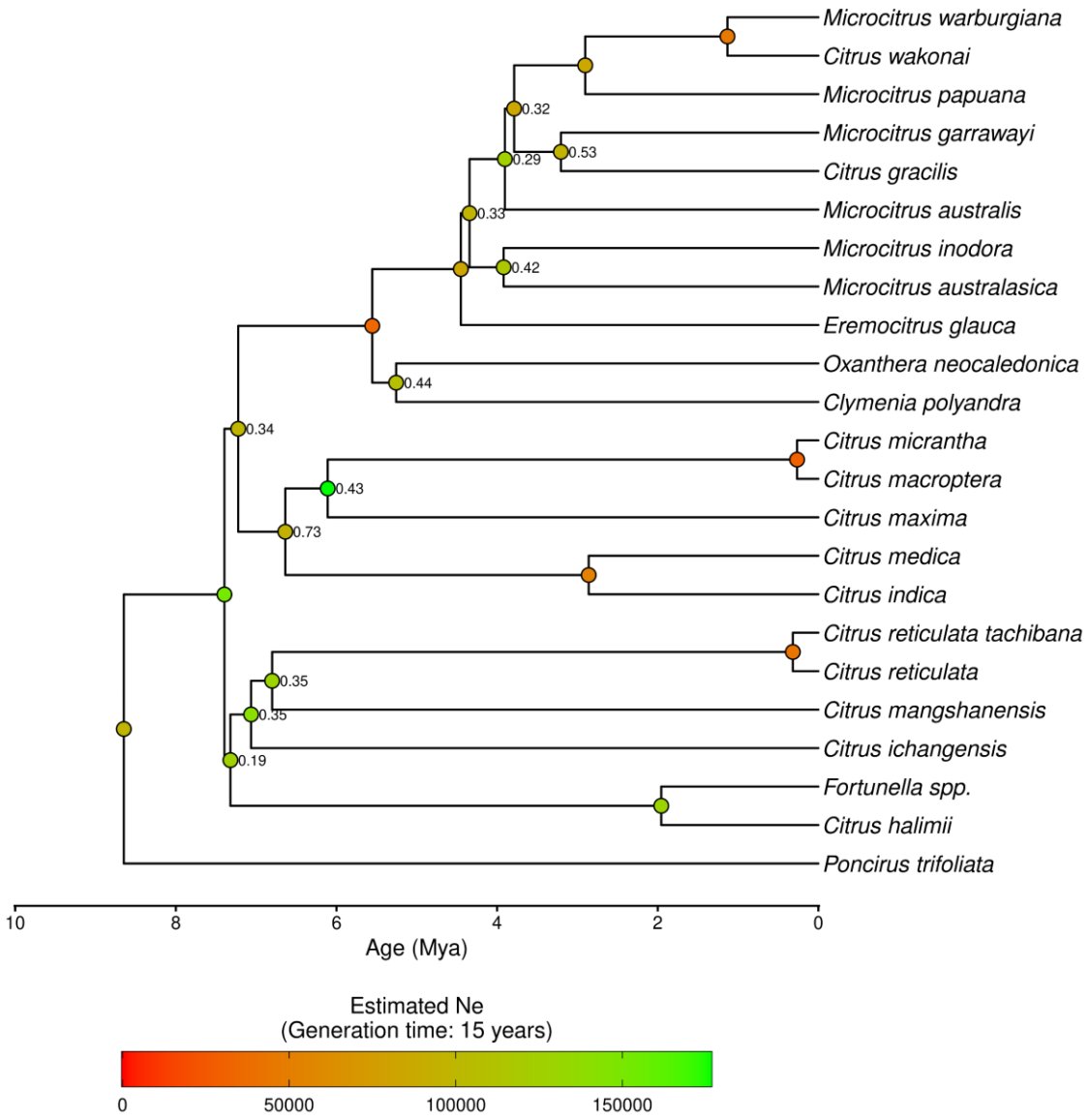


Figure 7: Phylogenetic tree of the genus *Citrus* inferred by StarBeast2. Maximum clade credibility tree for the combined posterior of the five converged StarBeast2 analysis. Posterior probability values below 0.9 are shown next to each node. The node colors represent the estimated population size for each node. Population sizes are scaled by generation times, implying that the values among branches cannot be directly compared.

DISCUSSION

The placement of the genus *Citrus* in the Aurantioideae

Chloroplast-derived sequences have been extensively used to infer the evolutionary history of distant species (Li *et al.*, 2019*b*), given their low mutation rate and conserved synteny (Ravi *et al.*, 2008). However, specific evolutionary events such as hybridizations or non-maternal chloroplast inheritance generate discordances in the chloroplast tree topologies (Bouillé *et al.*, 2011; Bruun-Lund *et al.*, 2017; Lee-Yaw *et al.*, 2018), which can therefore differ from the true species tree (Walker *et al.*, 2019). For instance, chloroplast-based phylogenies of *Citrus* directly conflict with nuclear phylogenies (Pfeil and Crisp, 2008; Carbonell-Caballero *et al.*, 2015; Wu *et al.*, 2018), possibly due to chloroplast capture events (Nagano *et al.*, 2018). Hence, the usage of an alternative data source, not restricted to the chloroplast genome, may provide a more complete perspective of the species phylogeny.

We have used a k-mer based-approach to infer the overall genetic distance among several members of Aurantioideae subfamily, considering the complete genomic space (Supplementary Figure 1). The resulting tree topology (Figure 2) generally agrees with previous estimates (Bayer *et al.*, 2009; Nagano *et al.*, 2018). Our placement of *Citrus* and *Poncirus* agrees with the nuclear tree topology, where *Poncirus* appears as an outgroup of the full *Citrus* clade (Wu *et al.*, 2018). The speciation time of the Aurantioideae crown was previously dated at roughly 20 million years ago and the *Citrus-Poncirus* clade (including the Australian *Citrus*), at 7.1 million years ago (Pfeil and Crisp, 2008). By using a fixed divergence time of 8 million years for *Citrus* and 27 million years for *Clausena* (Pan, 2010; Xie *et al.*, 2013), we dated the Aurantioideae crown at 32 Mya and the *Citrus-Poncirus* split at 10 Mya (Figure 2). In the previous figure, it can be observed that the ancestors of the most relevant clades of the Aurantioideae subfamily were generated during the Oligocene epoch, that extends from about 34 to 23. We report higher age estimates than Pfeil and Crisp (2008), possibly produced by the addition of a *Clausena* fossil as a calibration point, which was unavailable at the time of their study and limits the minimum age of this genus to at least 27 Mya.

We also observe a polyphyly within the genus *Clausena*, as *Clausena smyrelliana* diverges from other species of the genus before *Glycosmis pentaphylla*. *C. smyrelliana* is an endangered species of Queensland, with few known living individuals (Forster, 2000). Reduced population sizes strengthen the effects of genetic drift, as the chance for a neutral allele to become fixated into the population is greater (Woolfit, 2009; Lanfear *et al.*, 2014), which may inflate the genetic distance observed between *C. smyrelliana* and other clausenas. Alignment-free strategies make strong assumptions on the evolutionary history of the studied sequences and fail to model complex evolutionary events, including population bottlenecks (Bogusz and Whelan, 2017; Zielezinski *et al.*, 2019). Therefore, we cannot rule out the possibility that the placement of *C. smyrelliana* is the result of a methodological artifact.

Aurantioideae have been traditionally divided in two tribes: Clauseneae and Citreae (Swingle and Reece, 1967). We found, however, that Clauseneae a was paraphyletic tribe and that the genus *Murraya* appears as a sister clade to Citreae. In contrast, Citreae formed a monophyletic, well-supported clade (Pfeil and Crisp, 2008), with a series of short branches at its crown giving rise to three main clades: the bael clade (*Afraegle*, *Aegle*, *Aeglopsis* and *Balsamocitrus*), the genus *Luvunga* and a third clade including *Citrus*, *Poncirus*, *Severinia* and other far-related genera such as *Citropsis*, *Hesperethusa*, *Feroniella* or *Swinglea* (Figure 2). The existence of short branches near 25 Mya in the Citreae tribe may suggest a rapid radiation at that time, coinciding with the Oligocene-Miocene boundary.

The center of origin of the Aurantioideae is assumed to be located in the South East Asia region. Our results indicate the existence of at least two recent clades that include species solely found in Africa and Asia: the bael clade and the *Hesperethusa-Citropsis* clade, whose crowns are dated at 8 and 14 Mya, respectively (Figure 2). *Murraya* and *Citrus* also display a wide native area, from India or Pakistan to Japan and Australia, despite both diverging less than 10 Mya. The genus *Clausena* alone harbors species natively distributed in tropical Africa, Asia and Australia. Long distance dispersion has been used to explain the vast distribution ranges of some of these genera (Wu *et al.*, 2018; Nguyen *et al.*, 2019), even though it is considered infrequent (Jordano, 2017). Nevertheless, long distance dispersal from Asia has been described for other tree species as well. For example, in the last 15 million years hazels dispersed for Eastern Asia to North America

and Europe (Helmstetter *et al.*, 2019), where they spread at a rate of about 1500 meters per year (Boccacci and Botta, 2009). Another relevant example is that of the genus *Bridelia*, composed by 45 tree species distributed between Africa, Asia and Australia. In this case, at least two independent dispersion events from Asia to Africa, and two other dispersions to Australia, have occurred in the last 10 million years (Li *et al.*, 2009). We hypothesize that the geographical distribution of the different Aurantioideae species are compatible with several independent long distance dispersals from Asia to Africa and Australia in the las 15 million years.

A rapid radiation at the base of the genus *Citrus*

The phylogeny of the genus *Citrus* has been debated for a long time due to the inconsistencies between the reported tree topologies (Nicolosi *et al.*, 2000; Pang *et al.*, 2007; Bayer *et al.*, 2009; Ramadugu *et al.*, 2013; Carbonell-Caballero *et al.*, 2015; Oueslati *et al.*, 2016). The genome-wide phylogeny of *Citrus* partially settled the issue by considering SNPs across the complete genome, fossil calibrations and the ability of admixed individuals to influence the phylogenetic inference (Wu *et al.*, 2018). This revealed the existence of two independent rapid radiations in the *Citrus* phylogeny, in Asia and Australia, approximately 8 and 4 Mya, respectively.

In rapid radiations, high levels of incomplete lineage sorting can be expected, as the number of loci failing to coalesce before speciation increases with short branch lengths and high population sizes (Maddison, 1997). A succession of short branches can also put some of them in the anomaly zone (Degnan and Rosenberg, 2006; Rosenberg, 2013), where the majority of the gene trees do not reproduce the true species tree. In the anomaly zone, the concatenation strategy fails to recover the true species tree, as it recovers the average gene tree topology, which differs from the species tree (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007). Some studies suggest that given enough sites, the concatenation approach might even display high support values for a wrong tree topology, especially near the anomaly zone (Xi *et al.*, 2014; Liu *et al.*, 2015; Mendes and Hahn, 2018; Jiang *et al.*, 2020).

In this work we have used one alignment free analysis and three different phylogeny reconstruction tools, which produced different topologies despite using the exact same data. The concatenation tree reported a highly supported, well-resolved tree (Figure 4).

However, when ILS was taken into consideration using the summary species tree method ASTRAL, the retrieved topology displayed much shorter branches and lower support values for these nodes (Figure 5). Similar results were obtained when gene trees and the species trees were co-estimated using StarBeast2 (Figure 7). The observation that the non-matching tree topologies arise from the different branching order of these nodes indicate that very likely several *Citrus* species emerged almost simultaneously, in a short period of time in an evolutionary timescale. In previous studies, some authors have argued that rapid radiations are accompanied by "hard" polytomies, implying that the true evolutionary history of the clade tree involves a polytomy, produced by concurrent speciation events that generate a truly multifurcating tree (Suh *et al.*, 2015; Dillenberger and Kadereit, 2017; Koenen *et al.*, 2020). In hard polytomies, strictly bifurcating trees fail to capture the underlying evolutionary history of that clade, producing spurious topologies (Bapteste *et al.*, 2013; Hahn and Nakhleh, 2016). In contrast, "soft" polytomies might arise from the lack of resolving power of either the method or the data (Maddison, 1989).

We have tested the existence of polytomies in the *Citrus* phylogeny and found that the *Citrus* basal radiation and the Australian radiation are not statistically different from a polytomy (Figure 5). However, distinguishing between hard and soft polytomies requires a considerable effort, although it has been argued that this might be trivial if the speciation events took place few thousand years apart in an evolutionary timescale (Rokas and Carroll, 2006). Given the convulse climatic history of South East Asia in the Late Miocene (Herbert *et al.*, 2016; Holbourn *et al.*, 2018; Tanner *et al.*, 2020), the existence of other radiations at this time (Wen *et al.*, 2014; Favre *et al.*, 2015) and the many inconsistent citrus topologies reported in this and in other studies (Nicolosi *et al.*, 2000; Pang *et al.*, 2007; Bayer *et al.*, 2009; Ramadugu *et al.*, 2013; Carbonell-Caballero *et al.*, 2015; Oueslati *et al.*, 2016), it appears plausible that the true *Citrus* phylogeny includes two hard polytomies, the natural reflection of a sudden and fast radiation. The phylogenetic network analysis that we performed further support this hypothesis, since the nodes corresponding to these radiations appear as two entangled knots with multiple species arising in parallel.

The colonization of Oceania

Despite the existence of a basal polytomy, that might very well represent the true species tree topology, some clades such as the Oceanic one were consistently retrieved regardless of the approach (Figure 7). This clade includes all the *Citrus* species natively found in New Guinea, Australia and the surrounding Pacific islands: the Australian and New Guinean limes (*Eremocitrus* and *Microcitrus*), *Clymenia* and *Oxanthera*. *Clymenia* was botanically classified as a "true citrus" by Swingle (1967) and is native from the Bismarck Archipelago and the Admiralty Islands, two archipelagos located North of New Guinea. In contrast, the genus *Oxanthera*, endemic of New Caledonia, was initially assigned to a distant group (Swingle and Reece, 1967) and only molecular analyses draw it closer to the genus *Citrus* (Pfeil and Crisp, 2008; Oueslati *et al.*, 2016; Nagano *et al.*, 2018). We found that these two genera are no more distant to the Asian Citrus than the Australian limes, currently considered members of the genus *Citrus* even though the traditional nomenclature, i.e. *Eremocitrus* or *Microcitrus*, is generally preferred for clarity (Talon *et al.*, 2020). According to our data, *Clymenia* and *Oxanthera* diverged from the Oceanic ancestor earlier than the Australian and New Guinea limes, which form a well-supported monophyletic clade.

Within the Australian and New Guinea limes, the desert lime *Eremocitrus glauca* generally appears as a sister clade from the other limes. *Eremocitrus* is a pronounced xerophyte living in the semiarid regions of North and East Australia (Mabberley, 1998). This contrasts with the other Australian limes, most of them included in *Microcitrus*, which are generally found in rainforest margins in the Australian Eastern coast (Salvin, 2008) or dry grasslands on the North in the case of *C. gracilis* (Mabberley, 1998). The New Guinean limes thrive in rainforests or rainforest margins of the Papuan Peninsula and close islands (Forster and Smith, 2010; Lim, 2012). By clustering *E. glauca* in a sister clade of the rest of the limes, our phylogeny matches the botanical classification that differentiated *Eremocitrus* from *Microcitrus* (Swingle and Reece, 1967).

Based on the tree topology, we propose a plausible scenario for the native habitats and geographical distribution of the Oceanic *Citrus*. The common ancestor of the Oceanic *Citrus* arrived in New Guinea, either via long range dispersion from mainland Asia or through Sundaland, an emerged landmass that existed during the Late Miocene and intermittently during the Pliocene, that connected mainland Asia with the area nowadays

occupied by the islands of Sumatra and Borneo (Hall, 2012; Morley, 2018). Many plant species of mainland Asia dispersed through Sundaland, reaching South East Asian islands (Yang *et al.*, 2018) and even the Australian coast, contributing to the generation of the Australian rainforests (Crayn *et al.*, 2015; Yap *et al.*, 2018). From New Guinea, *Clymenia* and *Oxanthera* reached their current distributions, the latter very possibly through long range dispersions or via island hopping across the Solomon Islands and Vanuatu, as many other plants and animals (Nattier *et al.*, 2017). The orogeny of the New Guinean Central Range, which divided the island from West to East starting in the Late Miocene and Early Pliocene (Hall, 2009, 2012), might have imposed a physical barrier isolating *Clymenia* and *Oxanthera* ancestors from that of the limes.

We hypothesize that the ancestor of the Oceanic limes spread through New Guinea. It appears plausible, given the phylogeny and the biology of these limes, that at least two independent migration events occurred from New Guinea to Australia. The first event would have produced the current *E. glauca*, while the second brought the Australian *Microcitrus* to the Eastern Coast, splitting them from the New Guinean *Microcitrus*. Several migrations have been reported between Australia and New Guinea (Mitchell *et al.*, 2014; Tallowin *et al.*, 2020), two territories included in the same single biogeographic area called Sahul, which was connected by land at the time (Hall, 2009; Van Welzen *et al.*, 2011). In the tree genus *Aglaia*, the wild species of the Eastern and the Western Australian coasts arrived to their current locations via two separate migration tracks (Joyce *et al.*, 2021). We believe that two independent migration events of the Oceanic *Citrus* from New Guinea to Australia might explain our results.

The expansion across South East Asia

We consistently recovered a clade containing citrons, pummelos, *C. indica* and two different Papedas: *C. micrantha* and *C. macroptera* (Figure 7). The existence of a monophyletic clade clustering citrons, pummelos and *C. micrantha* has been already reported (Wu *et al.*, 2018), and our results added *C. macroptera* to this clade as a sister taxa of *C. micrantha*. The location of the so-called Indian wild orange *C. indica* close to citrons was hinted based on chloroplast data (Pfeil and Crisp, 2008; Oueslati *et al.*, 2016), although other studies suggested that this species might have a hybrid origin, including introgressions of citron, papeda and mandarin (Garcia-Lor *et al.*, 2015). The considerably

low heterozygosity of the three *C. indica* samples here analyzed rules out this hypothesis and suggests that it should be considered a pure, independent *Citrus* species.

Based on their phenotype, *C. indica* was initially considered a mandarin species (Tanaka, 1954; Swingle and Reece, 1967), but our results clearly indicate that it is closely related to citrons, as the *C. medica – C. indica* split is very well-supported across all the inference methods tested. Since this clade diverged roughly 2.5 Mya, the extensive phenotypical differences that exist between *C. indica* and citrons must have appeared in a short period of time. In other tree species such as peach, most of the currently commercial traits were acquired in a similar timeframe, mostly due to the selection of edibility traits by herbivores (Yu *et al.*, 2018). It is tempting to suggest that the extensive coincidences on the size and color of *C. indica* and mandarins might not be coincidental but the result of a similar process, either via herbivores or by posterior human selection. Although convergent evolution in fruit color and shape have been reported (Pickersgill, 2018), these processes alone cannot explain the similarities in other phenotypical traits such as, for example, the morphology of the leaves.

Apart from citrons and *C. indica*, the South East Asian clade also included pummelos and two papedas, which are natively found in a wide area (Swingle and Reece, 1967). Pummelos are found in the wild in Indochina, the Malay peninsula and close islands, and even though wild populations exist in South China, the Yunnan province on South West China appears to be the center of diversity of the Chinese pummelos (Yu *et al.*, 2017*b*). Some papedas, such as *Citrus macroptera* or *Citrus hystrix*, have wide distributions including Borneo, Sulawesi, the Philippines and New Guinea, while others are restricted to specific islands such as Cebu and Bohol in the Philippine archipelago (*C. micrantha*) or Sulawesi (*Citrus celebica*). Notably, the two species here analyzed, *C. micrantha* and *C. macroptera*, display substantial differences in their heterozygosity profiles, as *C. micrantha* is considerably more homozygous (Figure 1a). The endemism of *C. micrantha*, which is solely found in two Philippine islands, might explain the reduced genetic diversity, as described for other Philippine taxa (Brown *et al.*, 2013; Orsini *et al.*, 2013; Hamabata *et al.*, 2019).

Pummelos, papedas and the citron – *C. indica* clade all emerged very rapidly in a node that we cannot distinguish from a true polytomy. We suggest that these three clades appeared almost simultaneously, with citrons and *C. indica* becoming isolated in the

Eastern Himalaya hills, pummelos colonizing Indochina and the papedas migrating further South, reaching the Philippines and the Indonesian islands from where they dispersed towards New Guinea, long before the ancestor of the Australian limes arrived.

The expansion in South and Central China

In contrast with the Oceanic and South East Asian clades, the existence of a Chinese *Citrus* clade is not supported by our data. Despite displaying low support values, the ASTRAL and the StarBeast2 phylogenies roughly agree in grouping all the Chinese species together (Figure 5 and 7). A similar result was retrieved by Wu *et al*. (2018) except for the Mangshanyegan and *C. ichangensis*, which they placed outside of the main *Citrus* crown. *C. ichangensis* inhabits West and South West China, living in isolated populations given the landscape of the regions (Yang *et al.*, 2017). In contrast, mandarins are native from the Nanling mountains of South China (Wang *et al.*, 2018*a*), although some reached Japan in the last few million years, possibly aided by lowered sea levels during the Pleistocene (Wu *et al.*, 2018). *C. mangshanensis* also inhabits the Nanling mountains of South China, establishing a geographical connection with pure mandarins. Members of *Fortunella* are found mostly in mountainous regions of coastal South China, the island of Hainan and the Malay Peninsula (Deng *et al.*, 2020). Notably, the "mountain citron" *Citrus halimii*, first described at high altitude in Malaysia and later in Borneo, consistently appears as a neighbor taxon of *Fortunella. C. halimii* was initially described as an intermediate species between *Citrus* and *Fortunella* (Stone *et al.*, 1973), and the few molecular studies including *C. halimii* confirmed this hypothesis (Bayer *et al.*, 2009; Oueslati *et al.*, 2016), some even suggesting a hybrid origin for this species (Ramadugu *et al.*, 2013). However, as in the case of *C. indica*, the observed heterozygosity of *C. halimii* suggests that it should be better considered a pure species.

The distribution of these South Chinese species might reflect a single dispersion event from the *Citrus* center of origin, but the lack of support for this clade and the polytomy at its base hinders the formulation of more solid hypotheses. However, if the Oceanic *Citrus* stemmed from those of South China as reported by Wu *et al.* (2018), then the arrival into New Guinea might be better explained by long distance dispersal, possibly from mainland China. Some Australian species reached the island via long distance dispersal, when the Sunda and the Sahul shelves were further apart, covering distances of above 400 km (Crayn *et al.*, 2015; Huang *et al.*, 2016*a*). A plausible migration track could be from

Taiwan to the Philippines and then to New Guinea: dispersals between Taiwan and the Philippines (Tsai *et al.*, 2015), and between the Philippines and New Guinea (Dong *et al.*, 2018) have been already reported. Dispersals through the Sunda plate and then to New Guinea have also been described for other plants (Tsai *et al.*, 2020).

CONCLUDING REMARKS

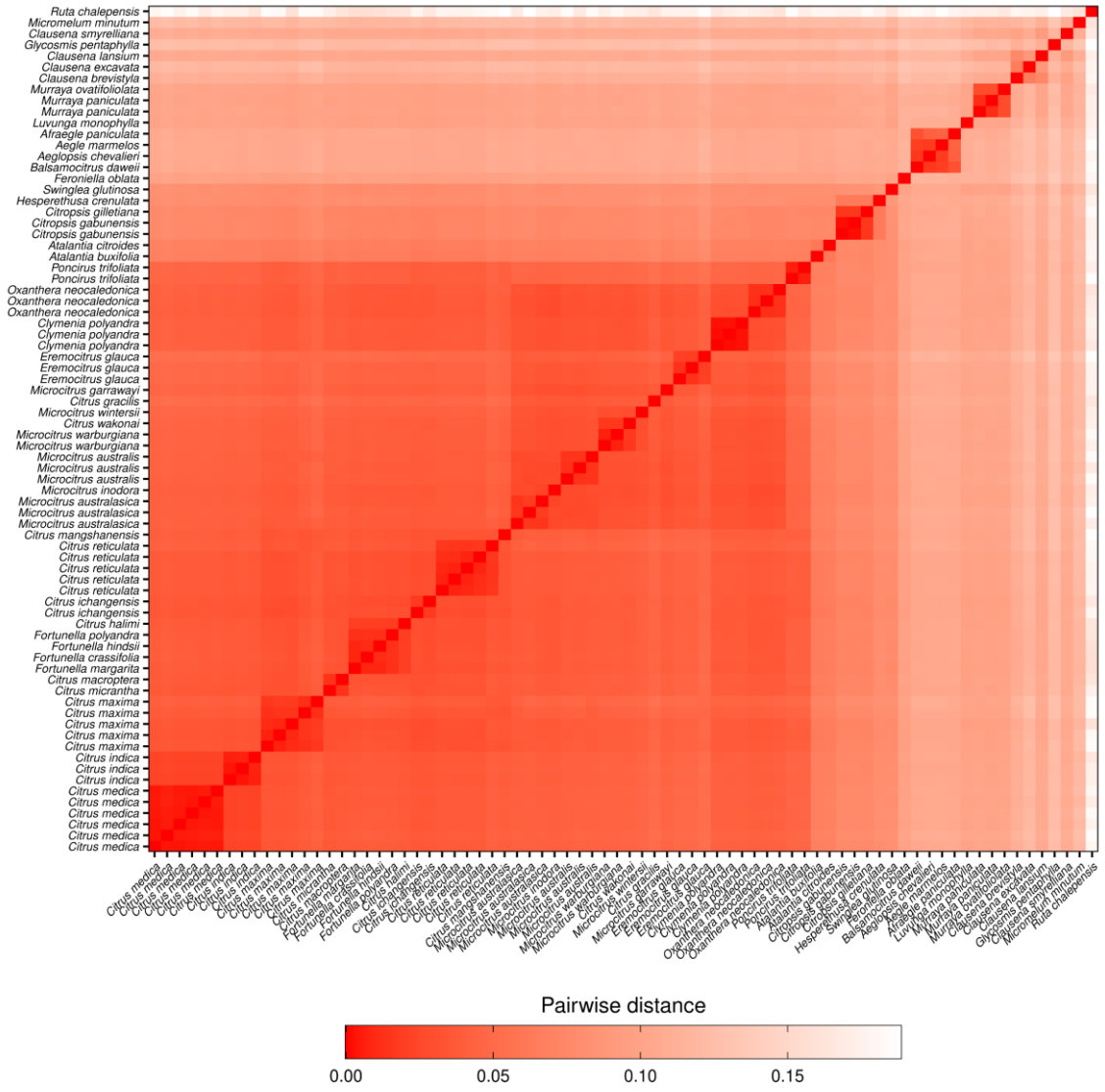
In this study we have analyzed the phylogenetic relationships between the several species of the genus Citrus. First of all, using several relevant genera of the Aurantioideae, we have generated the first genome-wide phylogeny of this subfamily to precisely anchor the citrus crown. According to our data, the Aurantioideae, that expands over three different continents, emerged in the boundaries of the Oligocene-Eocene epochs, some 32 Mya, while the ancestors of the relevant genera diversified mostly during the Oligocene. The Aurantioideae experienced several independent long-distance dispersals that occurred during the last 10 million years, suggesting a highly dynamic range distribution.

The phylogeny of the genus *Citrus* has been for a long time a matter of major controversy since the several analyses published are in general not totally congruent and in some instances even contradictory. The cause of these discrepancies appears to be the rapid speciation of the citrus ancestor that differentiated in a very short time, during the cooling period of the Late Miocene that occurred around 8 Mya. We have used robust methods to infer a consistent species tree under the multispecies coalescent model, a procedure more adequate to the study of rapid radiations. Based on these approaches we propose the occurrence of a true polytomy at the origin of this genus, a suggestion that explains the controversial phylogeny of citrus. We have also found that several genera traditionally defined as "related to citrus" such *Clymenia* or *Oxanthera* are certainly true citrus, thus enlarging the concept of citrus and modifying the definition and boundaries of this genus. These new insights on the tree topology and divergence time estimates allow us to visualize and reconstruct the paleogeographic migration paths for the major *Citrus* species.

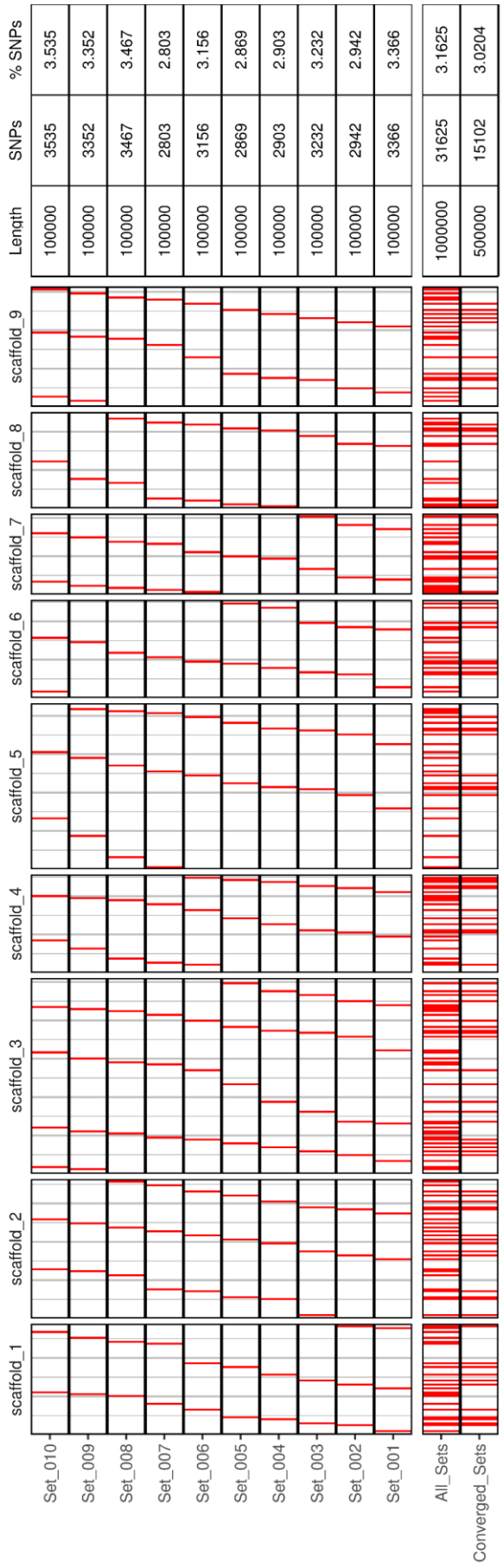
In this work we have only considered pure species, as admixed individuals distort the phylogenetic inference. The analysis of pure species is necessary to understand the evolutionary events that eventually gave birth to the genus *Citrus*, but we cannot disregard

the myriad of hybrids, admixtures and somatic clones that represent the cultivated varieties, which largely outnumber the pure species and are necessarily an essential part of the genus *Citrus*. We have established here the phylogenetic relationships between pure *Citrus* species, but further research is required to shed some light on the domestication processes that gave rise to the commercial citrus, generated in the last few thousand years as a result of human action.

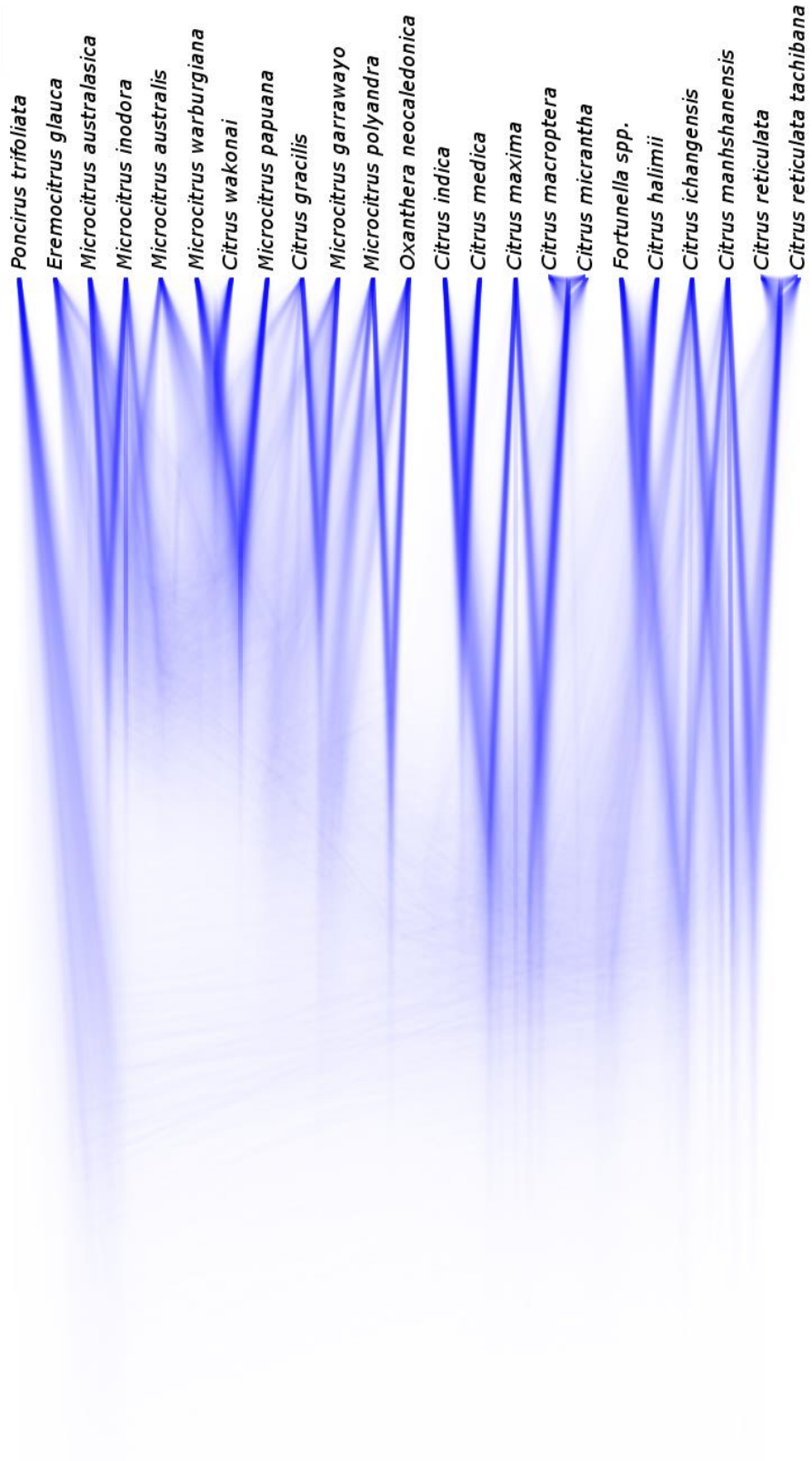
SUPPLEMENTARY DATA



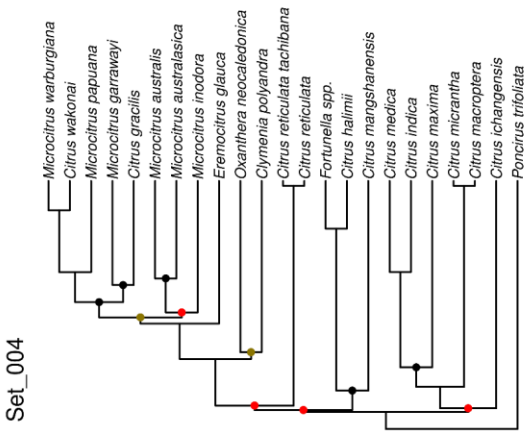
Supplementary Figure 1: Pairwise distance matrix between samples. The genetic distance between each sample pair was calculated in each of the 100 mash sketch replicates and the average pairwise distance was calculated.

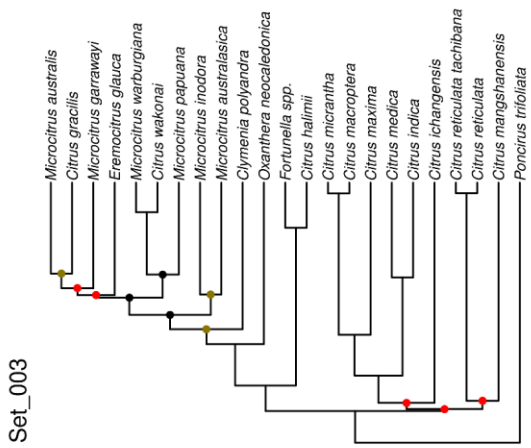


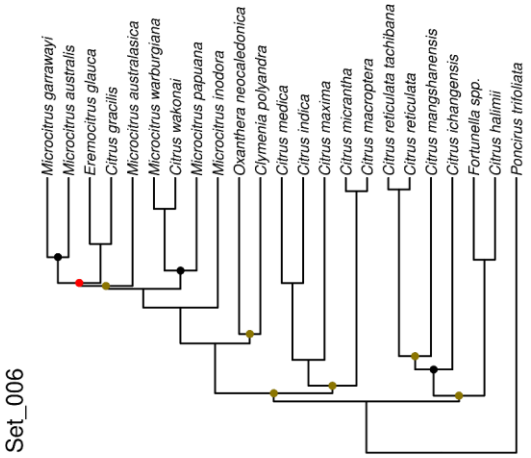
Supplementary Figure 2: Genomic distribution of phylogenetically informative regions. Map of the 200 regions used for the generation of phylogenetic trees based on concatenation and summary species approaches. These regions span across the nine major scaffold of the *Citrus clementina* reference and were split into 10 sets to perform the phylogenetic inference using StarBeast2. The red lines show in descending order the genomic position of these regions for each set, for the combination of all sets and for those sets that reach convergence in the StarBeast2 analysis. In all cases, the total length, number of SNPs and percentage of variable sites for each set are shown on the right.

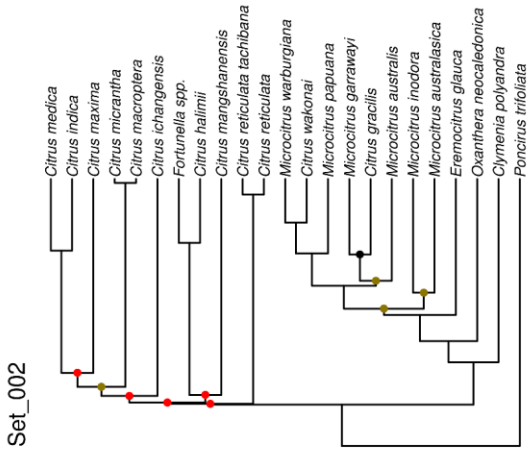


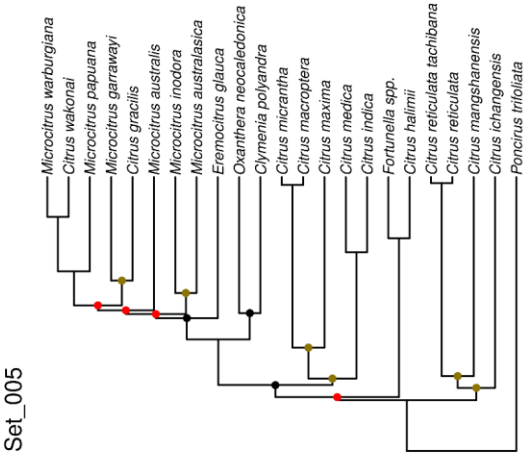
Supplementary Figure 3: Consensus trees of five converged sets combined. Combination of the consensus trees for the combination of the five converged analyses of StarBeast2 after discarding 250 million generations of burn-in. For each converged analysis, the two independent runs are considered. The tree was generated using Densitree.


Set_004
Microcitrus warburgiana
Citrus wakonai
Microcitrus papuana
Microcitrus garrawayi
Citrus gracilis
Microcitrus australis
Microcitrus australasica
Microcitrus inodora
Eremocitrus glauca
Oxanthera neocaledonica
Clymenia polyandra
Citrus reticulata tachibana
Citrus reticulata
Fortunella spp.
Citrus halimii
Citrus mangshanensis
Citrus medica
Citrus indica
Citrus maxima
Citrus micrantha
Citrus macroptera
Citrus ichangensis
Poncirus trifoliata


Set_003
Microcitrus australis
Citrus gracilis
Microcitrus garrawayi
Eremocitrus glauca
Microcitrus warburgiana
Citrus wakonai
Microcitrus papuana
Microcitrus inodora
Microcitrus australasica
Clymenia polyandra
Oxanthera neocaledonica
Fortunella spp.
Citrus halimii
Citrus micrantha
Citrus macroptera
Citrus maxima
Citrus medica
Citrus indica
Citrus ichangensis
Citrus reticulata tachibana
Citrus reticulata
Citrus mangshanensis
Poncirus trifoliata


Set_006
Microcitrus garrawayi
Microcitrus australis
Eremocitrus glauca
Citrus gracilis
Microcitrus australasica
Microcitrus warburgiana
Citrus wakonai
Microcitrus papuana
Microcitrus inodora
Oxanthera neocaledonica
Clymenia polyandra
Citrus medica
Citrus indica
Citrus maxima
Citrus micrantha
Citrus macroptera
Citrus reticulata tachibana
Citrus reticulata
Citrus mangshanensis
Citrus ichangensis
Fortunella spp.
Citrus halimii
Poncirus trifoliata


Set_002
Citrus medica
Citrus indica
Citrus maxima
Citrus micrantha
Citrus macroptera
Citrus ichangensis
Fortunella spp.
Citrus halimii
Citrus mangshanensis
Citrus reticulata tachibana
Citrus reticulata
Microcitrus warburgiana
Citrus wakonai
Microcitrus papuana
Microcitrus garrawayi
Citrus gracilis
Microcitrus australis
Microcitrus inodora
Microcitrus australasica
Eremocitrus glauca
Oxanthera neocaledonica
Clymenia polyandra
Poncirus trifoliata


Set_005
Microcitrus warburgiana
Citrus wakonai
Microcitrus papuana
Microcitrus garrawayi
Citrus gracilis
Microcitrus australis
Microcitrus inodora
Microcitrus australasica
Eremocitrus glauca
Oxanthera neocaledonica
Clymenia polyandra
Citrus micrantha
Citrus macroptera
Citrus maxima
Citrus medica
Citrus indica
Fortunella spp.
Citrus halimii
Citrus reticulata tachibana
Citrus reticulata
Citrus mangshanensis
Citrus ichangensis
Poncirus trifoliata

Supplementary Figure 4: Phylogenetic tree inferred from five independent sets. Maximum clade credibility of the five converged analyses of StarBeast2 after discarding 250 million generations of burn-in. For each converged analysis, the two independent runs are considered. Node colors represent bootstrap support values: black for nodes with 80% - 95% support, yellow for nodes with 50% - 80% support and red for nodes with < 50% support.

Supplementary Table 1. Sequence origin

Accession	Species	Sequence source*	Sample source**
ivia_000	*Citrus clementina*	SRA: SRX371962	-
ivia_004	*Citrus sinensis*	SRA: SRX372703	-
ivia_011	*Citrus maxima*	SRA: SRX372688	-
ivia_014	*Citrus aurantium*	SRA: SRX372786	-
ivia_017	*Citrus limon*	SRA: SRX3298457	-
ivia_020	*Poncirus trifoliata*	SRA: SRX3298456	-
ivia_1018	*Citrus indica*	This work: Nuclear DNA	IVIA
ivia_1024	*Microcitrus warburgiana*	This work: Nuclear DNA	IVIA
ivia_1025	*Clymenia polyandra*	This work: Nuclear DNA	IVIA
ivia_1051	*Citrus medica*	This work: Nuclear DNA	IVIA
ivia_1053	*Citrus ichangensis*	This work: Nuclear DNA	IVIA
ivia_1060	*Fortunella polyandra*	This work: Nuclear DNA	IVIA
ivia_1068	*Citrus halimi*	This work: Nuclear DNA	IVIA
ivia_107	*Microcitrus australasica*	This work: Nuclear DNA	IVIA
ivia_1073	*Fortunella crassifolia*	This work: Nuclear DNA	IVIA
ivia_1074	*Fortunella hindsii*	This work: Nuclear DNA	IVIA
ivia_1079	*Fortunella polyandra*	This work: CTAB	IVIA
ivia_1081	*Clymenia polyandra*	This work: CTAB	M. Smith
ivia_1083	*Eremocitrus glauca*	This work: CTAB	M. Smith
ivia_1084	*Oxanthera neocaledonica*	This work: CTAB	S. Lebegin
ivia_1085	*Oxanthera neocaledonica*	This work: CTAB	S. Lebegin
ivia_1086	*Oxanthera neocaledonica*	This work: CTAB	S. Lebegin
ivia_1089	*Eremocitrus glauca*	SRA: SRX3298476	-
ivia_1091	*Citrus indica*	SRA: SRX1973509	-
ivia_112	*Citrus medica*	This work: Nuclear DNA	IVIA
ivia_114	*Poncirus trifoliata*	This work: Nuclear DNA	IVIA
ivia_1159	*Microcitrus australasica*	SRA: SRX3298479	-
ivia_1160	*Microcitrus australis*	SRA: SRX3298478	-
ivia_1163	*Citrus indica*	This work: CTAB	M. Smith
ivia_1164	*Clymenia polyandra*	This work: CTAB	M. Smith
ivia_1166	*Microcitrus papuana*	This work: CTAB	M. Smith
ivia_1168	*Microcitrus garrawayi*	This work: CTAB	M. Smith
ivia_1172	*Microcitrus australasica*	This work: CTAB	M. Smith
ivia_1173	*Citrus gracilis*	This work: CTAB	M. Smith

Table S1 (continued)

Accession	Species	Sequence source*	Sample source**
ivia_1174	*Eremocitrus glauca*	This work: CTAB	M. Smith
ivia_1175	*Citrus wakonai*	This work: CTAB	M. Smith
ivia_1176	*Citrus macroptera*	This work: CTAB	M. Smith
ivia_1177	*Microcitrus inodora*	This work: CTAB	M. Smith
ivia_1178	*Microcitrus warburgiana*	This work: CTAB	M. Smith
ivia_1179	*Microcitrus australis*	This work: CTAB	M. Smith
ivia_1209	*Citrus maxima*	This work: CTAB	JBV
ivia_1219	*Fortunella margarita*	This work: CTAB	IVIA
ivia_135	*Citrus micrantha*	SRA: SRX3298460	-
ivia_217	*Citrus medica*	SRA: SRX3298463	-
ivia_317	*Citrus medica*	SRA: SRX3298462	-
ivia_319	*Citrus ichangensis*	SRA: SRX3298467	-
ivia_320	*Citrus medica*	SRA: SRX3298468	-
ivia_322	*Citrus medica*	SRA: SRX3298470	-
ivia_323	*Microcitrus australasica*	SRA: SRX3298471	-
ivia_324	*Microcitrus australis*	SRA: SRX3298472	-
ivia_326	*Citrus maxima*	SRA: SRX372702	-
ivia_327	*Citrus maxima*	From Huazong University	-
ivia_328	*Citrus maxima*	From Huazong University	-
ivia_329	*Citrus mangshanensis*	From Huazong University	-
ivia_5121	*Citrus reticulata*	SRA: SRX1901417	-
ivia_5123	*Citrus reticulata*	SRA: SRX1901407	-
ivia_5124	*Citrus reticulata*	SRA: SRX1901265	-
ivia_5132	*Citrus reticulata*	SRA: SRX3298473	-
ivia_5135	*Citrus reticulata*	SRA: SRX3298464	-
ivia_5137	*Citrus reticulata*	SRA: SRX2977586	-
ivia_1012	*Aegle marmelos*	This work: Nuclear DNA	IVIA
ivia_1013	*Aeglopsis chevalieri*	This work: Nuclear DNA	IVIA
ivia_1014	*Afraegle paniculata*	This work: Nuclear DNA	IVIA
ivia_1015	*Atalantia citroides*	This work: Nuclear DNA	IVIA
ivia_1016	*Balsamocitrus daweii*	This work: Nuclear DNA	IVIA
ivia_1017	*Citropsis gilletiana*	This work: Nuclear DNA	IVIA
ivia_1026	*Feroniella oblata*	This work: Nuclear DNA	IVIA
ivia_1027	*Glycosmis pentaphylla*	This work: Nuclear DNA	IVIA
ivia_1028	*Hesperethusa crenulata*	This work: Nuclear DNA	IVIA
ivia_1029	*Murraya paniculata*	This work: Nuclear DNA	IVIA
ivia_1057	*Clausena excavata*	This work: Nuclear DNA	IVIA
ivia_1058	*Clausena lansium*	This work: Nuclear DNA	IVIA
ivia_1082	*Citropsis gabunensis*	This work: CTAB	M. Smith
ivia_110	*Atalantia buxifolia*	SRA: SRX3298461	-
ivia_1165	*Citropsis gabunensis*	This work: CTAB	M. Smith
ivia_1167	*Swinglea glutinosa*	This work: CTAB	M. Smith
ivia_1169	*Clausena smyrelliana*	This work: CTAB	M. Smith

Table S1 (continued)

Accession	Species	Sequence source*	Sample source**
ivia_1170	*Clausena brevistyla*	This work: CTAB	M. Smith
ivia_1171	*Murraya paniculata*	This work: CTAB	M. Smith
ivia_1180	*Luvunga monophylla*	This work: CTAB	M. Smith
ivia_1181	*Murraya ovatifoliolata*	This work: CTAB	M. Smith
ivia_1182	*Micromelum minutum*	This work: CTAB	M. Smith
ivia_1220	*Ruta chalepensis*	This work: CTAB	JBV

*Nuclear DNA extraction protocol is described in Terol *et al.,* 2015. CTAB protocol is described in Webb and Knapp, 1990. The remaining samples were retrieved from SRA or from the Huazong Agricultural University website: http://citrus.hzau.edu.cn/orange/

**IVIA: Germplasm resources of the Instituto Valenciano de Investigaciones Agrarias. JBV: Jardí Botànic de València (Valencian Botanical Garden.). S. Lebegin: Stéphane Lebegin from the Institut Agronomique néo-Calédonien. M. Smith: Malcolm Smith from the the Bundaberg Research Station.

Chapter 2

Reprogramming of retrotransposon activity during speciation of the genus *Citrus*

Genome Biology and Evolution (2019), 11 (2) 3478–3495, DOI: 10.1093/gbe/evz246

My contribution in this work was key to its publication. I contributed with most of the analysis and a major part of the manuscript writing, as well as some minor conceptualization points.

ABSTRACT

Speciation of the genus *Citrus* from a common ancestor has recently been established to begin approximately 8 Mya during the Late Miocene, a period of major climatic alterations. Here, we report the changes in activity of *Citrus* LTR retrotransposons during the process of diversification that gave rise to the current *Citrus* species. To reach this goal, we analyzed four pure species that diverged early during *Citrus* speciation, three recent admixtures derived from those species and an outgroup of the *Citrus* clade. More than thirty thousand retrotransposons were grouped in 10 linages. Estimations of LTR insertion times revealed that retrotransposon activity followed a species-specific pattern of change that could be ascribed to one of three different models. In some genomes, the expected pattern of gradual transposon accumulation was suddenly arrested during the radiation of the ancestor that gave birth to the current *Citrus* species. The individualized analyses of retrotransposon lineages showed that in each and every species studied, not all lineages follow the general pattern of the species itself. For instance, in most of the genomes, the retrotransposon activity of elements from the SIRE lineage reached its highest level just before *Citrus* speciation while for Retrofit elements it has been steadily growing. Based on these observations we propose that C*itrus* retrotransposons may respond to stressful conditions driving speciation as a part of the genetic response involved in adaptation. This proposal implies that the evolving conditions of each species interacts with the internal regulatory mechanisms of the genome controlling the proliferation of mobile elements.

Key words: Genomic evolution, insertion time, LTR retrotransposon, speciation, structural variations

INTRODUCTION

LTR retrotransposons are widespread mobile DNA detected in virtually every genome studied to date (Bao *et al.*, 2015). They are found in great numbers due to their ability to replicate, as a new copy of each element is generated after transposition event. It is well known that in their transposition mechanism three main motifs are involved (a reverse transcriptase, an RNase H and an integrase, abbreviated RT, RH and IN), whose order has been recurrently used to classify LTR retrotransposons in two main groups: *Copia* and *Gypsy* (Boeke and Corces, 1989). Flanking the complete retrotransposon, two Target Site Duplications (TSDs) produced by the element insertion are also found.

LTR retrotransposons are named after the two long terminal repeats flanking the element core, that are identical upon insertion. Subsequently, each LTR accumulates mutations independently, an aspect that has been often used to date retrotransposon insertions (Pereira, 2004; Hu *et al.*, 2011; Xu and Du, 2014; Liu *et al.*, 2019). The homology between the LTRs of a single element also constitutes one of the main actors during the element excision, that generally involves recombination. Unequal recombination (UR) between homologous LTRs from the same element leaves a single LTR surrounded by TSDs (soloLTR) (Devos *et al.*, 2002). In contrast, when UR occurs between LTRs of different retrotransposons, one of the possible outcomes is a single LTR without flanking TSDs (Devos *et al.*, 2002). Similarly, illegitimate recombination (IR) between non-homologous elements is also relevant during retrotransposon purge, as it produces, among others, truncated elements with a single LTR and no TSDs (Devos *et al.*, 2002; Vitte and Bennetzen, 2006). LTRs produced by this mechanism are unpaired, but their formation mechanism is different from that of true soloLTRs; to differentiate both types of unpaired LTRs in this work, we will refer to LTRs produced by IR as nonsoloLTRs. Furthermore, the ratios between paired LTRs and soloLTRs have also been used to estimate retrotransposon purge rates in multiple studies (Vitte *et al.*, 2007; Hawkins *et al.*, 2009; Yin *et al.*, 2015; Lyu *et al.*, 2018).

Since their discovery, retrotransposons have proved their relevance in genome evolution, especially in repeat-rich plant genomes (Sanmiguel and Bennetzen, 1998; Bousios *et al.*, 2012). The effect of retrotransposons in plant evolution has been already described (Brookfield, 2005; Hanada *et al.*, 2009; Du *et al.*, 2009; Sela *et al.*, 2010; Butelli *et al.*,

2012) highlighting their importance in adaptive processes (Vicient and Casacuberta, 2017). Changes in retrotransposon activity have also been reported after drastic genomic events such as hybridization (Paz *et al.*, 2015) and polyploidization (Parisod *et al.*, 2009; Bardil *et al.*, 2015; Mhiri *et al.*, 2019) under the hypothesis of genomic shock (McClintock, 1984), although other authors have found evidences against it (Göbel *et al.*, 2018). It is also well accepted that environmental stresses may induce transposition, as well as the expression of genes neighboring residing transposons (Beguiristain *et al.*, 2001; Kimura *et al.*, 2001; Butelli *et al.*, 2012; Dubin *et al.*, 2018). The above premises strongly suggest that LTR retrotransposons might play a role in the evolutionary processes giving birth to distinct species. Associations between LTR retrotransposon activity and speciation have been certainly reported in rice and wheat (Zhang and Gao, 2017; Mascagni *et al.*, 2017), providing first insights on these connections. However, the recent establishment of solid phylogenies in several plant genera, such as in *Citrus* for instance (Wu *et al.*, 2014, 2018), may allow these relationships to be explored in detail. Actually, retrotransposon activity in *Citrus* is a matter of increasing interest (Rico-Cabanas and Martínez-Izquierdo, 2007; Du *et al.*, 2018; Liu *et al.*, 2019). The first retrotransposons found in *Citrus* were the *Copia*-like elements of sweet orange (Tao *et al.*, 2005). Subsequent reports showed an enhancement on the CLCoy1 transposon activity under stress conditions in *Citrus limon* (De Felice, 2009). Later, the expression of the *Ruby* gene, a major actor of the anthocyanin accumulation in blood oranges, was found to be regulated by a transposon promoter (Butelli *et al.*, 2012, 2017). It has also been reported that the Mutator-like DNA transposon CitMULE1 is responsible of the rearrangement of large genomic fragments in the genome of clementine mandarin and therefore a major source of new clementine genotypes and hence of new commercial varieties (Terol *et al.*, 2015).

While most of these works have focused on either a single genome or a reduced number of mobile elements, the growing interest of *Citrus* retrotransposons have led to the recent publication of two genome-wide surveys describing the retrotransposon landscape in different *Citrus* species, setting the background for deeper analysis. In the first study, LTR retrotransposons of *C. clementina* were mined and their phylogeny and distribution over the genome was described (Du *et al.*, 2018). Later, the mobilomes of six species corresponding to five *Citrus* genomes of reference (Ichang papeda, pummelo, citron, clementine and sweet orange) and a relatively close related genome (Chinese box orange)

were the subject of a study, mainly focused in the MITE landscape of each genome (Liu *et al.*, 2019). The authors also analyzed the phylogeny of the LTR retrotransposons, reaching results complementing those presented in (Du *et al.*, 2018) and in addition estimated their average insertion times and half-life across the six genomes.

In this study we expand these previous insights investigating LTR retrotransposon activity of the genus *Citrus* from an evolutionary context. To this end we have used all *Citrus* reference genomes available today, corresponding to the six genomes previously used in (Liu *et al.*, 2019) plus two additional genomes of recent accessibility. Thus, the analyses included four true *Citrus* species: *C. ichangensis* (Ichang papeda), *C. maxima* (pummelo), *C. medica* (citron) (Wang *et al.*, 2017*b*) and *C. reticulata* (mandarin) (Wang *et al.*, 2018*a*), and three different admixtures *of C. maxima* and *C. reticulata*, namely, *C. clementina* (clementine mandarin) (Wu *et al.*, 2014), *C. unshiu* (satsuma mandarin) (Shimizu *et al.*, 2017) and *C. sinensis* (sweet orange) (Xu *et al.*, 2013) in addition to *Severinia buxifolia* (Chinese box orange) (Wang *et al.*, 2017*b*). Out of these eight genomes, four of them consisted of thousands of scaffolds generated directly from Illumina sequencing (citron, Ichang papeda, Chinese box orange and mandarin). However, those of sweet orange, pummelo and satsuma and clementine mandarins are all resolved up to the pseudomolecule scale, including nine main scaffolds corresponding to the nine *Citrus* chromosomes.

Citrus taxonomy and phylogeny have been a matter of controversy during the last century due to an unusually high number of interspecific hybrids that hinders the identification of pure species and prevents the inference of a reliable phylogeny. *Citrus* pure species reproduce through sexual crosses between members of the same species and therefore are generally free of introgression events. In contrast, most commercial or domesticated *Citrus* are derived from interspecific crosses followed by successive backcrosses, producing in this way characteristic admixture patterns that contain genomic regions from different pure species (Wu *et al.*, 2014). Furthermore, commercial varieties are in general clonally propagated via grafting, which have allowed the admixture patterns that were generated many generations ago to reach our time. While there are no clear evidences on the origin of the first admixed genomes, there are records of sweet oranges (an admixture between pummelo and mandarin) dated 2300 years ago (Xu *et al.*, 2013), which might situate the origin of the first *Citrus* admixtures in the last few thousand years.

Of particular relevance for our goals are the comparative genomic analyses presented in Wu *et al*. (2014, 2018), that allowed the discrimination of pure and admixed *Citrus* germplasm and inferred the phylogeny, genealogy and chronology of the *Citrus* speciation. According to (Wu *et al.*, 2018), the phylogenetic relationship between the pure species of *Citrus* included in the current work is as follows. The Chinese box orange (*Severina buxifolia*), an outgroup of the *Citrus* clade, diverged from the *Citrus* group around 13 million years ago (Mya; (Pfeil and Crisp, 2008). The *Citrus* last common ancestor lived in continental Southeast Asia about 8 Mya, during the Late Miocene. This was a period of major climate changes characterized by a global carbon diozide level decline (Holbourn *et al.*, 2018) that brought about a worldwide cooling epoch resulting in extensive weakening of monsoons and aridity enhancement of the subtropical regions (Herbert *et al.*, 2016). In Southeast Asia, this marked climate alteration caused major changes in biota including rapid radiations of various plant lineages (see references in Wu *et al.*, 2018) including *Citrus*. Ichang papeda diverged at the very beginning of *Citrus* speciation and apparently migrated to Central China. Shortly thereafter, two main clades separated about 7-6 Mya: citrons and pummelos (India, Indochina and the Malay Archipelago) in one of them and mandarins (East and South China and Japan) in the other. The three *Citrus* admixtures of *C. maxima* and *C. reticulata* studied here harbor different proportions of pummelo introgression in the mandarin genome [*C. clementina* (12%), *C. unshiu* (24%) and *C. sinensis* (42%)] and were generated at different historic times, at most few thousand years ago, from different genetic backgrounds.

Since variations in retrotransposon activity have been repeatedly related to environmental stresses in multiple plants, we found very tempting to analyze their fluctuations during *Citrus* speciation, a process most likely stimulated by a dramatic climate change, to elucidate if those environmental changes left any recognizable signature or imprint in their genomes. Thus, the goal of this study was first to describe the LTR retrotransposon landscape of the genus *Citrus* and then report the changes in their pattern of accumulation during the process of diversification that gave rise to the current *Citrus* species.

MATERIALS AND METHODS

Genomic data

All the genomic data were retrieved from public repositories. Eight reference genomes were used: four true pure *Citrus* species including *Citrus reticulata* (wild mandarin), *Citrus ichangensis* (Ichang papeda), *Citrus maxima* (pummelo) and *Citrus medica* (citron), two admixed (*Citrus reticulata* x *Citrus sinensis*) commercial mandarins (*Citrus clementina* and *Citrus unshiu*, clementine and satsuma mandarins, respectively), one admixed (*Citrus maxima* x *Citrus reticulata*) commercial sweet orange (*Citrus sinensis*) and a close relative to the *Citrus* clade, *Severinia buxifolia* (Chinese box orange).

The reference genomes and the gene annotation data of *S. buxifolia*, *C. reticulata*, *C. maxima*, *C. medica*, *C. sinensis* and *C. ichangensis* were downloaded from http://citrus.hzau.edu.cn/. The *C. unshiu* genome and annotation data were downloaded from http://www.citrusgenome.jp/. The *C. clementina* reference genome and its annotation data were downloaded from Phytozome (*Citrus clementina* v1.0).

Paired-end Illumina reads for the structural variant analysis were retrieved from the NCBI Sequence Read Archive. The codes and equivalence of each accession are available in the Supplementary Table 1.

Detection and classification of LTR retrotransposon cores

Putative LTR retrotransposons were found and validated in *C. clementina* reference genome using an integrated detection pipeline, LocaTR (Mason *et al.*, 2016), which combines the results from several LTR retrotransposon detection tools (McCarthy and McDonald, 2003; Sperber *et al.*, 2007; Ellinghaus *et al.*, 2008). Results from LTR_FINDER (Xu and Wang, 2007) were also incorporated following the user manual of LocaTR to generate a comprehensive set of LTR retrotransposons.

A curated retrotransposon database, Gypsy Database (Llorens *et al.*, 2011), was searched to retrieve protein and DNA sequences of three LTR retrotransposon domains (IN, RT and RH) of every GyDB element annotated. To retrieve DNA sequences from the core retrotransposon domains, BLASTX analyses were performed using as queries each of the *C. clementina* and GyDB retrotransposon DNA sequences against a custom GyDB core

domain protein sequences. Only hits with an e-value below $1 \cdot 10^{-20}$ and containing the three core domains (IN + RT + RH, regardless of the order) in the *C. clementina* putative retrotransposons were selected. Each *C. clementina* element was classified as *Gypsy* or *Copia* depending on the order of their domains: RT-RH-IN as *Gypsy* and IN-RT-RH as *Copia*.

The *C. clementina* retrotransposon core collection was used as query in a BLASTN analysis against eight reference genomes: *C. clementina*, *C. ichangensis*, *C. reticulata*, *C. unshiu*, *C. maxima*, *C. medica*, *C. sinensis* and *S. buxifolia*. Only hits covering over 80% of the query and with an e-value lower than $1 \cdot 10^{-25}$ were selected, and overlapping hits were merged. Hits produced by *Copia C. clementina* elements were classified as belonging to the *Copia* superfamily, and the same was done with the *Gypsy* superfamily.

Retrotransposon cores sharing over 80% of sequence identity in at least 80% of the genome, with a minimum of 80 bp covered were independently clustered in each genome using a modified mean shift algorithm implemented in MeShClust (James *et al.*, 2018), and each cluster was assigned to a new retrotransposon family following the system of (Wicker *et al.*, 2007). The longest sequence of each family was selected as a cluster representative. Family representatives from *Copia* and *Gypsy* superfamilies were aligned with a GyDB pre-aligned profile. Both alignments were performed using MAFFT L-INS-I algorithm (Katoh and Standley, 2013). A maximum likelihood phylogenetic tree was built with FastTree (Price *et al.*, 2010) and the tree topology was explored using R and ggtree (Yu *et al.*, 2017*a*; R Core Team, 2018).

Citrus LTR and retrotransposon distribution

Each reference genome was split in non-overlapping windows of up to 1 Mb and each retrotransposon was associated to one of them, together with the gene content of each window. For scaffolds above 100 kb but below 1 Mb, the complete scaffold was used as a single window. Scaffolds below 100 kb were discarded. The median genic content among the windows of *Citrus clementina* was estimated and used to roughly locate the pericentromeric regions.

While the LocaTR pipeline is capable of detecting large amounts of LTR retrotransposons, it does not separately annotate LTRs. One of the tools integrated in LocaTR, LTR_Harvest, was used to detect paired LTRs. To do so, each LTR

retrotransposon core and 30 kb of flanking sequences were used as queries for LTR_Harvest. The representativity of the new LTR_Harvest dataset of the original dataset found by homology search was manually verified by checking if the proportions of retrotransposons found in each lineage and species are roughly conserved across the two datasets (Supplementary Figure 1). As every LTR defined by LTR_Harvest must have a pair, the two LTRs of each LTR_Harvest detected element were aligned using MAFFT (Katoh and Standley, 2013), and the Kimura-2-parameters distance was assessed for each alignment using DiStats (Astrin *et al.*, 2016). The conversion of Kimura-2-parameters distance to time was calculated using as mutation rate $4 \cdot 10^{-9}$ and $5 \cdot 10^{-9}$ substitutions per year, as previously reported (De La Torre *et al.*, 2017), multiplied by a factor of two as in (Vitte *et al.*, 2007; Hu *et al.*, 2011).

A BLASTN search was used to find sequences similar to the paired LTRs identified by LTR_Harvest, selecting hits with an identity of over 80% across 90% of the query (hits closer than 100 bp were merged). For each hit, a dot plot was performed against 30 kb of their flanking sequence using YASS (one seed to consider a hit and an Xdrop threshold score of 100 were used, the remaining parameters were left as by default) (Noe and Kucherov, 2005). Hits flanked with at least one similar (a hit extending over 90% of the sequence) copy of themselves were classified as paired LTRs. The remaining hits were considered unpaired LTRs (unpaired LTRs). Unpaired LTRs were then searched for TSDs to classify them in true solo-LTRs or nonsolo-LTRs. To do so, the 20 bp flanking both sides of each unpaired LTR were searched for identical kmers of lengths from 4 to 7 nucleotides using inhouse scripts. If a kmer was found in the two 20-nucleotide flanking sequences, it was defined as a TSD and the unpaired LTR was classified as a solo-LTR. In any other case, the unpaired LTR was classified as a nonsolo-LTR. Every LTR regardless of its type was associated to position-based windows as in the case of genes and complete retrotransposon cores.

Determination of unpaired LTRs closest relatives

Each unpaired LTR (soloLTR or nonsoloLTR) was used as a query in a BLASTN analysis against a database including all the LTRs found (paired and unpaired). The best hit for each sequence (excluding the sequence itself) was recorded provided it covered at least 90% of the query with 90% of identity. Only reciprocal best hits (A's best hit is B and

B's best hit is A) were selected, and the reference genomes of the query sequence and the hit were recorded.

Determination of transposition events via structural variant detection

Illumina paired-end reads from 43 mandarin accessions (Supplementary Table 1) were retrieved from SRA. Reads with over 30% of their bases showing a quality score below 30 were discarded, and the remaining were aligned against the *C. clementina* reference genome using bwa-mem (Li, 2013).

Structural variants were discovered using Lumpy 0.2.13 and SVTyper 0.1.3 (Layer *et al.*, 2014; Chiang *et al.*, 2015). Deletions with a size below 100 kb and with a reciprocal coverage of 80% between them and any complete LTR retrotransposon found by LTR_Harvest (at least 80% of the deletion annotated as a retrotransposon and vice versa) were selected and assigned as retrotransposon-induced deletions. This process was independently applied to each sample. Deletions supported by at least 20% and 80% of the reads were considered hemizygous and homozygous, respectively.

Statistical analyses and data representation

Correlation tests were performed using the non-parametrical Spearman rank correlation test implemented in R stats package (v3.5.1). Phylogenetic trees were plotted using ape, ggplot and ggtree (Wickham, 2016; Yu *et al.*, 2017*a*; Paradis and Schliep, 2019). The remaining plots were created using ggplot.

RESULTS

LTR retrotransposon detection and classification

Using a combined detection approach, 2666 putative LTR retrotransposons were found in the *Citrus clementina* haploid reference genome. Of them, 2376 contained exactly one copy of each of the three core motifs (integrase, RNAse H and reverse transcriptase) of the LTR retrotransposons and were consequently annotated as LTR retrotransposons. These LTR retrotransposons were then used as queries to identify similar elements in

eight reference genome sequences (*Severinia buxifolia*, *Citrus ichangensis*, *Citrus maxima*, *Citrus medica*, *Citrus reticulata*, *Citrus clementina*, *Citrus unshiu* and *Citrus sinensis*), retrieving a total of 32506 retrotransposon cores, which were classified in the *Gypsy* or *Copia* superfamilies depending on their motif order (Table 1).

All cores within each genome were grouped in families. The number of LTR retrotransposon families detected among the eight genomes varied between 316 and 446, accounting for 2974 families in total (Table 1). The longest sequence of each family was aligned with a representative set of sequences from GyDB and two independent phylogenetic trees were built for *Gypsy* (Figure 1a) and *Copia* (Figure 1b) retrotransposons. Every *Citrus* retrotransposon family was classified in one of the following plant retrotransposon lineages: Retrofit, Oryco, SIRE or Tork lineages for *Copia* retrotransposons, and CRM, Reina, Del, Galadriel, Athila or Tat lineages for *Gypsy* retrotransposons.

To study the *de novo* acquisition and loss of retrotransposon families the topology of each phylogenetic tree was explored. As retrotransposon families were independently defined in each genome, those shared by several genomes are clustered together in the phylogenetic tree as a clade containing multiple nodes, and with at least one member per genome. In contrast, family gains and losses are defined by clades whose families were present in many but not all the genomes. All clades harboring more than 20 terminal nodes were analyzed, and those missing one or more reference genomes among their nodes were identified (Figure 1). While most of the 20-node clades comprise a sequence from each reference genome, a small number of clades (8 in *Copia* and 9 in *Gypsy* trees) harbored families missing in some species. Out of these 17 clades, 5 of them were missing a representative in the reference genome of *S. buxifolia*, the most distant genome included in this work.

Table 1. Citrus LTR retrotransposons elements and families[a].

Superfamily	Lineages	*Citrus clementina*	*Citrus sinensis*	*Citrus unshiu*	*Citrus maxima*	*Citrus medica*	*Citrus ichangensis*	*Citrus reticulata*	*Severinia buxifolia*	Total
Copia	**Tork**	1001 [87]	556 [69]	721 [65]	1072 [86]	1294 [74]	1453 [110]	1073 [59]	864 [74]	**8034 [624]**
	SIRE	538 [15]	340 [12]	600 [9]	786 [18]	926 [16]	99 [27]	424 [19]	260 [9]	**3973 [125]**
	Oryco	123 [17]	92 [18]	102 [14]	69 [22]	191 [17]	128 [22]	121 [18]	118 [23]	**944 [151]**
	Retrofit	483 [58]	685 [61]	429 [80]	497 [56]	227 [47]	581 [56]	495 [53]	284 [57]	**3681 [468]**
	Total Copia	**2145 [177]**	**1673 [160]**	**1852 [168]**	**2424 [182]**	**2638 [154]**	**2261 [215]**	**2113 [149]**	**1526 [163]**	**16632 [1368]**
Gypsy	**Tat**	386 [36]	212 [33]	246 [27]	351 [51]	402 [21]	271 [31]	309 [34]	131 [29]	**2308 [262]**
	Athila	1510 [72]	830 [37]	768 [56]	1974 [58]	1355 [30]	886 [47]	1068 [36]	608 [28]	**8999 [364]**
	Galadriel	108 [21]	88 [21]	80 [28]	131 [26]	107 [20]	85 [23]	111 [19]	72 [26]	**782 [184]**
	Del	84 [23]	66 [13]	84 [24]	91 [25]	70 [15]	112 [20]	19 [15]	81 [17]	**607 [152]**
	CRM	209 [34]	148 [24]	205 [30]	336 [40]	213 [31]	268 [33]	203 [22]	234 [20]	**1816 [234]**
	Reina	163 [42]	128 [53]	360 [48]	141 [64]	157 [47]	157 [51]	118 [41]	138 [64]	**1362 [410]**
	Total Gypsy	**2460 [228]**	**1472 [181]**	**1743 [213]**	**3024 [264]**	**2304 [164]**	**1779 [205]**	**1828 [167]**	**1264 [184]**	**15874 [1606]**
Total LTR		**4605 [405]**	**3145 [341]**	**3595 [381]**	**5448 [446]**	**4942 [318]**	**4040 [420]**	**3941 [316]**	**2790 [347]**	**32506 [2974]**

[a] Family numbers are shown in brackets

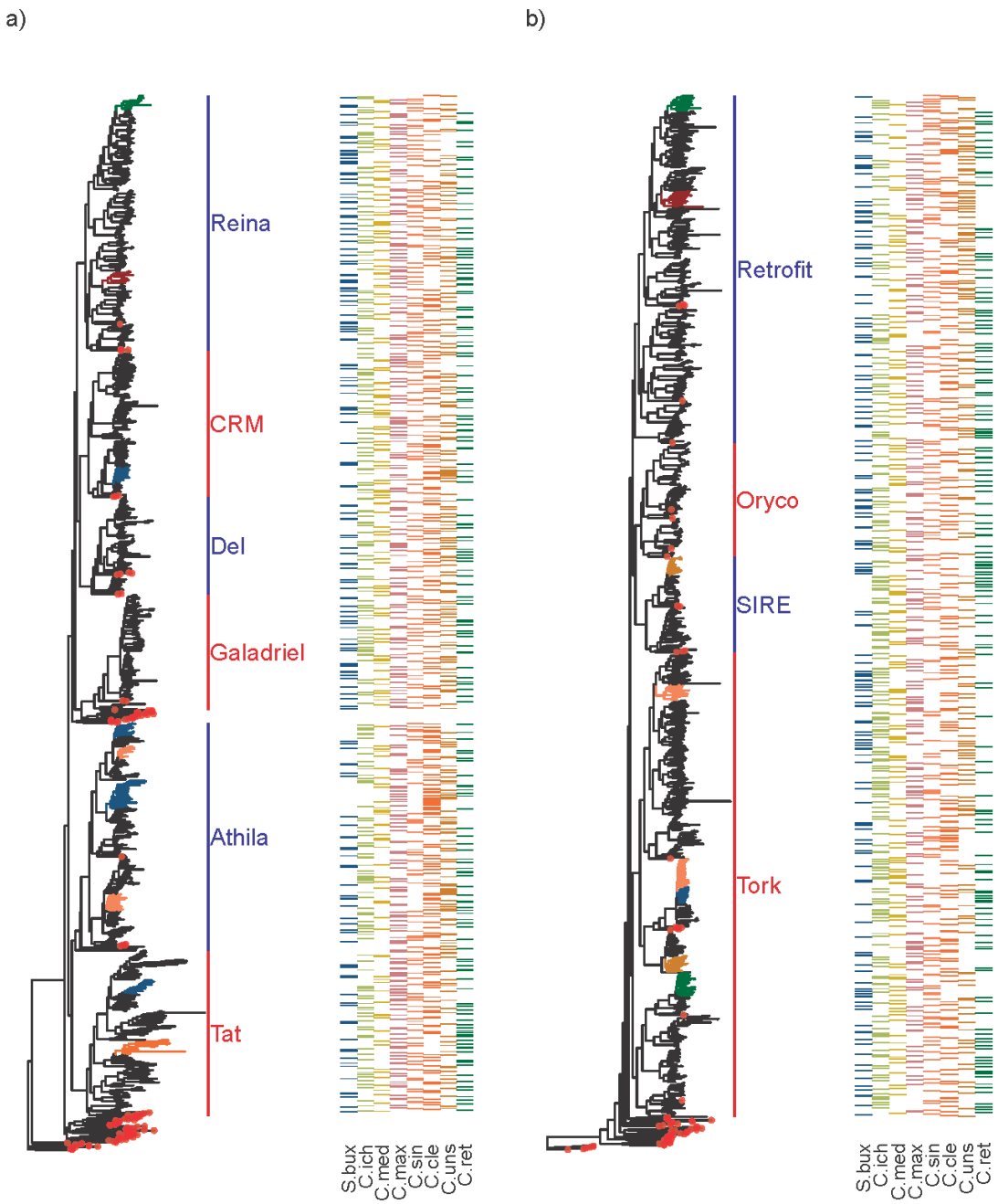


Figure 1: Citrus LTR retrotransposon phylogenetic trees and presence across species. Phylogenetic trees of LTR retrotransposon families belonging to *Gypsy* (a) and *Copia* (b) superfamilies are shown. Next to each tree a heatmap indicate the species of origin for each family (terminal node). Red dots mark terminal nodes belonging to sequences from the curated transposon database GyDB. Colored branches represent clades with over 20 terminal nodes not harboring families from the eight references studied. The color legend is the same as that of the heatmap, with clades missing two or more references highlighted in dark red. The following naming convention is used to refer to the reference genomes: S.bux = *Severinia buxifolia*, C.ret = *Citrus reticulata*, C.ich = *Citrus ichangensis*, C.max = *Citrus maxima*, C.med = *Citrus medica*, C.sin = *Citrus sinensis*, C.uns = *Citrus unshiu*, C.cle = *Citrus clementina*.

Accumulation patterns and dating of complete LTR retrotransposons

The genomic position of each LTR retrotransposon core of the *C. clementina* reference was used to study the retrotransposon core accumulation patterns along the genome. When the distribution of the LTR retrotransposon cores of *C. clementina* was studied (Figure 2a), a negative correlation between gene content and LTR retrotransposon abundance was found (p-value < 0.05). This association was also independently observed for each genome (Supplementary Table 2). In contrast, retrotransposon activity hotspots, characterized by a higher frequency of retrotransposon-induced deletions, were mostly found in genic regions of *C. clementina* (Figure 2a), as further discussed in subsequent sections of this work.

Paired LTRs were found flanking 3102 out of the 4605 similarity-found retrotransposon cores in clementine, allowing for the determination of complete elements, with an average length of 8701 bp. Considering the eight genomes, a total of 18630 complete retrotransposons with a global average of 8208 bp in length were detected (Table 2). The average genome proportion of LTR retrotransposons was calculated per species considering in each case the species average element length, the number of elements and the total genome length. These proportions ranged from 3.60% to 9.97% among the different species but are most probably an underestimation of the real values, as they are solely based on full-length LTR retrotransposons with well-defined LTRs, disregarding a considerable amount of retroelements. By considering each retrotransposon core as part of a complete element, the maximum LTR retrotransposon content was calculated per species (assigning to each core the genome-specific average length), which yielded a retrotransposon proportion ranging from 6.87 to 15.93% in the eight genomes studied (Table 2).

The genetic distance between both paired LTRs of each element was then used to estimate its insertion time (Hu *et al.*, 2011). The oldest LTR retrotransposons were generally found in pericentromeric regions where they were visibly more abundant, although this differential distribution was progressively less evident as younger elements were considered (Figure 2b). Elements containing two identical LTRs (distance equals 0) have been previously defined as newly inserted elements (Xu and Du, 2014). In *Citrus*

clementina, 87 of these new elements were found all across the genome in a distribution which was not dependent on the genic content (Supplementary Table 2, Figure 2b), which might indicate an unbiased insertion along the genome for the most recent *C. clementina* retrotransposons. Retrotransposon insertion times were then calculated for each species, and the same lack of correlation was observed when all species were considered except in the case of *C. maxima* and *C. sinensis*, in which new LTR retrotransposons were significantly less common in genic regions possibly indicating a biased insertion (Supplementary Table 2).

Table 2: *Citrus* LTR retrotransposon length, number and coverage

Organism	LTR-TE length and number		Genome coverage (%)		
	Cores length and number [a]	Complete elements length and number [a]	LTR-TE cores	Complete LTR-TE	Max. LTR-TE [b]
Citrus clementina	2650 [4605]	8701 [3102]	4.00	8.84	13.13
Citrus sinensis	2469 [3145]	7860 [1531]	3.20	4.95	10.17
Citrus unshiu	2564 [3595]	8097 [1777]	2.53	3.95	7.99
Citrus maxima	2627 [5448]	8940 [3410]	4.68	9.97	15.93
Citrus medica	2600 [4942]	8137 [2863]	3.16	5.73	9.89
Citrus ichangensis	2595 [4040]	8057 [2357]	2.93	5.31	9.10
Citrus reticulata	2587 [3941]	8087 [2129]	2.95	4.97	9.21
Severinia buxifolia	2563 [2790]	7792 [1461]	2.26	3.60	6.87
All species	**2590 [32506]**	**8308 [18630]**	**3.18**	**5.85**	**10.21**

[a] Number of elements is shown in brackets

[b] Considering the total core number and the complete element length

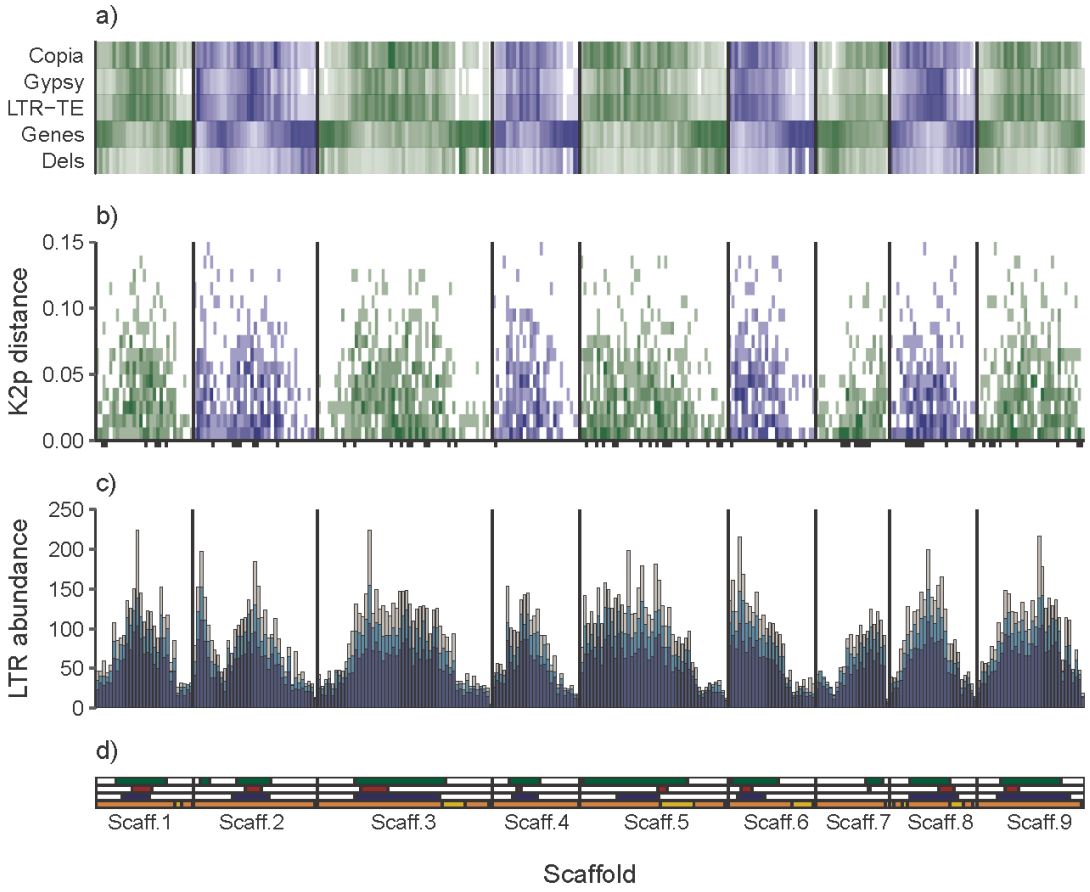


Figure 2: LTR retrotransposon abundance, age and activity in the clementine reference genome. Only the nine main scaffolds of the clementine reference are shown. All results are summarized in 1 Mb windows. a) Distribution of LTR retrotransposons (LTR-TE) disaggregated into *Copia*, *Gypsy* and total elements. Below, the per Mb genic content is shown. On the lowermost row, a per-window average of the transposon-associated deletions across 43 mandarin genomes is shown, the full data can be found in Supplementary Figure 2. The intensity of each bin is proportional to the percentage of bases covered per window, with the maximum intensity normalized to the maximum value in each row. b) LTR-based dating retrotransposons in *C. clementina*. The relative age was calculated as the Kimura-2-parameters genetic distance (K2p) (Hu *et al.*, 2011) between LTR pairs. Each LTR retrotransposon was classified in an age interval (windows of 0.01 distance units) and genomic position. The coordinates of each bin are given by the genomic position of each element and its age, and the intensity is proportional to the number of transposons included in the bin. Elements with identical LTRs (K2p distance equals 0) are marked as black ticks under the x axis. c) Total number of soloLTR (purple), nonsoloLTR (blue) and pairedLTRs (gray) across the *C. clementina* reference genome, shown as a stacked bar plot. Total LTR (totalLTRs) counts are given by the total height of each bar. d) Genomic features of the *C. clementina* reference genome. On top, the centromeres predicted in this work based on the genic content (green), together to those of (Aleza *et al.*, 2015) (red) and (Wu *et al.*, 2014) (blue). The last row shows the admixture map of the *C. clementina* haploid reference genome:

genomic fragments coming from mandarin and pummelo are shown in orange and yellow, respectively, while fragments with unknown precedence are shown in gray. The data were obtained as explained in (Wu *et al.*, 2014).

Genomes were divided in windows of 1 Mb that were assigned to one of six categories regarding their gene content (from 0% to 60% of the window covered by genes, in 10% bins). Each retrotransposon was assigned to one genomic region based on their position in the genome, and the age distribution per gene-content bin and per species was calculated (Figure 3). Among all the studied genomes, the correlation between the genic content and the LTR retrotransposon age distribution was not consistent. In *C. clementina*, young elements were present along the genome regardless of the gene content, while older elements became progressively less common as the genic content dropped. This results in an age distribution with an abundance peak becoming more prominent as the genic content increases (Figure 3). Similar but less pronounced patterns were also found in *C. ichangensis*, *C. sinensis*, *C. reticulata* and *C. unshiu*. On the other hand, *C. maxima* and *C. medica* showed a more uniform age distribution across different gene content levels. Finally, *S. buxifolia* followed a different distribution, without visible changes except for the last category (comprising the highest gene density) that reveals a very recent accumulation of young elements in genic regions.

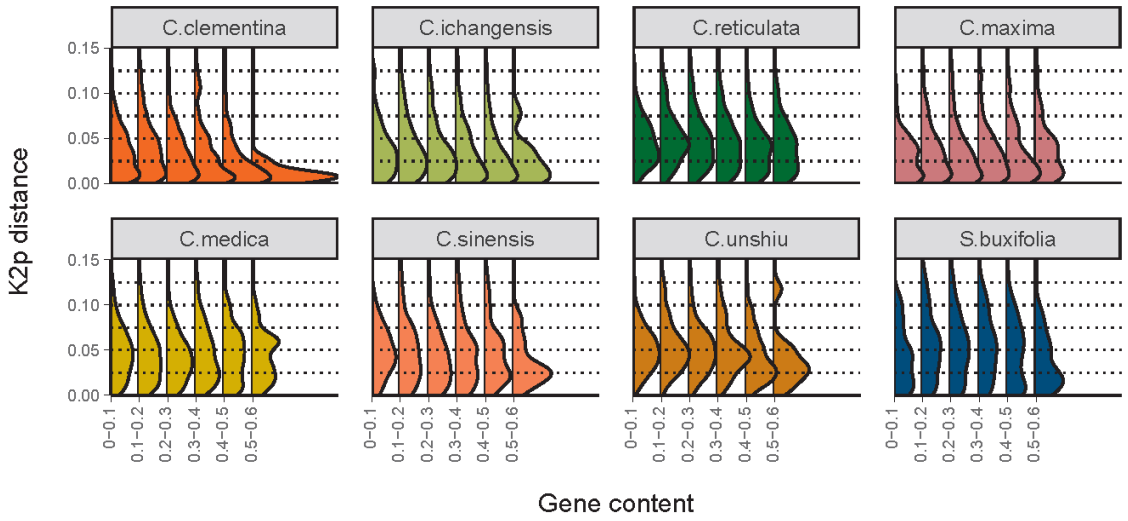


Figure 3: Relative age distribution of paired LTRs per species and gene density. Panels show the eight reference genomes and contain six retrotransposon age distributions each, one per genic-content bin. In each distribution, the height of the curve represents has been normalized to represent the proportion of elements with a given pairwise distance between their LTRs.

Moreover, the purge rate of LTR retrotransposons in *Citrus clementina* was determined studying the proportion of soloLTR, nonsoloLTR and paired LTR across the genome (Figure 2c). Based on these proportions we conclude that the retrotransposon elimination in *Citrus clementina* occurs at a faster rate in genic regions (see below).

Finally, the location of pericentromeric regions in the *C. clementina* genome was calculated. The overall median genic content across the whole *C. clementina* genome was determined to be 23%. Up to ten 1 Mb-windows were assigned as pericentromeric regions along the 9 main scaffolds as their genic content fell below that threshold (Figure 2d). Consistently, the centromere locations correlated with retrotransposon abundance, their aging and the presence of activity hotspots.

Retrotransposon activity patterns among mandarins

An indicator of retrotransposon recent activity in re-sequenced genomes is the presence of retrotransposon-induced deletions that are easily evidenced after comparison with the reference genome. Deletions could be generated by either a true deletion of the element in the re-sequenced cultivar via one of the methods mentioned above, or through an insertion of that element in the reference genome after its divergence from the re-sequenced genome (Rahman *et al.*, 2015).

In principle, the strategy followed in this work could certainly detect novel element insertions since it is expected that these elements would be completely missing in the re-sequenced genome. For retroelement true deletions, the observed deletion would span across most of the retrotransposon, except for the LTRs that consequently remain in both, the re-sequenced and the reference genomes. Unfortunately, reads mapped within a retrotransposon (such as those that would support these deletions) are usually unreliable due to the repetitive nature of mobile elements. For this reason, deletions reciprocally spanning over 80% of an element (see Methods) were assigned as either insertions or deletions, without distinguishing between them.

The distribution of retrotransposon-induced deletions across 43 mandarin varieties (Supplementary Table 1) was studied to identify retrotransposon activity hotspots across the clementine genome. A total of 15388 deletions spanning over LTR retrotransposons were annotated (see Methods) with an average of 358 deletions per sample, all of them ranging from 2515 bp to 15378 bp (the average length was 7818 bp). Their genomic

coordinates were used to study the retrotransposon activity across the genome, which was significantly higher in genic regions (Figure 2a and Supplementary Figure 2).

Cross-homology of unpaired LTRs among *Citrus*

Each unpaired LTR was queried against the total LTR collection to find its closest relative, and the genome harboring it was recorded in each case (Figure 4). *C. clementina* unpaired LTR closest relatives were mostly found in *C. sinensis*, *C. reticulata* and *C. unshiu*, all of them containing great amounts of mandarin genome as they are either mandarin admixtures (*C. sinensis*, *C. clementina* and *C. unshiu*) or a pure mandarin itself (*C. reticulata*). The remaining clementine unpaired LTR relatives were found mainly in the other pure species involved in clementine's admixture, *C. maxima*, followed by more distant *Citrus* species such as *C. ichangensis* and *C. medica*. A small proportion of the clementine unpaired LTRs showed a significant homology to those of *S. buxifolia*. It is worth highlighting that *C. clementina* unpaired LTR have by definition their pairs excised and therefore the number of closely related unpaired LTR within the same genome should be lower than that of closely related admixtures, in which the generation of an unpaired LTR from the same retrotransposon has not taken place necessarily.

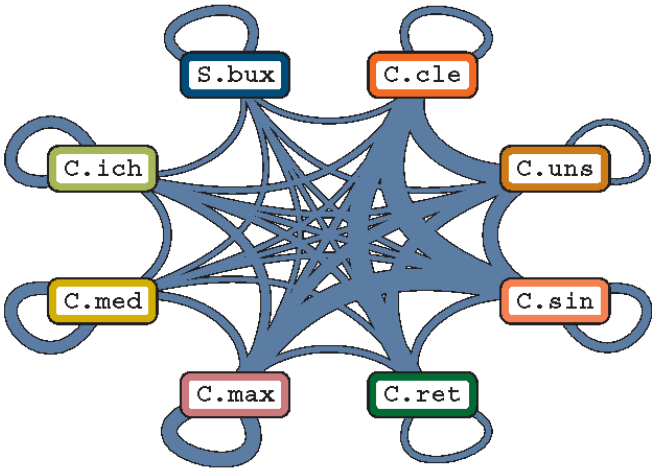


Figure 4: Unpaired LTR relatedness network. The width of the line between every pair of species is proportional to the number of shared soloLTRs and nonsoloLTRs. Loops indicate elements whose closest relative is found in the same genome. Only reciprocal hits were considered, and hence, no directionality is required. The same naming convention as that of Figure 1 is used.

For the remaining admixtures, a similar pattern was found, in which the majority of unpaired LTR had their closest relatives in either other admixtures or the pure species that gave rise to them. In contrast, in the pure species *C. medica*, *C. ichangensis* and *S. buxifolia*, most unpaired LTR found their closest relatives within the same genome, probably because they correspond to multiple insertions of similar elements. The case of *S. buxifolia* is especially remarkable, with 65% of its unpaired LTR having their closest relative within the same genome and only 35% of them being more similar to elements found in the *Citrus* genomes.

Accumulation patterns of Long Terminal Repeats across the genome

In the clementine genome, a total of 31221 LTRs (total LTR or total LTRs) were found by similarity with those detected by LTR_Harvest (Figure 2c). Of them, 9826 were paired LTRs, that is, they have at least one similar LTR in their flanking 30 kb. Of the remaining unpaired LTRs, 15471 were identified as true soloLTRs as they were flanked by a 4 to 7 bp long TSD. Finally, 5924 LTRs were found unpaired and lacking any TSD signature, thus being marked as nonsoloLTRs probably produced by IR or inter-element UR. The remaining 4 LTRs showed no homology with themselves, probably due to a misassignment as complete LTRs, and were discarded for further analysis. The pairedLTR:soloLTR:nonsoloLTR ratio was 1:1.57:0.60.

When the same methodology was applied to the set of species analyzed, a similar proportion of paired LTRs, soloLTRs and nonsoloLTRs were found. In this case, 96381 paired LTRs were detected. The number of soloLTR and nonsoloLTR was 123743 and 54009, respectively. 22 LTRs were discarded for the same reasons as above, and the final pairedLTR:soloLTR:nonsoloLTR ratio was 1:1.28:0.56.

By considering in a per-window basis the genic content, the number of paired, solo and non-solo LTR and their proportion related to the total number of LTRs, the correlation between purge rate and gene content was established (Supplementary Table 2). A negative correlation between total LTRs and genes was found in all genomes. When genic content was compared with the proportion of soloLTRs over total LTRs, a positive correlation was detected, indicating that soloLTR are more common in gene-rich regions. In contrast, nonsoloLTRs showed a positive correlation with the genic content in *C.*

medica, but also a negative correlation in *C. ichangensis* and *C. unshiu*. Finally, the proportion of paired LTRs, which should be a proxy of the complete retrotransposon abundance, was negatively correlated with the genic content in all but *C. ichangensis* genomes.

Evolution of retrotransposon activity among *Citrus* genomes

The distribution of the number of LTR retrotransposons dated at a certain age was used as a proxy of the activity of elements belonging to a specific lineage or superfamily at that given age (Figure 5a and 5b).

The number of retrotransposons dated at each age evolved similarly over time within each genome in both *Copia* and *Gypsy* superfamilies. However, when different species were compared, this similitude was no longer observed (Figure 5a). In the leftmost part of each plot, representing the oldest retrotransposons, the number of elements steadily increased with the age following a gradual rise in all eight species. However, starting from 0.06 K2p distance units, this pattern was no longer maintained among species (Figure 5a). Instead, from this point the age distribution in each species followed one of three different models: a) in the case of *C. clementina*, *C. maxima* and *C. ichangensis*, it increased progressively over time following an almost exponential pattern of growth; b) in *C. medica*, *C. reticulata* and *C. unshiu*, it was first arrested and then reduced, either slightly or considerably; c) in *C. sinensis* and *S. buxifolia* it followed a third pattern similar to the previous model b) except for a final recent burst.

When LTR retrotransposon superfamilies were disaggregated into lineages, their differences became more noticeable. In each of the species analyzed, different retrotransposon lineages followed distinct patterns that often differed from the species-specific patterns (Figure 5b). In 32 out of 46 reliable histograms (those including at least 100 elements), the retrotransposon age distribution resembled that of the species (Figure 5a). In some cases, a general trend in all lineages on a single species (or vice versa) was found, but every time some exceptions arose. For example, all lineages on *C. maxima* and *C. clementina* genomes were exponentially growing, except for SIRE and Reina elements. Conversely, Retrofit elements seemed to grow exponentially in all species except in *C. unshiu*, *C. reticulata* and *S. buxifolia*; meanwhile, SIRE element distribution peaked at

some point in the past in every genome except in *Severinia*, and its activity started to decay since then.

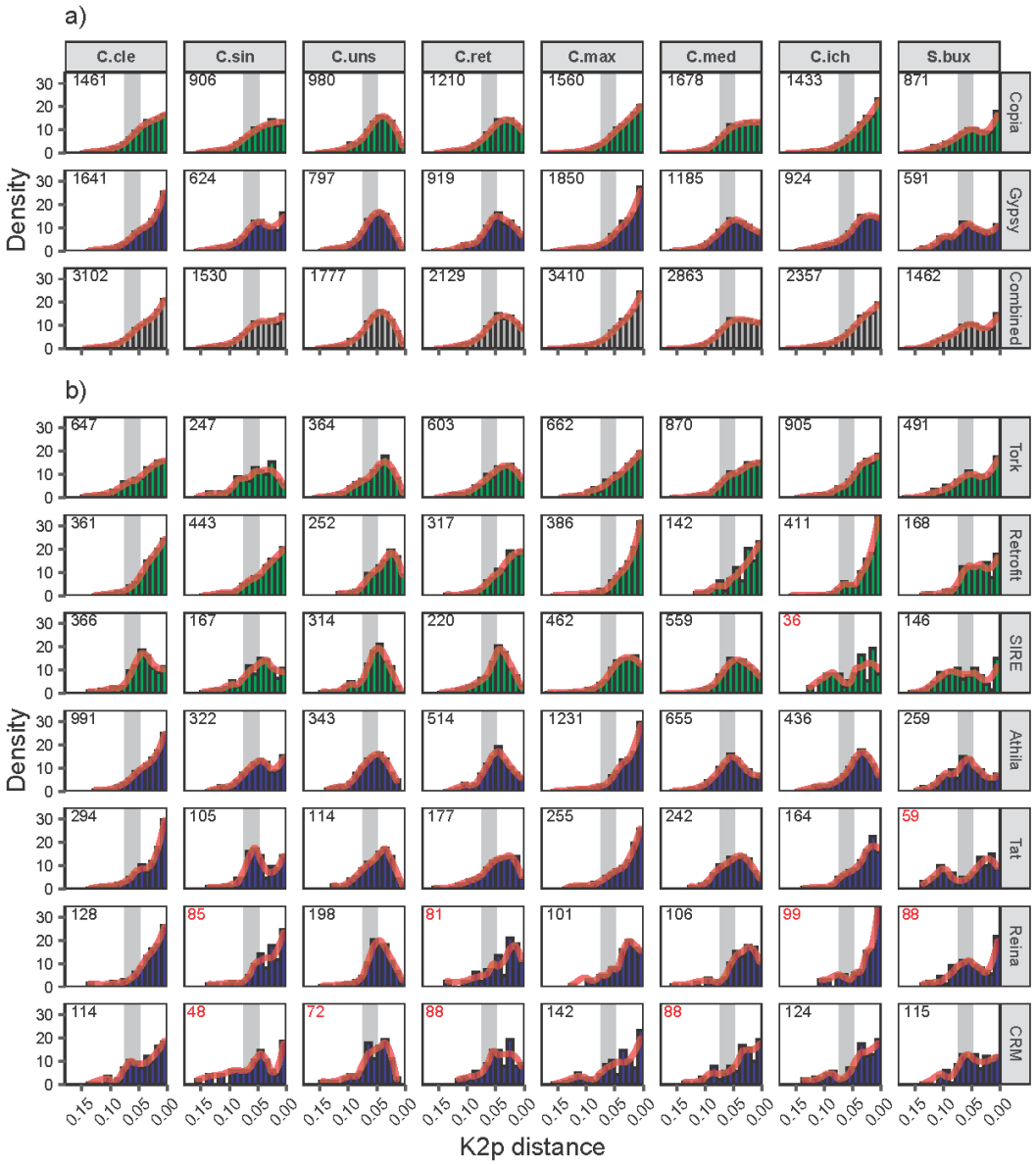


Figure 5: Retrotransposon activity pattern per species and lineage. Retrotransposon activity evolution over time. For each species, retrotransposons were grouped either in a) superfamilies or b) lineages. The proportion of retrotransposons falling in each specific age bin is shown, the total transposon numbers per each species and superfamily or lineage is shown in the top left corner. Histograms containing less than 100 observations had this number in red. Members from *Gypsy* and *Copia* superfamilies are colored green and blue, respectively. In gray, the proposed date for the *Citrus* radiation giving rise to the species studied (7.5-6.0 Mya) converted to distance units (0.075-0.048 K2p units) (Hu *et al.*, 2011) is shown. Species naming convention are as in Figure 1.

DISCUSSION

The retrotransposon landscape in *Citrus*

Citrus retrotransposons have recently seen a growing interest, especially since the publication of several reference genomes that have enabled high throughput retrotransposon surveys to be performed. The results presented above generally agree with two previous descriptive works reporting the retrotransposon landscape in different *Citrus* genomes (Du *et al.*, 2018; Liu *et al.*, 2019). We have found 32506 retrotransposon cores in eight genomes, and approximately half of them were annotated as full-length elements since they were flanked by two LTRs (the presence of other retrotransposon features such as a polypurine tract or a primer binding site was not verified). The average length of these complete retroelements, calculated both from the LTR-Harvest results and from the retrotransposon-induced deletions in *C. clementina*, was slightly above 8 kb per LTR retrotransposon, a length roughly conserved in the eight reference genomes (Table 2) and in agreement with the two abovementioned reports (Du *et al.*, 2018; Liu *et al.*, 2019). The average retrotransposon length was used to estimate the percentage of the genome covered by complete retrotransposons, that ranged from 3% to 10% of the genome (Table 2). These proportions were higher in the two better resolved genomes (*C. clementina* and *C. maxima*), possibly due to the difficulties in the detection of retrotransposons in Illumina-generated references. The retrotransposon abundances found for the different genomes largely agree with those of clementine (Du *et al.*, 2018) but are not in concordance with the results published by Liu *et al*. (Liu *et al.*, 2019), that reported values around 30% in six of the eight genomes studied in this work. These discrepancies might arise due to an overestimation of the retrotransposon collection, especially if fragmented retrotransposons were taken into consideration. In general, big genomes tend to contain higher proportions of mobile elements than smaller ones, as observed in maize (> 2 Gb genome size, 75% LTR retrotransposons) (Baucom *et al.*, 2009) and Arabidopsis (160 Mb, 6%) (Pereira, 2004), although rice for instance (390 Mb, 35%) (Sasaki, 2005) exhibits an intermediate situation.

Retrotransposon cores were grouped in families that could be classified in ten plant retrotransposon lineages, as reported in *C. clementina* (Du *et al.*, 2018). Our results are also comparable with those reported in (Liu *et al.*, 2019), even though the use of a

different retrotransposon lineage nomenclature hinders a direct comparison, an issue already encountered by other authors (Neumann *et al.*, 2019). Overall, the data show that only these ten retrotransposon lineages can be found across the multiple *Citrus* genomes. Interestingly, the great majority of the retrotransposon families of *Citrus* are present in all the genomes analyzed (Figure 1) and even in the distant species *S. buxifolia* that diverged from *Citrus* 13 Mya (Pfeil and Crisp, 2008), suggesting that most retrotransposon families were already hosted by the common ancestor of both. We also identified 17 families that were absent in some species and among them, five were not detected in *S. buxifolia*. Failure to detect every member of a family of LTR retrotransposons in a species is unexpected to occur due to technical limitations because these families are in general composed of numerous members inserted in different genomic positions. The absence of a given family in a specific species might be the result of insertions or deletions of retroelements, such as the colonization of a specific genome after its divergence with the remaining species (Piednoël *et al.*, 2013) or the depletion of a whole family previous to their proliferation, when the copy number remains low in the genome (Rahman *et al.*, 2015). An alternative explanation for undetected retrotransposon families is the process of incomplete lineage sorting, that can generate inconsistent genetic signals when alleles not fixed in a population are studied. Incomplete linage sorting has been considered in the field of plant phylogenetics (Strickler *et al.*, 2015; Zhou *et al.*, 2017) and has also been proposed as an explanation to unexpected retrotransposon presence/absence patterns in animals (Suh *et al.*, 2015; Kuritzin *et al.*, 2016; Doronina *et al.*, 2017). Since only one sampled individual per species was analyzed in this work, we cannot reject the possibility that some of the missing clades are produced by this process. Finally, de novo acquisition of families via hybridization or horizontal transfer, events already described in plants, may also be considered (Roulin *et al.*, 2009; El Baidouri *et al.*, 2014). While any of the above mechanisms may in principle cause the apparent loss of these 17 families, the 5 retrotransposon families missing *S. buxifolia* presumably colonized the *Citrus* genomes after their divergence with the genus *Severinia*.

We further investigated the relatedness between the retrotransposons present in the distinct species by estimating the degree of LTR sharing (Figure 4). In most pure species, the closest relative to each unpaired LTR was found in the same genome. This was expected, since retrotransposition events intrinsically generate copies of the same element and, before the first transposition within a genome, the closest relative of each LTR must

be generally found on the same genome. Oppositely, admixed genomes showed a completely different behavior: since admixtures are recent events, most retrotransposons have not yet replicated in the admixed genome, and therefore the transferred unpaired LTRs are more closely related to those present in the original species or in other admixtures derived from these species. These results highlight the importance of admixtures in the generation of novel LTRs combinations (and potentially retrotransposons) by combining haplotypes from different origins, a hypothesis proposed in one of the earliest transposon studies (Suoniemi *et al.*, 1998). While most LTRs followed the abovementioned trend, some of them found their closest relatives in distant species (for instance, clementine's LTRs whose closest relative was detected in *S. buxifolia* or *C. ichangensis*). Although this observation may certainly pinpoint to a failure in the detection of their closest homologues, the occurrence of closely related LTRs in highly divergent species supports the idea that they can indeed persist over long periods of time even when the retrotransposon itself is no longer present (Ma and Bennetzen, 2004; Hawkins *et al.*, 2009).

Mechanisms of retrotransposon accumulation in *Citrus*

Regarding the retrotransposon distribution across the genome, we first focused on the *Citrus clementina* genome. The genic content per genomic window was used to roughly estimate the location of pericentromeric regions in the different chromosomes (Figure 2d), that was generally in accordance with previously reported centromere locations (Wu *et al.*, 2014; Aleza *et al.*, 2015). Pericentromeric regions were indeed enriched in LTR retrotransposons while the genic abundance was low (Figure 2a), a pattern conserved in all genomes analyzed (Supplementary Table 2) in line with previous findings in *Citrus* (Du *et al.*, 2018; Liu *et al.*, 2019) and other species (Paterson *et al.*, 2009; Xu and Du, 2014). It is generally accepted that this pattern may arise to either a purifying selection against gene-disrupting retrotransposon insertions (Pereira, 2004) or an increased unequal recombination rate in uncondensed regions (Tian *et al.*, 2009), two processes that would reduce retrotransposon half-life in gene-rich regions and produce a preferential accumulation of recently inserted elements in them, as observed in Figure 2b. However, both hypotheses are not mutually exclusive, and their combination actually might better explain the accumulation pattern observed in this work. Consequently, the patterns of

retrotransposon insertion, accumulation and purge were analyzed to determine their effects on shaping the studied genomes.

To understand whether UR has a decisive effect in the retrotransposon distribution, UR rates across each genome were estimated. Considering that the paired LTR to soloLTR conversion is unidirectional, the soloLTR to total LTR proportion was taken as a proxy of the soloLTR generation frequency (Cossu *et al.*, 2017; Liu *et al.*, 2019), which equals the intra-element UR rate. We found UR to be consistently more frequent in the genic regions of every genome analyzed (Supplementary Table 2), in agreement with previous works in Arabidopsis (Pereira, 2004), providing an explanation for the accumulation of complete LTR retrotransposons in pericentromeric regions. This hypothesis is further supported by the position of the retrotransposon activity hotspots found in mandarins (Figure 2a and Supplementary Figure 2), that were primarily located in genic regions, as observed for the tomato genome (Xu and Du, 2014).

We also studied the rate of generation of nonsoloLTR to determine the sum of the inter-element UR and IR rates, and found no significant or consistent variations between genic and non-genic regions in most of the genomes (Figure 2c and Supplementary Table 2). This inconsistency together with the low number of nonsoloLTRs found in all genomes (only 30% of the unpaired LTR) may suggest that the combined effect of UR and IR is not determinant in the LTR accumulation patterns observed.

On the other hand, the increase in the retrotransposon purge rate (the sum of UR and IR purge) in the genic regions appears to account for the retrotransposon age distribution found in six out of the eight species analyzed (Figure 3), as has been described in Arabidopsis and tomato (Pereira, 2004; Xu and Du, 2014). In these genomes, old retrotransposons are preferentially accumulated in the pericentromeric regions, that show a reduced transposon deletion rate which in turn slows the transposon turnover while increasing their half-life (Tian *et al.*, 2009; Pellicer *et al.*, 2018). In citrons and pummelos, however, other different mechanisms must operate since the retrotransposon age distribution in genic and pericentromeric regions are very similar. In pummelos, new retrotransposons are preferentially inserted in pericentromeric regions leading to uniform age distributions along the chromosome but with a much larger number of retrotransposons in non-genic regions. Currently, there is not a general agreement on whether or not retrotransposons preferentially insert in some regions of the genome since

evidences have been found for centromeric (Tsukahara *et al.*, 2012) and euchromatic (Wei *et al.*, 2016; Nakashima *et al.*, 2018) preferential insertions, or even for a completely unbiased distribution (Levin and Moran, 2011).

Apart from these mechanisms, the effect of purifying selection has been suggested to become relevant in gene-rich regions, where insertion has higher chances of reducing the overall fitness of the individuals favoring the selection of transposon-free alleles (Pereira, 2004; Xu and Du, 2014) without requiring recombination or leaving any detectable signature on the genome. In *Citrus*, the total LTR count is significantly higher in pericentromeric regions even if insertion is generally unbiased. This observation strongly suggests that purifying selection is playing an important role in shaping the retrotransposon landscape of *Citrus*, since that count, i.e., the number of paired LTRs plus twice the number of unpaired LTRs (soloLTR and nonsoloLTR), is not constant across the genome (Figure 2c), as expected when insertion is uniformly distributed.

While multiple studies have reported the accumulation of complete LTR retrotransposons in pericentromeric regions, here we extend this concept and propose that the total LTR count is an indicator of retrotransposon purge through mechanisms other than recombination, provided the occurrence of unbiased insertion. It is worth to mention that differences in the selective pressure could modulate the reduction of the number of young elements in the genic regions, shifting the distribution towards older ages to distinct levels. Thus, an increased selective pressure might produce, for instance, the pattern depicted for *C. medica* in Figure 3. Therefore, our results suggest that the retrotransposon accumulation pattern found in the eight genomes analyzed might be explained by the combination of UR purge and purifying selection, whose combined effect permits the pericentromeric regions of *Citrus* and *Severinia* genomes to behave as safe havens for retrotransposons, as described in many plants (Pereira, 2004; Levin and Moran, 2011).

Regulation of retrotransposon activity during *Citrus* speciation

It is generally accepted that retrotransposon insertion rate continuously increases over time while the purge rate remains constant. Based on these premises, LTR age distribution has been suggested to follow an exponential growth curve, as modelled in multiple species including *Citrus* (Wicker and Keller, 2007; Hawkins *et al.*, 2009; Liu *et al.*, 2019). While retrotransposon removal is in principle an unspecific process derived from

recombination, retrotransposon activity appears to be a clearer target for differential regulation. Consequently, the number of elements detected in each bin has been repeatedly used as a proxy to date retrotransposons in several works (Hu *et al.*, 2011; Bousios *et al.*, 2012; Zhang and Gao, 2017). However, some authors suggest that the commonly observed ever-growing profile of retrotransposon activity might be indeed produced by retrotransposon removal process, that steadily deletes elements (Dai *et al.*, 2018). This vision implies that the old elements that are detected in current genomes are those that survived by chance all this time, while the deleted elements are systematically disregarded as they are no longer present in the genome. Under these circumstances, the age distribution is not exactly comparable with the insertion history, but rather a proxy that underestimates the insertion rate values, especially in older age bins. However, as long as the deletion rate does not abruptly change among species, the age distribution shape in the most recent times should resemble that of the insertion history.

In this work, retrotransposons were independently dated in every superfamily, lineage of retrotransposons and *Citrus* species (Figure 5). Within a given species, activity of both *Copia* and *Gypsy* superfamilies followed similar patterns, although each species developed a specific pattern of change. The results show that the species-specific patterns of transposon activity detected in the *Citrus* genomes can be basically grouped in three models: a) exponential or continuous increase over time (*C. clementina, C. maxima* and *C. ichangensis*), b) initial continuous increase followed by a sudden arrest and a final phase of gradual reduction (*C. unshiu*, *C. reticulata* and *C. medica*) and c) initial increase, sudden arrest, reduction and a final period of regrowth (*C. sinensis* and *S. buxifolia*).

The observation that genomes from pure *Citrus* species sharing a recent common ancestor (*C. maxima* and *C. medica* diverged about 6 Mya (Wu *et al.*, 2018)) exhibit different patterns of activity suggests that such activity may evolve independently in species with a common ancestor and therefore, that the phylogenetic relatedness of the genomes is not necessarily associated with their activity pattern. The same conclusion can be inferred from the comparison of other pure species pairs such as *C. maxima* and *C. ichangensis* (that shared their last common ancestor 8 Mya (Wu *et al.*, 2018)) since both followed the same activity pattern type a. These evidences highlight the different transposon activity profiles that can be found even in relatively close genomes, as previously suggested (Hawkins *et al.*, 2009; Zhang and Gao, 2017). In general, transposon activity among

similar species tend to evolve in parallel (Kim *et al.*, 2017) while more distant species do not present analogous activity trends (Wicker and Keller, 2007; Xu and Du, 2014), although this is not always the case (Estep *et al.*, 2013).

Remarkably, the patterns of activity change in *Citrus* show two observations of relevance that are apparently connected. One is that the speed of change among the different *Citrus* species is extremely fast when compared to those published up to date in other plants (Estep *et al.*, 2013; Piednoël *et al.*, 2013; Xu and Du, 2014; Kim *et al.*, 2017). Moreover, in three out of the five pure species analyzed (*C. reticulata*, *C. medica* and *S. buxifolia*) the increase of transposon abundance is strikingly arrested at similar K2p distance units (0.06-0.04). A rate of $4 \cdot 10^{-9}$ to $5 \cdot 10^{-9}$ silent base-pair substitution per year (De La Torre *et al.*, 2017), multiplied by a factor of two to correct for the LTR increased substitution rate (Ma and Bennetzen, 2004; Hu *et al.*, 2011), was used to date the element insertions. These calculations revealed that the turning point dating the arrest of activity took place 7.5-4.0 Mya (using the widest intervals). Interestingly, the radiation originating the foundational *Citrus* species studied in here has been reported to occur 7.5-6.0 Mya during the Late Miocene in continental Southeast Asia (Wu *et al.*, 2018), a period and region characterized by deep environmental changes. A causal connection of environmental changes and reprograming of retrotransposon activity would require further studies, but it is nevertheless very tempting to suggest that *Citrus* retrotransposons may also respond to the stressful conditions driving speciation, as a part of the genetic machinery responsible of adaptation. It is also worth to mention that the pattern of change of retrotransposon activity previous to the speciation processes is practically identical among all *Citrus* species analyzed (Figure 5) as theoretically expected, since these by definition come all from a common ancestor.

Furthermore, our results also suggest that the evolution of retrotransposon activity is, in principle, associated with the genealogic proximity, as observed in the three *Citrus* admixtures *C. sinensis* (sweet orange), *C. unshiu* (satsuma mandarin), and *C. clementina* (clementine mandarin). Actually, next generation sequencing has revealed that most important domesticated *Citrus* cultivars are in fact admixtures of true species, that are popularly recognized as oranges, mandarins and lemons (Wang *et al.*, 2017*b*; Wu *et al.*, 2018). These admixtures had distinct recent origins, but a similar genomic background composed of combinations of *C. reticulata* and *C. maxima.* Sweet oranges, that contain

pummelo chloroplasts, are grouped under the binomial name of *C. sinensis*, while the term "mandarin" comprises a very heterogenic collection of genomes including pure mandarin species (*C. reticulata*) and genotypes with different proportions of pummelo introgression (i.e., *C. unshiu*, *C. clementina*, *C. deliciosa*, etc.) in a maternal mandarin genome. Our data indicate that the genome of the satsuma mandarin *C. unshiu*, for instance, that contains a high proportion of pure *C. reticulata* (86 %,) showed resembling or parallel changes (model b) to those of the pure mandarin. Similarly, transposon activity in the orange *C. sinensis* (42 % of *C. reticulata*) appears to follow a pattern (model c) intermediate between *C. maxima and C. reticulata*.

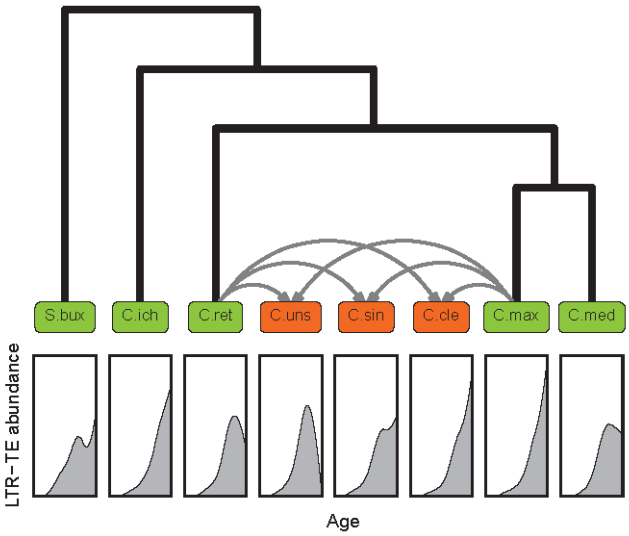


Figure 6: Retrotransposon activity and *Citrus* phylogeny. Cladogram representing the phylogeny of the eight species analyzed in this study associated with the pattern of retrotransposon activity found in each one of them. Pure species are framed in green boxes while admixtures are framed in orange boxes, with gray arrows indicating their pure species progenitors. The overall retrotransposon activity evolution over time is presented below each species name. Species codes are as in Figure 1

The activity pattern (model a) of *C. clementina*, an admixture of the orange *C. sinensis* (*C. maxima x C. reticulata*) and the mandarin *C. deliciosa* (*C. reticulata* x *C. maxima*), was similar to that of *C. maxima* (Figure 6), although the contribution of pummelo to the clementine genome is only of 12 % (Wu *et al.*, 2018). These observations suggest that *C. deliciosa* mandarin, whose reference genome is not available, must carry highly active retrotransposons to produce the profile observed in clementine and that the mandarin haplotype included in *C. deliciosa* neither is the same that contains the *C. unshiu* mandarin nor is directly associated with the genome of the pure *C. reticulata* sequenced

(Wang *et al.*, 2018*a*) and used in the current work. This last assumption is derived from this previous study (Wang *et al.*, 2018*a*) that divided domesticated mandarins in two different clades, one evolving through the north of the Nanling Mountains, which included *C. unshiu*, and the other expanding to the south of this mountain range and harboring *C. deliciosa*. Nanling Mountains in Southern China separate south and central subtropical zones. It is worth to mention that not only *C. unshiu* and *C. clementina* arose from different mandarin genomic backgrounds but at least four different pummelo haplotypes are also found into the genomes of these two mandarin admixtures.

Another set of interesting data come from the individualized analyses of the different retrotransposon lineages that evidences how in every species studied, some lineages did not follow the general pattern of activity of the species itself. For example, the increase in activity of SIRE elements was the highest in the past just before the beginning of the *Citrus* speciation, i.e., the abundance of SIRE elements was progressively reduced in all *Citrus* analyzed, but not in *Severinia*. This together with their abundance (they rank 3rd or 4th) suggests among other possible explanations, that these elements have not been able to counteract the genomic mechanisms implicated in their silencing process in *Citrus*. On the contrary, Retrofit elements have continuously been growing over time in most of the genomes, including some of those showing different models in the general tendency, such as *C. reticulata* (model b) or *C. sinensis* (model c). Retrofit elements, therefore, show an elevated ability to overcome hosts regulation, as described previously for other lineages (Hernández-Pinzón *et al.*, 2012; Fu *et al.*, 2013; Lu *et al.*, 2017*a*). This is not a surprise since different behaviors of transposon lineages and families within a single genome have been already reported (Piegu *et al.*, 2006; Bousios *et al.*, 2012) and recent studies have also observed great variations on transposon activity in groups of closely related species (Estep *et al.*, 2013; Quadrana *et al.*, 2016; Zhang and Gao, 2017; Carpentier *et al.*, 2019)

The detailed analyses of the activity of each retrotransposon lineage revealed that only in two genomes, *C. unshiu* (model b) and *S. buxifolia* (model c), all lineages showed the same pattern. As mentioned above, *C. clementina and C. reticulata* followed models a and b, except for the SIRE and Retrofit families. There were two lineages that escaped to the general tendencies found in *C. sinensis* (model c), *C. medica* (model b) and *C. ichangensis* (model a). These were Tork and Retrofit in the first two genomes and Athila and Tat in the papeda. Finally, Reina, CRM and SIRE retrotransposon families showed

evolutionary trends dissimilar to the pivotal patterns of gradual growth found in *C. maxima*. Overall, these results indicate that mobile element activity in each *Citrus* genome follows a characteristic and recognizable pattern of change although very often a few retrotransposon lineages evolve independently following a different trend. Except for the SIRE elements that in *Citrus* always show a tendency of type b, all lineages show patterns that follow either models of type a or b, while many lineages of the *Gypsy* superfamily in addition exhibit models of type c.

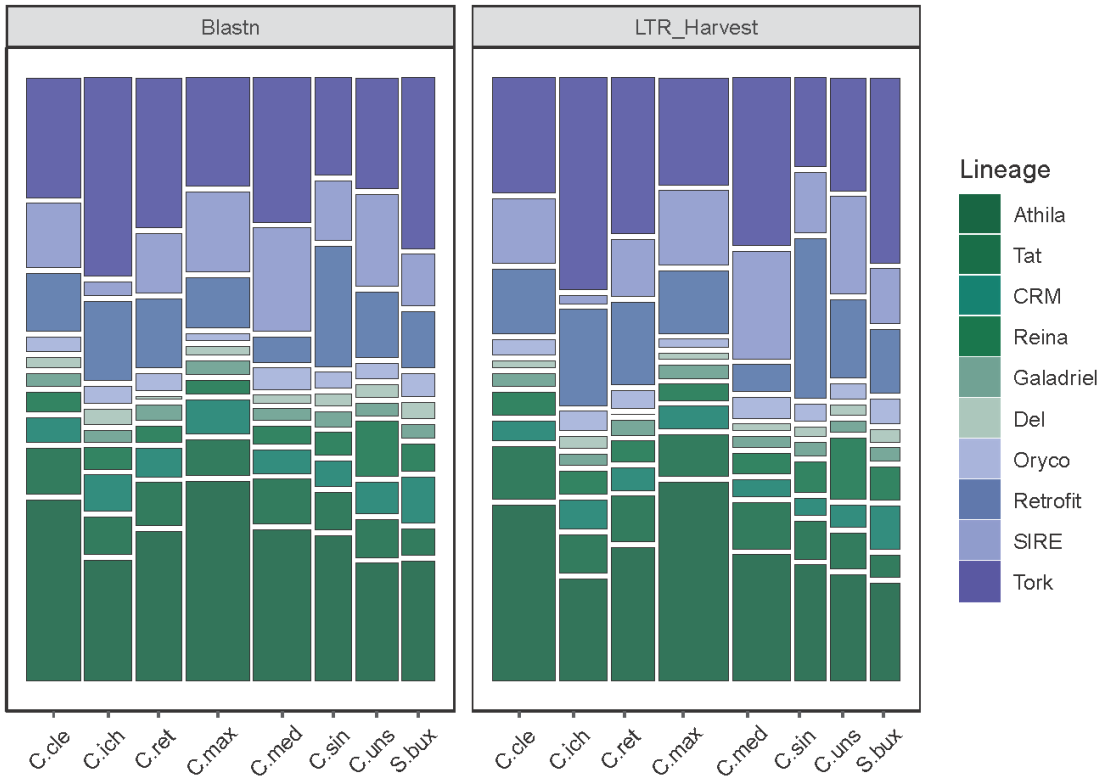
In conclusion, our results show that in *Citrus*, retrotransposon activity in a given species or admixture is not clearly related to any fundamental genomic or phylogenetic factor. Although the pattern of activity of the *Citrus* admixtures is originally associated with the genealogic proximity of their genomes, the drastic changes in the activity that each species experiences over time appear to be mainly driven by the evolutive history of its particular genome. Interestingly, in some genomes the expected pattern of gradual transposon accumulation is strikingly arrested shortly after the radiation of the *Citrus* genus, coinciding with a geological era characterized by dramatic climate changes. Overall, our results may suggest that the retrotransposon evolutionary landscape is largely governed by the individual past of each species or population, a hypothesis compatible with the changing environmental scenarios and evolving conditions that occurred during *Citrus* speciation. Based on these observations we propose that *Citrus* retrotransposons might respond to those stressful conditions driving speciation, as a part of the genetic machinery responsible of adaptation. This proposal implies that the evolving conditions of each species may interact with the internal regulatory mechanisms of the genome regulating proliferation of the mobile elements and that this interaction may be very subtle since it discriminates between different lineages of retrotransposons.

DATA AVAILABILITY

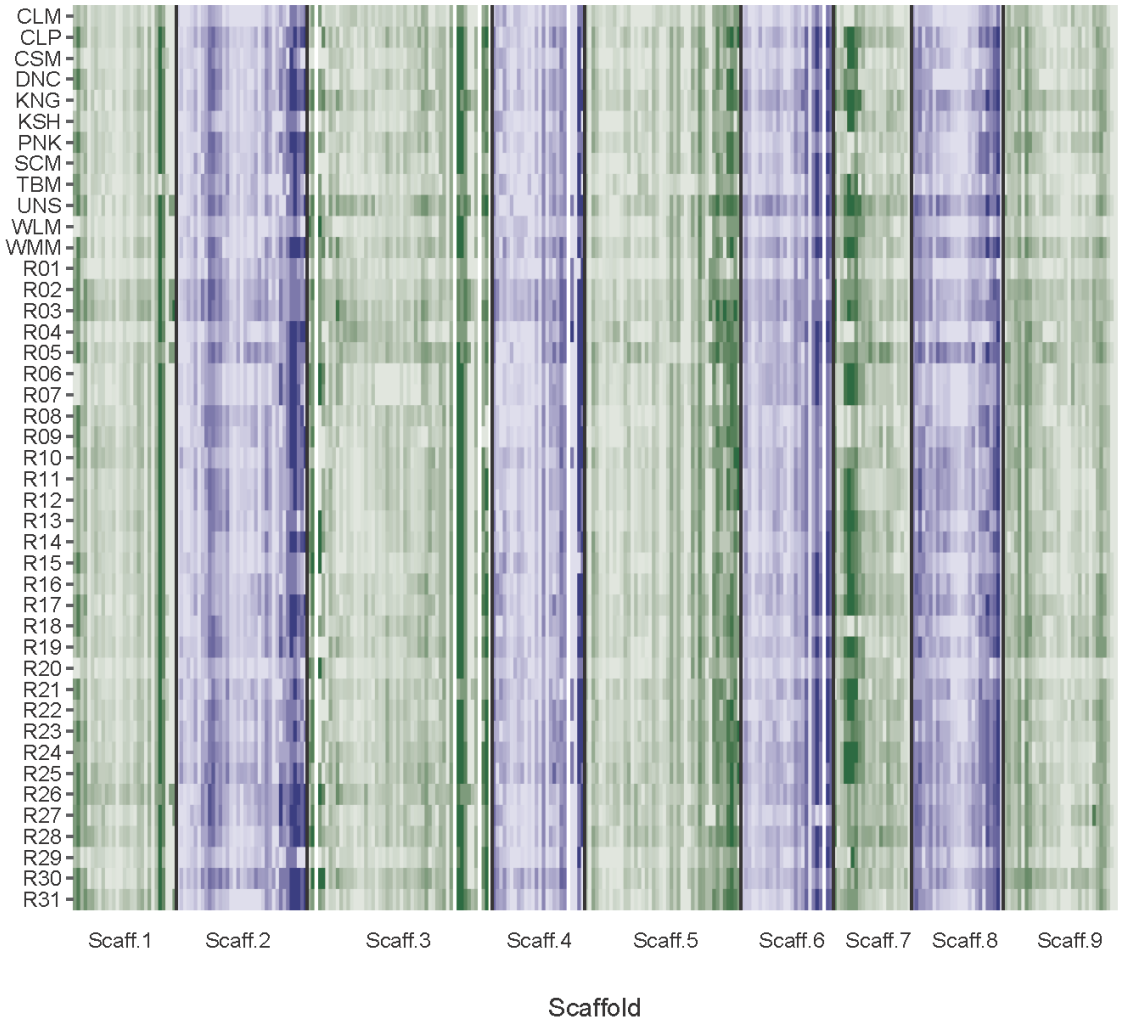
The reference genomes used in the current study are available in the NCBI Assembly repository. The GeneBank assembly accession for each reference genomes are the following: *C. clementina* (GCA_000493195.1), *C. sinensis* (GCA_000317415.1), *C. unshiu* (GCA_002897195.1), *C. reticulata* (GCA_003258625.1), *C. maxima* (GCA_002006925.1), *C. medica* (GCA_002013955.2), *C. ichangensis* (GCA_002013975.2), *S. buxifolia* (GCA_002013935.1). The Illumina-sequenced

mandarin genomes are available from the Sequence Read Archive and their identifiers are provided in the Supplementary Table 1. The genomic locations of every LTR retrotransposon (either complete elements or retrotransposon cores) are listed in the Supplementary Table 3 based on the coordinated of the reference genomes used in this work (see Methods).

SUPPLEMENTARY DATA



Supplementary Figure 1: LTR retrotransposon abundance per genome, lineage and detection method. For each detection method (homology and LTR_Harvest), the transposon proportion per genome and clade was calculated and plotted as a mosaic plot. The area of each rectangle is proportional to the number of elements for this category. The lineages are divided by color, with *Copia* elements in blue tones and *Gypsy* elements in green tones. The genome names are abbreviated as in Figure 1.



Supplementary Figure 2: Transposon-induced deletions across mandarins. Name codes can be found in the Supplementary Table 1. Color intensity is proportional to the number of LTR retrotransposons deleted in the *C. clementina* reference genome.

Supplementary Table 1. Sequence Read Archive identifiers of the Illumina-sequenced mandarin genomes.

SRA accession code	Sample abbreviation	Common name
SRX371962	CLM	Clementine mandarin
SRX3298480	CLP	Cleopatra mandarin
SRX3298481	CSM	Changsha mandarin
SRX3298482	DNC	Dancy mandarin
SRX3298483	KNG	King mandarin
SRX3298465	KSH	Kishu mandarin
SRX372665	PNK	Ponkan mandarin
SRX3298473	SCM	Sun Chun Sha mandarin
SRX3298464	TBM	Tachibana mandarin
SRX3298475	UNS	Satsuma mandarin
SRX372685	WLM	Willowleaf mandarin
SRX372687	WMM	W. Murcott mandarin
SRX2178448	R01	Chinese mandarin 1
SRX2177849	R02	Chinese mandarin 2
SRX2177806	R03	Chinese mandarin 3
SRX1923226	R04	Chinese mandarin 4
SRX1922157	R05	Chinese mandarin 5
SRX1922136	R06	Chinese mandarin 6
SRX1922109	R07	Chinese mandarin 7
SRX1906045	R08	Chinese mandarin 8
SRX1905992	R09	Chinese mandarin 9
SRX1905979	R10	Chinese mandarin 10
SRX1904603	R11	Chinese mandarin 11
SRX1904247	R12	Chinese mandarin 12
SRX1904234	R13	Chinese mandarin 13
SRX1904214	R14	Chinese mandarin 14
SRX1904177	R15	Chinese mandarin 15
SRX1903372	R16	Chinese mandarin 16
SRX1901508	R17	Chinese mandarin 17
SRX1901493	R18	Chinese mandarin 18
SRX1901484	R19	Chinese mandarin 19
SRX1901458	R20	Chinese mandarin 20
SRX1901417	R21	Chinese mandarin 21
SRX1901408	R22	Chinese mandarin 22
SRX1901407	R23	Chinese mandarin 23
SRX1901265	R24	Chinese mandarin 24
SRX1901202	R25	Chinese mandarin 25
SRX3030196	R26	Chinese mandarin 26
SRX3030172	R27	Chinese mandarin 27
SRX3030222	R28	Chinese mandarin 28

Table S1 (continued)

SRA accession code	Sample abbreviation	Common name
SRX2977418	R29	Chinese mandarin 29
SRX3030557	R30	Chinese mandarin 30
SRX3032880	R31	Chinese mandarin 31

Data was retrieved from the SRA using the codes specified in the table. The naming convention and the sample accession common name are shown. Wild chinese mandarins without a known common name were noted as Chinese mandarin (1 - 31).

Supplementary Table 2. Spearman rank correlation test.

Spearman rank correlation test	LTR-RT cores abundance and genic content	New LTR-RT insertions and genic content	sLTR to tLTR proportion and genic content	nsLTR to tLTR proportion and genic content	pLTR to tLTR proportion and genic content	TotalLTR abundance and genic content
Citrus clementina	**-0,899 ***	-0,026	**0,297 ***	-0,056	**-0.281 ***	**-0,801 ***
Citrus ichangensis	**-0,617 ***	0,035	**0,148 ***	**-0,122 ***	-0,06	**-0,334 ***
Citrus reticulata	**-0,708 ***	0,048	**0,194 ***	-0,085	**-0.146 ***	**-0,316 ***
Citrus maxima	**-0,691 ***	**-0,216 ***	**0,302 ***	0,101	**-0.332 ***	**-0,762 ***
Citrus medica	**-0,411 ***	0,034	**0,298 ***	**0,093 ***	**-0.323 ***	**-0,451 ***
Citrus sinensis	**-0,874 ***	**-0,265 ***	**0,242 ***	-0,054	**-0.256 ***	**-0,925 ***
Citrus unshiu	**-0,601 ***	0,054	**0,207 ***	**-0,073 ***	**-0.166 ***	**-0,201 ***
Severinia buxifolia	**-0,527 ***	0,060	**0,137 ***	-0,057	**-0.128 ***	**-0,296 ***

* p-value < 0.05

Supplementary Table 3. Genomic locations of complete LTR retrotransposon and retrotransposon cores. Table S3 can be found, in the online version, at https://doi.org/10.1093/gbe/evz246.

Chapter 3

Effects of domestication on gene expression in ripening citrus fruits

ABSTRACT

Citrus comprises hundreds of commercial varieties with a striking phenotypical diversity, especially in their fruits, which are most appreciated because of their taste, bright colors and health benefits. Despite the importance of this fruit crop in the global market, little is known about the domestication mechanisms generating the current *Citrus* diversity. To better understand the process of citrus domestication, the fruit transcriptomes of seven citrus species, representing wild species and domesticated varieties, have been analyzed. The admixed nature of the samples has been considered in order to determine the genomic regions involved in the domestication of the genus *Citrus*. This genus-wide study allowed the extension of previous hypotheses and the proposal of new mechanisms determining some of the commercially relevant traits mentioned above. The transcriptomic analysis revealed a consistent overexpression of vacuolar ATPases in the acidic citron and lemon pulps compared with other species. We also suggest a role for the carotenoid cleavage dioxygenase *CCD4a* in determining carotenoid content, despite its low expression levels in colored citrus fruits. The results also highlight the existence of a chalcone synthase *CHS* highly expressed in mandarins and their admixtures but not in citron and pummelo, which appears to be strongly related to the accumulation and diversification of flavonoids in mandarin peels. Finally, this work provides evidence supporting that citrus domestication was mostly shaped by early interspecific hybridizations and subsequent selection, with the desired traits being maintained across generations by the clonal propagation of the admixed cultivars.

Keywords: Citrus, domestication, RNAseq, allele specific expression, fruit ripening

INTRODUCTION

Citrus are among the main fruit crops worldwide, with oranges, grapefruits, lemons and mandarins as the most economically relevant cultivars. Most of these commercial citrus are not pure species but interspecific hybrids, also known as admixtures, that harbor genomic fragments from citron, pummelo and mandarin, considered pure species (Wu *et al.*, 2018). Pure or wild mandarins are not edible, while the cultivated varieties are appreciated by their palatability, that is associated with several pummelo introgressions in the original mandarin genome (Wu *et al.*, 2014, 2018). Sweet oranges and grapefruits are also mandarin/pummelo admixtures harboring larger and more frequent pummelo introgressions, with some genomic regions displaying two pummelo alleles (Oueslati *et al.*, 2017). Sour oranges share the pummelo/mandarin ancestry, but as direct hybrids of a mandarin x pummelo cross, their genome display two complete parental haplotypes. Lemons, that resulted from of a cross of sour orange and citron, have one complete citron haplotype, while the other one shows the admixture produced by the mandarin and pummelo ancestries (Curk *et al.*, 2016).

Genomic analyses suggest that the specific admixture patterns of each citrus cultivar largely determines the phenotype, and might imply human participation (Curk *et al.*, 2014; Oueslati *et al.*, 2017; Wu *et al.*, 2018). The generation of such admixtures is a complex process that requires crosses between pure species followed by backcrosses and/or crosses with other admixtures. This process has been related to the domestication of citrus species, together with the selection and propagation of the admixture with desirable traits. The complex relatedness network shared by mandarins, oranges and grapefruits, suggesting that they all share some recent common ancestors, would support this hypothesis (Wu *et al.*, 2018).

Human selection during crop domestication has resulted in remarkable transformations of plant phenotypes, and progress in advanced molecular technologies allowed the study of the genetic architecture of novel plant traits. These advances revealed a diversity of factors affecting phenotypes important in plant domestication, including novel gene expression patterns. Human selection unknowingly targeted structural and regulatory genes, with results that propagate through the transcriptome as well as to other levels in the biosynthetic and morphogenetic networks (Olsen and Wendel, 2013).

Domestication have profound effects in gene expression, reshaping the transcriptome at a global level, and enhancing the differential expression of genes associated with the agronomical traits targeted by the domestication process. This way, comparative transcriptomic revealed patterns of selection in domesticated tomato (Koenig *et al.*, 2013; Sauvage *et al.*, 2017), and RNA-seq performed at the population level showed how artificial selection greatly shaped the tomato transcriptome, altering the fruit sugar content and resistance to abiotic and biotic stresses (Liu *et al.*, 2020*b*). Reshaping of the maize transcriptome by domestication has been also analyzed by expression profiling analyses, documenting alterations in the maize transcriptome following domestication and identifying several genes that may have contributed to the evolution of maize (Swanson-Wagner *et al.*, 2012).

Knowledge of the genetic changes that occurred during the domestication and improvement of perennial trees at the transcriptomic level is limited, although RNA sequencing analysis of wild, landrace, and improved cultivars of pear (*Pyrus pyrifolia*), revealed specific patterns of domestication and improvement, many of them highly associated with important fruit traits (Li *et al.*, 2019*a*). Evolutionary transcriptomics has been also used to reveal the origins of olives and the genomic changes associated with their domestication, showing how the domestication of olives resulted in only moderate genomic consequences and that the domestication syndrome is mainly related to changes in gene expression, consistent with the olive tree evolutionary history and life history traits (Gros-Balthazard *et al.*, 2019)

Citrus fruits display a wide variability in size (from very small to very large), shape (from round to cylindrical), color (from green to orange) and flavor (from very acid to very sweet), which make them a favorite of the markets. Despite this broad phenotypic diversity, genomic studies have reported a highly conserved genome, both in structure and gene content. Thus, all the *Citrus* analyzed genomes are organized in 9 chromosomes, that show an almost perfect synteny (Shimizu *et al.*, 2017; He *et al.*, 2020), as well as a very similar number of highly conserved genes (Shimizu *et al.*, 2017; Peng *et al.*, 2020). Therefore, the variability found in citrus must rely in other factors, and changes at the expression level might appear as some of the influential ones that could be ultimately associated to the domestication of *Citrus* species.

As fruit quality is a direct consequence of the ripening process, much effort has been made to analyze maturation at different levels, and recently several studies have used transcriptomic approaches to unveil the genetic mechanisms controlling citrus fruit ripening. These works focused on the study of the regulation of sugar content and acidity (Lu *et al.*, 2016; Huang *et al.*, 2016*b*) as well as on the accumulation of relevant metabolites such as flavonoids (Wang *et al.*, 2017*c*) and terpenoids (Lu *et al.*, 2018). These works are only based in one-to-one comparisons between somatic mutants against their original variety, or between closely related varieties, but no genus-wide study has been carried out so far.

As mentioned above, the main assets of citrus fruits are their characteristic flavors and bright colors, so we also directed our efforts to study them in the context of the evolutionary transcriptomic analysis carried out in this work. Flavor is mainly given by the acidity to sweetness balance, which sets towards the end of the ripening process. Acidity is mostly determined by pH and titratable acid concentration (Chaimanee and Suntornwat, 1994; Da Conceicao Neta *et al.*, 2007; Hussain *et al.*, 2017), while sweetness depends on the total sugar concentration, measured as total soluble solid content (TSS). While TSS has been directly linked to fruit taste in many commercial fruits (Fellers, Paul J, 1991; Kuhn *et al.*, 2014; Ikegaya *et al.*, 2019), in some acidic citrus the extreme juice acidities might mask the sugar content and dominate the flavor perception (Strazzer *et al.*, 2019).

In these acidic citrus, the vacuolar lumen in the pulp vesicles reaches pH values as low as 2, more than five points below the cytoplasmic pH (Müller *et al.*, 1996). The steep pH gradient is promoted by the citrate vacuolar intake (Brune *et al.*, 1998), which buffers the vacuolar lumen and allows a continuous proton intake that maintains the low vacuolar pH (Shimada *et al.*, 2006). Despite its central role in this process, citrate biosynthesis does not appear to be directly correlated with its accumulation in citrus pulp (Chen *et al.*, 2013; Lin *et al.*, 2015; Lu *et al.*, 2016; Guo *et al.*, 2016). Instead, citrate accumulation seems to depend on other processes, notably its degradation (Cercós *et al.*, 2006; Hu *et al.*, 2015) and storage in the cell vacuoles (Guo *et al.*, 2016; Strazzer *et al.*, 2019). In non-acidic commercial citrus such as sweet oranges and clementines, sugar accumulation is a major trait for fruit quality. This accumulation is the result of a metabolic change from sucrose utilization to its storage (Tadeo *et al.*, 2008), which increases sucrose concentration in the

pulp (Komatsu *et al.*, 2002). As the fruit ripens, the expression levels of sugar invertases (*INV*) drop, while sucrose phosphate synthases (*SPS*), sucrose phosphatases (*SPP*) and sucrose synthases (*SuSy*), all involved in sucrose synthesis, increase their expression (Komatsu *et al.*, 1999, 2002; Katz *et al.*, 2011), which result in the accumulation of sucrose and its derivatives in the pulp during the late ripening stages (Hussain *et al.*, 2020).

Another major agronomical trait of citrus fruits is linked to their bright colors, produced mostly by the accumulation of carotenoids. During color break, peel chlorophylls are hydrolyzed (Jacob-Wilk *et al.*, 1999) while carotenoid biosynthesis is promoted in the chromoplasts (Kato *et al.*, 2004) and, in red-colored fruits, the carbon flux is redirected towards the production of β-carotene and its derivatives (Zhang *et al.*, 2012). The differential accumulation profile of carotenoids and apocarotenoids generates the broad range of colors observed in *Citrus* species (reviewed in Rodrigo *et al.*, 2013*a*). Specifically, the bright red color found in many citrus is produced by the accumulation of C30 apocarotenoids such as β-citraurin, while other carotenoids such as violaxanthin also contribute to the final color (Oberholster *et al.*, 2001). In contrast, yellow and non-colored fruits such as pummelos, citrons and lemons display lower carotenoid contents (Xu *et al.*, 2006; Wang *et al.*, 2008; Ikoma *et al.*, 2014). In the pulp vesicles of most citrus species, the accumulation of carotenoid derivatives follows a similar process, although it starts earlier and the final carotenoid content is lower (Lu *et al.*, 2017*b*; Lux *et al.*, 2019).

Besides their organoleptic properties, citrus fruits are also appreciated by their well-known health effects (Yamada *et al.*, 2011; Mulvihill *et al.*, 2016; Cirmi *et al.*, 2017), which are linked with the presence of bioactive compounds, including flavonoids. Recent studies have highlighted the vast flavonoid diversity existing in *Citrus* species and tissues, especially that of polymethoxylated flavonoids and O-glycosylated flavonoids on the fruit flavedo, where their concentration is higher (Wang *et al.*, 2017*c*; Elkhatim *et al.*, 2018). Moreover, flavonoid profiles in different citrus are extremely variable between species and admixtures, even allowing their clustering based on such profiles that greatly resembles the phylogenetic tree of the genus *Citrus* (Zhao *et al.*, 2017).

In this work, we use the RNA-seq technology to analyze global changes of the transcriptome in order to address the influence of hybridization and admixing of wild *Citrus* species in shaping gene expression, a process that gave rise to the attractive

commercial varieties we enjoy today. While citrus ripening has been thoroughly studied, most of the published works have focused on either somatic mutants or closely related cultivars (Huang *et al.*, 2016*b*; Wang *et al.*, 2017*a*; Lu *et al.*, 2018). We have performed a comprehensive transcriptomic analysis of flavedo and pulp of fruits of seven different citrus cultivars at the time of color break. Three of them are pure species and belong to the main taxonomic groups (citrons, pummelos and mandarins), while the remaining four are commercial varieties with varying admixture levels, extending previous studies of *Citrus* ripening. Using a novel approach that involves the analysis of genomic unbalance and allele-specific expression, we provide new insights of the effects of citrus hybridizations and domestication on gene expression during ripening.

MATERIALS AND METHODS

Plant material

Plant material was provided by the germplasm resources at the Instituto Valenciano de Investigaciones Agrarias (IVIA): SunChuSha Kat mandarin (*Citrus reticulata*), Chandler pummelo (*Citrus maxima*), Diamante citron (*Citrus medica*), Seville sour orange (*Citrus aurantium*), Salustiana sweet orange (*Citrus sinensis*), Willowleaf mandarin (*Citrus deliciosa*) and Eureka lemon (*Citrus limon*). Accession numbers of each genotype are shown in Table 1.

Phenotypical data collection

Fresh fruit samples were collected every three weeks from mid-September to January. Peel color was measured on field using a hand colorimeter Konica Minolta CR400. For each sample analyzed, color was measured in four different fruits performing three technical replicates on each. Fruits were then collected and processed the same day.

Fruits were squeezed and the titratable acid content of the juice was measured by titration with 0.1 M sodium hydroxide and a phenolphthalein indicator. Juice total soluble sugar content was measured in Brix degrees using a table refractometer ATAGO PR-1. Brix and acidity were analyzed on the pooled juice performing two technical replicates.

Table 1. Characteristics of studied citrus fruits.

Accession number	Species name	Binomial name	Abbreviation	Admixure origin	Final acid content	Final color
B483	SunChuSha Kat Mandarin	*Citrus reticulata*	SCM	Pure mandarin	Medium	Red
B207	Chandler Pummelo	*Citrus maxima*	CHP	Pure pummelo	Low	Yellow
B560	Diamante Citron	*Citrus medica*	DIA	Pure citron	High	Yellow
B154	Willowleaf Mandarin	*Citrus deliciosa*	WLM	Mandarin x Pummelo	Low	Red
B031	Sweet Orange	*Citrus sinensis*	SWO	Mandarin x Pummelo	Low	Red
B117	Sour Orange	*Citrus aurantium*	SSO	Mandarin x Pummelo	High	Red
B297	Eureka Lemon	*Citrus limon*	EUR	Mandarin x Pummelo x Citron	High	Yellow

RNA extraction, library preparation and sequencing

For each species, three representative samples were collected at color break. Flavedo and pulp tissues were manually separated from each fruit and treated independently. Tissues were grinded frozen and total RNA was extracted using the acid phenol extraction coupled with lithium chloride precipitation as described in Ecker 1987 (Ecker and Davis, 1987). RNA-seq library preparation and sequencing were carried out by Novogene Company. Briefly, RNA samples were enriched in mRNA using oligo (dT) beads and the mRNA was randomly fragmented. cDNAs were then synthesized from mRNA using random hexamers, followed by adapter ligation, size selection and PCR enrichment. Samples were sequenced in a NovaSeq 6000 platform, delivering 150 bp pair ended reads with an insert size of approximately 250 bp.

RNA-seq read mapping and DEG analysis

Illumina reads were mapped against the *Citrus clementina* reference genome (Wu *et al.*, 2014) using STAR 2.7.2 (Dobin *et al.*, 2013). *C. clementina* genome annotation was downloaded from the NCBI and reads mapped to each genomic feature were counted using featureCounts 2.0 (Liao *et al.*, 2014). Read counts were normalized using a variance stabilizing transformation implemented in R (Anders and Huber, 2010); these *pseudocounts* were used for the sample clustering for quality control and downstream analysis. Differential gene expression analyses were performed using the R package DESeq2 1.26 (Love *et al.*, 2014) following the author's recommendations. Pulp and flavedo data were analyzed independently, performing pairwise comparisons among every species pair, as well as pairwise comparisons of pooled samples of citron and lemon against the rest and citron, lemon and pummelo against the rest. Differentially expressed genes (DEG, log2 fold change expression > 1, false sign or smaller rate < 0.01) were detected using the model implemented in apeglm (Zhu *et al.*, 2019*a*). Genes annotated into admixed regions (Wu *et al.*, 2018) were used to assess the admixture effect in gene expression.

KEGG enrichment analysis

A GO enrichment analysis was carried out for the comparison of citron and lemon pulp against the other analyzed samples. GO enrichment was performed using the R package clusterProfiler, KEGG data was accessed using AnnotationHub (Yu *et al.*, 2012; Morgan, 2019). To account for multiple hypothesis testing, p-values were corrected using the Bonferroni-Holmes method (FDR < 0.05).

Confirmation of RNA-seq data by RT-qPCR.

To validate the RNA-seq analyses, one-step RT-qPCR of a set of genes was carried out. Reverse transcription was performed by incubating the RNA samples with the reverse transcriptase MultiScribe (Invitrogen) at 48 ºC, 30 minutes. RNAse activity was inhibited using RNAse Inhibitor (Applied Biosystems). Real-time qPCR was performed using the LightCycler FastStart DNA Master Plus SYBR Green I kit in a LightCycler 2.0 Instrument. Two technical replicates were performed for each reaction. Amplification specificity was verified by the presence of a single peak in the melting curve analysis. Oligonucleotides used for each reaction can be found in the Supplementary Table 1.

Relative quantification of the gene expression was expressed as a log 2-fold change expression compared with a housekeeping gene, *CitUBC1* (Merelo *et al.*, 2017), using the ΔΔCt method.

DNA extraction, sequencing and mapping

To find diagnostic SNPs and further validate the genomic structure of the studied genes, whole genome sequencing data was used. For the already published data, raw reads were retrieved from the Sequence Read Archive database (the SRA accession numbers are available in Supplementary Table 2). The Diamante citron genome was sequenced in this work using Illumina whole genome sequencing. In short, high molecular weight DNA was extracted using an in-house protocol. Whole genome sequencing library preparation and sequencing were carried out by the Centro Nacional de Análisis Genómico (CNAG). Briefly, libraries were constructed using the Illumina TruSeq DNA Sample Prep protocol, selecting for an insert size of 500 bp. Paired-end sequencing was performed on a HiSeq 2000 instrument.

Allele-specific expression

Allele-specific expression of different genes was studied with the following workflow: first, the two phases of every gene were established based on DNA sequencing using diagnostic heterozygous SNPs. Then, the mapped RNA-seq reads were scanned and those displaying different alleles of each heterozygous SNPs were counted independently, which allowed the expression quantification at the allele level.

To achieve this, the Illumina DNA reads (Supplementary Table 2) were mapped to the clementine reference genome using bwa mem (Li, 2013) to generate one BAM file per sample, and SNPs were called in each sample using GATK 4.1.1 HaplotypeCaller in GVCF mode (Van der Auwera *et al.*, 2013). SNPs were hard-filtered following GATK best practices, and only SNPs showing a genotype quality (GQ) over 20 were selected.

Since pure species in *Citrus* show low heterozygosity, the two pseudophases could not be established based on phased SNPs. For admixed varieties, the admixture pattern of each genomic region was retrieved from previous works (Wu *et al.*, 2018). Then, for each gene within an admixed region, heterozygous SNPs were selected. Among them, diagnostic SNPs were defined as those sharing each allele with a pure species contributing to the admixture, being that species homozygous in that specific position. For example, an A/T position in sour orange would only be considered diagnostic if pummelo was A/A and mandarin was T/T for that position. Diagnostic SNPs were used for allelic phasing as previously described (Wu *et al.*, 2014, 2018) in order to obtain one phase for each pure species.

The total number of RNA-seq reads sequenced from each phase, both in pure species and admixtures, was assessed using the samjdk utility of the jvarkit toolset (Lindenbaum and Redon, 2018), merging the counts of the three RNA-seq biological replicates together.

Detection of runs of homozygosity

The distribution of highly homozygous regions in the genome was studied to assess the prevalence on inbreeding in the palatable admixtures. To do so, all the heterozygous SNPs along the complete genome were retrieved from the two genomes, using the SNP set generated in the above section. The reference genome was split into non-overlapping 200

kb windows and those displaying a heterozygosity below 0.1% (less than 1 SNP per kb) were considered as runs of homozygosity.

Analysis of the *chalcone synthase* promoter region

The upstream region of a *chalcone synthase* (*CHS*) LOC18042808 was extracted from the *C. clementina* reference genome based on its genomic coordinates. The ortholog regions in other reference genomes of *Citrus* and related genera were obtained by similarity search using BLASTN 2.7.1 (Camacho *et al.*, 2009), querying published reference genomes (Xu *et al.*, 2013; Wu *et al.*, 2014; Shimizu *et al.*, 2017; Wang *et al.*, 2017*b*, 2018*a*; Zhu *et al.*, 2019*b*; Peng *et al.*, 2020). The genomic structure of the region was also manually curated with the Integrative Genome Viewer (IGV) browser software (Robinson *et al.*, 2011) with the same DNA-based alignments used for phasing.

The promoter region was amplified in pure mandarin, citron and pummelo by conventional PCR using the forward and reverse primers described in Supplementary Table 1. Each PCR product was sequenced using Sanger sequencing, and the sequences obtained were aligned using MAFFT (Katoh and Standley, 2013). After manual curation of the Sanger sequencing and comparison with the reference genomes, DNA conserved motifs were searched using the online tool PlantCARE (Lescot *et al.*, 2002).

RESULTS

Physiological characterization of ripening

Fruit acidity and sugar content of the seven selected species were measured during the ripening process. Major changes in juice acidity were only observed in the two mandarins, whose acidity decreased considerably as ripening progressed. Citron, lemon and sour orange presented a titratable acid (TA) content above 5% during the whole period; conversely, sweet orange and pummelo showed a constant, low level of TA content (Figure 1a). Sugar content (measured in Brix degrees, or ºBrix) increased considerably in the two mandarins, while it remained invariably high in pummelo, very low in lemon and citron, and in intermediate values in the two oranges (Figure 1b). The color break and the fruit color at the time of sampling is shown in Figure 1c.

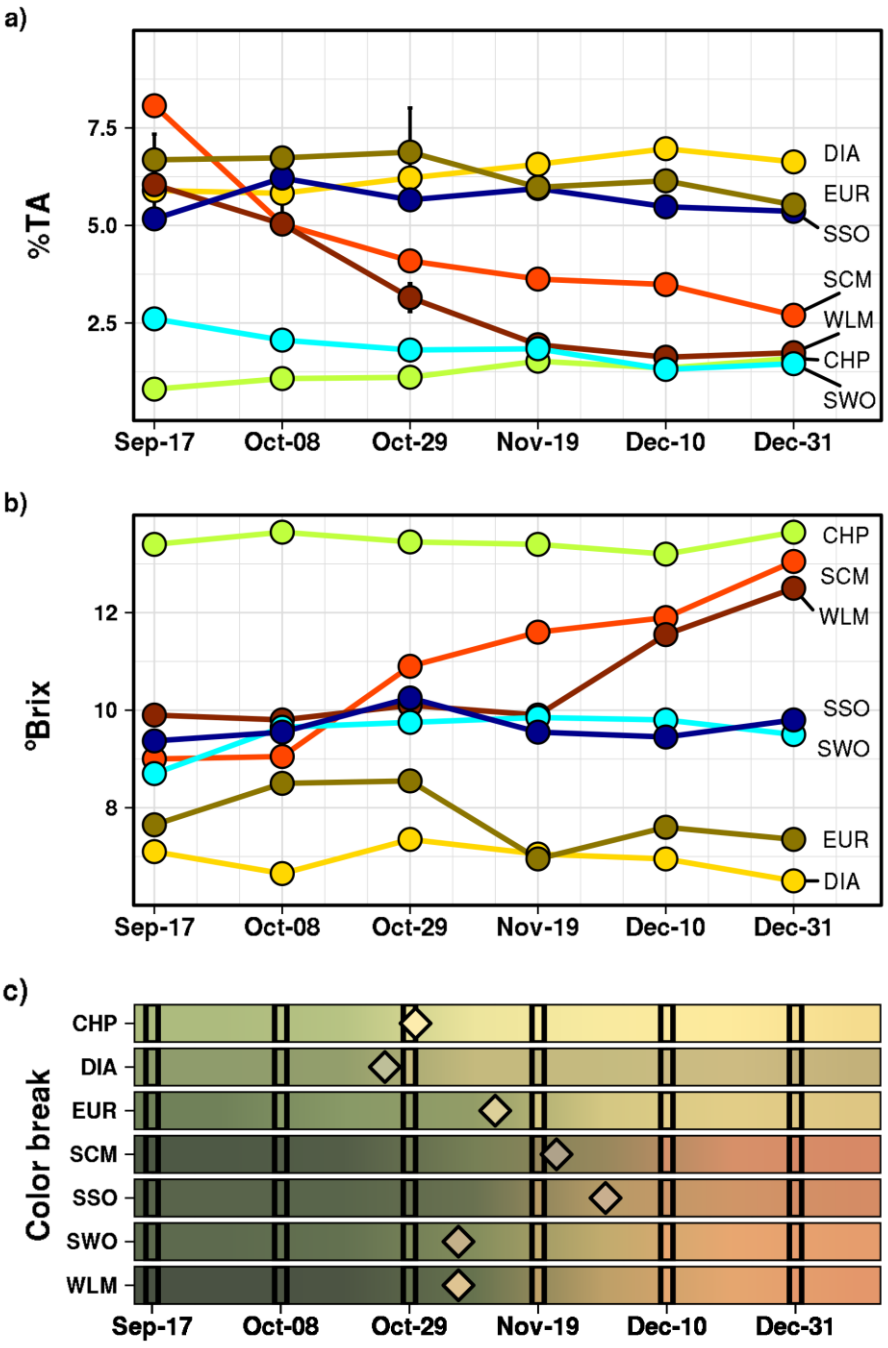


Figure 1. Phenotyping of the selected accessions across three months in six different timepoints. a) Average titratable acid content per sample across three replicates. Vertical bars represent the standard error for each measurement. b) Average Brix degrees per sample. c) Average color per sample at six measuring times, indicated by the sampling date. Color was calculated from the L*a*b values provided by the Minolta colorimeter, applying a correcting factor of 1.2 to the luminosity value L. The color values between measurements were interpolated. Diamonds mark the color and date at which samples were collected and processed for RNA-seq sequencing. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

RNA-seq read mapping and sample clustering

An average of 22 million reads per sample were obtained, 80% of which mapped to the 26944 annotated features of the reference genome.

To analyze the variability of the samples, a principal component analysis was performed based on transformed read counts (see methods). As expected, the three replicates from each sample clustered together in all cases, supporting the reproducibility of the results (Figure 2). Pulp and flavedo samples formed two completely independent clusters regardless of the species considered and consequently, samples from each tissue were treated independently in all the analyses performed. Pure species were the most different samples in both tissues, as indicated by their dispersion in the PCA plot. In contrast, admixture samples like sweet and sour oranges distributed in between their ancestral pure species (Figure 2).

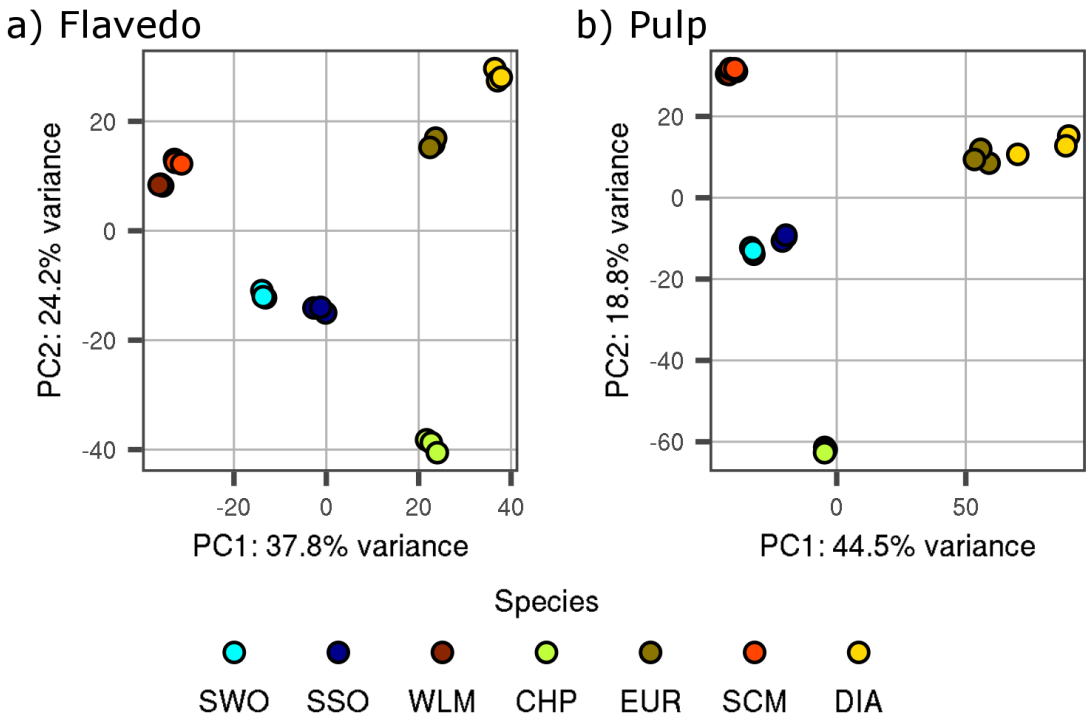


Figure 2: Principal component analysis based on expression data. Read counts for each gene were normalized using a variance stabilizing transformation (see Methods). Each replicate is shown as an independent dot colored by sample. Flavedo and pulp samples were analyzed independently and are displayed in a) and b), respectively. Only the principal components PC1 and PC2 are displayed in each case. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

Hierarchical clustering of pulp samples grouped together lemon and citron apart from the other species. A second cluster only included pummelo, and the third one grouped the two mandarins (pure and commercial), and the two oranges (sour and sweet). With the flavedo data, samples formed two main clusters depending on their peel coloration, separating species with red and yellow fruits into two clusters (Supplementary Figure 1).

Differentially expressed genes during fruit ripening

In order to find relevant differentially expressed genes, an all versus all strategy was used, so genic expression was compared between every species pair. The number of differentially expressed genes (at least two-fold expression change, s-value < 0.01) varied between 574 and 7773 among the pairwise comparisons. Roughly 20% of the DEGs corresponded to uncharacterized loci and genes of unknown function, matching the uncharacterized proportion of the total genic space of *C. clementina*. Four differentially expressed genes and a housekeeping one were selected and qRT-PCR analyses were carried out to validate the differential expression analysis: overall, the log fold changes obtained from the RNA-seq data matched those calculated via the ΔΔCt method (Supplementary Table 3), validating the RNA-seq results.

Based on the NCBI annotation, we selected genes associated with several metabolic pathways involved in citrus ripening: glycolysis and tricarboxylic acid cycle, vacuolar proton intake and secondary metabolism including carotenoid and flavonoid biosynthesis. Differentially expressed genes belonging to any of these pathways were selected for further analyses.

Citrus fruit flavor depends on citrate and sugar accumulation and the vacuolar pH of the pulp. Notably, many genes involved in the glycolysis and the tricarboxylic acid (TCA) cycle from lemon and citron pulp showed a consistently differential expression pattern (Supplementary Figure 2 and Supplementary Figure 3). This way, most of the genes involved in the glycolytic process displayed a lower expression level in lemon and citron when compared to the remaining samples, including those genes involved in the TCA cycle and the γ-aminobutyric acid (GABA) cycle (Figure 3). In contrast, most subunits of the vacuolar ATPase were overexpressed in lemon and citron (Figure 3).

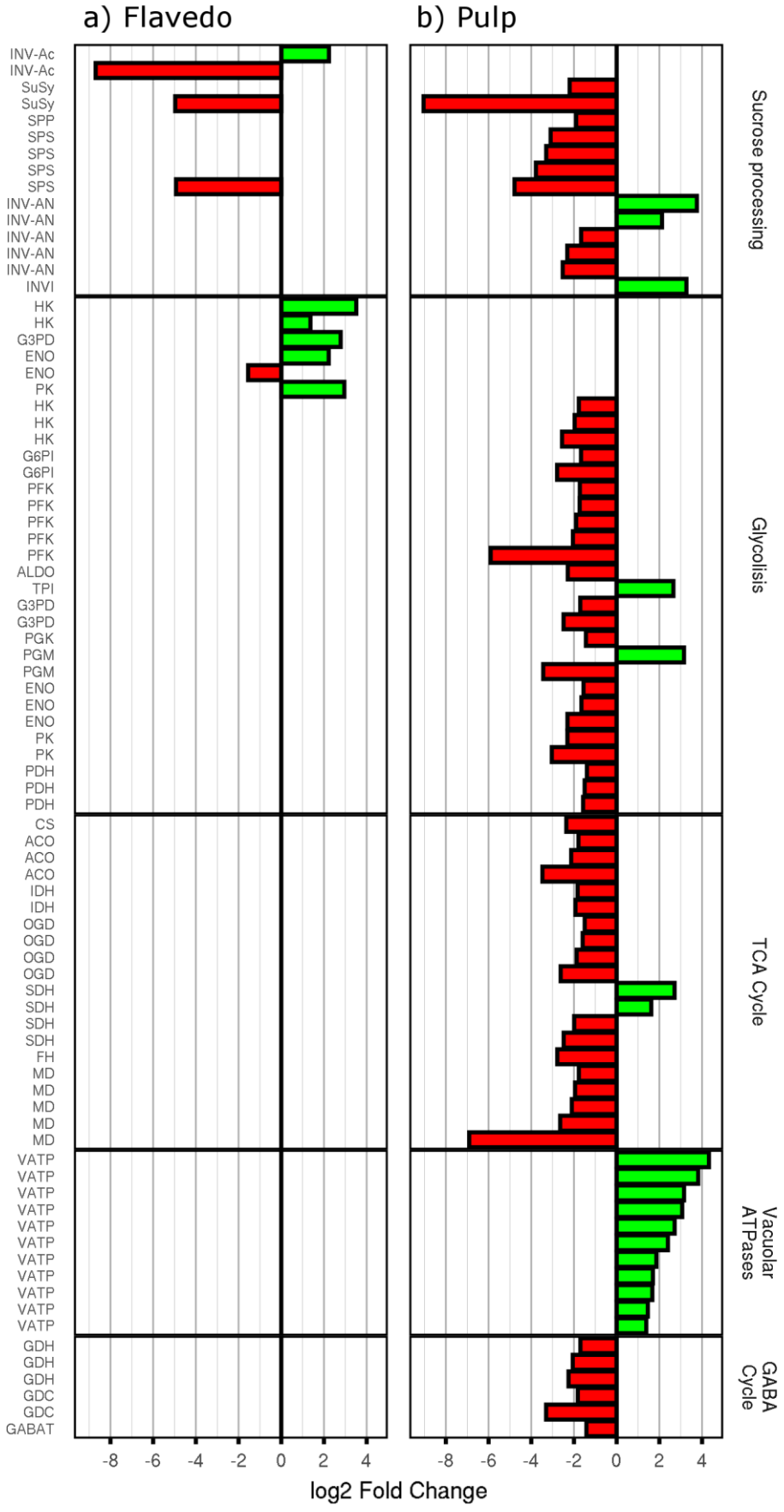

a) Flavedo
b) Pulp
INV-Ac
INV-Ac
SuSy
SuSy
SPP
SPS
SPS
SPS
SPS
INV-AN
INV-AN
INV-AN
INV-AN
INV-AN
INVI
HK
HK
G3PD
ENO
ENO
PK
HK
HK
HK
G6PI
G6PI
PFK
PFK
PFK
PFK
PFK
ALDO
TPI
G3PD
G3PD
PGK
PGM
PGM
ENO
ENO
ENO
PK
PK
PDH
PDH
PDH
CS
ACO
ACO
ACO
IDH
IDH
OGD
OGD
OGD
OGD
SDH
SDH
SDH
SDH
FH
MD
MD
MD
MD
MD
VATP
VATP
VATP
VATP
VATP
VATP
VATP
VATP
VATP
VATP
VATP
GDH
GDH
GDH
GDC
GDC
GABAT
-8
-6
-4
-2
0
2
4
log2 Fold Change
Sucrose processing
Glycolisis
TCA Cycle
Vacuolar ATPases
GABA Cycle

Figure 3: Differentially expressed genes involved in sugar processing and glycolytic activities. Flavedo (a) and pulp (b) are shown independently. The fold changes are obtained from comparing lemon and citron against the remaining five samples. Only genes with a statistically significant differential expression (LFC > 1, s-value < 0.01) are displayed. Equivalences for the gene names are displayed in the Supplementary Table 4.

We could not find these patterns in the flavedo samples, where the number of DEGs was considerably lower when compared to the pulp ones (Supplementary Figure 2, and Supplementary Figure 3). Functional annotation and enrichment analysis of the DEGs found in citron and lemon pulp samples showed an enrichment in terms related with organic acid metabolism and ion transport via ATP hydrolysis (Supplementary Table 5).

When the expression profile of the genes involved in carotenogenesis was analyzed in detail, a different landscape was observed, as the clustering of the pulp and flavedo samples produced groups that did not correlate with the fruit color (Supplementary Figure 4). Furthermore, when a two-way comparison of red (sweet and sour orange, wild and commercial mandarins) against yellow (lemon, citron and pummelo) fruits was performed, only a few genes appeared differentially expressed (Figure 4). Nevertheless,

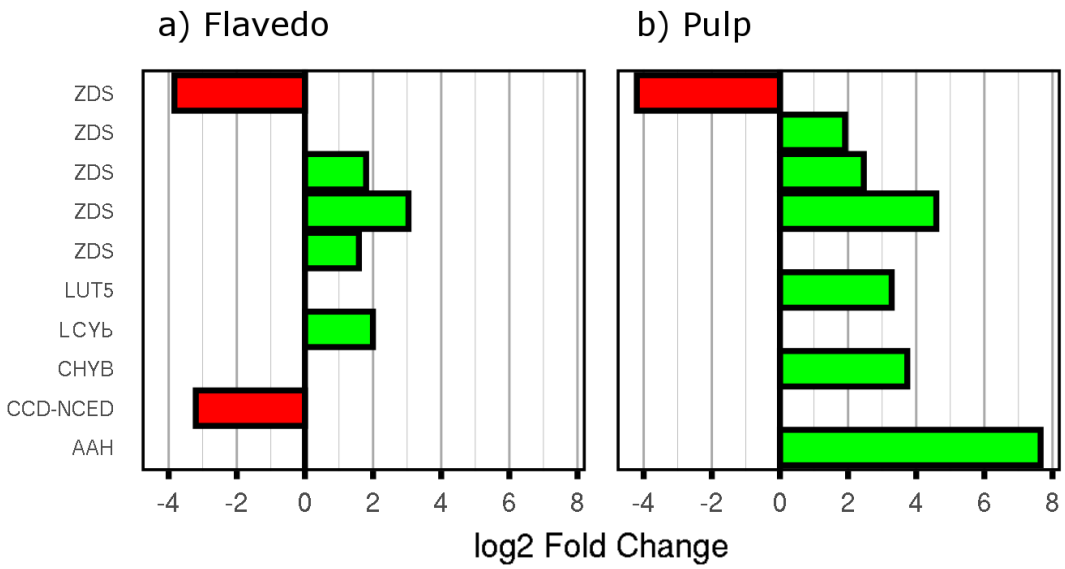


Figure 4: Differentially expressed genes involved in carotenoid biosynthesis Flavedo (a) and pulp (b) are shown independently. The fold changes are obtained from comparing red (pure mandarin, commercial mandarin, sweet orange and sour orange) against yellow cultivars. Only genes with a statistically significant differential expression (LFC > 1, s-value < 0.01) are displayed. Equivalences for the gene names are displayed in the Supplementary Table 4.

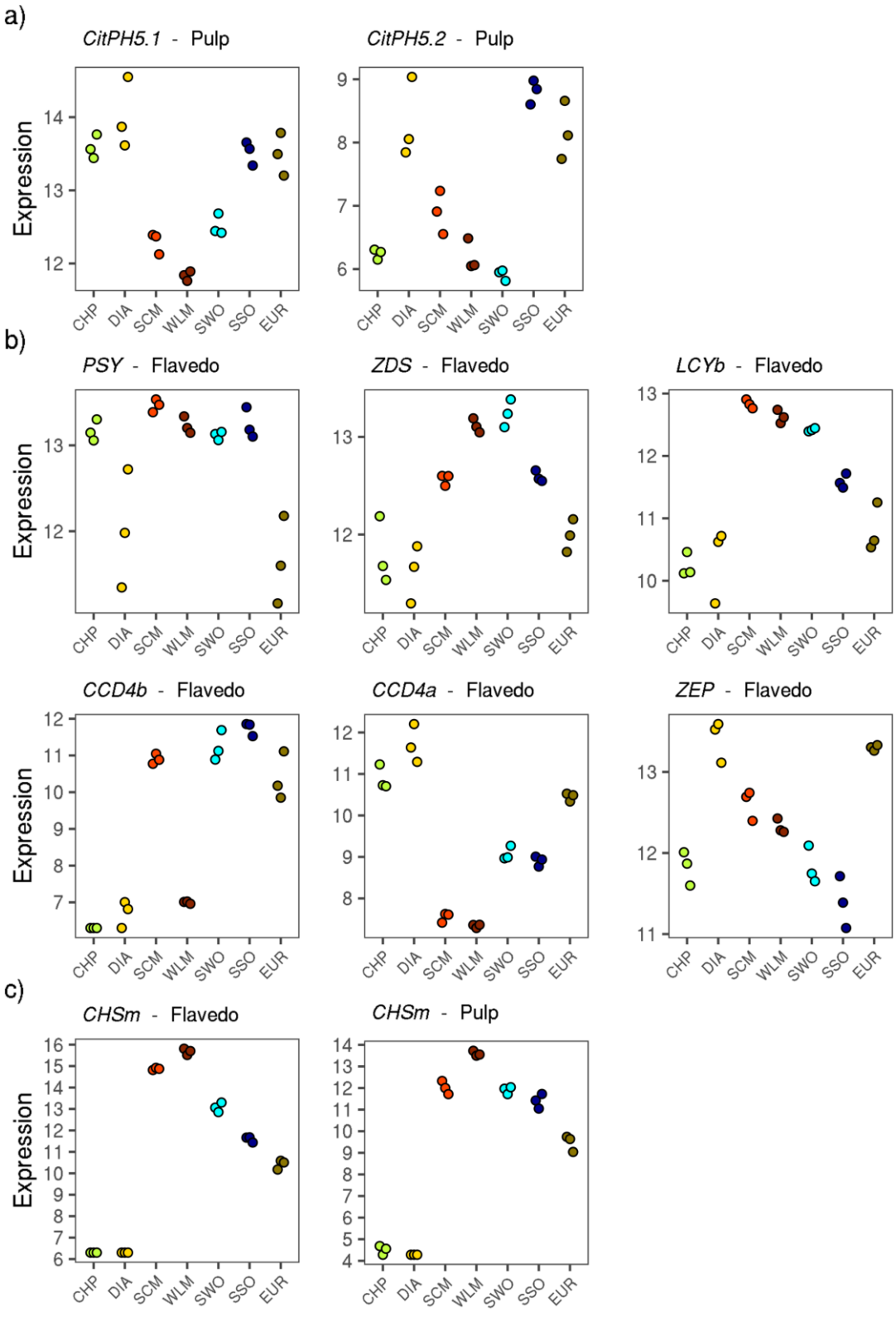


Figure 5: Expression levels of ten genes based on normalized counts. The selected genes are involved in a) pulp acidity, b) carotenoid accumulation, or c) flavonoid accumulation. Each color represents a different sample as in Figure PCA. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

comparisons between sample pairs yielded several differentially expressed genes, although they were not consistent across all the samples with the same fruit color (Figure 5b). One of these genes coded for a *phytoene synthase* (*PSY* LOC18039146), one of the first enzymes in the carotenogenic pathway, which was significantly downregulated in citron and lemon. The citrus *zeta-carotene desaturase* gene (*ZDS* LOC112098231) was more expressed in red fruits than in yellow ones, although not all comparisons showed significant differences. Another gene coding for a key enzyme involved in the synthesis of *β-β*-carotenes, a *β-lycopene cyclase* (*LCYb2* LOC18034834), was overexpressed in the red fruits when compared to the yellow ones. Several genes coding for carotenoid cleavage dioxygenases (*CCDs*), involved in carotenoid accumulation and color setting, were also differentially expressed. One of them, *CCD4a* (LOC18043465), displayed lower expression levels in mandarins and oranges, while *CCD4b* (LOC18043103) was more expressed in sweet and sour oranges, lemon and wild mandarin. The *zeaxanthin epoxidas*e gene (*ZEP* LOC18036737), involved in carotenoid degradation, was overexpressed in lemon and citron. It must be noted that these comparisons were performed on flavedo samples since color break takes place earlier in the pulp, although pulp samples presented similar expression patterns for most of these genes.

Finally, the flavonoid biosynthetic pathway was studied in order to address the variability of flavonoid derivatives found in *Citrus* flavedo (Wang *et al.*, 2017*c*). The diversity of flavonoid compounds correlates with the extreme diversity of expression patterns observed in a large number of genes involved in flavonoid modification, especially the flavonoid and phenylpropanoid O-methyltransferases (FOMTs) families. Many of these genes had no expression in at least one species, while showed considerably high expression levels in others (Figure 6). The most noticeable differential expression was displayed by a *chalcone synthase* gene (*CHS* LOC18042808) that was exclusively expressed at high levels in mandarin and their admixed species, lemon, sweet orange, sour orange and commercial mandarin (Figure 5c). We will refer to this locus as *CHSm* due to their mandarin-linked expression, which was confirmed in both pulp and flavedo samples.

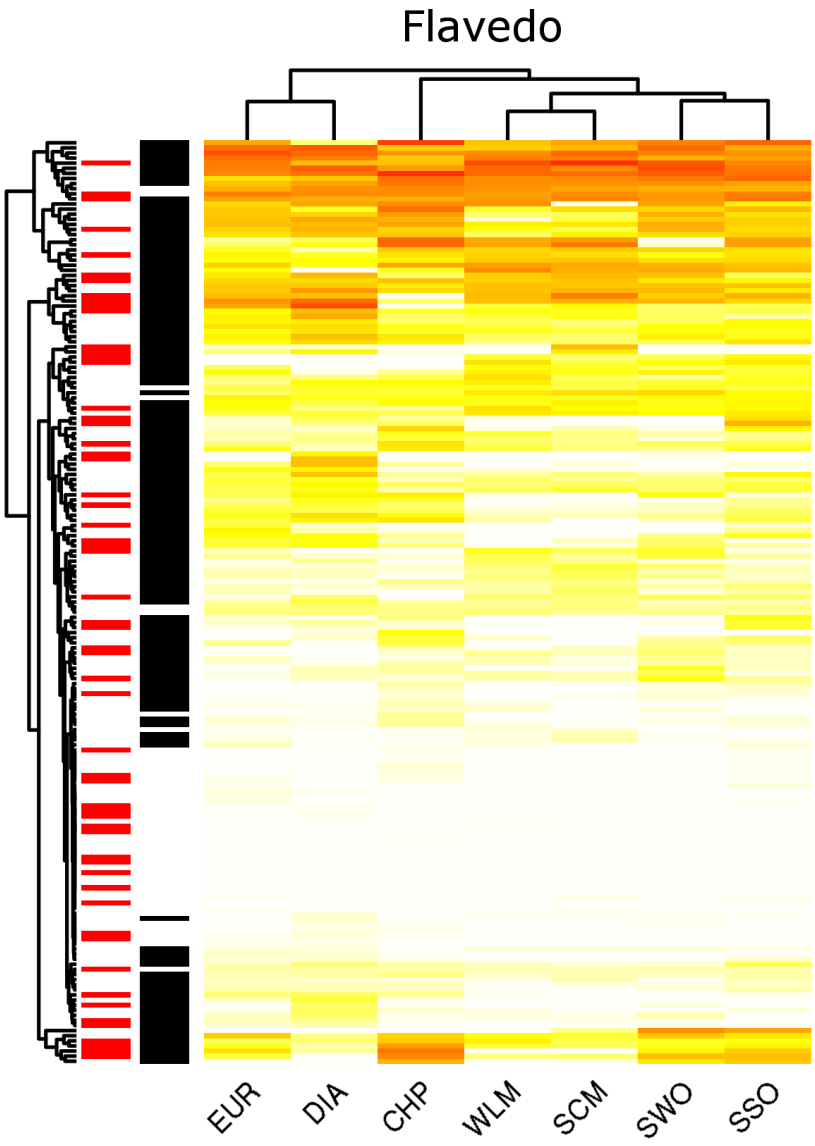


Figure 6: Flavonoid-related gene expression across samples. Heatmap representing the expression levels of genes involved in flavonoid modifications in flavedo tissues per sample and per gene. Color intensity represent expression levels based on normalized read counts. Black rectangles mark differentially expressed genes between at least two samples. Red rectangles denote flavonoid O-methyltransferases**.** Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

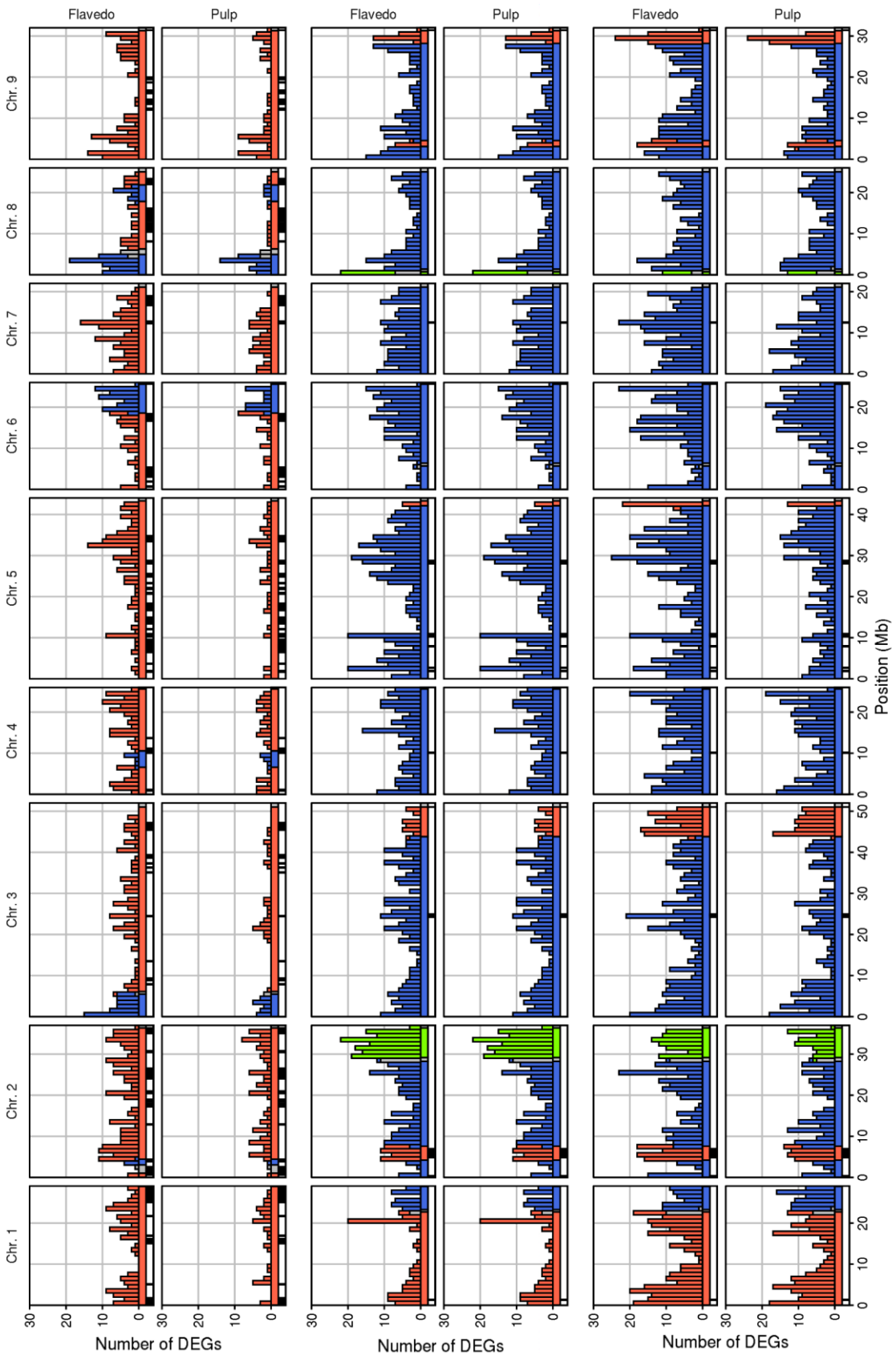

Flavedo
Pulp
Flavedo
Pulp
Flavedo
Pulp
Chr. 9
Chr. 8
Chr. 7
Chr. 6
Chr. 5
Chr. 4
Chr. 3
Chr. 2
Chr. 1
Position (Mb)
Number of DEGs
Number of DEGs
Number of DEGs

Figure 7: Distribution of differentially expressed genes along the genome of domesticated citrus cultivars. The total number of DEGs are represented along the genome for the comparison of wild and domesticated mandarin (a), and sweet orange with wild mandarin (b) and pummelo (c). The admixture patterns for each species is shown below each plot: red for mandarin/mandarin (M/M) regions, blue for pummelo/mandarin (P/M) regions and green for pummelo/pummelo (P/P) regions. Runs of homozygosity are shown below the admixture patterns as black rectangles. Abbreviations: Chr: chromosome, DEGs: differentially expressed genes.

Segmental ancestry patterns and differential gene expression

To assess the effects of the different introgressions in the global gene expression and domestication, we compared the number of DEGs between the two palatable admixtures, sweet orange and commercial mandarin, and the pure species that originated them, the pummelo and the wild mandarin. Using the admixed regions provided by Wu *et al*. (2018) we assigned an admixture pattern to each individual gene and compared the distribution of DEGs across the genome.

In the case of the commercial mandarin, the number of DEGs against wild mandarin on the pummelo/mandarin admixed regions was considerably higher than in the rest of the genome, an observation that was not held for the highly homozygous regions of the commercial mandarin genome. This was especially notorious in the case of the end of the chromosome 6 and the beginning of the chromosomes 3 and 8 (Figure 7a).

In sweet orange, two independent comparisons were made, one against wild mandarin and the other against pummelo. The number of DEGs in the admixed regions was not especially different in each comparison, but those of the pure regions were considerably different. Specifically, the major pummelo/pummelo region of the sweet orange genome, located at the end of the chromosome 2, was amongst the regions with a higher amount of DEGs against mandarin (Figure 7b), while it was not particularly enriched when compared with pummelo. The mandarin/mandarin regions of the sweet orange genome, especially those at the end of the chromosomes 3 and 9, displayed a large amount of DEGs against pummelo, while showing considerably less when compared with mandarin (Figure 7c).

The number of DEGs was also studied across the existing runs of homozygosity of the commercial mandarin and sweet orange genomes. The runs of homozygosity, here

defined as genomic windows with a heterozygosity below 0.1%, were scarce in sweet orange, with only 1.6% of its genome being composed by runs of homozygosity. In the case of the commercial mandarin, this percentage ascended up to 16.5%. In none of the two species DEGs were particularly enriched in the runs of homozygosity.

Allele-specific expression in *Citrus* ripening genes

In pure species the allele imbalance in RNA-seq reads was based on the number of reference and alternative reads in the heterozygous sites. A total of 261845, 243454 and 109258 heterozygous SNPs were found in pummelo, wild mandarin and citron, respectively. In the admixed species, allele expression imbalance was calculated having into account the species-specific phasing for each genomic region. This strategy can properly phase the two alleles in admixed species, but only in the genomic regions displaying the 2 ancestral haplotypes, called from now on admixed regions. As the extent of admixed regions differs greatly between species, the total number of SNPs was not directly comparable. Therefore, the number of diagnostic SNPs per admixed kb was estimated, allowing comparisons between all species, that ranging from 6 to 10 SNPs per kb of admixed region (Table 2). Then, genes displaying allele-specific expression (ASE) were detected counting the frequency of reads carrying these diagnostic SNPs. The number of genes displaying ASE in any of the two analyzed tissues ranged from 363 in commercial mandarin to 3952 in lemon. It has to be taken into account that the admixed regions range from a mere 10% of the genome in commercial mandarin to the whole genome in lemon and sour orange. Although the number of ASE genes was roughly comparable among cultivars when considering their admixture proportions, an exception was found in lemon. The number ASE genes in the citron/mandarin admixed regions of lemon, which represent around 50% of the genomic space, was as high as that found in the pummelo/mandarin regions in sour orange, which account for the whole genome of this species (Figure 8). This was especially pronounced in pulp samples, which were found to harbor more ASE genes than the flavedo ones in all the analyzed species. In the lemon pulp about 10% of the total genic content (2734 genes) displayed ASE, while it was around 6.5% in the case of sour orange (1772 genes), and the other two admixtures, where similar values were found. Some notable examples showing allele-specific differential expression were found. For example, the *phytoene synthase PSY* gene

Table 2. Diagnostic SNPs for allele-specific expression analyses

Sample	Type of diagnostic SNP	Number of diagnostic SNPs	Total admx region	Genic admx	SNPs/kb (admx)	SNPs/kb (genic admx)
Sweet Orange	Alternate fixed between pummelo and mandarin	421474	238273500	68024255	1,7689	6,1959
Sour Orange	Alternate fixed between pummelo and mandarin	560651	287956647	85816733	1,9470	6,5331
Willowleaf Mandarin	Alternate fixed between pummelo and mandarin	68602	26835800	10383511	2,5564	6,6068
Eureka Lemon	Alternate fixed between pummelo and citron	236523	124636400	33940319	1,8977	6,9688
Eureka Lemon	Alternate fixed between mandarin and citron	506678	158743347	50615776	3,1918	10,0103

(LOC18039146) preferentially expressed the mandarin allele in sweet orange and commercial mandarin. The *hexokinase* gene (*HK*, LOC18035909) expressed preferentially the citron allele in lemon, while in sweet orange, the pummelo one was prevalent. Most remarkably, the mandarin exclusive *chalcone synthase CHSm* gene only expressed the mandarin allele in lemon, sweet orange and sour orange (Supplementary Table 6).

Promoter structure of the chalcone synthase *CHSm*

The allele-specific expression analysis showed that *CHSm* was not expressed in citron and pummelo, and that only the mandarin allele was expressed in the other species. In order to understand the cause of the lack of expression of *CHSm* in citron and pummelo, the promoter region of *CHSm* was studied. A sequence of 750 bp upstream of the *CHS* locus was retrieved from the *C. clementina* reference genome, and orthologs in other available citrus genome sequences (Xu *et al.*, 2013; Wu *et al.*, 2014; Shimizu *et al.*, 2017; Wang *et al.*, 2017*b*, 2018*a*; Zhu *et al.*, 2019*b*; Peng *et al.*, 2020) were identified by

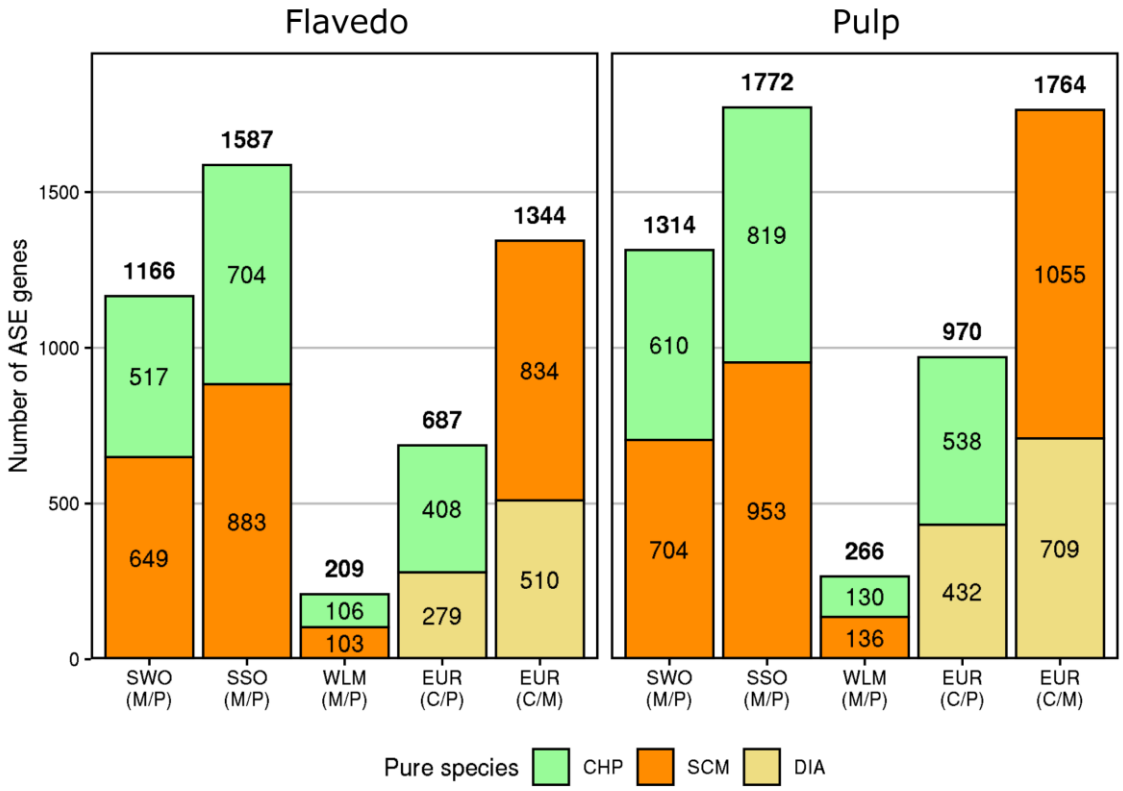


Figure 8: Distribution of allele-specific expressed genes across tissues and cultivars. The number of ASE genes preferentially expressing the citron, pummelo and mandarin alleles is shown for each sample, split in admixed regions: M/P (mandarin/pummelo), C/P (citron/pummelo) and C/M (citron/mandarin). Since for each admixed region only two haplotypes exist, the number of genes overexpressing one of the alleles coincides with those underexpressing the alternate allele. The total number of ASE genes is shown on top of each bar. Flavedo and pulp samples are shown independently in a) and b), respectively. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

similarity. Besides pummelo, citron and mandarin, wild mandarin (*C. reticulata*), satsuma mandarin (*Citrus unshiu*), sweet orange (*C. sinensis*), Ichang papeda (*Citrus ichangensis*), the Hong Kong kumquat (*Fortunella hindsii*) and two *Citrus* outgroups, *Poncirus trifoliata* and *Severinia buxifolia*, were included in the study. The alignment of the matching regions of the 10 genomes revealed three different alleles 70 bp upstream of the transcription start site: the archaic allele, present solely in the outgroup species, the pummelo allele, present in citron, pummelo and their admixtures, and the mandarin allele, present in mandarins, kumquats and the Ichang papeda (Figure 9). It is noteworthy that

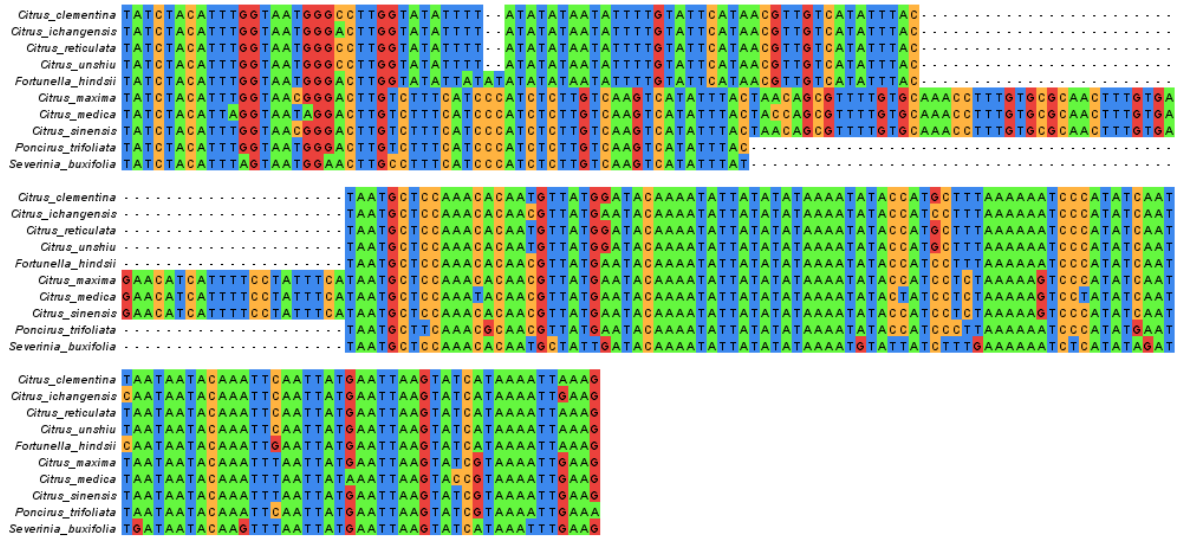


Figure 9: *CHSm* promoter sequence. Promoter region of the *CHSm* locus was retrieved from the 10 publicly available reference genomes of *Citrus* and related genera. The varying region that differentiates the three *CHSm* alleles is marked in red. The archaic, citron/pummelo and mandarin alleles are ordered from bottom to top.

the archaic and the pummelo alleles are more similar to each other, although it is longer in pummelo. In contrast, the mandarin allele is completely different, with a very low GC content, down to 16%, compared with the average 34% of the *Citrus* genomes. These sequences were searched for specific regulatory motifs, and an ethylene responsive element (ERE) was found exclusively in those species carrying the pummelo allele: citron, pummelo and sweet orange (Figure 10a). This region was Sanger sequenced in citron, pummelo and mandarin, confirming the presence of the citron/pummelo and mandarin alleles (see Supplementary Table 1 for primer details).

An *in silico* study of the presence or absence of these alleles was performed, using DNA sequencing reads from the different genomes in the IGV browser. It was found that the citron/pummelo allele was homozygous in these two species; that lemon, sour orange and sweet orange were hemizygous, and that the commercial mandarin was homozygous for the mandarin allele (Figure 10b).

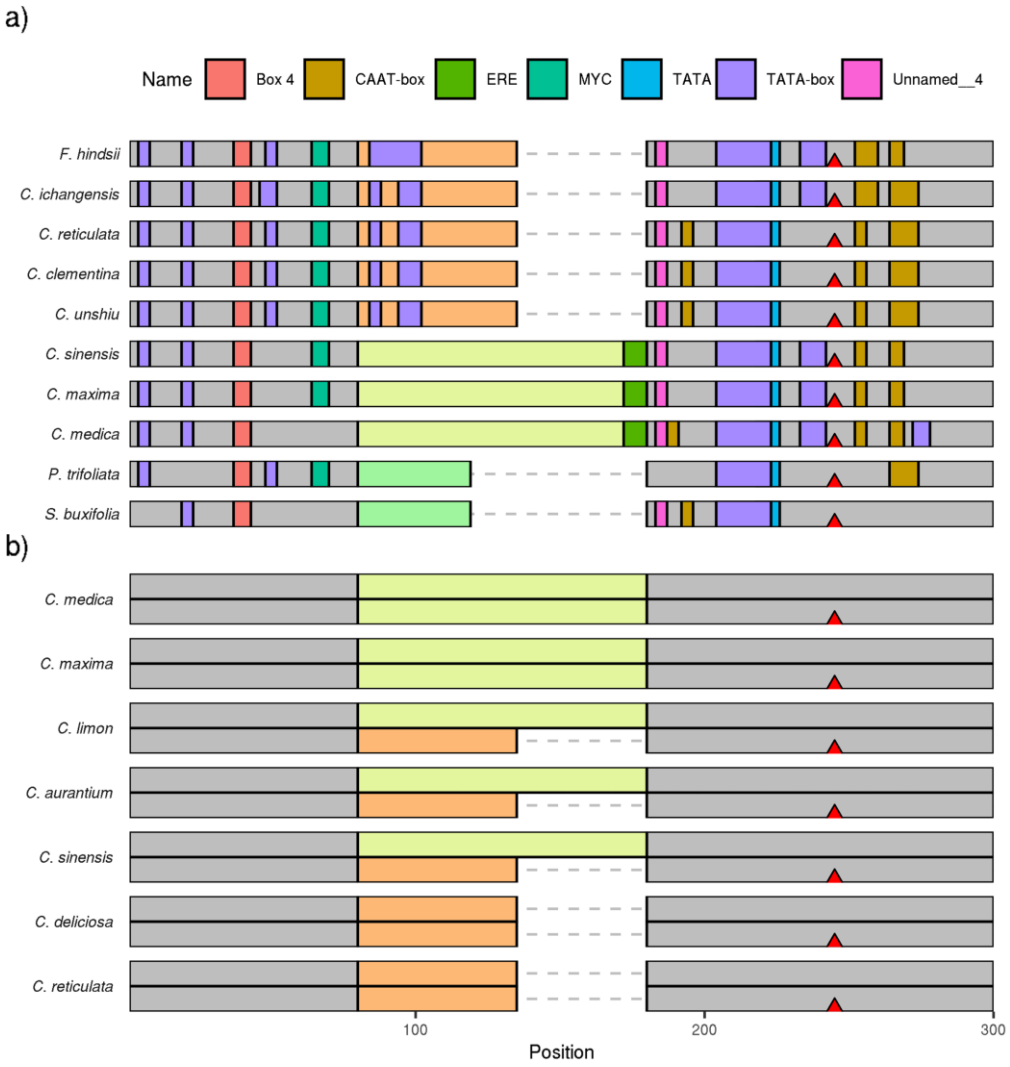


Figure 10: Genomic structure of *CHSm* across *Citrus* and close genera. a) Motif distribution across 10 reference genomes of Citrus and close relatives. In grey, the invariable regions of the promoter. In the middle, the three alleles are represented in different colors: orange for the mandarin allele, yellow for the pummelo allele and green for the ancestral allele. The transcription start site, as annotated in the *C. clementina* reference genome, is shown as a red triangle. b) Allele distribution across the seven genomes studied by RNA-seq. Motifs are not depicted for clarity. The structure has been inferred from whole genome sequencing data via hand curation. The two alleles for each sample are depicted to allow for the representation of admixed individuals.

DISCUSSION

Admixture patterns and their role in shaping gene expression

Recent genomic studies of commercial citrus species have shown extensive introgressions in their genomes from three pure species: citrons, pummelos and mandarins, so that principal component analysis (PCA) based on sequence variability depicts the three pure species in the vertices of a triangle, with the admixtures scattered somewhere in the middle (Wu *et al.*, 2018). A phenotype-based classification using 146 citrus botanical traits produced a similar figure (Barrett and Rhodes, 1976). In our study, the PCA based on gene expression generated a highly similar distribution (Figure 2). Although these results could be expected, it is worth noting the completely different types of the data giving rise to the same distribution, which might link the genomic, transcriptomic and phenotypical levels.

At the genomic level, the differences among the studied commercial varieties are caused mostly by their specific admixture patterns and the evolutionary history of the pure species. The different genome structures were generated during the interspecific crosses of *Citrus*, that possibly took place during the domestication of the genus (Wu *et al.*, 2018), and would have been maintained since by vegetative propagation.

In plant interspecific hybrids, gene expression can reach extreme values when compared with both parentals, in a process called transgressive gene expression (Dickinson *et al.*, 2003). This phenomenon is explained by inter-loci epistatic relations and complementarity between loci, among others (Mao *et al.*, 2011), and can sometimes result in improved phenotypes compared with the parental species (Zhou *et al.*, 2012; Kitazumi *et al.*, 2018). However, our results show that, in *Citrus*, admixed transcriptomes show expression levels that in general are an average of the ones from their ancestor pure species. While specific genes or traits present transgressive segregation, our results suggest that gene expression levels can be usually explained by those of the specific ancestors of each admixture.

Hybrid crops can also display ASE in specific genes (Shao *et al.*, 2019; Cai *et al.*, 2020), which can increase the fitness of hybrid species by granting a higher genomic plasticity (Botet and Keurentjes, 2020). Many possible causes for allele-specific expression have

been suggested so far, with transposon insertions within specific genes being one of them (Zhang *et al.*, 2020). Taking into account how *Citrus* pure species display considerable differences in terms of transposable element number and activity (Borredá *et al.*, 2019; Liu *et al.*, 2019), and the interspecific nature of commercial *Citrus* cultivars, we analyzed possible effects of domestication by comparing gene expression in their admixed regions compared with pure species.

We first studied the abundance of genes displaying ASE across the four admixed cultivars. Allele-specific expression was more prevalent in all the pulp samples when compared with the flavedo ones (Figure 8). In the lemon pulp the amount of ASE genes represented about 10% of the total gene number, dropping to 6.5% in the other analyzed samples. This value is comparable with the results found in hybrid rice, in which nearly 6% of the genic space displays some sort of ASE (Shao *et al.*, 2019). However, other crops display higher proportions of ASE genes, such as tomato (Albert *et al.*, 2018) or maize (Springer and Stupar, 2007), reaching 20% and 50% of the total number of genes. Overall, our data indicate that ASE is not very common in *Citrus*, suggesting that the phenotypic differences between cultivars might be better explained by quantitative changes in the expression levels.

To this end, we studied the distribution of differentially expressed genes along the genomes of the two palatable admixtures included in this study, the commercial mandarin and the sweet orange. DEGs between the commercial and the pure mandarin are considerably more concentrated in the admixed regions of the commercial mandarin genome, while the runs of homozygosity of the genome do not show any significant increase (Figure 7). Interestingly, the start of chromosome 8 showed one of the highest accumulations of DEGs in both the pulp and the flavedo samples. A previous study proposed this region as a major domestication target for mandarins, based on the high prevalence on this introgression in commercial mandarins, as well as its significant association with fruit quality traits found in a genome wide association study (Wu *et al.*, 2018). Although the authors suggested an *isocitrate dehydrogenase* gene located on that region as a putative determinant of fruit acidity, we could not find differential expression of this gene between any two species, suggesting that other genes on this region might be involved in the domestication process. Notably, a previous association study also linked this region with fruit weight (Minamikawa *et al.*, 2017). A correlation between fruit size

and amount of pummelo introgression exist in *Citrus* (Wu *et al.*, 2018), which might explain its significance in the association study of Minamikawa *et al.* (2017) and might be linked to the large amount of genes that are differentially expressed in the domesticated mandarin with respect the wild one in this admixed region. Similarly, the beginning of chromosome 3, which is significantly associated with fruit hardness and easy peeling (Minamikawa *et al.*, 2017), also displayed a high amount of DEGs between wild and commercial mandarins, especially in the flavedo.

We also compared the number of DEGs between sweet orange and its two ancestors, pure mandarin and pummelo. The presence of pummelo/pummelo (P/P) and mandarin/mandarin (M/M) regions along the sweet orange genome allows to study the differences in gene expression between P/P and M/M regions in different genomic backgrounds: a pure pummelo, a pure mandarin and an admixed genome such as sweet orange. Notably, we found that a large number of genes at the P/P region towards the end of chromosome 2 are differentially expressed compared with the wild mandarin, but not with pummelo (Figure 7). The reverse observation was made in some but not all the M/M regions, as for example the end of chromosomes 3 and 9, where many more genes are differentially expressed compared with pummelo than with mandarin (Figure 7). It is worth noting that the end of chromosomes 2 and 3 were also significantly associated to fruit weight in the GWAS mentioned above (Minamikawa *et al.*, 2017).

Several studies have shown that the domestication process causes the depletion of genetic diversity in annual crops like rice, wheat or barley (Civáň *et al.*, 2015; Pankin *et al.*, 2018; Maccaferri *et al.*, 2019), and even in fruit tree species like peach (Cao *et al.*, 2014). The loss of diversity is generally explained by a population bottleneck, generated by the initial selection of a reduced number of wild individuals followed by recurrent inbreeding of elite cultivars, although new evidences revealed that these processes are not universal (Meyer and Purugganan, 2013; Smith *et al.*, 2019). In the case of *Citrus*, previous studies suggest that wild mandarin populations went through at least two bottlenecks, that took place 1 million and 100000 years ago (Wang *et al.*, 2018*a*), and therefore would be unrelated to human domestication as they occurred much before the development of agriculture in the Neolithic period, about 10000 years ago. The existence of several runs of homozygosity in some *Citrus* species revealed a considerable degree of inbreeding in

the mandarin germplasm, suggesting multiple events of hybridization and selection during the domestication of wild mandarins (Wu *et al.*, 2018).

Our results in commercial mandarins show that the largest differences at the transcriptomic level compared with the wild ones are more frequent in genes located on the small fraction of P/M admixed regions of the genome. On the other hand, in sweet orange, that has a highly admixed genome with a few non-admixed regions, the major differences at the expression level were found in the last ones. This might suggest that the specific admixture patterns of each commercial citrus cultivar have a great effect at the transcriptomic level, which would explain why the results obtained in the principal component analyses were remarkably similar to those obtained with the genomic (Wu *et al.*, 2018), and phenotypical (Barrett and Rhodes, 1976) data.

The importance of interspecific hybridizations between genetically distant individuals or species has been repeatedly reported as a major force in the domestication of many tree species including olive (Julca *et al.*, 2020), date palm (Flowers *et al.*, 2019), apple (Duan *et al.*, 2017) or grapevine (Myles *et al.*, 2011). Even in annual crops such as maize, admixed regions have proven their relevance in driving the adaptation of this species from the tropical latitudes of Mexico to Northern latitudes, affecting the flowering time and cold resistance of the European and North American landraces (Brandenburg *et al.*, 2017). In tomato, a single interspecific introgression can regulate the expression of multiple genes, even if they are located far from the introgressed region itself (Koenig *et al.*, 2013).

Our study suggests that during mandarin domestication the selection of desirable traits targeted introgressed regions of the genome, as previously suggested for mandarins (Wu *et al.*, 2018). In the case of sweet oranges and likely grapefruits, produced by backcrosses of mandarins with pummelos, the domestication process generated the current distribution of mandarin and pummelo homozygous regions in their genomes, which very possibly carry the desired traits. Given the ease of vegetative propagation of *Citrus* and the apomictic nature of many commercial cultivars, the domestication process of these species might be explained by an initial selection of specific admixture patterns by crossing species, followed by the selection of somatic mutants of the clonally propagated progeny.

Sugarless pulps and the reorganization of the glycolytic pathway

In this work, differential gene expression in ripening fruit from seven citrus species has been studied. Analyses of gene expression in the acid fruits from sour orange, citron and lemon, showed in the last two species a large number of DEGs involved in carbohydrate metabolism in pulp (Figure 3). Some of these genes coded for enzymes involved in hexose mobilization, like sucrose synthases (*SuSy*), sucrose phosphate synthases (*SPS*) and sucrose phosphatases (*SPP*). We also found that in lemon, many glycolytic genes, including one coding for a hexokinase, expressed preferentially the citron allele. Hexokinases catalyze a rate-limiting step at the beginning of the glycolytic pathway, and their expression patterns have been linked with the total sugar accumulation in fruits such as pear or apple (Li *et al.*, 2016; Zhao *et al.*, 2019). In *Citrus*, most of the sugars found in the fruits are actually synthetized elsewhere, and then translocated into the fruit (Sadka *et al.*, 2019); sink strength has been linked with the hydrolysis of sucrose by several enzymes including *SuSy* (Baxter *et al.*, 2005; Ntoukakis *et al.*, 2017; Sadka *et al.*, 2019). In our study, sugar content remained invariably low across the whole ripening process in lemon and citron samples, which might be in relation with the reduced expression of the *SuSy* and *hexokinase* genes.

A remarkable observation is the consistently lower expression in lemon and citron of several genes involved in carbohydrates and organic acids metabolism, while the expression of genes coding for several V-ATPase subunits was significantly increased (Figure 3). Furthermore, many differentially expressed genes in citron and lemon pulp are involved in organic acids metabolism and ATP-dependent molecular transport of several molecules (Supplementary Table 5). Some studies have suggested that citrate accumulation is not determined by its synthesis (Lin *et al.*, 2015; Guo *et al.*, 2016), but rather by its degradation and accumulation in the vacuolar lumen, a process that requires a steep pH gradient (Cercós *et al.*, 2006; Guo *et al.*, 2016). Recently, a plasma membrane ATPase *CitPH5* has been identified as a relevant factor in determining the vacuolar proton gradient (Strazzer *et al.*, 2019); other studies report that V-ATPases can fulfill the same role, complementing each other (Shi *et al.*, 2015, 2018). We found that one of the subunits of *CitPH5* was indeed overexpressed in citron, lemon and sour orange (Figure 5a), in agreement with previous results (Strazzer *et al.*, 2019), although the other subunit was

also highly expressed in the acidless pummelo. The consistent overexpression of V-ATPases in citron and lemon suggests that the two mechanisms might be working in these species. Our phenotypic results support this idea, since the acidity of citron, lemon and sour orange is maintained at a high level throughout the whole ripening process (Figure 1).

Despite the evidence of some allele imbalance affecting rate-limiting glycolytic enzymes on pulp lemon samples, we believe that the differential expression of genes involved in many steps of the main glycolytic pathway might be the main cause of the different acidity of the studied samples. Our results support that *CitPH5* is relevant for citrus acidity, since it is indeed overexpressed in the three most acidic samples analyzed here: citron, lemon and sour orange (Figure 5b). However, our data suggest the possibility of an independent mechanism besides *CitPH5*, that would involve the accumulation of citrate in the pulp of citron and lemon caused by an increased V-ATPase activity (Figure 3). The overall reduction in sugar accumulation in these fruits would also contribute to increase their sourness.

Citrus pigmentation from a genus-wide perspective

Although we observed several changes in the expression of key genes involved in carotenoid biosynthesis, we could not find a distinctive pattern differentiating red and yellow fruit species (Figure 4). Our data show expression changes in specific genes, suggesting that citrus coloration does not depend on a single master gene, but instead depends on the additive effect of several genes acting independently (Figure 5c).

For example, a *PSY* gene was downregulated in citron and lemon fruits, which show low carotenoid levels when ripe (Alquézar *et al.*, 2008; Zhu *et al.*, 2020). Phytoene synthase activity has been related with carotenoid content in some citrus species (Tao *et al.*, 2007; Zhang *et al.*, 2009) and, although data from lemon and citron would be in agreement with the previous studies, we did not find a reduction of *PSY* expression in pummelo, which also produces yellow fruits, suggesting that an alternative mechanism might be taking place in this species.

Another differentially expressed gene coding for a *zeta-carotene desaturase* showed increased expression levels in the flavedo of sweet orange and the commercial mandarin, both producing red fruits. *ZDS* is essential for the red coloration of tomatoes, as it is

required for the production of lycopene and β-carotene derivatives (McQuinn *et al.*, 2020); *ZDS* has been also associated to carotenoid biosynthesis in carrot (Flores-Ortiz *et al.*, 2020). In *Citrus*, the sweet orange Pinalate mutant, that produces yellow fruits, was initially thought to be a *ZDS* defective mutant (Rodrigo *et al.*, 2003), linking ZDS activity with the red coloration of sweet oranges. Further studies revealed that the defective gene was a zeta-carotene isomerase, and not a desaturase (Rodrigo *et al.*, 2019). The high expression levels of *ZDS* gene that we found in sweet oranges and commercial mandarin might suggest that *ZDS* might after all be involved in the red pigmentation of oranges and mandarins.

One of the main branching points in carotenoid biosynthesis is the lycopene cyclization, carried out by the *lycopene β-cyclase LCYb*, which funnels the carbon flux towards the *β*-*β*-carotene production (Zhang *et al.*, 2012). Indeed, *LCYb* gene expression has been shown to increase during color break of mandarin and orange fruits, suggesting its role in this process (Alquézar *et al.*, 2008; Terol *et al.*, 2019). In this work we show that thr *LCYb2* gene was consistently overexpressed in the flavedo of all red fruits, when compared with the yellow ones from citron, pummelo and lemon, both in the pairwise and in the red against yellow fruits comparisons (Figure 4). The role of *LCYb2* directing the carbon flux of the carotenoid pathway towards *β*-carotene and its derivatives has already been suggested in *Citrus* (Zhang *et al.*, 2012; Rodrigo *et al.*, 2013*a*) and other species such as sweet potato (Kang *et al.*, 2018) or carrot (Moreno *et al.*, 2013). Overall, our results suggest that *LCYb2* activity might be involved with the fruit red coloration in different species from the genus *Citrus*.

The gene coding for a zeaxanthin epoxidase also presented differential expression patterns, being overexpressed in citron and lemon when compared with the remaining species, although only in pummelo, sweet orange and sour orange were statistically significant. In Arabidopsis, *ZEP* defective mutants accumulate *β*-carotene, *β*-cryptoxanthin and zeaxanthin due to a metabolic blockage in carotenoid degradation (Gonzalez-Jorge *et al.*, 2016); a similar observation was made in potato, where reduced *ZEP* expression resulted into the accumulation of zeaxanthin (Wolters *et al.*, 2010). In maize, specific *ZEP* alleles have been identified as reliable predictors of total carotenoid content, highlighting their crucial role in this process (Owens *et al.*, 2014). The increased

expression of the *ZEP* gene we found in citron and lemon could be the cause of the lower carotene accumulation described in the flavedo of these species (Kato *et al.*, 2004).

We also found significant alterations in the expression of genes coding for carotenoid cleavage dioxygenases, including *CCD4b*, that has been postulated as the major enzyme involved in the production of the predominant red carotenoids in mandarins and oranges by cleaving *β*-carotene, *β*-cryptoxanthin and zeaxanthin into C30-apocarotenoids (Rodrigo *et al.*, 2013*b*; Zheng *et al.*, 2019). Among our samples, *CCD4b* was significantly overexpressed in some but not all of the red-colored flavedo samples when compared against the others. Commercial mandarin presented a low *CCD4b* expression, with levels comparable to those of the yellow citron and pummelo. Conversely, *CCD4b* expression in the lemon flavedo was high, reaching values similar to those of the red fruits. *CCD4a* is a paralog of *CCD4b*, and some studies suggest that the latter possibly went through a neofunctionalization process that ultimately produced a gene involved in the degradation of carotenoids into apocarotenoids (Zheng *et al.*, 2019). *CCD4a* was found to be differentially expressed among red and yellow species. Although *CCD4a* has been considerably less studied in *Citrus* since its expression in mandarin and orange peel negligible (Rodrigo *et al.*, 2013*b*), more recent studies have reported that it is actually expressed in the flavedo of yellow fruits (Zheng *et al.*, 2015). According to our results, *CCD4a* is hardly expressed in red fruits, but its expression is higher in the yellow fruits pummelo, citron and lemon. *CCD4a* is involved in the degradation of colored carotenoids in *Chrysanthemun* and *Petunia* petals (Yoshioka *et al.*, 2012; Kishimoto *et al.*, 2018; Phadungsawat *et al.*, 2020), where an impairment in its expression results in an accumulation of carotenoids in the flower petals. Based on these observations and in the reduced expression in citrus red peels, we suggest that *CCD4a* might be responsible of the increased catabolism of carotenoids during citrus ripening in yellow fruits, as it has been described in other plants, hence degrading pigmented compounds.

Our results would not support the existence of a master gene controlling carotenoid accumulation, but rather suggest that this trait would depend on the additive effects of several genes involved in this process. This idea is supported by the large number of somatic mutants that display an altered fruit color, most of which have been linked with mutations affecting genes all along the carotenoid biosynthetic pathway, which ultimately produces the mutant phenotype (Liu *et al.*, 2007; Alquézar *et al.*, 2008; Alós *et al.*, 2008;

Rodrigo *et al.*, 2019; Lana *et al.*, 2020). According to our results, some of these genes would have a more determinant role in of the red or yellow fruits. This way, *CCD4a* and *ZEP* would be relevant in the pummelo, citron and lemon yellow fruits, determining their carotenoid content with their catabolic activity. *LCYb2* appears as a potential candidate to funnel carbon flux towards the *β-β*-carotenoid branch (Zhang *et al.*, 2012). Indeed, substrate availability appears to play an important role in determining enzymatic activity (Baldermann *et al.*, 2010; Rodrigo *et al.*, 2013*b*); suggesting that alterations in the upstream enzymes of the pathway, including *PSY* or *ZDS*, could also cause large differences in the final carotenoid accumulation.

Stepwise evolution of flavonoid accumulation profiles in mandarins

Citrus peels accumulate an immense range of flavonoids and flavonoid derivatives (Wang *et al.*, 2017*c*), with mandarin and orange peel displaying the highest concentrations (Chen *et al.*, 2020). One of the first steps in flavonoid biosynthesis is the synthesis of naringenin chalcone, which is carried out by a chalcone synthase *CHS* (Dao *et al.*, 2011). This is a rate-limiting enzyme and acts as a major regulatory step in flavonoid production, as described in several plants including *Citrus* (Dao *et al.*, 2011; Chaudhary *et al.*, 2016; Wang *et al.*, 2018*b*). In our work we found a chalcone synthase gene *CHSm* which was solely expressed in in pure mandarin and mandarin derived species (Figure 5c). ASE analysis revealed that only the mandarin allele was expressed in sweet and sour orange, as 99% of the reads were from the mandarin haplotype, indicating that the pummelo allele was silenced. The case of lemon is more complex: the genomic coordinates of the *CHSm* locus were previously assigned to a citron/pummelo (C/P) admixed region, hence a mandarin allele should not be present, as the closest citron/mandarin (C/M) admixed region is located about 60 kb away from *CHSm* locus (Wu *et al.*, 2018). However, a manual analysis based on diagnostic SNPs in the *CHSm* locus revealed that only the mandarin allele was expressed in lemon for that locus, and indeed the mandarin allele was found in the *CHSm* locus in the genomic sequencing as well. The assignment of this region as a citron/pummelo admixed region might be erroneous, possibly due to the proximity of a true citron/mandarin region and especially considering that the methodology used to determine the segmental ancestry in *Citrus* is more error-prone near admixed region boundaries, where the local ancestry can be ambiguous (Wu *et al.*, 2018).

The analysis of the promoter region of *CHSm* locus revealed several differences between the pure species. Less than 100 bp upstream the transcription start site, citron and pummelo displayed a completely different sequence to that from clementine that distinguished the citron/pummelo allele from the mandarin one (Figure 9). The different alleles could be found in the admixed species in accordance with their admixture patterns (Figure 10b). The analysis of the regulatory target sites showed an ethylene responsive element in the citron/pummelo allele, that was absent in the mandarin and the ancestral alleles (Figure 10a). EREs have been already described as recurrent elements in promoter region of the *CHS* genes in eggplant (Wu *et al.*, 2020), indicating that they could regulate *CHS* expression. Similar results were obtained in grapevine, where *CHS* expression increased under ethylene treatment (El-Kereamy *et al.*, 2003). In Arabidopsis, an *erf6* (*ethylene responsive factor 6*) defective mutant increased *CHS* expression to a roughly 6 fold, compared to wild type plants (Sewelam *et al.*, 2013). In *Citrus*, transcription factors belonging to the AP/ERF family have been shown to regulate the expression of a *chalcone isomerase* gene (Zhao *et al.*, 2020).

The analyses of the available genome sequences of *Citrus* species and wild relatives showed that the archaic allele, present in *S. buxifolia* and *P. trifoliata*, is a shorter version of the citron/pummelo allele. The mandarin allele is present in the wild species *C. reticulata*, as well as in *C. ichangensis* and *F. hindsii*, that are not directly related to mandarins. Commercial varieties showed an allele composition that correlated with their admixture patterns around the *CHSm* locus: *C. clementina* and *C. unshiu* showed only the mandarin allele. *C. sinensis* showed the pummelo allele in the reference genome, but the resequencing data revealed the presence of the mandarin allele as well, since the *CHSm* locus in sweet orange is located in a pummelo/mandarin admixed region.

Mandarins have gone through an intensive domestication process which implied the selection of beneficial traits. However, the *CHSm* mandarin allele appears to be present in other pure species such as *F. hindsii* and *C. ichangensis*, that diverged from the mandarin clade millions of years before the domestication started (Wu *et al.*, 2018). As *CHS* is a rate-limiting enzyme in the flavonoid biosynthetic pathway, the presence of the mandarin allele would increase the expression level of *CHS* compared with pummelo or citron, ultimately leading to a greater flavonoid accumulation.

The study of the expression of the genes involved in the flavonoid biosynthetic pathway has found a significant variability in those coding for flavonoid-modifying enzymes (Figure 6), and especially for the flavonoid-O-methyltransferases. Despite this variability, only a few FOMT genes presented allele-specific expression, suggesting that the flavonoid diversity found in *Citrus* arises from interspecific changes in FOMT expression levels. Methylated flavonoids, and most notably polymethoxylated flavonoids, are a diverse family of compounds fulfilling multiple biological functions. Previous studies have assigned a broad substrate specificity to *Citrus* FOMTs (Itoh *et al.*, 2016; Liu *et al.*, 2020*a*), whose expression is also extremely variable across tissues and development stages, with up to 58 different genes being expressed in specific conditions (Liu *et al.*, 2016). This is partly explained by the recent expansion of FOMT gene families observed in *Citrus* when compared with other plant lineages (Gonzalez-Ibeas *et al.*, 2021, in press). It is well known that gene family expansion paves the way for neofunctionalization by providing with extra copies of genes belonging to the family, therefore allowing for a better adaptation to new environments. Within the genus *Citrus*, FOMTs are further expanded in mandarins and, to a lesser extent, in *C. ichangensis* and *Fortunella spp.*.

It is interesting to note that the FOMT gene family is specifically expanded in those species with the mandarin *CHSm* allele (Gonzalez-Ibeas *et al.*, 2021, in press). An increased *CHSm* expression in the fruits of these species could in principle generate greater amounts of flavonoid precursors, which could be further modified by the expanded FOMT family to produce a broader range of compounds. Considering that the mandarin *CHSm* allele is found in several pure species across the *Citrus* genus, while the FOMT expansion is more pronounced in mandarins and their derivatives, we believe that these two processes might have occurred in different time periods. It is possible that the mandarin *CHSm* allele conferred an adaptive advantage, becoming widespread along different *Citrus* species, while the FOMT expansion responded to a posterior process, possibly during the mandarin early domestication, or even sometime in between the evolutionary and domestication processes.

CONCLUDING REMARKS

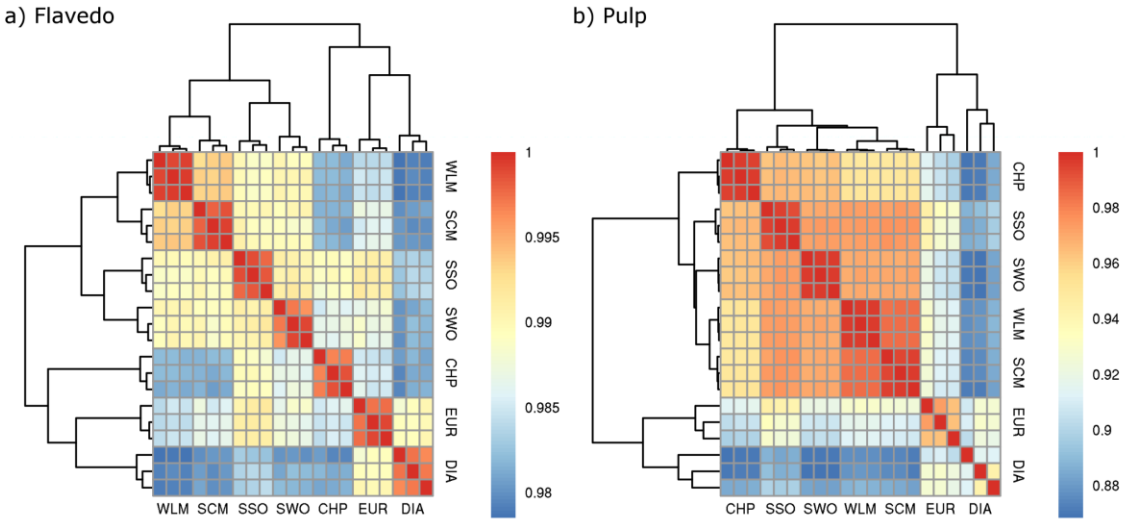
In this work we have performed a genus wide RNA-seq analysis in fruits from seven citrus species, in order to assess the impact of the complex evolutionary and

domestication histories of the commercial citrus varieties in gene expression. We report for the first time the effects of the segmental ancestry of specific citrus species in the expression patterns of the genes contained within. Our results reveal that the different admixed regions of the commercial mandarin and sweet orange genomes harbor a great number of genes that are expressed differently from their wild progenitors. This observation highlights the importance of introgressions during the early domestication of the genus *Citrus* and might help us to understand the process that gave birth to the currently existing species.

The broader scope of this work has allowed us to describe the extensive alteration of the glycolytic pathway in citron and lemon and the involvement of *CCD4a* in setting up *Citrus* fruit color, modifications that could have not been detected analyzing only clonal varieties. This advocates for the relevance of studying these processes from an evolutionary perspective, especially in a genus like *Citrus*, as most of the commercial cultivars are mosaics of the pure species that form this genus.

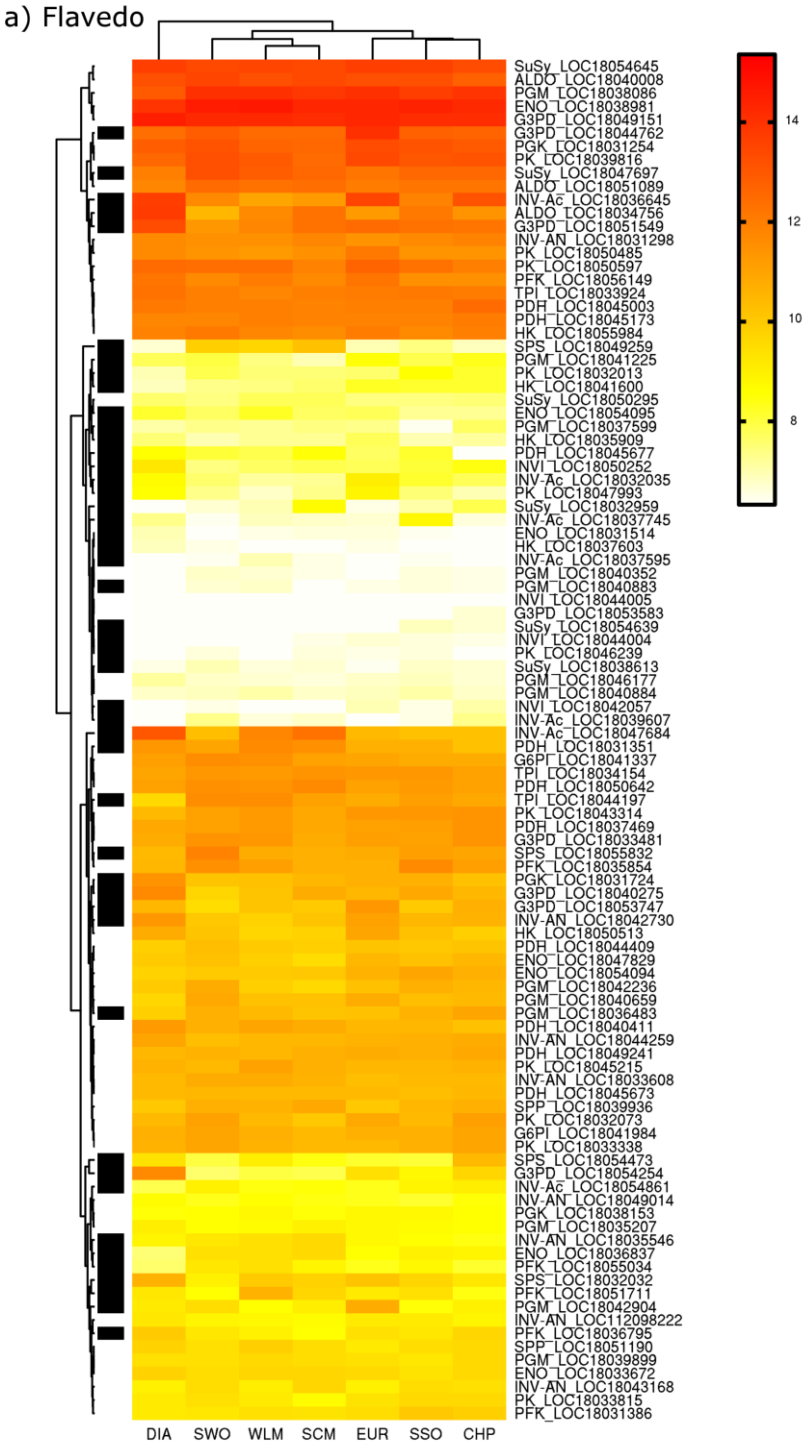
Expression data analyzed from a genus-wide perspective also allowed us to suggest the evolutionary history of the chalcone synthase gene *CHSm*. Our results indicate that the particular expression of this locus, the differences in its promoter region and the recent expansion of the flavonoid O-methyltransferase family could be related, so that the current diversity of polymethoxylated flavonoids found in *Citrus*, and especially in mandarins, might have evolved in a stepwise manner. The specific expression of *CHSm* in the fruits of mandarins and close species may have provided with flavonoid precursors. Then, the pronounced expansion of flavonoid O-methyltransferases in mandarins would have facilitated the neofunctionalization of the new loci. Thus, the increased production of flavonoid precursors and the specialization of the mandarin flavonoid O-methyltransferases could very possibly explain the high concentration and broad diversity of flavonoid derivatives described in mandarins so far.

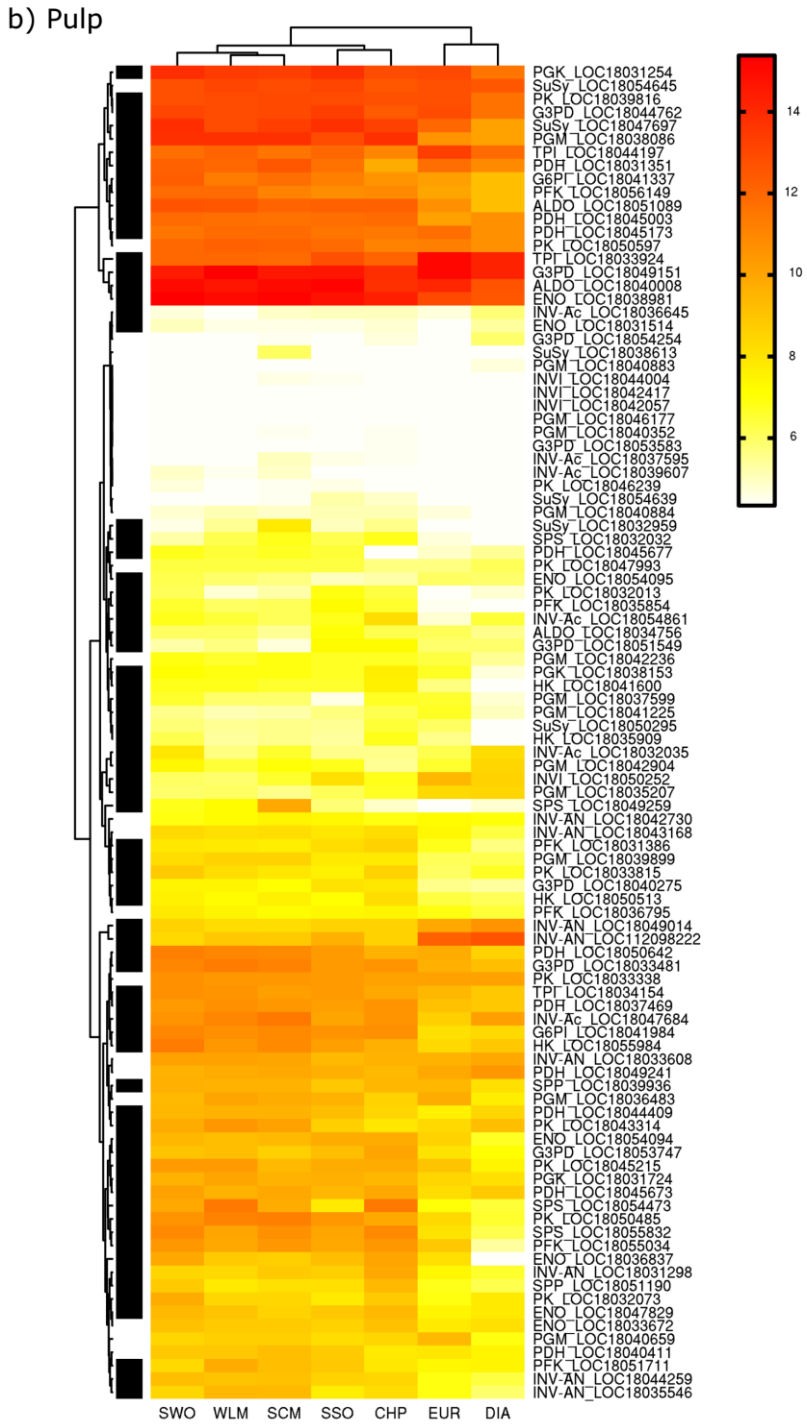
SUPPLEMENTARY DATA



Supplementary Figure 1: Sample clustering based on transcriptomic reads. Euclidean distances were calculated between every samples pair using normalized read counts. Flavedo (a) and pulp (b) were analyzed independently. The color scale for each figure is depicted on the right. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

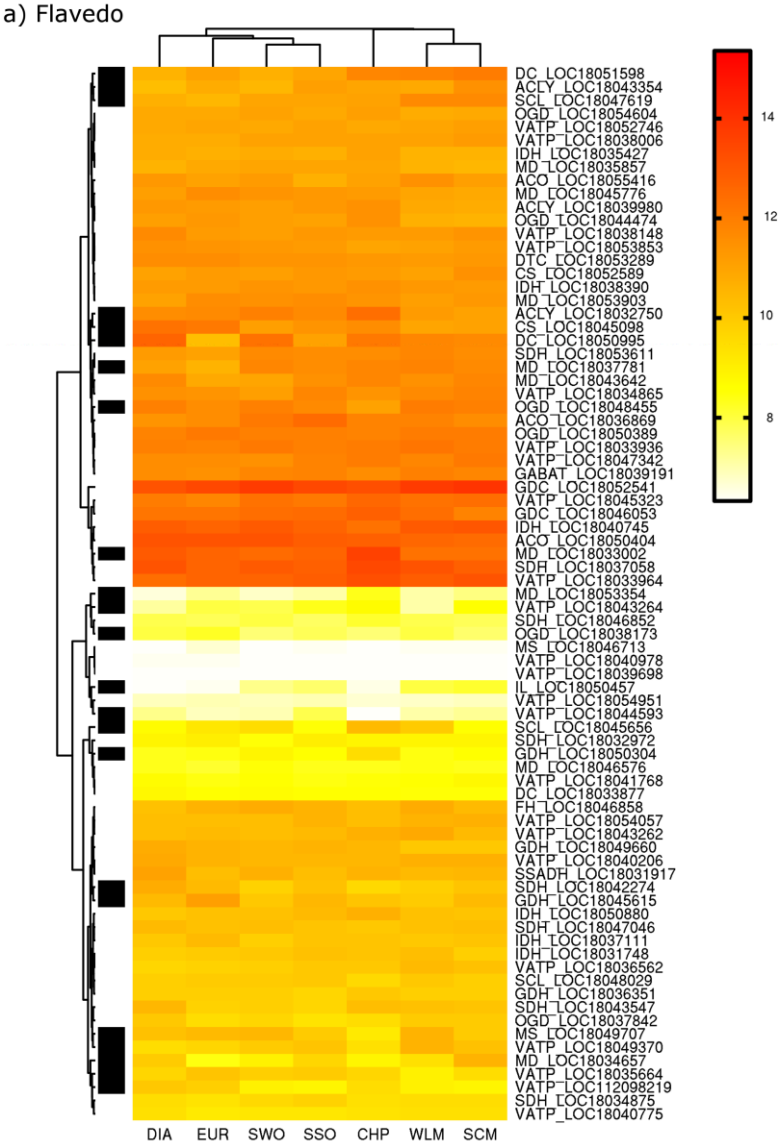
a) Flavedo

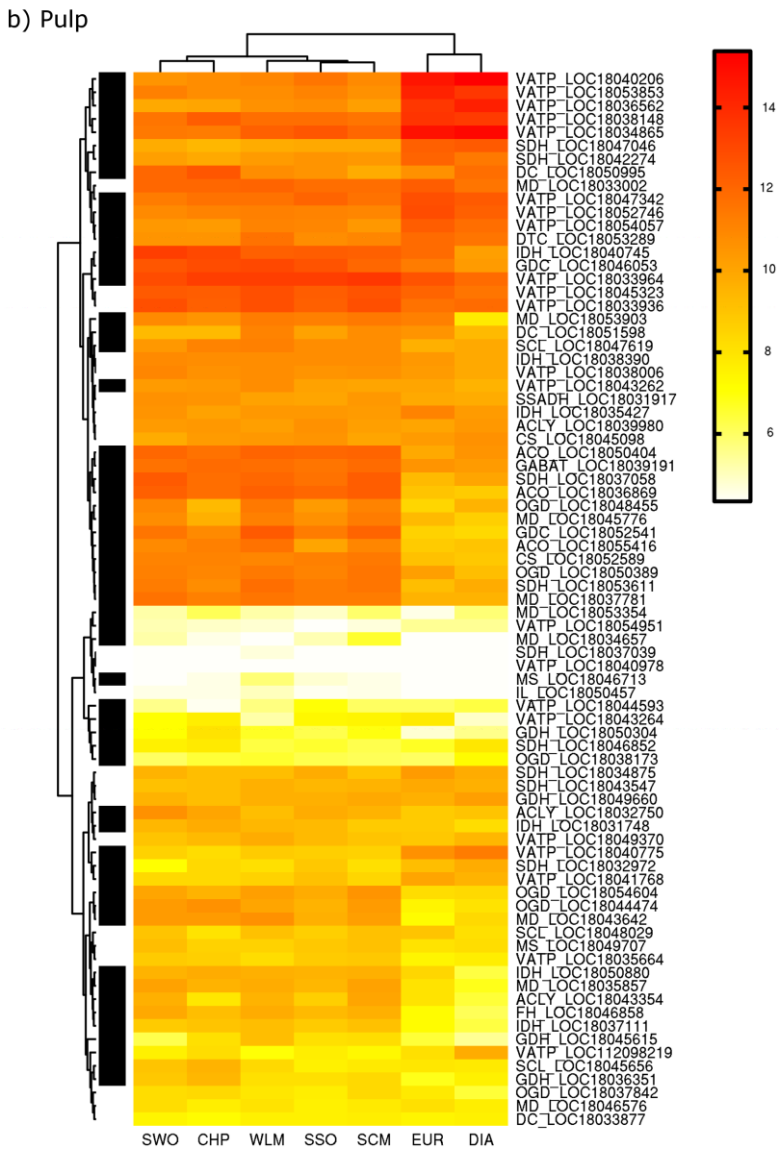

SuSy_LOC18054645
ALDO_LOC18040008
PGM_LOC18038086
ENO_LOC18038981
G3PD_LOC18049151
G3PD_LOC18044762
PGK_LOC18031254
PK_LOC18039816
SuSy_LOC18047697
ALDO_LOC18051089
INV-Ac_LOC18036645
ALDO_LOC18034756
G3PD_LOC18051549
INV-AN_LOC18031298
PK_LOC18050485
PK_LOC18050597
PFK_LOC18056149
TPI_LOC18033924
PDH_LOC18045003
PDH_LOC18045173
HK_LOC18055984
SPS_LOC18049259
PGM_LOC18041225
PK_LOC18032013
HK_LOC18041600
SuSy_LOC18050295
ENO_LOC18054095
PGM_LOC18037599
HK_LOC18035909
PDH_LOC18045677
INVI_LOC18050252
INV-Ac_LOC18032035
PK_LOC18047993
SuSy_LOC18032959
INV-Ac_LOC18037745
ENO_LOC18031514
HK_LOC18037603
INV-Ac_LOC18037595
PGM_LOC18040352
PGM_LOC18040883
INVI_LOC18044005
G3PD_LOC18053583
SuSy_LOC18054639
INVI_LOC18044004
PK_LOC18046239
SuSy_LOC18038613
PGM_LOC18046177
PGM_LOC18040884
INVI_LOC18042057
INV-Ac_LOC18039607
INV-Ac_LOC18047684
PDH_LOC18031351
G6PI_LOC18041337
TPI_LOC18034154
PDH_LOC18050642
TPI_LOC18044197
PK_LOC18043314
PDH_LOC18037469
G3PD_LOC18033481
SPS_LOC18055832
PFK_LOC18035854
PGK_LOC18031724
G3PD_LOC18040275
G3PD_LOC18053747
INV-AN_LOC18042730
HK_LOC18050513
PDH_LOC18044409
ENO_LOC18047829
ENO_LOC18054094
PGM_LOC18042236
PGM_LOC18040659
PGM_LOC18036483
PDH_LOC18040411
INV-AN_LOC18044259
PDH_LOC18049241
PK_LOC18045215
INV-AN_LOC18033608
PDH_LOC18045673
SPP_LOC18039936
PK_LOC18032073
G6PI_LOC18041984
PK_LOC18033338
SPS_LOC18054473
G3PD_LOC18054254
INV-Ac_LOC18054861
INV-AN_LOC18049014
PGK_LOC18038153
PGM_LOC18035207
INV-AN_LOC18035546
ENO_LOC18036837
PFK_LOC18055034
SPS_LOC18032032
PFK_LOC18051711
PGM_LOC18042904
INV-AN_LOC112098222
PFK_LOC18036795
SPP_LOC18051190
PGM_LOC18039899
ENO_LOC18033672
INV-AN_LOC18043168
PK_LOC18033815
PFK_LOC18031386
DIA
SWO
WLM
SCM
EUR
SSO
CHP
14
12
10
8



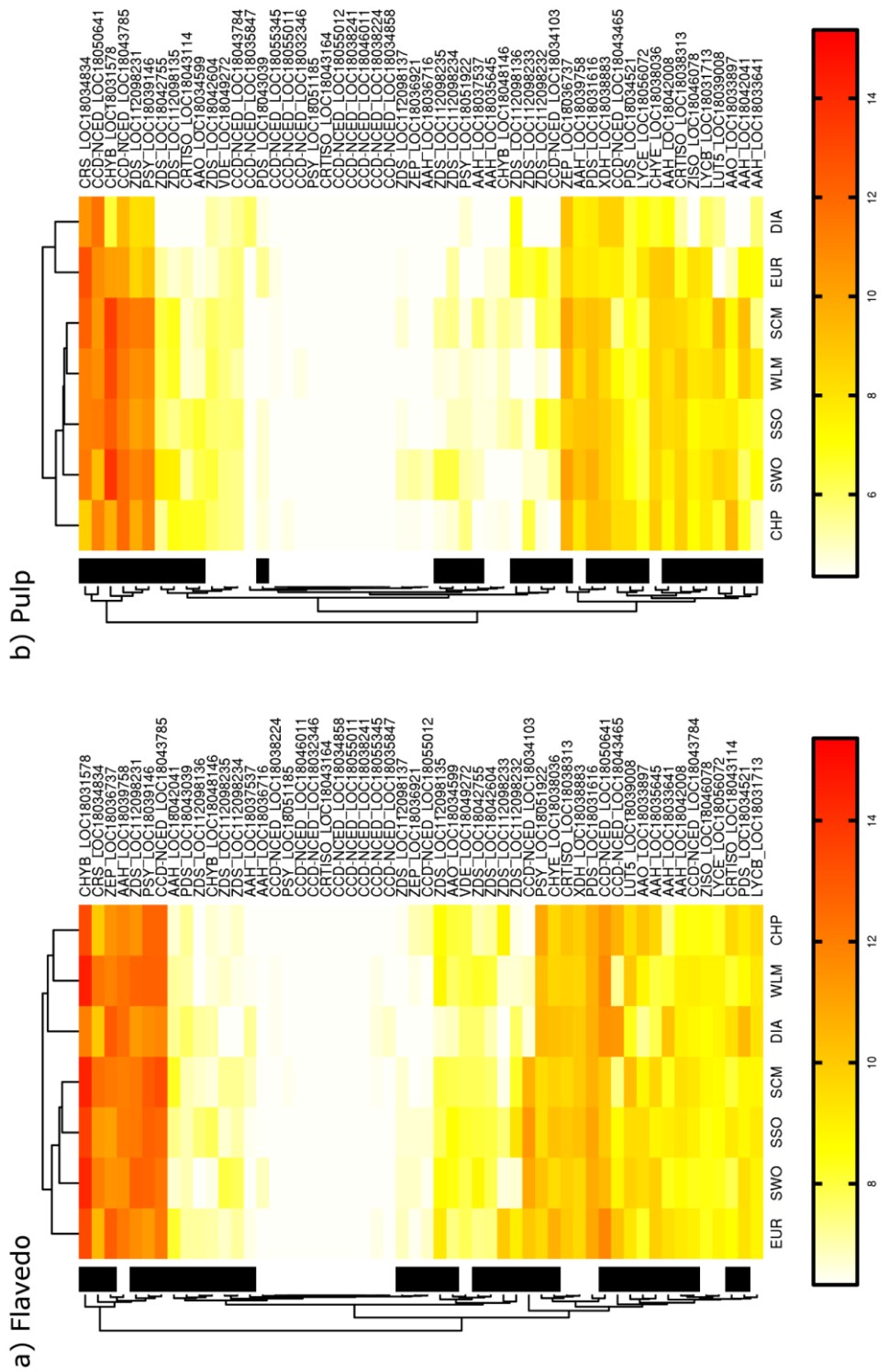
Supplementary Figure 2: Heatmap representing the expression levels on glycolysis metabolism. Color intensity represent read abundance based on normalized read counts. Black rectangles mark genes with significant differences in their expression between at least two samples. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

a) Flavedo


DC_LOC18051598
ACLY_LOC18043354
SCL_LOC18047619
OGD_LOC18054604
VATP_LOC18052746
VATP_LOC18038006
IDH_LOC18035427
MD_LOC18035857
ACO_LOC18055416
MD_LOC18045776
ACLY_LOC18039980
OGD_LOC18044474
VATP_LOC18038148
VATP_LOC18053853
DTC_LOC18053289
CS_LOC18052589
IDH_LOC18038390
MD_LOC18053903
ACLY_LOC18032750
CS_LOC18045098
DC_LOC18050995
SDH_LOC18053611
MD_LOC18037781
MD_LOC18043642
VATP_LOC18034865
OGD_LOC18048455
ACO_LOC18036869
OGD_LOC18050389
VATP_LOC18033936
VATP_LOC18047342
GABAT_LOC18039191
GDC_LOC18052541
VATP_LOC18045323
GDC_LOC18046053
IDH_LOC18040745
ACO_LOC18050404
MD_LOC18033002
SDH_LOC18037058
VATP_LOC18033964
MD_LOC18053354
VATP_LOC18043264
SDH_LOC18046852
OGD_LOC18038173
MS_LOC18046713
VATP_LOC18040978
VATP_LOC18039698
IL_LOC18050457
VATP_LOC18054951
VATP_LOC18044593
SCL_LOC18045656
SDH_LOC18032972
GDH_LOC18050304
MD_LOC18046576
VATP_LOC18041768
DC_LOC18033877
FH_LOC18046858
VATP_LOC18054057
VATP_LOC18043262
GDH_LOC18049660
VATP_LOC18040206
SSADH_LOC18031917
SDH_LOC18042274
GDH_LOC18045615
IDH_LOC18050880
SDH_LOC18047046
IDH_LOC18037111
IDH_LOC18031748
VATP_LOC18036562
SCL_LOC18048029
GDH_LOC18036351
SDH_LOC18043547
OGD_LOC18037842
MS_LOC18049707
VATP_LOC18049370
MD_LOC18034657
VATP_LOC18035664
VATP_LOC112098219
SDH_LOC18034875
VATP_LOC18040775
DIA
EUR
SWO
SSO
CHP
WLM
SCM
14
12
10
8



Supplementary Figure 3: Heatmap representing the expression levels on TCA metabolism and V-ATPases. Color intensity represent read abundance based on normalized read counts. Black rectangles mark genes with significant differences in their expression between at least two samples. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.



Supplementary Figure 4: Heatmap representing the expression levels on carotenoid metabolism. Color intensity represent read abundance based on normalized read counts. Black rectangles mark genes with significant differences in their expression between at least two samples. Abbreviations: CHP: *C. maxima*, DIA: *C. medica*, EUR: *C. limon*, SCM: *C. reticulata*, SSO: *C. aurantium*, SWO: *C. sinensis*, WLM: *C. deliciosa*.

Supplementary Table 1. Primers used for qPCR analysis.

Name	Sequence	NCBI Annotation
CHS_PuM	F - CCACATCTAATAATGTAGTCATGGGAC R - CATGGGACTCACGTTTCATCC	LOC18042808
CHS_Cit	F - GATATGGGACTTTTTAGAGGATGGT R - TGGGACTTTTTAGAGGATGGTATATTT	LOC18042808
CHS_Man	F - TGTTCCAGGAGATAAGAACAAACAAG R - CCACATCTAATAATGTAGTCATGGGAC	LOC18042808
UBC	F - GTGCAGCGAGAGAAATCAGC R - ACTTGTGGAGGTTGCAGAGG	LOC18055321
VATPaseG	F - GCTGGCTGCTGAACAAGAAG R - AGCCTCTTCACATTCGCACC	LOC18038148
CHSm	F - ACTTGTGGGCGTAGACATGC R - CTGTGGCGCCAATGTAACAG	LOC18042808
CCD4b	F - CTACAACACCAAATCCGCGC R - CGGTTAGAGAGTCCGGAAGC	LOC18034103
CCD4a	F - TGGCGTCCATAACCAGGAAC R - ATTGCATCGTGGCTACCAGG	LOC18043465

Supplementary Table 2. Sequence Read Archive accession codes of the studied species.

Accession Number	Species name	Binomial name	SRA Accession code
B483	SunChuSha Kat Mandarin	*Citrus reticulata*	SRX3298473
B207	Chandler Pummelo	*Citrus maxima*	ERR466631
B560	Diamante Citron	*Citrus medica*	Sequenced in this work
B154	Willowleaf Mandarin	*Citrus deliciosa*	ERR466627
B031	Sweet Orange	*Citrus sinensis*	ERR466624
B117	Sour Orange	*Citrus aurantium*	ERR466633
B297	Eureka Lemon	*Citrus limon*	ERR466636

Supplementary Table 3. qPCR validation of the differential expression analysis. Table S3 is only partially displayed given its size. The complete version is available upon demand**.**

Tissue	Sample A	Sample B	Abbreviation	Locus ID	ΔΔCt (qPCR)	2log fold change (DESEq2)
Flavedo	DIA	SWO	*CCD4b*	LOC18034103	-3.53	-7.82
			V-ATPase G	LOC18038148	-0.14	0.61
			CHSm	LOC18042808	-5.83	-16.96
			CCD4a	LOC18043465	3.08	3.04
			UBC	LOC18055321	0	-0.15
Flavedo	DIA	SSO	*CCD4b*	LOC18034103	-3.67	-8.32
			V-ATPase G	LOC18038148	0.01	0.42
			CHSm	LOC18042808	-4.06	-15.28
			CCD4a	LOC18043465	3.58	3.28
			UBC	LOC18055321	0	-0.08
Flavedo	SCM	WLM	*CCD4b*	LOC18034103	2.14	6.5
			V-ATPase G	LOC18038148	-0.02	0.15
			CHSm	LOC18042808	-0.43	-0.79
			CCD4a	LOC18043465	-1.53	0.49
			UBC	LOC18055321	0	0.06
Flavedo	SCM	CHP	*CCD4b*	LOC18034103	3.86	13.76
			V-ATPase G	LOC18038148	-0.57	0.21
			CHSm	LOC18042808	7.37	18.01
			CCD4a	LOC18043465	-2.19	-4.78
			UBC	LOC18055321	0	0.38
Flavedo	SCM	EUR	*CCD4b*	LOC18034103	0.07	0.41
			V-ATPase G	LOC18038148	-1.52	0.22
			CHSm	LOC18042808	2.51	4.59
			CCD4a	LOC18043465	-2.73	-4.27
			UBC	LOC18055321	0	0.44

Supplementary Table 4. Equivalences for the gene abbreviation names.

Gene name	Abbreviation	Gene function
Sucrose synthase	*SuSy*	Sucrose processing
Sucrose-phosphatase	*SPP*	Sucrose processing
Sucrose-phosphate synthase	*SPS*	Sucrose processing
Acid invertase	*INV-Ac*	Sucrose processing
Alkaline/neutral invertase	*INV-AN*	Sucrose processing
Invertase inhibitor	*INVI*	Sucrose processing
Hexokinase	*HK*	Glycolisis
Glucose-6-phosphate isomerase	*G6PI*	Glycolisis
ATP-dependent 6-phosphofructokinase	*PFK*	Glycolisis
Fructose-bisphosphate aldolase	*ALDO*	Glycolisis
Triosephosphate isomerase	*TPI*	Glycolisis
Glyceraldehyde-3-phosphate dehydrogenase	*G3PD*	Glycolisis
Phosphoglycerate kinase	*PGK*	Glycolisis
Phosphoglycerate mutase	*PGM*	Glycolisis
Enolase	*ENO*	Glycolisis
Pyruvate kinase	*PK*	Glycolisis
Pyruvate dehydrogenase	*PDH*	Glycolisis
ATP-citrate lyase	*ACLY*	Citrate degradation
Citrate synthase	*CS*	TCA
Aconitate hydratase	*ACO*	TCA
Isocitrate dehydrogenase	*IDH*	TCA
2-oxoglutarate dehydrogenase	*OGD*	TCA
Succinate-CoA ligase	*SCL*	TCA
Succinate dehydrogenase	*SDH*	TCA
Fumarate hydratase 1, mitochondrial	*FH*	TCA
Malate dehydrogenase	*MD*	TCA
Isocitrate lyase	*IL*	Glyoxylate cycle
Malate synthase	*MS*	Glyoxylate cycle
V-type proton ATPase subunit	*VATP*	Vacuolar proton intake

Table S4 (continued)

Gene name	**Abbreviation**	**Gene function**
Glutamate dehydrogenase	*GDH*	GABA Cycle
Glutamate decarboxylase	*GDC*	GABA Cycle
Gamma aminobutyrate transaminase 3, chloroplastic	*GABAT*	GABA Cycle
Succinate-semialdehyde dehydrogenase, mitochondrial	*SSADH*	GABA Cycle
Tonoplast dicarboxylate transporter	*DC*	Vacuolar citrate intake
Mitochondrial dicarboxylate/tricarboxylate transporter	*DTC*	Mitochondrial carboxylate transport
Phytoene synthase	*PSY*	Carotenoid pathway
Phytoene dehydrogenase	*PDS*	Carotenoid pathway
Zeta-carotene isomerase	*ZISO*	Carotenoid pathway
Zeta-carotene desaturase	*ZDS*	Carotenoid pathway
Prolycopene isomerase	*CRTISO*	Carotenoid pathway
Lycopene epsilon cyclase	*LCYE*	Carotenoid pathway
Protein LUTEIN DEFICIENT 5, chloroplastic	*LUT5*	Carotenoid pathway
Carotene epsilon-monooxygenase	*CHYE*	Carotenoid pathway
Lycopene beta cyclase	*LCYB*	Carotenoid pathway
Capsanthin/capsorubin synthase	*CRS*	Carotenoid pathway
Beta-carotene 3-hydroxylase	*CHYB*	Carotenoid pathway
Zeaxanthin epoxidase	*ZEP*	Carotenoid pathway
Violaxanthin de-epoxidase	*VDE*	Carotenoid pathway
Carotenoid cleavage dioxygenase	*CCD-NCED*	Carotenoid pathway
Xanthoxin dehydrogenase	*XDH*	Carotenoid pathway
Abscisic-aldehyde oxidase	*AAO*	Carotenoid pathway
Abscisic acid 8'-hydroxylase	*AAH*	Carotenoid pathway

Supplementary Table 5. qPCR validation of the differential expression analysis.

ID	GO type	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
GO:0016874	MF	ligase activity	77/4386	98/9007	9,91E-10	3,57E-07	2,98E-07	77
GO:0016818	MF	hydrolase activity, acting on acid anhydrides, in	240/4386	382/9007	9,74E-09	1,75E-06	1,47E-06	240
GO:0016817	MF	hydrolase activity, acting on acid anhydrides	242/4386	387/9007	1,55E-08	1,87E-06	1,56E-06	242
GO:0017111	MF	nucleoside-triphosphatase activity	226/4386	359/9007	2,08E-08	1,88E-06	1,57E-06	226
GO:0016462	MF	pyrophosphatase activity	234/4386	376/9007	4,88E-08	3,51E-06	2,94E-06	234
GO:0140098	MF	catalytic activity, acting on RNA	114/4386	165/9007	7,07E-08	4,24E-06	3,55E-06	114
GO:0003723	MF	RNA binding	295/4386	495/9007	3,63E-07	1,87E-05	1,56E-05	295
GO:0004812	MF	aminoacyl-tRNA ligase activity	32/4386	39/9007	1,72E-05	6,87E-04	5,75E-04	32
GO:0016875	MF	ligase activity, forming carbon-oxygen bonds	32/4386	39/9007	1,72E-05	6,87E-04	5,75E-04	32
GO:0051082	MF	unfolded protein binding	40/4386	52/9007	2,82E-05	1,02E-03	8,50E-04	40
GO:0016879	MF	ligase activity, forming carbon-nitrogen bonds	29/4386	36/9007	8,33E-05	2,73E-03	2,28E-03	29
GO:0008270	MF	zinc ion binding	213/4386	367/9007	1,55E-04	4,65E-03	3,89E-03	213

Table S5 (continued)

ID	GO type	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
GO:0004386	MF	helicase activity	38/4386	52/9007	2,95E-04	8,18E-03	6,84E-03	38
GO:0004222	MF	metalloendopeptidase activity	26/4386	33/9007	3,83E-04	9,84E-03	8,23E-03	26
GO:0008237	MF	metallopeptidase activity	34/4386	46/9007	4,32E-04	1,04E-02	8,66E-03	34
GO:0140101	MF	catalytic activity, acting on a tRNA	48/4386	70/9007	5,83E-04	1,31E-02	1,10E-02	48
GO:0015662	MF	ATPase activity, coupled to transmembrane movement of	17/4386	20/9007	8,71E-04	1,84E-02	1,54E-02	17
GO:0015399	MF	primary active transmembrane transporter	50/4386	75/9007	1,22E-03	2,32E-02	1,94E-02	50
GO:0015405	MF	P-P-bond-hydrolysis-driven transmembrane transporter	50/4386	75/9007	1,22E-03	2,32E-02	1,94E-02	50
GO:0003774	MF	motor activity	32/4386	45/9007	1,88E-03	3,10E-02	2,59E-02	32
GO:0042626	MF	ATPase activity, coupled to transmembrane movement of	46/4386	69/9007	1,89E-03	3,10E-02	2,59E-02	46
GO:0043492	MF	ATPase activity, coupled to movement of substances	46/4386	69/9007	1,89E-03	3,10E-02	2,59E-02	46
GO:0005388	MF	calcium-transporting ATPase activity	11/4386	12/9007	2,41E-03	3,45E-02	2,89E-02	11
GO:0001882	MF	nucleoside binding	97/4386	162/9007	2,56E-03	3,45E-02	2,89E-02	97

Table S5 (continued)

ID	GO type	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
GO:0043021	MF	ribonucleoprotein complex binding	13/4386	15/9007	2,72E-03	3,45E-02	2,89E-02	13
GO:0003899	MF	DNA-directed 5'-3' RNA polymerase activity	25/4386	34/9007	2,83E-03	3,45E-02	2,89E-02	25
GO:0001883	MF	purine nucleoside binding	94/4386	157/9007	2,97E-03	3,45E-02	2,89E-02	94
GO:0005525	MF	GTP binding	94/4386	157/9007	2,97E-03	3,45E-02	2,89E-02	94
GO:0019001	MF	guanyl nucleotide binding	94/4386	157/9007	2,97E-03	3,45E-02	2,89E-02	94
GO:0032550	MF	purine ribonucleoside binding	94/4386	157/9007	2,97E-03	3,45E-02	2,89E-02	94
GO:0032561	MF	guanyl ribonucleotide binding	94/4386	157/9007	2,97E-03	3,45E-02	2,89E-02	94
GO:0032549	MF	ribonucleoside binding	96/4386	161/9007	3,21E-03	3,61E-02	3,02E-02	96
GO:0003777	MF	microtubule motor activity	23/4386	31/9007	3,45E-03	3,65E-02	3,06E-02	23
GO:0003924	MF	GTPase activity	67/4386	108/9007	3,45E-03	3,65E-02	3,06E-02	67
GO:0004707	MF	MAP kinase activity	10/4386	11/9007	4,58E-03	4,71E-02	3,94E-02	10
GO:0008757	MF	S-adenosylmethionine-dependent methyltransferase	31/4386	45/9007	4,83E-03	4,83E-02	4,04E-02	31
GO:0043021	MF	ribonucleoprotein complex binding	13/4386	15/9007	2,72E-03	3,45E-02	2,89E-02	13

Table S5 (continued)

ID	GO type	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
GO:0043038	BP	amino acid activation	34/2355	41/4550	3,02E-05	0,01078892	0,00981389	34
GO:0043039	BP	tRNA aminoacylation	34/2355	41/4550	3,02E-05	0,01078892	0,00981389	34
GO:0006418	BP	tRNA aminoacylation for protein translation	32/2355	39/4550	7,92E-05	0,0188394	0,01713682	32
GO:0006457	BP	protein folding	61/2355	86/4550	0,00020285	0,0362091	0,03293678	61
GO:0006082	BP	organic acid metabolic process	222/2355	367/4550	0,00028008	0,03999504	0,03638057	222
GO:0019752	BP	carboxylic acid metabolic process	219/2355	363/4550	0,00038286	0,04084754	0,03715603	219
GO:0043436	BP	oxoacid metabolic process	220/2355	365/4550	0,00040047	0,04084754	0,03715603	220

Supplementary Table 6. Relevant genes displaying allele-specific expression.

Tissue	Gene	Locus ID	Sample	Allele balance	Dominant species
Flavedo	*CHS*	LOC18042808	SWO	0	Mandarin
			SSO	0.006	Mandarin
	HK	LOC18035909	SWO	0.903	Pummelo
			SSO	0.891	Pummelo
			EUR	0.971	Citron
	PSY	LOC18039146	SWO	0.323	Mandarin
			WLM	0.21	Mandarin
Pulp	*CHS*	LOC18042808	SWO	0	Mandarin
			SSO	0	Mandarin
	HK	LOC18035909	SWO	0.89	Pummelo
			EUR	1	Citron
	PSY	LOC18039146	SWO	0.258	Mandarin
			WLM	0.114	Mandarin
			EUR	0.809	Pummelo

General discussion

GENERAL DISCUSSION

The genus *Citrus* encompasses thousands of cultivars of great relevance for humans, displaying a phenotypical diversity that has long intrigued botanists and other researchers, generating numerous debates regarding their taxonomy and evolution (Tanaka, 1954; Swingle and Reece, 1967; Ollitrault *et al.*, 2020). Pure *Citrus* species are found in the wild in most of South East Asia, South China, North East India, Japan, Australia and neighboring islands. However, the most popular citrus, including oranges, lemons and most mandarins are interspecific hybrids of *Citrus* pure species (Wu *et al.*, 2014), and as such, should be included in the concept of the genus *Citrus*. The prevalence of these admixtures across cultivated varieties has hindered for a long time the inference of a solid phylogenetic tree for *Citrus*. Recently, however, the backbone of the *Citrus* phylogeny has been revealed, discriminating pure species from admixtures *via* genome-wide analyses (Wu *et al.*, 2018). These authors studied representative species of the major groups comprising the genus *Citrus*, although the analyses neither included other intriguing species such as *Citrus indica*, or *Citrus halimii*, nor other taxa traditionally classified as "citrus-related genera". Overall, Wu *et al*. (2018) showed that most of the current *Citrus* species appeared in a rapid radiation starting 8 Mya, very possibly caused by the sudden cooling and aridification that took place in South East Asia at that time (Herbert *et al.*, 2016; Holbourn *et al.*, 2018; Tanner *et al.*, 2020).

In the current work, the Aurantioideae phylogeny has been studied to anchor the genus *Citrus* on this subfamily and the relationships between its members have been explored in detail. An alignment-free method, based on whole genome sequencing data, has been employed to infer the genetic distances between species and generate a phylogenetic tree of the complete Aurantioideae subfamily. Previous studies have addressed the same question by studying specific chloroplast markers or full chloroplast sequences (Pfeil and Crisp, 2008; Carbonell-Caballero *et al.*, 2015; Oueslati *et al.*, 2016), sometimes including a reduced number of nuclear markers (Ramadugu *et al.*, 2013). A major issue with the phylogenies based on chloroplast sequences in the genus *Citrus* is that they invariably cluster together citrons and the Australian species, forming an outgroup apart from the other *Citrus* species and *Poncirus*. This clustering is in direct conflict with nuclear phylogenies and botanical classifications, and some authors have suggested the rare event

of chloroplast capture as a plausible explanation for this behavior (Nagano *et al.*, 2018). With the alignment-free method, the first genome-wide Aurantioideae phylogeny is presented, in order to provide a comprehensive answer to the phylogenetic placement of *Citrus* on this subfamily. Two different fossil calibrations have been used (Pan, 2010; Xie *et al.*, 2013) to date the divergence times of the different Aurantioideae taxa. The tree topology obtained here agrees in general with previous studies (Pfeil and Crisp, 2008; Ramadugu *et al.*, 2013; Nagano *et al.*, 2018), although slightly older divergence times than those of other authors were found (Pfeil and Crisp, 2008). Briefly, the results of the current work support the previous hypotheses that classified the Clauseneae tribe as a paraphyletic clade, while the tribe Citreae remains as a monophyletic clade with the genus *Murraya* as a sister taxon. As this work used whole genome sequencing data, the results presented here are highly coincidental with those of Nagano *et al.* (2018), which used RAD-seq instead of focusing on the chloroplast genome. The inclusion of *Ruta chalepensis* in the current work further allowed the rooting of this tree, providing a precise estimate of the speciation order of the Aurantioideae subfamily.

The geographical distribution and divergence times of two Aurantioideae clades, one comprising *Aeglopsis*, *Afraegle*, *Balsamocitrus* and *Aegle* and the other comprising *Hesperethusa* and *Citropsis*, indicate that they migrated from Asia to Africa during the last 10 million years, in the Late Miocene. Even though long distance dispersals are considered infrequent (Jordano, 2017), multiple independent dispersions from Asia to other continents have been reported in other plants (Li *et al.*, 2009; Baker and Couvreur, 2013; Huang *et al.*, 2019; Helmstetter *et al.*, 2019), including the Aurantioideae genera (Pfeil and Crisp, 2008; Wu *et al.*, 2018; Nguyen *et al.*, 2019). Taken together, the results suggest that long distance dispersals from Asia to Africa and Oceania were relatively frequent in the case of Aurantioideae, especially during the last ten million years, although the dispersal method remains unknown.

Once the phylogenetic placement of *Citrus* within Aurantioideae was established, the analysis focused on the phylogeny of the species within this genus. The diversification of the genus *Citrus*, according to the phylogeny generated by Wu *et al*. (2018), generated two main clades: one including citrons, pummelos and papedas, and the other including mandarins, *Fortunella* and a clade including the Australian limes. The two Chinese *Citrus* species *Citrus ichangensis* and *Citrus mangshanensis* appeared as sister taxa to the genus

Citrus. Despite the existence of two main clades and two sister taxa, the authors also suggest that most of these species diverged in a rapid radiation, from 8 to 6 million years ago (Wu *et al.*, 2018).

The phylogenetic study here presented includes several *Citrus* species not analyzed in the study mentioned above, as well as the *Clymenia* and *Oxanthera* genera. The results reveal that both genera *Clymenia* and *Oxanthera* are actually nested within the genus *Citrus*, hence expanding the concept of the genus. It is well known that the process of incomplete lineage sorting can significantly interfere with the phylogenetic inference and produce spurious results, especially when studying rapid radiations (Maddison, 1997; Liu *et al.*, 2015; Jiang *et al.*, 2020). To minimize the effects of incomplete lineage sorting, a comprehensive genome-wide phylogeny of the genus *Citrus* was performed, including some disregarded species and using evolutionary models that integrate incomplete lineage sorting as a major source of gene tree discordance. The results indicate that the *Citrus* radiation is statistically indistinguishable from a series of simultaneous speciation events, and thus that the true *Citrus* phylogeny might be better explained as a multifurcating tree at the origin of the crown. Previous studies have revealed the existence of hard polytomies in the base of other plant radiations, where the increase of the amount of data analyzed fails to converge into a unique solution (Carlsen *et al.*, 2018; Koenen *et al.*, 2020; Larson *et al.*, 2020). Given the drastic change in the climatic conditions that occurred in South East Asia in the Late Miocene (Herbert *et al.*, 2016; Holbourn *et al.*, 2018; Tanner *et al.*, 2020), it is hypothesized that the basal polytomy at the *Citrus* crown might represent the true speciation history of the genus, which would in turn explain the historical inconsistencies regarding its phylogeny (Nicolosi *et al.*, 2000; Ramadugu *et al.*, 2013; Oueslati *et al.*, 2016).

The biogeography of the different *Citrus* species in light of the obtained phylogenetic tree was also explored. Several independent dispersal paths for the *Citrus* ancestors inhabiting the Eastern Himalayas are in principle plausible. Most of these imply short distance dispersals through mainland South East Asia and the Sunda plate, which was emerged at that time and might have acted as a land bridge between most of the islands in maritime South East Asia (Morley, 2018). This is compatible with the proposal that the Oceanic *Citrus* arrived first to New Guinea via long distance dispersal from Asia, as previously suggested (Wu *et al.*, 2018), though a dispersal via the Sunda plate cannot be completely

ruled out. The orogeny of the Central Range of New Guinea, which took place during the Late Miocene and Early Pliocene (Hall, 2009), might have created a physical barrier that isolated the northern *Clymenia* from other New Guinean *Citrus*. The arrival of *Citrus* to Australia from New Guinea probably occurred at least twice, with one first event producing the desert limes and a second one giving rise to the Australian limes of the East Coast of Australia, that are more related to the New Guinean limes. Plant and animal exchanges among New Guinea, Australia and Sunda have been, in fact, repeatedly reported (Mitchell *et al.*, 2014; Crayn *et al.*, 2015; Yap *et al.*, 2018; Tallowin *et al.*, 2020). On the other hand, the integration of *Oxanthera* inside *Citrus* has also implications in the biogeography of this genus. Since *Oxanthera* is endemic of New Caledonia, its inclusion in *Citrus* forcibly implies that its arrival to the island must have occurred in the last few million years, via long distance dispersal. Although New Caledonia is considered a refuge for ancient taxa, a recent study revealed that many of the island biota arrived there in the last few million years (Nattier *et al.*, 2017)

Despite the importance of *Citrus* radiation in shaping the current genus diversity, its effects in the genomes of the involved species have not been assessed. Many rapid radiations implied a pervasive positive selection (Kapralov *et al.*, 2013; Nevado *et al.*, 2019), including the case of *Citrus*, where adaptive selection and in tandem gene duplications have contributed to a large degree to shape their genomic space (Gonzalez-Ibeas *et al.*, 2021*a*,*b*, in press). Current *Citrus* species still present an overall highly conserved synteny (He *et al.*, 2020), suggesting that major chromosome rearrangements have not been the main drivers of citrus variability. Mobile elements represent an independent source of diversity that may have effects comparable or even greater than those of SNPs (Sanseverino *et al.*, 2015; Domínguez *et al.*, 2020). Transposons can be activated under stressful conditions (Lee *et al.*, 2017; Benoit *et al.*, 2019), and their insertions can induce profound changes in the genomes of plants, shaping their evolutionary history (Zhang and Gao, 2017; Mascagni *et al.*, 2017). Mobile elements, therefore, play important roles in generating the variability needed during adaptive radiations (Schrader and Schmitz, 2019), as they can produce extreme mutations targeting any kind of genes (Quadrana *et al.*, 2019). In this regard, the retrotransposon landscape in eight different reference genomes including seven *Citrus* species and one species of *Severinia* has been characterized, expanding previous retrotransposon surveys carried out in *Citrus* species (Du *et al.*, 2018; Liu *et al.*, 2019) with the inclusion of all the available

genome sequences in public databases. Given the small size of *Citrus* genomes, retrotransposons are moderately abundant, and accumulate mainly in the pericentromeric regions, as has been reported in many other plants. (Beulé *et al.*, 2015; Anderson *et al.*, 2019). The pericentromeric regions contain low gene densities and show reduced recombination rates, allowing non-detrimental insertions and increased transposon longevity. (Xu and Du, 2014). The estimation of the retrotransposon insertion and excision rates across different genomic regions revealed that the distribution of recent insertions is uniform along the genome of several *Citrus* species, and although similar observations have been made in other plant species, this pattern is not universal (Levin and Moran, 2011; Tsukahara *et al.*, 2012; Nakashima *et al.*, 2018). In contrast, the excision rate was found to be higher in genic regions, an observation possibly related to the elevated recombination rate, as described in tomato (Xu and Du, 2014). Since the increased excision rate in genic regions is not enough to explain the observed differences in transposon abundances, it is postulated that purifying selection against novel insertions in genic regions might contribute to the observed patterns, at least in some of the analyzed species.

Interestingly, the great majority of the *Citrus* retrotransposon families are also present in *Severinia*. This finding may be interpreted as a sign that only a few new families have been acquired *de novo*, as reported in other species (Piednoël *et al.*, 2013), and that therefore the existing families have not changed in the recent past. Despite this conserved diversity, the results here obtained reveal that the retrotransposon accumulation rates were strikingly different among species. In some species this rate grew exponentially while being severely halted in others. A similar observation was made regarding the different retrotransposon lineages: some of them increased their activity over time while in others it peaked at some point in the past and declined since them. Other lineages behaved in a species-dependent way, displaying one of both patterns depending on the host species. Remarkably, the observed differences among both lineages and species started around 5.5 million years ago, that is, shortly after the *Citrus* radiation, suggesting that the activity of these mobile elements may be associated to the process of speciation in citrus.

It is generally accepted that the activity of mobile elements is in part controlled by epigenetic silencing via DNA methylation (Tsukahara *et al.*, 2009; Cheng *et al.*, 2015). Genome-wide epigenetic modifications are in turn linked to environmental stresses (Lira-

Medeiros *et al.*, 2010; Wibowo *et al.*, 2016), an observation recently confirmed in *Citrus* trees growing under drought stress (Neves *et al.*, 2017). In this study, the authors found different responses in the stress methylation patterns of Sunki mandarin, an almost pure mandarin, and Rangpur lime, a direct citron x mandarin hybrid (Wu *et al.*, 2018). These results may imply that, upon abiotic stresses, each *Citrus* species might suffer different epigenomic alterations. Considering the clear links between abiotic stresses, epigenomic variation and transposon activity, it is reasonable to propose that the *Citrus* radiation, the climate changes characterizing the Late Miocene and the retrotransposon activity differences among *Citrus* species are very likely connected. Further studies, especially those focusing on the genus-wide epigenome of *Citrus*, might allow for a causal relation to be established.

The interspecific *Citrus* variability, amplified by the differential retrotransposon activity among species, is responsible for the wide range of phenotypes currently found across *Citrus* wild species. However, like in other tree species (Duan *et al.*, 2017; Julca *et al.*, 2020), the domestication of the genus *Citrus* was profoundly shaped by interspecific crosses that, in parallel with the selection of desirable traits, gave rise to the current admixtures and commercial cultivars. The great phenotypic diversity found within *Citrus* possibly paved the way for these interspecific hybridizations. In order to study how the admixture patterns affected the domestication process, the transcriptomic profiles of ripening fruits from seven different *Citrus* cultivars were analyzed. This way, three *Citrus* pure species (citron, pummelo and mandarin) and four admixtures (sweet orange, sour orange, commercial mandarin and lemon) were selected to capture the diversity within the *Citrus* commercial varieties and assess the effects of the different introgression patterns in modulating gene expression.

Despite the great degree of admixture of the analyzed species, the amount of genes preferentially expressing the allele of one species over the other was relatively low when compared with other plant crops (Springer and Stupar, 2007; Albert *et al.*, 2018; Shao *et al.*, 2019). This suggests that allele-specific expression in citrus does not play an essential role in determining the fruit traits as it does in other fruit crops such as apple (Sun *et al.*, 2020) or tomato (Yuste-Lisbona *et al.*, 2020). In contrast, the distribution of differentially expressed genes across the genome was clearly dependent upon the admixture patterns inherent to each cultivar. For example, the transcriptomic changes between the wild and

domesticated mandarins here studied accumulated in the pummelo introgressions, and especially in two of them which overlap with two genomic regions previously linked with pulp acidity (Wu *et al.*, 2018) and fruit size (Minamikawa *et al.*, 2017). However, in sweet orange, a highly admixed commercial citrus, the transcriptomic differences with the pure parental species in general accumulated in the non-admixed regions of the genome. Specifically, the number of genes in the pummelo/pummelo regions showing differential expression compared with those from pure mandarin was well above average, while in the mandarin/mandarin regions a similar trend was observed compared with pummelo. Again, some of these regions were significantly associated with fruit weight (Minamikawa *et al.*, 2017), a finding that is in agreement with the reported correlation between the percentage of pummelo introgressions in *Citrus* genomes and fruit size (Wu *et al.*, 2018). It is widely accepted that homozygous regions of the genome are possible domestication targets, since this process is generally linked to the loss of genetic diversity, as shown in several annual cereals (Pankin *et al.*, 2018; Maccaferri *et al.*, 2019) and some fruit crops (Cao *et al.*, 2014). However, these regions were scarce in the sweet orange genome, while in the domesticated mandarin, where runs of homozygosity regions were more frequent, increased proportion of genes differentially expressed between wild and domesticated mandarins were neither observed. These results indicate that, despite the great degree of relatedness that exists among commercial mandarins (Wu *et al.*, 2018), the major transcriptomic differences between wild and domesticated mandarins accumulate in the admixed regions of the genome. The role of interspecific hybridization as a major driver in domestication has been thoroughly confirmed in several tree species (Myles *et al.*, 2011; Duan *et al.*, 2017; Flowers *et al.*, 2019; Julca *et al.*, 2020), including *Citrus* (Wu *et al.*, 2018). In this species, the dominant domestication mechanism appears to have been fundamentally dependent on the prevalent asexual propagation of this genus. In this scenario, the desirable traits were very likely obtained via interspecific crosses, including rearranged genomic introgressions, while the improved cultivars were maintained over time either via grafting or apomictic seed dispersal.

Two major determinants of citrus fruit quality such as sweetness and sourness (Lado *et al.*, 2018) were studied by analyzing in detail the expression of genes involved in sugar metabolism and citrate accumulation, respectively. The results of the analyses show that a large number of genes linked to sucrose processing, the glycolytic pathway and the tricarboxylic acid cycle were consistently less expressed in the two extremely acidic

species, citron and lemon, which share a complete haplotype (Curk *et al.*, 2016; Wu *et al.*, 2018). In contrast, most of the subunits composing the vacuolar V-ATPase were highly expressed in these acidic varieties. Higher expression of the P-type vacuolar ATPase *CitPH5*, a pivotal player in controlling citrus fruit acidity (Strazzer *et al.*, 2019), was detected also in the other acidic sample sour orange, in addition to citron and lemon. In general, the expression of vacuolar proton pumps has been associated with an enhanced vacuolar citrate intake and, ultimately, with pulp sourness (Shimada *et al.*, 2006; Shi *et al.*, 2015, 2018; Guo *et al.*, 2016). These results reveal that both the V-ATPase and the *CitPH5* mechanisms are in play in different *Citrus* species, maybe having an additive effect. Furthermore, the overall reduced expression in citron and lemon fruits across most of the sugar-processing genes, especially those linked with sucrose breakdown, correlates with the low sugar accumulation observed in these organs which, together with their high acidity, characterize both species.

The accumulation of carotenoids in the peel, another process of great relevance in *Citrus* also displayed considerable changes among the studied cultivars. The results obtained in this work indicate that the red pigmentation in the peel of citrus fruits, which has been so far assigned mainly to the activity of *CCD4b* (Rodrigo *et al.*, 2013*b*; Zheng *et al.*, 2015), might not only depend upon this gene, as in the lemon flavedo, which does not accumulate red carotenoids, this specific gene displays a high expression comparable to those of sweet oranges or the wild mandarin. On the contrary, the isoform *CCD4a*, usually discarded as a key player in these processes due to its reduced expression in red citrus peels, might be a determinant contributor to citrus fruit color by degrading apocarotenoid precursors into colorless derivatives, as described in other plant species (Yoshioka *et al.*, 2012; Kishimoto *et al.*, 2018; Phadungsawat *et al.*, 2020). These analyses also show that red fruits overexpressed *LCYB*, which can redirect the carbon flux towards the synthesis of β-carotenoids (Zhang *et al.*, 2012; Rodrigo *et al.*, 2013*a*). Several species-specific changes in the expression of many other genes involved in carotenoid accumulation of many plant species, including *Citrus* (Zhang *et al.*, 2009; Gonzalez-Jorge *et al.*, 2016; Rodrigo *et al.*, 2019) were also detected. Taken together, these observations suggest that *Citrus* red peel coloration is not a consequence of a single key gene but rather the result of the combined effect of many independent genes. This statement is supported by the multiple pigmentation mutants reported in *Citrus*, affecting a many different loci (Liu *et al.*, 2007; Alquézar *et al.*, 2008; Alós *et al.*, 2008; Rodrigo *et al.*, 2019; Lana *et al.*, 2020).

The current study also revealed the expression pattern of the chalcone synthase *CHSm*, that catalyzes a rate-limiting step in the flavonoid biosynthesis (Wang *et al.*, 2018*b*). Transcripts of this gene were only found in mandarins but absent in pummelo and citron fruits. Furthermore, in the admixed species only the mandarin haplotype was expressed, while the copy from the other ancestor remained silenced. The analysis of the *CHSm* promotor region in 10 species revealed three different alleles, an archaic one found in *Citrus* outgroups, the pummelo and citron allele, and the mandarin one. Whether the archaic allele can express *CHSm* could not be determined, but the pummelo and citron variant is not expressed, and the mandarin *CHSm* is, in contrast, highly expressed in ripening fruits. The elevated *CHSm* levels found in mandarin correlate with the fact that mandarin peel is rich in flavonoids and flavonoid derivatives, both in terms of compound diversity and total flavonoid concentration (Zhao *et al.*, 2017; Wang *et al.*, 2017*c*). Notably, a wide variety in the expression patterns of flavonoid O-methyltransferases, was also observed, as it had been previously reported in *Citrus* (Liu *et al.*, 2016). These enzymes, which are responsible to a great extent of the flavonoid diversity found in *Citrus*, present a broad substrate specificity (Itoh *et al.*, 2016; Liu *et al.*, 2020*a*), and may therefore generate a wide range of products. The flavonoid O-methyltransferase gene family is expanded in *Citrus* when compared with other plants, and this expansion is more pronounced in *Fortunella*, *Citrus ichangensis* and especially in mandarins (Gonzalez-Ibeas *et al.*, 2021, in press). Thus, the O-methyltransferase expansion might be associated to the appearance of the *CHSm* mandarin allele, since the species carrying this allele are the ones displaying a more pronounced enlargement of this gene family. These results suggest a stepwise evolution process for the flavonoid-rich mandarin flavedo. Initially, the appearance of the *CHSm* mandarin allele might have conferred some adaptive advantage, likely linked to photoprotection, as the accumulation of flavonoids appears to be stimulated upon UV radiation (Sytar *et al.*, 2018; Yamaga and Hamasaki, 2020). Therefore, the species carrying this allele, mandarins, *C. ichangensis* and *Fortunella*, could have migrated to more sun exposed areas eastwards from Yunnan, as has previously proposed (Wu *et al.*, 2018). Finally, the increased availability of flavonoid precursors opened the room for new functions to be explored, which might have triggered the expansion of the O-methyltransferase family. Other similar evolutionary processes have been recently described in *Citrus*, as is the case of the *CCD4b* gene in mandarins (Zheng *et al.*, 2019).

In this doctoral thesis, I have analyzed different aspects of the evolution, diversification and domestication of the genus *Citrus*, expanding and enriching the existing knowledge in the field of *Citrus* genomics by applying an evolutionary view to major relevant processes of the *Citrus* biology. This broader perspective, which takes into consideration the genomic complexity of the members of this genus, allowed me to provide novel and original hypotheses regarding some of major processes that shaped the *Citrus* genome and produced the different wild and cultivated *Citrus* that we enjoy today.

Conclusions

CONCLUSIONS

1. The results suggest that the Aurantioideae subfamily, that includes the genus *Citrus*, emerged during the Early Oligocene 32 Mya and diversified during the Oligocene, with a rapid radiation taking place 25 Mya coinciding with the Oligocene-Miocene boundary. During the Late Miocene, several Aurantioideae clades dispersed by multiple long-distance migrations from Asia to either Africa or Oceania.

2. The *Citrus* phylogeny adjusts more precisely to a multifurcating topology rather than to a strictly binary tree, a vision that implies the occurrence a polytomy at the base of the citrus crown. This suggestion resolves the incongruences presented in previous works and settles the associated debate about the true phylogeny of *Citrus*.

3. The genera *Oxanthera* and *Clymenia* belong to the *Citrus* clade, which enlarges the current boundaries of the genus *Citrus*. The consideration of these genera and other *Citrus* species allowed the generation of the most comprehensive *Citrus* phylogeny presented up to date.

4. The *Citrus* LTR retrotransposon landscape is largely governed by the individual past of each species and can be completely different even among closely related species.

5. *Citrus* retrotransposons may respond to stressful conditions driving speciation as a part of the genetic response involved in adaptation. This proposal implies that the evolving conditions of each species interact with the internal regulatory mechanisms of the genome controlling the proliferation of mobile elements.

6. The study of the fruit transcriptomes of wild and domesticated citrus supports the hypothesis that interspecific hybridizations played a pivotal role during citrus domestication. *Citrus* asexual propagation allowed the expansion and dispersal of the admixed genomes, perpetuating the varieties carrying desirable traits such as reduced acidity or increased peel color.

7. Non-edible acidic wild and domesticated citrus display a consistent overexpression of vacuolar ATPases, which might have been early domestication targets. The results also suggest that *CCD4a* and *LCYb* could be important genes controlling carotenoid content and peel coloration.

8. A *chalcone synthase* gene *CHSm* expressed in mandarins and their admixtures but not in citron and pummelo appears to be related to the accumulation and diversification of flavonoids characterizing the peel of mandarins.

References

REFERENCES

Albert E, Duboscq R, Latreille M, *et al.* 2018. Allele-specific expression and genetic determinants of transcriptomic variations in response to mild water deficit in tomato. The Plant Journal **96**, 635–650.

Aleza P, Cuenca J, Hernández M, Juárez J, Navarro L, Ollitrault P. 2015. Genetic mapping of centromeres in the nine *Citrus clementina* chromosomes using half-tetrad analysis and recombination patterns in unreduced and haploid gametes. BMC Plant Biology **15**, 80.

Alonge M, Wang X, Benoit M, *et al.* 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell **182**, 145-161.e23.

Alós E, Roca M, Iglesias DJ, Mínguez-Mosquera MI, Damasceno CMB, Thannhauser TW, Rose JKC, Talón M, Cercós M. 2008. an evaluation of the basis and consequences of a stay-green mutation in the navel negra citrus mutant using transcriptomic and proteomic profiling and metabolite analysis. Plant Physiology **147**, 1300–1315.

Alquézar B, Rodrigo MJ, Zacarías L. 2008. Regulation of carotenoid biosynthesis during fruit maturation in the red-fleshed orange mutant Cara Cara. Phytochemistry **69**, 1997–2007.

Alquézar B, Zacarías L, Rodrigo MJ. 2009. Molecular and functional characterization of a novel chromoplast-specific *lycopene β-cyclase* from *Citrus* and its relation to lycopene accumulation. Journal of Experimental Botany **60**, 1783–1797.

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel metropolis coupled Markov chain Monte Carlo for bayesian phylogenetic inference. Bioinformatics **20**, 407–415.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biology **11**, R106.

Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, O'Connor CH, Hirsch CD, Ross-Ibarra J, Hirsch CN, Springer NM. 2019. Transposable elements contribute to dynamic genome content in maize. The Plant Journal **100**, 1052–1065.

Aprile A, Federici C, Close TJ, De Bellis L, Cattivelli L, Roose ML. 2011. Expression of the H^+-ATPase AHA10 proton pump is associated with citric acid accumulation in lemon juice sac cells. Functional & Integrative Genomics **11**, 551–563.

Arumuganathan K, Earle E. 1991. Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter **9**, 208–218.

Astrin JJ, Höfer H, Spelda J, *et al.* 2016. Towards a DNA barcode reference database for spiders and harvestmen of Germany. PLOS ONE **11**, e0162624.

Atchadé YF, Roberts GO, Rosenthal JS. 2011. Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo. Statistics and Computing **21**, 555–568.

Aziz RK, Breitbart M, Edwards RA. 2010. Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Research **38**, 4207–4217.

Baker WJ, Couvreur TLP. 2013. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. I. Historical biogeography. Journal of Biogeography **40**, 274–285.

Baldermann S, Kato M, Kurosawa M, Kurobayashi Y, Fujita A, Fleischmann P, Watanabe N. 2010. Functional characterization of a carotenoid cleavage dioxygenase 1 and its relation to the carotenoid accumulation and volatile emission during the floral development of *Osmanthus fragrans* Lour. Journal of Experimental Botany **61**, 2967–2977.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA **6**, 11.

Bapteste E, van Iersel L, Janke A, *et al.* 2013. Networks: expanding evolutionary thinking. Trends in Genetics **29**, 439–441.

Bardil A, Tayalé A, Parisod C. 2015. Evolutionary dynamics of retrotransposons following autopolyploidy in the Buckler Mustard species complex. The Plant Journal **82**, 621–631.

Barrett HC, Rhodes AM. 1976. A numerical taxonomic study of affinity relationships in cultivated citrus and its close relatives. Systematic Botany **1**, 105–136.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genetics **5**, e1000732.

Baxter CJ, Carrari F, Bauke A, Overy S, Hill SA, Quick PW, Fernie AR, Sweetlove LJ. 2005. Fruit carbohydrate metabolism in an introgression line of tomato with increased fruit soluble solids. Plant & Cell Physiology **46**, 425–437.

Bayer RJ, Mabberley DJ, Morton C, Miller CH, Sharma IK, Pfeil BE, Rich S, Hitchcock R, Sykes S. 2009. A molecular phylogeny of the orange subfamily (Rutaceae: Aurantioideae) using nine cpDNA sequences. American Journal of Botany **96**, 668–685.

Beattie GAC, Holford P, Haigh AM, Broadbent P. 2009. On the origins of *Citrus*, huanglongbing, *Diaphorina citri* and *Trioza erytreae*. Proceedings of the international research conference on huanglongbing. Orlando, USA, 23-56.

Beattie GAC, Holford P, Mabberley DJ, Haigh AM, Bayer R, Broadbent P. 2006. Aspects and insights of Australia-Asia collaborative research on huanglongbing. Proceedings of the international workshop for the prevention of citrus greening disease in severely infected areas. Tokyo, Japan, 47–64.

Beguiristain T, Grandbastien MA, Puigdomènech P, Casacuberta JM. 2001. Three *Tnt1* subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant physiology **127**, 212–21.

Benoit M, Drost HG, Catoni M, Gouil Q, Lopez-Gomollon S, Baulcombe D, Paszkowski J. 2019. Environmental and epigenetic regulation of Rider retrotransposons in tomato. PLOS Genetics **15**, e1008370.

Beulé T, Agbessi MD, Dussert S, Jaligot E, Guyot R. 2015. Genome-wide analysis of LTR-retrotransposons in oil palm. BMC Genomics **16**, 795.

Boccacci P, Botta R. 2009. Investigating the origin of hazelnut (*Corylus avellana* L.) cultivars using chloroplast microsatellites. Genetic Resources and Crop Evolution **56**, 851–859.

Boeke JD, Corces VG. 1989. transcription and reverse transcription of retrotransposons. Annual Review of Microbiology **43**, 403–434.

Boeke JD, Garfinkel DJ, Styles CA, Fink GR. 1985. Ty elements transpose through an RNA intermediate. Cell **40**, 491–500.

Bogusz M, Whelan S. 2017. phylogenetic tree estimation with and without alignment: new distance methods and benchmarking. Systematic Biology **66**, 218–231.

Borredá C, Pérez-Román E, Ibanez V, Terol J, Talon M. 2019. Reprogramming of retrotransposon activity during speciation of the genus *Citrus*. Genome Biology and Evolution, 11, 3478–3495.

Botet R, Keurentjes JJB. 2020. The Role of transcriptional regulation in hybrid vigor. Frontiers in Plant Science **11**, 410.

Bouckaert R, Vaughan TG, Barido-Sottani J, *et al.* 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology **15**, e1006650.

Bouillé M, Senneville S, Bousquet J. 2011. Discordant mtDNA and cpDNA phylogenies indicate geographic speciation and reticulation as driving factors for the diversification of the genus *Picea*. Tree Genetics & Genomes **7**, 469–484.

Bousios A, Kourmpetis YAI, Pavlidis P, Minga E, Tsaftaris A, Darzentas N. 2012. The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. The Plant Journal **69**, 475–488.

Brandenburg JT, Mary-Huard T, Rigaill G, Hearne SJ, Corti H, Joets J, Vitte C, Charcosset A, Nicolas SD, Tenaillon MI. 2017. Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLOS Genetics **13**, e1006666.

Brookfield JFY. 2005. The ecology of the genome — mobile DNA elements and their hosts. Nature Reviews Genetics **6**, 128–136.

Brown RM, Siler CD, Oliveros CH, *et al.* 2013. evolutionary processes of diversification in a model island archipelago. Annual Review of Ecology, Evolution, and Systematics **44**, 411–435.

Brune A, Gonzalez P, Goren R, Zehavi U, Echeverria E. 1998. Citrate uptake into tonoplast vesicles from acid lime (*Citrus aurantifolia*) juice cells. Journal of Membrane Biology **166**, 197–203.

Bruun-Lund S, Clement WL, Kjellberg F, Rønsted N. 2017. First plastid phylogenomic study reveals potential cyto-nuclear discordance in the evolutionary history of *Ficus* L. (Moraceae). Molecular Phylogenetics and Evolution **109**, 93–104.

Bryant D, Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution **21**, 255–265.

Budiarto R, Poerwanto R, Santosa E, Efendi D, Agusta A. 2019. production, post-harvest and marketing of kaffir lime (*Citrus hystrix* DC) in Tulungagung, Indonesia. Journal of Tropical Crop Science **6**, 138–143.

Bureau TE, Wessler SR. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. The Plant Cell **4**, 1283–1294.

Butelli E, Garcia-Lor A, Licciardello C, *et al.* 2017. changes in anthocyanin production during domestication of *Citrus*. Plant physiology **173**, 2225–2242.

Butelli E, Licciardello C, Ramadugu C, Durand-Hulak M, Celant A, Reforgiato Recupero G, Froelicher Y, Martin C. 2019. Noemi controls production of flavonoid pigments and fruit acidity and illustrates the domestication routes of modern *Citrus* Varieties. Current Biology **29**, 158-164.e2.

Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. 2012. retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. The Plant Cell **24**, 1242–1255.

Cai M, Lin J, Li Z, Lin Z, Ma Y, Wang Y, Ming R. 2020. Allele specific expression of Dof genes responding to hormones and abiotic stresses in sugarcane. PLOS ONE **15**, e0227716.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics **10**, 421.

Cao K, Zheng Z, Wang L, *et al.* 2014. Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. Genome Biology **15**, 415.

Carbonell-Caballero J, Alonso R, Ibanez V, Terol J, Talon M, Dopazo J. 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus*. Molecular Biology and Evolution **32**, 2015–2035.

Carlsen MM, Fér T, Schmickl R, Leong-Škorničková J, Newman M, Kress WJ. 2018. Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: pushing the limits of genomic data. Molecular Phylogenetics and Evolution **128**, 55–68.

Carpentier M-C, Manfroi E, Wei F-J, Wu H-P, Lasserre E, Llauro C, Debladis E, Akakpo R, Hsing Y-I, Panaud O. 2019. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nature Communications **10**, 24.

Carrier G, Cunff LL, Dereeper A, Legrand D, Sabot F, Bouchez O, Audeguin L, Boursiquot JM, This P. 2012. Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. PLOS ONE **7**, e32973.

Caruso M, Las Casas G, Scaglione D*, et al.* 2019. Detection of natural and induced mutations from next generation sequencing data in sweet orange bud sports. Acta Horticulturae, 119–124.

Caruso M, Smith MW, Froelicher Y, Russo G, Gmitter FG. 2020. Chapter 7 - Traditional breeding. The Genus Citrus. Woodhead Publishing, 129–148.

Castle WS. 2010. A career perspective on citrus rootstocks, their development, and commercialization. HortScience **45**, 11–15.

Cercós M, Soler G, Iglesias DJ, Gadea J, Forment J, Talón M. 2006. Global analysis of gene expression during development and ripening of citrus fruit flesh. A Proposed Mechanism for Citric Acid Utilization. Plant Molecular Biology **62**, 513–527.

Chaimanee P, Suntornwat O. 1994. Changes in carbohydrate content during fruit ripening - a new approach of teaching of carbohydrate chemistry in biochemistry course. Biochemical Education **22**, 101–102.

Chaudhary PR, Bang H, Jayaprakasha GK, Patil BS. 2016. Variation in key flavonoid biosynthetic enzymes and phytochemicals in 'Rio Red' grapefruit (*Citrus paradisi* Macf.) during fruit development. Journal of Agricultural and Food Chemistry **64**, 9022–9032.

Chávez-González ML, López-López LI, Rodríguez-Herrera R, Contreras-Esquivel JC, Aguilar CN. 2016. Enzyme-assisted extraction of citrus essential oil. Chemical Papers **70**, 412–417.

Chen Q, Wang D, Tan C, Hu Y, Sundararajan B, Zhou Z. 2020. Profiling of flavonoid and antioxidant activity of fruit tissues from 27 chinese local citrus cultivars. Plants **9**, 196.

Chen M, Xie X, Lin Q, Chen J, Grierson D, Yin X, Sun C, Chen K. 2013. Differential expression of organic acid degradation-related genes during fruit development of navel oranges (*Citrus sinensis*) in two habitats. Plant Molecular Biology Reporter **31**, 1131–1140.

Cheng C, Tarutani Y, Miyao A, Ito T, Yamazaki M, Sakai H, Fukai E, Hirochika H. 2015. Loss of function mutations in the rice chromomethylase *OsCMT3a* cause a burst of transposition. The Plant Journal **83**, 1069–1081.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature Methods **12**, 966–968.

Cirmi S, Maugeri A, Ferlazzo N, Gangemi S, Calapai G, Schumacher U, Navarra M. 2017. Anticancer potential of *Citrus* juices and their extracts: a systematic review of both preclinical and clinical studies. Frontiers in Pharmacology **8**, 420.

Civáň P, Craig H, Cox CJ, Brown TA. 2015. Three geographically separate domestications of Asian rice. Nature Plants **1**, 15164.

Clift PD, Wan S, Blusztajn J. 2014. Reconstructing chemical weathering, physical erosion and monsoon intensity since 25Ma in the northern South China Sea: a review of competing proxies. Earth-Science Reviews **130**, 86–102.

Cornejo OE, Yee MC, Dominguez V, *et al.* 2018. Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. Communications Biology **1**, 167.

Cossu RM, Casola C, Giacomello S, Vidalis A, Scofield DG, Zuccolo A. 2017. LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. Genome Biology and Evolution **9**, 3449–3462.

Crayn DM, Costion C, Harrington MG. 2015. The Sahul–Sunda floristic exchange: dated molecular phylogenies document Cenozoic intercontinental dispersal dynamics. Journal of Biogeography **42**, 11–24.

Curk F, Ancillo G, Garcia-Lor A, Luro F, Perrier X, Jacquemoud-Collet JP, Navarro L, Ollitrault P. 2014. Next generation haplotyping to decipher nuclear genomic interspecific admixture in *Citrus* species: analysis of chromosome 2. BMC Genetics **15**, 152.

Curk F, Ollitrault F, Garcia-Lor A, Luro F, Navarro L, Ollitrault P. 2016. Phylogenetic origin of limes and lemons revealed by cytoplasmic and nuclear markers. Annals of Botany **117**, 565–583.

Da Conceicao Neta ER, Johanningsmeier SD, McFeeters RF, Neta ERDC, Johanningsmeier SD, McFeeters RF. 2007. The chemistry and physiology of sour taste - a review. Journal of Food Science **72**, 33–38.

Dai X, Wang H, Zhou H, Wang L, Dvořák J, Bennetzen JL, Müller HG. 2018. Birth and death of LTR-retrotransposons in *Aegilops tauschii*. Genetics **210**, 1039–1051.

Danecek P, Bonfield JK, Liddle J, *et al.* 2021. Twelve years of SAMtools and BCFtools. GigaScience **10**, giab008.

Dao TTH, Linthorst HJM, Verpoorte R. 2011. Chalcone synthase and its functions in plant resistance. Phytochemistry Reviews **10**, 397–412.

De Felice B. 2009. Transposable sequences in *Citrus* genome: role of mobile elements in the adaptation to stressful environments. Tree and Forestry Science and Biotechnology **3**, 79–86.

De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK. 2017. contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. Molecular Biology and Evolution **34**, 1363–1377.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics **2**, e68.

Deng X, Yang X, Yamamoto M, Biswas MK. 2020. Chapter 3 - Domestication and history. The Genus Citrus. Woodhead Publishing, 33–55.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome research **12**, 1075–1079.

Dickinson HG, Hiscock SJ, Crane PR, Rieseberg LH, Widmer A, Arntz AM, Burke B. 2003. The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. Philosophical Transactions of the Royal Society of London B **358**, 1141–1147.

Dillenberger MS, Kadereit JW. 2017. Simultaneous speciation in the European high mountain flowering plant genus *Facchinia* (*Minuartia* s.l., Caryophyllaceae) revealed by genotyping-by-sequencing. Molecular Phylogenetics and Evolution **112**, 23–35.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. Sequence analysis STAR: ultrafast universal RNA-seq aligner. **29**, 15–21.

Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, Colot V, Quadrana L. 2020. The impact of transposable elements on tomato diversity. Nature Communications **11**, 4058.

Dong J, Kergoat GJ, Vicente N, Rahmadi C, Xu S, Robillard T. 2018. Biogeographic patterns and diversification dynamics of the genus *Cardiodactylus* Saussure (Orthoptera, Grylloidea, Eneopterinae) in Southeast Asia. Molecular Phylogenetics and Evolution **129**, 1–14.

Dong Y, Wang YZ. 2015. Seed shattering: from models to crops. Frontiers in Plant Science **6**, 476.

Doronina L, Churakov G, Kuritzin A, Shi J, Baertsch R, Clawson H, Schmitz J. 2017. Speciation network in Laurasiatheria: retrophylogenomic signals. Genome research **27**, 997–1003.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLOS Biology **4**, e88.

Du D, Du X, Mattia MR, Wang Y, Yu Q, Huang M, Yu Y, Grosser JW, Gmitter FG. 2018. LTR retrotransposons from the *Citrus x clementina* genome: characterization and application. Tree Genetics & Genomes **14**, 43.

Du C, Fefelova N, Caronna J, He L, Dooner HK. 2009. The polychromatic Helitron landscape of the maize genome. Proceedings of the National Academy of Sciences **106**, 19916–19921.

Duan N, Bai Y, Sun H, *et al.* 2017. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. Nature Communications **8**, 249.

Dubin MJ, Mittelsten Scheid O, Becker C. 2018. Transposons: a blessing curse. Current Opinion in Plant Biology **42**, 23–29.

Ecker JR, Davis RW. 1987. Plant defense genes are regulated by ethylene. Proceedings of the National Academy of Sciences of the United States of America **84**, 5202–5206.

El Baidouri M, Carpentier -C, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. Genome research **24**, 831–838.

El-Kereamy A, Chervin C, Roustan JP, *et al.* 2003. Exogenous ethylene stimulates the long-term expression of genes related to anthocyanin biosynthesis in grape berries. Physiologia Plantarum **119**, 175–182.

Elkhatim KAS, Elagib RAA, Hassan AB. 2018. Content of phenolic compounds and vitamin C and antioxidant activity in wasted parts of Sudanese citrus fruits. Food Science & Nutrition **6**, 1214–1219.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9**, 18.

Estep MC, DeBarry JD, Bennetzen JL. 2013. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity **110**, 194–204.

Fang D, Krueger RR, Roose ML. 1998. Phylogenetic relationships among selected *Citrus* germplasm accessions revealed by inter-simple sequence repeat (ISSR) markers. Journal of the American Society for Horticultural Science **123**, 612–617.

FAO. 2016. Citrus fruit fresh and processed - Statistical Bulletin. Rome, Italy: Food and Agriculture Organization of the United Nations.

FAO. 2021. FAOSTAT statistical database. Rome, Italy: Food and Agriculture Organization of the United Nations.

Favre A, Päckert M, Pauls SU, Jähnig SC, Uhl D, Michalak I, Muellner-Riehl AN. 2015. The role of the uplift of the Qinghai-Tibetan Plateau for the evolution of Tibetan biotas. Biological Reviews **90**, 236–253.

Fellers, Paul J. 1991. The relationship between the ratio of degrees Brix to percent acid and sensory flavor in grapefruit juice. The relationship between the ratio of degrees Brix to percent acid and sensory flavor in grapefruit juice **45**, 68–75.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. Trends in Genetics **5**, 103–107.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytologist **183**, 557–564.

Flores-Ortiz C, Alvarez LM, Undurraga A, Arias D, Durán F, Wegener G, Stange C. 2020. Differential role of the two ζ-carotene desaturase paralogs in carrot (*Daucus carota*): *ZDS1* is a functional gene essential for plant development and carotenoid synthesis. Plant Science **291**, 110327.

Flowers JM, Hazzouri KM, Gros-Balthazard M, *et al.* 2019. Cross-species hybridization and the origin of North African date palms. Proceedings of the National Academy of Sciences **116**, 1651–1658.

Forster P. 2000. *Clausena smyrelliana* (Rutaceae: Aurantioideae), a new and critically endangered species from south-east Queensland. Austrobaileya **5**, 715–720.

Forster PI, Smith MW. 2010. *Citrus wakonai* P.I.Forst. & M.W.Sm. (Rutaceae), a new species from Goodenough Island, Papua New Guinea. Austrobaileya **8**, 133–138.

Froelicher Y, Mouhaya W, Bassene JB, Costantino G, Kamiri M, Luro F, Morillon R, Ollitrault P. 2011. New universal mitochondrial PCR markers reveal new information on maternal citrus phylogeny. Tree Genetics & Genomes **7**, 49–61.

Fu Y, Kawabe A, Etcheverry M, Ito T, Toyoda A, Fujiyama A, Colot V, Tarutani Y, Kakutani T. 2013. Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. The EMBO Journal **32**, 2407–2417.

Fujino K, Sekiguchi H, Kiguchi T. 2005. Identification of an active transposon in intact rice plants. Molecular genetics and genomics: MGG **273**, 150–157.

Fuller DQ, Denham T, Arroyo-Kalin M, Lucas L, Stevens CJ, Qin L, Allaby RG, Purugganan MD. 2014. Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. Proceedings of the National Academy of Sciences **111**, 6147–6152.

Gao C, Xiao M, Ren X, Hayward A, Yin J, Wu L, Fu D, Li J. 2012. Characterization and functional annotation of nested transposable elements in eukaryotic genomes. Genomics **100**, 222–230.

Garcia-Lor A, Luro F, Ollitrault P, Navarro L. 2015. Genetic diversity and population structure analysis of mandarin germplasm by nuclear, chloroplastic and mitochondrial markers. Tree Genetics & Genomes **11**, 123.

Gaut BS, Díez CM, Morrell PL. 2015. Genomics and the contrasting dynamics of annual and perennial domestication. Trends in Genetics **31**, 709–719.

Göbel U, Arce AL, He F, Rico A, Schmitz G, de Meaux J. 2018. Robustness of transposable element regulation but no genomic shock observed in interspecific *Arabidopsis* hybrids. Genome biology and evolution **10**, 1403–1415.

Goff SA, Ricke D, Lan TH, *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296**, 92–100.

Gonzalez-Ibeas D, Ibanez V, Perez-Roman E, Borredá C, Terol J, Talon M. 2021a. Shaping the biology of *Citrus* I: genomic determinants of evolution. The plant genome, in press.

Gonzalez-Ibeas D, Ibanez V, Perez-Roman E, Borredá C, Terol J, Talon M. 2021b. Shaping the biology of *Citrus* II: genomic determinants of evolution. The plant genome, in press.

Gonzalez-Jorge S, Mehrshahi P, Magallanes-Lundback M, Lipka AE, Angelovici R, Gore MA, DellaPenna D. 2016. *ZEAXANTHIN EPOXIDASE* activity potentiates carotenoid degradation in maturing seed. Plant Physiology **171**, 1837–1851.

González-Mas MC, Rambla JL, López-Gresa MP, Amparo Blázquez M, Granell A. 2019. Volatile compounds in citrus essential oils: a comprehensive review. Frontiers in Plant Science **10**, 12.

Gottwald TR, Graça JV da, Bassanezi RB. 2007. Citrus Huanglongbing: the pathogen and its impact. Plant Health Progress **8**, 31.

Greenblatt IM, Alexander Brink R. 1963. Transpositions of modulator in maize into divided and undivided chromosome segments. Nature **197**, 412–413.

Gros-Balthazard M, Besnard G, Sarah G, Holtz Y, Leclercq J, Santoni S, Wegmann D, Glémin S, Khadari B. 2019. Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. The Plant Journal **100**, 143–157.

Gulsen O, Roose ML. 2001. Chloroplast and nuclear genome analysis of the parentage of lemons. Journal of the American Society for Horticultural Science **126**, 210–215.

Guo C, Spinelli M, Ye C, Li QQ, Liang C. 2017. Genome-wide comparative analysis of miniature inverted repeat transposable elements in 19 *Arabidopsis thaliana* ecotype Accessions. Scientific Reports **7**, 2634.

Guo LX, Shi CY, Liu X, Ning DY, Jing LF, Yang H, Liu YZ. 2016. Citrate accumulation-related gene expression and/or enzyme activity analysis combined with metabolomics provide a novel insight for an orange mutant. Scientific Reports **6**, 29343.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution **70**, 7–17.

Hall R. 2009. Southeast Asia's changing palaeogeography. Blumea - Biodiversity, Evolution and Biogeography of Plants **54**, 148–161.

Hall R. 2012. Sundaland and Wallacea: geology, plate tectonics and palaeogeography. Biotic Evolution and Environmental Change in Southeast Asia. Cambridge University Press, 32–78.

Hamabata T, Kinoshita G, Kurita K, Cao PL, Ito M, Murata J, Komaki Y, Isagi Y, Makino T. 2019. Endangered island endemic plants have vulnerable genomes. Communications Biology **2**, 244.

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu SH, Jiang N. 2009. The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. The Plant Cell **21**, 25–38.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution **22**, 160–174.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proceedings of the National Academy of Sciences **106**, 17811–17816.

He L, Zhao H, He J, Yang Z, Guan B, Chen K, Hong Q, Wang J, Liu J, Jiang J. 2020. Extraordinarily conserved chromosomal synteny of *Citrus* species revealed by chromosome-specific painting. The Plant Journal **103**, 2225–2235.

Helmstetter AJ, Buggs RJA, Lucas SJ. 2019. Repeated long-distance dispersal and convergent evolution in hazel. Scientific Reports **9**, 16016.

Herbert TD, Lawrence KT, Tzanova A, Peterson LC, Caballero-Gill R, Kelly CS. 2016. Late Miocene global cooling and the rise of modern ecosystems. Nature Geoscience **9**, 843–847.

Hernández-Pinzón I, Cifuentes M, Hénaff E, Santiago N, Espinás ML, Casacuberta JM. 2012. The Tnt1 retrotransposon escapes silencing in tobacco, its natural host. PLoS ONE **7**, e33816.

Hodkinson TR, Chonghaile GN, Sungkaew S, Chase MW, Salamin N, Stapleton CMA. 2010. Phylogenetic analyses of plastid and nuclear DNA sequences indicate a rapid late Miocene radiation of the temperate bamboo tribe Arundinarieae (Poaceae, Bambusoideae). Plant Ecology & Diversity **3**, 109–120.

Holbourn AE, Kuhnt W, Clemens SC, Kochhann KGD, Jöhnck J, Lübbers J, Andersen N. 2018. Late Miocene climate cooling and intensification of southeast Asian winter monsoon. Nature Communications **9**, 1584.

Hu TT, Pattyn P, Bakker EG, *et al.* 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nature Genetics **43**, 476–481.

Hu XM, Shi CY, Liu X, Jin LF, Liu YZ, Peng SA. 2015. Genome-wide identification of citrus ATP-citrate lyase genes and their transcript analysis in fruits reveals their possible role in citrate utilization. Molecular Genetics and Genomics **290**, 29–38.

Huang X, Deng T, Moore MJ, Wang H, Li Z, Lin N, Yusupov Z, Tojibaev KSh, Wang Y, Sun H. 2019. Tropical Asian Origin, boreotropical migration and long-distance dispersal in Nettles (Urticeae, Urticaceae). Molecular Phylogenetics and Evolution **137**, 190–199.

Huang JF, Li L, van der Werff H, Li HW, Rohwer JG, Crayn DM, Meng HH, van der Merwe M, Conran JG, Li J. 2016*a*. Origins and evolution of cinnamon and camphor: A phylogenetic and historical biogeographical analysis of the *Cinnamomum* group (Lauraceae). Molecular Phylogenetics and Evolution **96**, 33–44.

Huang D, Zhao Y, Cao M, Qiao L, Zheng ZL. 2016*b*. Integrated systems biology analysis of transcriptomes reveals candidate genes for acidity control in developing fruits of sweet orange (*Citrus sinensis* L. Osbeck). Frontiers in Plant Science **7**, 486.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution **23**, 254–267.

Hussain SB, Guo LX, Shi CY, Khan MA, Bai YX, Du W, Liu YZ. 2020. Assessment of sugar and sugar accumulation-related gene expression profiles reveal new insight into the formation of low sugar accumulation trait in a sweet orange (*Citrus sinensis*) bud mutant. Molecular Biology Reports **47**, 2781–2791.

Hussain SB, Shi CY, Guo LX, Kamran HM, Sadka A, Liu YZ. 2017. Recent advances in the regulation of citric acid metabolism in citrus fruit. Critical Reviews in Plant Sciences **36**, 241–256.

Ikegaya A, Toyoizumi T, Ohba S, Nakajima T, Kawata T, Ito S, Arai E. 2019. Effects of distribution of sugars and organic acids on the taste of strawberries. Food Science & Nutrition **7**, 2419–2426.

Ikoma Y, Matsumoto H, Kato M. 2014. The characteristics of carotenoid biosynthesis in citrus fruit. Japan Agricultural Research Quarterly **48**, 9–16.

Itoh N, Iwata C, Toda H. 2016. Molecular cloning and characterization of a flavonoid-O-methyltransferase with broad substrate specificity and regioselectivity from *Citrus depressa*. BMC Plant Biology **16**, 180.

Jacob-Wilk D, Holland D, Goldschmidt EE, Riov J, Eyal Y. 1999. Chlorophyll breakdown by chlorophyllase: isolation and functional expression of the *Chlase1* gene from ethylene-treated *Citrus* fruit and its regulation during development. The Plant Journal **20**, 653–661.

James BT, Luczak BB, Girgis HZ. 2018. MeShClust: an intelligent tool for clustering DNA sequences. Nucleic Acids Research **46**, e83.

Jiang X, Edwards SV, Liu L. 2020. The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. Systematic Biology **69**, 795–812.

Jiao Y, Leebens-Mack J, Ayyampalayam S, *et al.* 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biology **13**, R3.

Jo C, Kim S. 2020. Transposition of a non-autonomous DNA transposon in the gene coding for a bHLH transcription factor results in a white bulb color of onions (*Allium cepa* L.). Theoretical and Applied Genetics **133**, 317–328.

Jordano P. 2017. What is long-distance dispersal? And a taxonomy of dispersal events. Journal of Ecology **105**, 75–84.

Joyce EM, Pannell CM, Rossetto M, Yap JYS, Thiele KR, Wilson PD, Crayn DM. 2021. Molecular phylogeography reveals two geographically and temporally separated floristic exchange tracks between Southeast Asia and northern Australia. Journal of Biogeography **48**, 1213-1227.

Julca I, Marcet-Houben M, Cruz F, Gómez-Garrido J, Gaut BS, Díez CM, Gut IG, Alioto TS, Vargas P, Gabaldón T. 2020. Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.). BMC Biology **18**, 148.

Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. Bioinformatics **26**, 1669–1670.

Kang C, Zhai H, Xue L, Zhao N, He S, Liu Q. 2018. A lycopene β-cyclase gene, *IbLCYB2*, enhances carotenoid contents and abiotic stress tolerance in transgenic sweetpotato. Plant Science **272**, 243–254.

Kapralov MV, Votintseva AA, Filatov DA. 2013. Molecular adaptation during a rapid adaptive radiation. molecular biology and evolution **30**, 1051–1059.

Kato M, Ikoma Y, Matsumoto H, Sugiura M, Hyodo H, Yano M. 2004. Accumulation of carotenoids and expression of carotenoid biosynthetic genes during maturation in citrus fruit. Plant Physiology **134**, 824–837.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution **30**, 772–780.

Katz LS, Griswold T, Morrison SS, Caravas JA, Zhang S, Bakker HC den, Deng X, Carleton HA. 2019. Mashtree: a rapid comparison of whole genome sequence files. Journal of Open Source Software **4**, 1762.

Katz E, Hwan Boo K, Youn Kim H, Eigenheer RA, Phinney BS, Shulaev V, Negre-Zakharov F, Sadka A, Blumwald E. 2011. Label-free shotgun proteomics and metabolite analysis reveal a significant metabolic shift during citrus fruit development. Journal of Experimental Botany **62**, 5367–5384.

Keidar D, Bariah I, Kashkush K. 2018. Genome-wide analyses of miniature inverted-repeat transposable elements reveals new insights into the evolution of the *Triticum-Aegilops* group. PloS one **13**, e0204972.

Kim S, Park J, Yeom SI, *et al.* 2017. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biology **18**, 210.

Kimura Y, Tosa Y, Shimada S, Sogo R, Kusaba M, Sunaga T, Betsuyaku S, Eto Y, Nakayashiki H, Mayama S. 2001. OARE-1, a Ty1-copia retrotransposon in oat activated by abiotic and biotic stresses. Plant and Cell Physiology **42**, 1345–1354.

Kishimoto S, Oda-Yamamizo C, Ohmiya A. 2018. Regulation of carotenoid pigmentation in corollas of petunia. Plant Molecular Biology Reporter **36**, 632–642.

Kitazumi A, Pabuayon ICM, Ohyanagi H, *et al.* 2018. Potential of *Oryza officinalis* to augment the cold tolerance genetic mechanisms of *Oryza sativa* by network complementation. Scientific Reports **8**, 16346.

Koenen EJM, Kidner C, Souza ÉR de, *et al.* 2020. Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the mimosoid legumes and reveals the polytomous origins of a large pantropical radiation. American Journal of Botany **107**, 1710–1735.

Koenig D, Jiménez-Gómez JM, Kimura S, *et al.* 2013. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proceedings of the National Academy of Sciences of the United States of America **110**, E2655-2662.

Komatsu A, Moriguchi T, Koyama K, Omura M, Akihama T. 2002. Analysis of sucrose synthase genes in citrus suggests different roles and phylogenetic relationships. Journal of Experimental Botany **53**, 61–71.

Komatsu A, Takanokura Y, Moriguchi T, Omura M, Akihama T. 1999. Differential expression of three sucrose-phosphate synthase isoforms during sucrose accumulation in citrus fruits (*Citrus unshiu* Marc.). Plant Science **140**, 169–178.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology **56**, 17–24.

Kuhn N, Guan L, Dai ZW, Wu BH, Lauvergeat V, Gomès E, Li SH, Godoy F, Arce-Johnson P, Delrot S. 2014. Berry ripening: recently heard through the grapevine. Journal of Experimental Botany **65**, 4543–4559.

Kuritzin A, Kischka T, Schmitz J, Churakov G. 2016. Incomplete lineage sorting and hybridization statistics for large-scale retroposon insertion data. PLOS Computational Biology **12**, e1004812.

Lado J, Gambetta G, Zacarias L. 2018. Key determinants of citrus fruit quality: metabolites and main changes during maturation. Scientia Horticulturae **233**, 238–248.

Lana G, Zacarias-Garcia J, Distefano G, Gentile A, Rodrigo MJ, Zacarias L. 2020. Transcriptional analysis of carotenoids accumulation and metabolism in a pink-fleshed lemon mutant. Genes **11**, 1294.

Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. Trends in Ecology & Evolution **29**, 33–41.

Larson DA, Walker JF, Vargas OM, Smith SA. 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. American Journal of Botany **107**, 773–789.

Larson G, Piperno DR, Allaby RG, *et al.* 2014. Current perspectives and the future of domestication studies. Proceedings of the National Academy of Sciences **111**, 6139–6146.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biology **15**, R84.

Lee J, Waminal NE, Choi HI, Perumal S, Lee SC, Nguyen VB, Jang W, Kim NH, Gao L, Yang TJ. 2017. Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus *Panax*. Scientific Reports **7**, 9045.

Lee-Yaw JA, Grassa CJ, Joly S, Andrew RL, Rieseberg LH. 2018. An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). New Phytologist, 12.

Legrand S, Caron T, Maumus F, *et al.* 2019. Differential retention of transposable element-derived sequences in outcrossing *Arabidopsis* genomes. Mobile DNA **10**, 30.

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. 2002. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Research **30**, 325–327.

Levin HL, Moran J V. 2011. Dynamic interactions between transposable elements and their hosts. Nature Reviews Genetics **12**, 615–627.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li Y, Dressler S, Zhang D, Renner SS. 2009. More miocene dispersal between Africa and Asia - the case of *Bridelia* (Phyllanthaceae). Systematic Botany **34**, 521–529.

Li M, Li D, Feng F, Zhang S, Ma F, Cheng L. 2016. Proteomic analysis reveals dynamic regulation of fruit development and sugar and acid accumulation in apple. Journal of Experimental Botany **67**, 5145–5157.

Li X, Liu L, Ming M, Hu H, Zhang M, Fan J, Song B, Zhang S, Wu J. 2019*a*. Comparative transcriptomic analysis provides insight into the domestication and improvement of pear (*P. pyrifolia*) fruit. Plant Physiology **180**, 435–452.

Li HT, Yi TS, Gao LM, *et al.* 2019*b*. Origin of angiosperms and the puzzle of the Jurassic gap. Nature Plants **5**, 461–470.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**, 923–930.

Lim TK. 2012. *Citrus wintersii*. Edible medicinal and non-medicinal plants: Volume 4, Fruits. Springer Netherlands, 739–741.

Lin Q, Wang C, Dong W, Jiang Q, Wang D, Li S, Chen M, Liu C, Sun C, Chen K. 2015. Transcriptome and metabolome analyses of sugar and organic acid metabolism in Ponkan (*Citrus reticulata*) fruit during fruit maturation. Gene **554**, 64–74.

Lindenbaum P, Redon R. 2018. bioalcidae, samjs and vcffilterjs: object-oriented formatters and filters for bioinformatics files. Bioinformatics **34**, 1224–1225.

Lira-Medeiros CF, Parisod C, Fernandes RA, Mata CS, Cardoso MA, Ferreira PCG. 2010. Epigenetic variation in mangrove plants occurring in contrasting natural environment. PloS One **5**, e10326.

Liu GF, He SW, Li WB. 1990. Two new species of *Citrus* in China. Acta Botanica Yunnanica **12**, 287–289.

Liu X, Luo Y, Wu H, Xi W, Yu J, Zhang Q, Zhou Z. 2016. Systematic analysis of O-methyltransferase gene family and identification of potential members involved in the formation of O-methylated flavonoids in *Citrus*. Gene **575**, 458–472.

Liu Y, Tahir ul Qamar M, Feng JW, Ding Y, Wang S, Wu G, Ke L, Xu Q, Chen L-L. 2019. Comparative analysis of miniature inverted–repeat transposable elements (MITEs) and long terminal repeat (LTR) retrotransposons in six Citrus species. BMC Plant Biology **19**, 140.

Liu X, Wang Y, Chen Y, Xu S, Gong Q, Zhao C, Cao J, Sun C. 2020*a*. Characterization of a flavonoid 3’/5’/7-O-methyltransferase from *Citrus reticulata* and evaluation of the *in vitro* cytotoxicity of its methylated products. Molecules **25**, 858.

Liu L, Xi Z, Davis CC. 2015. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. Molecular Biology and Evolution **32**, 791–805.

Liu Q, Xu J, Liu Y, Zhao X, Deng X, Guo L, Gu J. 2007. A novel bud mutation that confers abnormal patterns of lycopene accumulation in sweet orange fruit (*Citrus sinensis* L. Osbeck). Journal of Experimental Botany **58**, 4161–4171.

Liu D, Yang L, Zhang J, *et al.* 2020*b*. Domestication and breeding changed tomato fruit transcriptome. Journal of Integrative Agriculture **19**, 120–132.

Llorens C, Futami R, Covelli L, *et al.* 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research **39**, D70–D74.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**, 550.

Lu X, Cao X, Li F, Li J, Xiong J, Long G, Cao S, Xie S. 2016. Comparative transcriptome analysis reveals a global insight into molecular processes regulating citrate accumulation in sweet orange (*Citrus sinensis*). Physiologia Plantarum **158**, 463–482.

Lu L, Chen J, Robb SMC, Okumoto Y, Stajich JE, Wessler SR. 2017*a*. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. Proceedings of the National Academy of Sciences **114**, E10550–E10559.

Lu PJ, Wang CY, Yin TT, Zhong SL, Grierson D, Chen KS, Xu CJ. 2017*b*. Cytological and molecular characterization of carotenoid accumulation in normal and high-lycopene mutant oranges. Scientific Reports **7**, 761.

Lu S, Zhang Y, Zhu K, Yang W, Ye J, Chai L, Xu Q, Deng X. 2018. The citrus transcription factor *CsMADS6* modulates carotenoid metabolism by directly regulating carotenogenic genes. Plant Physiology **176**, 2657–2676.

Luro F, Curk F, Froelicher Y, Ollitrault P. 2018. Recent insights on *Citrus* diversity and phylogeny. AGRUMED: Archaeology and history of citrus fruit in the Mediterranean: acclimatization, diversifications, uses. Naples: Publications du Centre Jean Bérard.

Lux PE, Carle R, Zacarías L, Rodrigo MJ, Schweiggert RM, Steingass CB. 2019. Genuine carotenoid profiles in sweet orange [*Citrus sinensis* (L.) Osbeck cv. Navel] peel and pulp at different maturity stages. Journal of Agricultural and Food Chemistry **67**, 13164–13175.

Lyu H, He Z, Wu CI, Shi S. 2018. Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. New Phytologist **217**, 428–438.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences **101**, 12404–12410.

Mabberley D. 1998. Australian Citreae with notes on other Aurantioideae (Rutaceae). Telopea **7**, 333–344.

Maccaferri M, Harris NS, Twardziok SO, *et al.* 2019. Durum wheat genome highlights past domestication signatures and future improvement targets. Nature Genetics **51**, 885–895.

Macko-Podgórni A, Stelmach K, Kwolek K, Grzebelus D. 2019. Stowaway miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. Mobile DNA **10**, 47.

Maddi T, Pérez-Román E, Maiza-Benabdesselam F, Khettal B, Talon M, Ibanez-Gonzalez V. 2018. New *Citrus* chloroplast haplotypes revealed by molecular markers using Algerian and Spanish accessions. Genetic Resources and Crop Evolution **65**, 2199–2214.

Maddison W. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics **5**, 365–377.

Maddison WP. 1997. Gene trees in species trees. Systematic Biology **46**, 523–536.

Magalhães DM, Scholte LLS, Silva NV, Oliveira GC, Zipfel C, Takita MA, De Souza AA. 2016. LRR-RLK family from two *Citrus* species: genome-wide identification and evolutionary aspects. BMC Genomics **17**, 623.

Mao D, Liu T, Xu C, Li X, Xing Y. 2011. Epistasis and complementary gene action adequately account for the genetic bases of transgressive segregation of kilo-grain weight in rice. Euphytica **180**, 261–271.

Mao H, Wang H, Liu S, Li Z, Yang X, Yan J, Li J, Tran LSP, Qin F. 2015. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nature Communications **6**, 8326.

MAPA. 2021. Superficies y producciones anuales de cultivo. Madrid, Spain: Ministerio de Agricultura, Pesca y Alimentación.

Martinez-Godoy MA, Mauri N, Juarez J, Marques MC, Santiago J, Forment J, Gadea J. 2008. A genome-wide 20 K citrus microarray for gene expression analysis. BMC Genomics **9**, 318.

Mascagni F, Giordani T, Ceccarelli M, Cavallini A, Natali L. 2017. Genome-wide analysis of LTR-retrotransposon diversity and its impact on the evolution of the genus *Helianthus* (L.). BMC Genomics **18**, 634.

Mason AS, Fulton JE, Hocking PM, Burt DW. 2016. A new look at the LTR retrotransposon content of the chicken genome. BMC Genomics **17**, 688.

McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics **19**, 362–367.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences **36**, 344–355.

McClintock B. 1984. The significance of responses of the genome to challenge. Science **226**, 792–801.

McQuinn RP, Gapper NE, Gray AG, Zhong S, Tohge T, Fei Z, Fernie AR, Giovannoni JJ. 2020. Manipulation of *ZDS* in tomato exposes carotenoid- and ABA-specific effects on fruit development and ripening. Plant Biotechnology Journal **18**, 2210-2224.

Mendes FK, Hahn MW. 2018. Why concatenation fails near the anomaly zone. Systematic Biology **67**, 158–169.

Merelo P, Agustí J, Arbona V, *et al.* 2017. Cell wall remodeling in abscission zone cells during ethylene-promoted fruit abscission in citrus. Frontiers in Plant Science **8**, 126.

Meyer RS, DuVal AE, Jensen HR. 2012. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytologist **196**, 29–48.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics **14**, 840–852.

Mhiri C, Parisod C, Daniel J, Petit M, Lim KY, Dorlhac de Borne F, Kovarik A, Leitch AR, Grandbastien MA. 2019. Parental transposable element loads influence their dynamics in young Nicotiana hybrids and allotetraploids. New Phytologist **221**, 1619–1633.

Minamikawa MF, Nonaka K, Kaminuma E, *et al.* 2017. Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. Scientific Reports **7**, 4721.

Mitchell KJ, Pratt RC, Watson LN, *et al.* 2014. Molecular phylogeny, biogeography, and habitat preference evolution of marsupials. Molecular Biology and Evolution **31**, 2322–2330.

Moreno JC, Pizarro L, Fuentes P, Handford M, Cifuentes V, Stange C. 2013. Levels of *Lycopene β-Cyclase 1* modulate carotenoid gene expression and accumulation in *Daucus carota*. PLOS ONE **8**, e58144.

Morgan M. 2019. AnnotationHub: Client to access AnnotationHub resources.

Morley RJ. 2018. Assembly and division of the South and South-East Asian flora in relation to tectonics and climate change. Journal of Tropical Ecology **34**, 209–234.

Morton CM. 2009. Phylogenetic relationships of the Aurantioideae (Rutaceae) based on the nuclear ribosomal DNA ITS region and three noncoding chloroplast DNA regions, atpB-rbcL spacer, rps16, and trnL-trnF. Organisms Diversity & Evolution **9**, 52–68.

Mudge K, Janick J, Scofield S, Goldschmidt EE. 2009. A history of grafting. Horticultural Reviews. Hoboken, John Wiley & Sons, 437–493.

Müller NF, Bouckaert RR. 2020. Adaptive metropolis-coupled MCMC for BEAST 2. PeerJ **8**, e9473.

Müller ML, Irkens-Kiesecker U, Rubinstein B, Taiz L. 1996. On the mechanism of hyperacidification in lemon: Comparison of the vacuolar H^+-ATPase activities of fruits and epicotyls. Journal of Biological Chemistry **271**, 1916–1924.

Mulvihill EE, Burke AC, Huff MW. 2016. Citrus flavonoids as regulators of lipoprotein metabolism and atherosclerosis. Annual Review of Nutrition **36**, 275–299.

Myles S, Boyko AR, Owens CL, *et al.* 2011. Genetic structure and domestication history of the grape. Proceedings of the National Academy of Sciences **108**, 3530–3535.

Nagano Y, Mimura T, Kotoda N, Matsumoto R, Nagano AJ, Honjo MN, Kudoh H, Yamamoto M. 2018. Phylogenetic relationships of Aurantioideae (Rutaceae) based on RAD-Seq. Tree Genetics & Genomes **14**, 6.

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature **461**, 1130–1134.

Nakashima K, Abe J, Kanazawa A. 2018. Chromosomal distribution of soybean retrotransposon SORE-1 suggests its recent preferential insertion into euchromatic regions. Chromosome Research **26**, 199–210.

Nattier R, Pellens R, Robillard T, Jourdan H, Legendre F, Caesar M, Nel A, Grandcolas P. 2017. Updating the phylogenetic dating of new caledonian biodiversity with a meta-analysis of the available evidence. Scientific Reports **7**, 3705.

Neph S, Kuehn MS, Reynolds AP, *et al.* 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics **28**, 1919–1920.

Nesom GL. 2014. *Citrus trifoliata* (Rutaceae): Review of biology and distribution in the USA. Phytoneuron **46**, 14.

Neumann P, Novák P, Hoštáková N, Macas J. 2019. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA **10**, 1.

Nevado B, Wong ELY, Osborne OG, Filatov DA. 2019. Adaptive evolution is common in rapid evolutionary radiations. Current Biology **29**, 3081-3086.e5.

Neves DM, Almeida LA da H, Santana-Vieira DDS, Freschi L, Ferreira CF, Soares Filho W dos S, Costa MGC, Micheli F, Coelho Filho MA, Gesteira A da S. 2017. Recurrent water deficit causes epigenetic and hormonal changes in citrus plants. Scientific Reports **7**, 13684.

Nguyen CH, Beattie GAC, Haigh AM, Astuti IP, Mabberley DJ, Weston PH, Holford P. 2019. Molecular differentiation of the *Murraya paniculata* Complex (Rutaceae: Aurantioideae: Aurantieae). BMC Evolutionary Biology **19**, 236.

Nicolosi E, Deng ZN, Gentile A, La Malfa S, Continella G, Tribulato E. 2000. *Citrus* phylogeny and genetic origin of important species as investigated by molecular markers: Theoretical and Applied Genetics **100**, 1155–1166.

Nile SH, Park SW. 2014. Bioactive components and health-promoting properties of yuzu (*Citrus ichangensis* × *C. reticulate*). Food Reviews International **30**, 155–167.

Noda T, Daiou K, Mihara T, Nagano Y. 2020. Development of Indel markers for the selection of Satsuma mandarin (*Citrus unshiu* Marc.) hybrids that can be used for low-cost genotyping with agarose gels. Euphytica **216**, 115.

Noe L, Kucherov G. 2005. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research **33**, W540–W543.

Ntoukakis VE, Negm FB, Lovatt CJ. 2017. Sink activity in Washington navel orange fruit borne on leafy and leafless inflorescences. Citrus Research & Technology **38**.**Oberholster R, Cowan AK, Molnár P, Tóth Gy**. 2001. Biochemical basis of color as an aesthetic quality in *Citrus sinensis*. Journal of Agricultural and Food Chemistry **49**, 303–307.

Ogilvie HA, Bouckaert RR, Drummond AJ. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Molecular Biology and Evolution **34**, 2101–2114.

Ollitrault P, Curk F, Krueger R. 2020. Chapter 4 - Citrus taxonomy. The Genus Citrus. Woodhead Publishing, 57–81.

Olsen KM, Wendel JF. 2013. A bountiful harvest: genomic insights into crop domestication phenotypes. Annual Review of Plant Biology **64**, 47–70.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology **17**, 132.

Orsini L, Vanoverbeke J, Swillen I, Mergeay J, Meester LD. 2013. Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. Molecular Ecology **22**, 5983–5999.

Orth AM, Yu L, Engel KH. 2013. Assessment of dietary exposure to flavouring substances via consumption of flavoured teas. Part 1: occurrence and contents of monoterpenes in Earl Grey teas marketed in the European Union. Food Additives & Contaminants: Part A **30**, 1701–1714.

Oueslati A, Ollitrault F, Baraket G, Salhi-Hannachi A, Navarro L, Ollitrault P. 2016. Towards a molecular taxonomic key of the Aurantioideae subfamily using chloroplastic SNP diagnostic markers of the main clades genotyped by competitive allele-specific PCR. BMC Genetics **17**, 118.

Oueslati A, Salhi-Hannachi A, Luro F, Vignes H, Mournet P, Ollitrault P. 2017. Genotyping by sequencing reveals the interspecific *C. maxima* / *C. reticulata* admixture along the genomes of modern citrus varieties of mandarins, tangors, tangelos, orangelos and grapefruits. PLOS ONE **12**, e0185618.

Owens BF, Lipka AE, Magallanes-Lundback M, *et al.* 2014. A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. Genetics **198**, 1699–1716.

Pan AD. 2010. Rutaceae leaf fossils from the Late Oligocene (27.23 Ma) Guang River flora of northwestern Ethiopia. Review of Palaeobotany and Palynology **159**, 188–194.

Pang XM, Hu CG, Deng XX. 2007. Phylogenetic relationships within *Citrus* and its related genera as inferred from AFLP markers. Genetic Resources and Crop Evolution **54**, 429–436.

Pankin A, Altmüller J, Becker C, Korff M von. 2018. Targeted resequencing reveals genomic signatures of barley domestication. New Phytologist **218**, 1247–1259.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics **35**, 526–528.

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien MA, Ainouche M. 2009. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytologist **184**, 1003–1015.

Paterson AH, Bowers JE, Bruggmann R, *et al.* 2009. The Sorghum bicolor genome and the diversification of grasses. Nature **457**, 551–556.

Paz RC, Rendina González AP, Ferrer MS, Masuelli RW. 2015. Short-term hybridisation activates Tnt1 and Tto1 Copia retrotransposons in wild tuber-bearing *Solanum* species. Plant Biology **17**, 860–869.

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. 2018. Genome size diversity and its impact on the evolution of land plants. Genes **9**, 88.

Peng Z, Bredeson JV, Wu GA, *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*. The Plant Journal **104**, 1215–1232.

Pereira V. 2004. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biology **5**, R79.

Pfeil BE, Crisp MD. 2008. The age and biogeography of *Citrus* and the orange subfamily (Rutaceae: Aurantioideae) in Australasia and New Caledonia. American Journal of Botany **95**, 1621–1631.

Phadungsawat B, Watanabe K, Mizuno S, Kanekatsu M, Suzuki S. 2020. Expression of *CCD4* gene involved in carotenoid degradation in yellow-flowered *Petunia* × *hybrida*. Scientia Horticulturae **261**, 108916.

Pickersgill B. 2018. Parallel vs. convergent evolution in domestication and diversification of crops in the Americas. Frontiers in Ecology and Evolution **6**, 56.

Piednoël M, Carrete-Vega G, Renner SS. 2013*a*. Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. The Plant Journal **75**, 699–709.

Piegu B, Guyot R, Picault N, *et al.* 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Research **16**, 1262–1269.

Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. 2020. GenMap: ultra-fast computation of genome mappability. Bioinformatics **36**, 3687–3692.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS ONE **5**, e9490.

Purugganan MD, Fuller DQ. 2009. The nature of selection during plant domestication. Nature **457**, 843–848.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biology **20**, 38.

Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, Colot V. 2016. The *Arabidopsis thaliana* mobilome and its impact at the species level. eLife **5**, e15716.

Quadrana L, Etcheverry M, Gilly A, *et al.* 2019. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nature Communications **10**, 3421.

Quesneville H. 2020. Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. Mobile DNA **11**, 28.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 841–842.

R Core Team. 2018. R: a language and environment for statistical computing.

Rabiee M, Sayyari E, Mirarab S. 2019. Multi-allele species reconstruction using ASTRAL. Molecular Phylogenetics and Evolution **130**, 286–296.

Rahman R, Chirn G, Kanodia A, Sytnikova YA, Brembs B, Bergman CM, Lau NC. 2015. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. Nucleic Acids Research **43**, 10655–10672.

Ramadugu C, Pfeil BE, Keremane ML, Lee RF, Maureira-Butler IJ, Roose ML. 2013. A six nuclear gene phylogeny of *Citrus* (Rutaceae) taking into account hybridization and lineage sorting. PLOS ONE **8**, e68410.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in bayesian phylogenetics using Tracer 1.7. Systematic Biology **67**, 901–904.

Ravi V, Khurana JP, Tyagi AK, Khurana P. 2008. An update on chloroplast genomes. Plant Systematics and Evolution **271**, 101–122.

Reuther W, Batchelor L, Webber H. 1967. The Citrus industry revised: History World Distribution, Botany and varieties vol.1.

Rico-Cabanas L, Martínez-Izquierdo JA. 2007. CIRE1, a novel transcriptionally active Ty1-copia retrotransposon from *Citrus sinensis*. Molecular Genetics and Genomics **277**, 365–377.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nature Biotechnology **29**, 24–26.

Rodrigo MJ, Alquézar B, Alós E, Lado J, Zacarías L. 2013*a*. Biochemical bases and molecular regulation of pigmentation in the peel of *Citrus* fruit. Scientia Horticulturae **163**, 46–62.

Rodrigo MJ, Alquézar B, Alós E, Medina V, Carmona L, Bruno M, Al-Babili S, Zacarías L. 2013*b*. A novel carotenoid cleavage activity involved in the biosynthesis of *Citrus* fruit-specific apocarotenoid pigments. Journal of Experimental Botany **64**, 4461–4478.

Rodrigo MJ, Lado J, Alós E, Alquézar B, Dery O, Hirschberg J, Zacarías L. 2019. A mutant allele of *ζ-carotene isomerase* (*Z-ISO*) is associated with the yellow pigmentation of the "Pinalate" sweet orange mutant and reveals new insights into its role in fruit carotenogenesis. BMC Plant Biology **19**, 465.

Rodrigo MJ, Marcos JF, Alférez F, Mallent MD, Zacarías L. 2003. Characterization of Pinalate, a novel *Citrus sinensis* mutant with a fruit-specific alteration that results in yellow pigmentation and decreased ABA content. Journal of Experimental Botany **54**, 727–738.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLOS Biology **4**, e352.

Rosenberg NA. 2013. Discordance of species trees with their most likely gene trees: a unifying principle. Molecular Biology and Evolution **30**, 2709–2713.

Roulin A, Piegu B, Fortune PM, Sabot F, D'Hont A, Manicacci D, Panaud O. 2009. Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon *Route66* in *Poaceae*. BMC Evolutionary Biology **9**, 58.

Sadka A, Shlizerman L, Kamara I, Blumwald E. 2019. Primary metabolism in *Citrus* fruit as affected by its unique structure. Frontiers in Plant Science **10**.

Salvin S. 2008. The new crop industries handbook: native foods. Barton, A.C.T.: Rural Industries Research and Development Corporation.

Samuel R, Ehrendorfer F, Chase MW, Greger H. 2001. Phylogenetic analyses of Aurantioideae (Rutaceae) based on non-coding plastid DNA sequences and phytochemical features. Plant Biology **3**, 77–87.

Sanmiguel P, Bennetzen JL. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany **82**, 37–44.

Sanseverino W, Hénaff E, Vives C, Pinosio S, Burgos-Paz W, Morgante M, Ramos-Onsins SE, Garcia-Mas J, Casacuberta JM. 2015. Transposon insertions, structural variations, and snps contribute to the evolution of the melon genome. Molecular Biology and Evolution **32**, 2760–2774.

Sasaki T. 2005. The map-based sequence of the rice genome. Nature **436**, 793–800.

Sauvage C, Rau A, Aichholz C, Chadoeuf J, Sarah G, Ruiz M, Santoni S, Causse M, David J, Glémin S. 2017. Domestication rewired gene expression and nucleotide diversity patterns in tomato. The Plant Journal **91**, 631–645.

Sayyari E, Mirarab S. 2018. Testing for polytomies in phylogenetic species trees using quartet frequencies. Genes **9**, 132.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics **27**, 592–593.

Schnable PS, Ware D, Fulton RS, *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. Science **326**, 1112–1115.

Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution. Molecular Ecology **28**, 1537–1549.

Schwartz T, Nylinder S, Ramadugu C, Antonelli A, Pfeil BE. 2016. The origin of oranges: a multi-locus phylogeny of Rutaceae subfamily Aurantioideae. Systematic Botany **40**, 1053–1062.

Seker M, Tuzcu O, Ollitrault P. 2003. Comparison of nuclear DNA content of citrus rootstock populations by flow cytometry analysis. Plant Breeding **122**, 169–172.

Sela N, Kim E, Ast G. 2010. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. Genome Biology **11**, R59.

Sewelam N, Kazan K, Thomas-Hall SR, Kidd BN, Manners JM, Schenk PM. 2013. Ethylene Response Factor 6 is a regulator of reactive oxygen species signaling in *Arabidopsis*. PLOS ONE **8**, e70289.

Shao L, Xing F, Xu C, *et al.* 2019. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. Proceedings of the National Academy of Sciences **116**, 5653–5658.

Shi CY, Hussain SB, Guo LX, Yang H, Ning DY, Liu YZ. 2018. Genome-wide identification and transcript analysis of vacuolar-ATPase genes in citrus reveal their possible involvement in citrate accumulation. Phytochemistry **155**, 147–154.

Shi CY, Song RQ, Hu XM, Liu X, Jin LF, Liu YZ. 2015. Citrus *PH5*-like H^+-ATPase genes: identification and transcript analysis to investigate their possible relationship with citrate accumulation in fruits. Frontiers in Plant Science **6**, 135.

Shimada T, Endo T, Fujii H, Nakano M, Sugiyama A, Daido G, Ohta S, Yoshioka T, Omura M. 2018. MITE insertion-dependent expression of *CitRKD1* with a RWP-RK domain regulates somatic embryogenesis in citrus nucellar tissues. BMC Plant Biology **18**, 166.

Shimada T, Nakano R, Shulaev V, Sadka A, Blumwald E. 2006. Vacuolar citrate/H^+ symporter of citrus juice cells. Planta **224**, 472–480.

Shimizu T, Tanizawa Y, Mochizuki T, Nagasaki H, Yoshioka T, Toyoda A, Fujiyama A, Kaminuma E, Nakamura Y. 2017. Draft sequencing of the heterozygous diploid genome of satsuma (*Citrus unshiu* Marc.) using a hybrid assembly approach. Frontiers in genetics **8**, 180.

Siverio F, Marco-Noales E, Bertolini E, *et al.* 2017. Survey of huanglongbing associated with '*Candidatus Liberibacter'* species in Spain: analyses of citrus plants and *Trioza erytreae*. Phytopathologia Mediterranea **56**, 98–110.

Smith O, Nicholson WV, Kistler L, *et al.* 2019. A domestication history of dynamic adaptation and genomic deterioration in Sorghum. Nature Plants **5**, 369–379.

Sperber GO, Airola T, Jern P, Blomberg J. 2007. Automated recognition of retroviral sequences in genomic data-RetroTector. Nucleic acids research **35**, 4964–76.

Springer NM, Stupar RM. 2007. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. The Plant Cell **19**, 2391–2402.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**, 1312–1313.

Stone BC, Lowry JB, Scora RW, Jong K. 1973. *Citrus halimii*: a new species from Malaya and Peninsular Thailand. Biotropica **5**, 102–110.

Strazzer P, Spelt CE, Li S, Bliek M, Federici CT, Roose ML, Koes R, Quattrocchio FM. 2019. Hyperacidification of *Citrus* fruits by a vacuolar proton-pumping P-ATPase complex. Nature Communications **10**, 744.

Strickler SR, Bombarely A, Munkvold JD, York T, Menda N, Martin GB, Mueller LA. 2015. Comparative genomics and phylogenetic discordance of cultivated tomato and close wild relatives. PeerJ **3**, e793.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nature genetics **43**, 1160–1163.

Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLOS Biology **13**, e1002224.

Sun X, Jiao C, Schwaninger H, *et al.* 2020. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nature Genetics **52**, 1423–1432.

Suoniemi A, Tanskanen J, Schulman AH. 1998. Gypsy-like retrotransposons are widespread in the plant kingdom. The Plant Journal **13**, 699–705.

Swanson-Wagner R, Briskine R, Schaefer R, Hufford MB, Ross-Ibarra J, Myers CL, Tiffin P, Springer NM. 2012. Reshaping of the maize transcriptome by domestication. Proceedings of the National Academy of Sciences **109**, 11878–11883.

Swingle WT, Reece PC. 1967. The botany of *Citrus* and its wild relatives. The citrus industry. University of California, 190–430.

Sytar O, Zivcak M, Bruckova K, Brestic M, Hemmerich I, Rauh C, Simko I. 2018. Shift in accumulation of flavonoids and phenolic acids in lettuce attributable to changes in ultraviolet radiation and temperature. Scientia Horticulturae **239**, 193–204.

Tadeo FR, Cercós M, Colmenero-Flores JM, *et al.* 2008. Molecular physiology of development and quality of *Citrus*. Advances in Botanical Research **47**, 147–223.

Tallowin OJS, Meiri S, Donnellan SC, Richards SJ, Austin CC, Oliver PM. 2020. The other side of the Sahulian coin: biogeography and evolution of Melanesian forest dragons (Agamidae). Biological Journal of the Linnean Society **129**, 99–113.

Talon M, Gmitter FG. 2008. *Citrus* genomics. International Journal of Plant Genomics **2008**, e528361.

Talon M, Wu GA, Gmitter FG, Rokhsar DS. 2020. Chapter 2 - The origin of citrus. The Genus Citrus. Woodhead Publishing, 9–31.

Tanaka T. 1954. Species Problem in Citrus: A Critical Study of Wild and Cultivated Units of Citrus, Based Upon Field Studies in Their Native Homes. Japanese Society for the Promotion of Science.

Tanaka T. 1959. Origin of *Citrus* fruits with reference to Himalaya region. Journal of the Japanese Society for Horticultural Science **28**, 71–75.

Tanner T, Hernández-Almeida I, Drury AJ, Guitián J, Stoll H. 2020. Decreasing atmospheric CO_2 during the Late Miocene cooling. Paleoceanography and Paleoclimatology **35**, e2020PA003925.

Tanskanen JA, Sabot F, Vicient C, Schulman AH. 2007. Life without GAG: the BARE-2 retrotransposon as a parasite's parasite. Gene **390**, 166–174.

Tao N, Hu Z, Liu Q, Xu J, Cheng Y, Guo L, Guo W, Deng X. 2007. Expression of phytoene synthase gene (*Psy*) is enhanced during fruit ripening of Cara Cara navel orange (*Citrus sinensis* Osbeck). Plant Cell Reports **26**, 837–843.

Tao NG, Xu J, Cheng YJ, Hong L, Guo WW, Yi HL, Deng XX. 2005. Isolation and characterization of Copia-like retrotransposons from 12 sweet orange (*Citrus sinensis* Osbeck) cultivars. Journal of Integrative Plant Biology **47**, 1507–1515.

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biology and Evolution **3**, 219–229.

Terol J. 2020. A whole genome association study in mandarin hybrids. Plant and Animal Genome XXVIII Conference.

Terol J, Nueda MJ, Ventimilla D, Tadeo F, 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 Biology **19**, 47.

Terol J, Ibanez V, Carbonell-Caballero J, Alonso R, Estornell LH, Licciardello C, Gut IG, Dopazo J, Talon M. 2015. Involvement of a citrus meiotic recombination TTC-repeat motif in the formation of gross deletions generated by ionizing radiation and MULE activation. BMC Genomics **16**, 69.

Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, Jackson SA, Gaut BS, Ma J. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Research **19**, 2221–2230.

Trabut JL. 1902. L'hybridation des Citrus: une nouvelle tangérine: la Clémentine. Revue Horticole, 232–234.

Tsai CC, Chou CH, Wang HV, Ko YZ, Chiang TY, Chiang YC. 2015. Biogeography of the *Phalaenopsis amabilis* species complex inferred from nuclear and plastid DNAs. BMC Plant Biology **15**, 202.

Tsai CC, Liao PC, Ko YZ, Chen CH, Chiang YC. 2020. Phylogeny and Historical Biogeography of *Paphiopedilum* Pfitzer (Orchidaceae) Based on Nuclear and Plastid DNA. Frontiers in Plant Science **11**, 126.

Tsugane K, Maekawa M, Takagi K, Takahara H, Qian Q, Eun CH, Iida S. 2006. An active DNA transposon nDart causing leaf variegation and mutable dwarfism and its related elements in rice. The Plant Journal **45**, 46–57.

Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, Toyoda A, Fujiyama A, Tarutani Y, Kakutani T. 2012. Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. Genes & Development **26**, 705–713.

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. 2009. Bursts of retrotransposition reproduced in *Arabidopsis*. Nature **461**, 423–426.

Tuskan GA, Difazio S, Jansson S, *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**, 1596–1604.

Valcárcel V, Guzmán B, Medina NG, Vargas P, Wen J. 2017. Phylogenetic and paleobotanical evidence for late Miocene diversification of the Tertiary subtropical lineage of ivies (*Hedera* L., Araliaceae). BMC Evolutionary Biology **17**, 146.

Van der Auwera GA, Carneiro MO, Hartl C, *et al.* 2013. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics **11**, 11.10.1-11.10.33.

Van Welzen PC, Parnell J. N, Slik JWF. 2011. Wallace's Line and plant distributions: two or three phytogeographical areas and where to group Java? Biological Journal of the Linnean Society **103**, 531–545.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. Annals of Botany **120**, 195–207.

Vitte C, Bennetzen JL. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceedings of the National Academy of Sciences **103**, 17638–17643.

Vitte C, Panaud O, Quesneville H. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics **8**, 218.

Waki T, Hiki T, Watanabe R, Hashimoto T, Nakajima K. 2011. The *Arabidopsis* RWP-RK protein RKD4 triggers gene expression and pattern formation in early embryogenesis. Current Biology **21**, 1277–1281.

Walker JF, Walker-Hale N, Vargas OM, Larson DA, Stull GW. 2019. Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ **7**, e7747.

Wang YC, Chuang YC, Hsu HW. 2008. The flavonoid, carotenoid and pectin content in peels of citrus cultivated in Taiwan. Food Chemistry **106**, 277–284.

Wang L, He F, Huang Y, *et al.* 2018*a*. Genome of wild mandarin and domestication history of mandarin. Molecular Plant **11**, 1024–1037.

Wang W, Li H, Chen Z. 2014. Analysis of plastid and nuclear DNA data in plant phylogenetics—evaluation and improvement. Science China Life Sciences **57**, 280–286.

Wang J hui, Liu J jun, Chen K ling, Li H wen, He J, Guan B, He L. 2017*a*. Comparative transcriptome and proteome profiling of two *Citrus sinensis* cultivars during fruit development and ripening. BMC Genomics **18**, 984.

Wang X, Xu Y, Zhang S, *et al.* 2017*b*. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. Nature Genetics **49**, 765–772.

Wang S, Yang C, Tu H, Zhou J, Liu X, Cheng Y, Luo J, Deng X, Zhang H, Xu J. 2017*c*. Characterization and metabolic diversity of flavonoids in *Citrus* species. Scientific Reports **7**, 10549.

Wang Z, Yu Q, Shen W, El Mohtar CA, Zhao X, Gmitter FG. 2018*b*. Functional study of *CHS* gene family members in citrus revealed a novel *CHS* gene affecting the production of flavonoids. BMC Plant Biology **18**, 189.

Warren DL, Geneva AJ, Lanfear R. 2017. RWTY (R We There Yet): An R package for examining convergence of bayesian phylogenetic analyses. Molecular Biology and Evolution **34**, 1016–1020.

Webb DM, Knapp SJ. 1990. DNA extraction from a previously recalcitrant plant genus. Plant Molecular Biology Reporter **8**, 180.

Wei B, Liu H, Liu X, Xiao Q, Wang Y, Zhang J, Hu Y, Liu Y, Yu G, Huang Y. 2016. Genome-wide characterization of non-reference transposons in crops suggests non-random insertion. BMC Genomics **17**, 536.

Wen J, Zhang J, Nie ZL, Zhong Y, Sun H. 2014. Evolutionary diversifications of plants on the Qinghai-Tibetan Plateau. Frontiers in Genetics **5**, 4.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends in Ecology & Evolution **22**, 258–265.

Wibowo A, Becker C, Marconi G, *et al.* 2016. Hyperosmotic stress memory in *Arabidopsis* is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. eLife **5**, e13546.

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, De Oliveira R, Mayer KFX, Paux E, Choulet F. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biology **19**, 103.

Wicker T, Guyot R, Yahiaoui N, Keller B. 2003. CACTA transposons in triticeae. a diverse family of high-copy repetitive elements. Plant Physiology **132**, 52–63.

Wicker T, Keller B. 2007. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Research **17**, 1072–1081.

Wicker T, Sabot F, Hua-Van A, *et al.* 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics **8**, 973–982.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wolters AMA, Uitdewilligen JGAML, Kloosterman BA, Hutten RCB, Visser RGF, van Eck HJ. 2010. Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. Plant Molecular Biology **73**, 659–671.

Woolfit M. 2009. Effective population size and the rate and pattern of nucleotide substitutions. Biology Letters **5**, 417–420.

Wu GA, Prochnik S, Jenkins J, *et al.* 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nature Biotechnology **32**, 656–662.

Wu GA, Terol J, Ibanez V, *et al.* 2018. Genomics of the origin and evolution of *Citrus*. Nature **554**, 311–316.

Wu X, Zhang S, Liu X, Shang J, Zhang A, Zhu Z, Zha D. 2020. Chalcone synthase (*CHS*) family members analysis from eggplant (*Solanum melongena* L.) in the flavonoid biosynthetic pathway and expression patterns in response to heat stress. PLOS ONE **15**, e0226537.

Xi Z, Liu L, Rest JS, Davis CC. 2014. Coalescent versus concatenation methods and the placement of amborella as sister to water lilies. Systematic Biology **63**, 919–932.

Xie M, Chung CYL, Li MW, *et al.* 2019. A reference-grade wild soybean genome. Nature Communications **10**, 1216.

Xie S, Manchester SR, Liu K, Wang Y, Sun B. 2013. *Citrus linczangensis* sp. n., a leaf fossil of Rutaceae from the Late Miocene of Yunnan, China. International Journal of Plant Sciences **174**, 1201–1207.

Xing Y, Ree RH. 2017. Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot. Proceedings of the National Academy of Sciences **114**, E3444–E3451.

Xu Q, Chen LL, Ruan X, *et al.* 2013. The draft genome of sweet orange (*Citrus sinensis*). Nature Genetics **45**, 59–66.

Xu Y, Du J. 2014. Young but not relatively old retrotransposons are preferentially located in gene-rich euchromatic regions in tomato (*Solanum lycopersicum*) plants. The Plant Journal **80**, 582–591.

Xu CJ, Fraser PD, Wang WJ, Bramley PM. 2006. Differences in the carotenoid content of ordinary *Citrus* and lycopene-accumulating mutants. Journal of Agricultural and Food Chemistry **54**, 5474–5481.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research **35**, W265–W268.

Yamada T, Hayasaka S, Shibata Y, Ojima T, Saegusa T, Gotoh T, Ishikawa S, Nakamura Y, Kayaba K. 2011. Frequency of citrus fruit intake is associated with the incidence of cardiovascular disease: the jichi medical school cohort study. Journal of Epidemiology **21**, 169–175.

Yamaga I, Hamasaki S. 2020. Seasonal effect of ultraviolet irradiation on polymethoxyflavone and hesperidin content in ponkan and tachibana flavedo. HortScience **55**, 1078–1082.

Yang CK, Chiang YC, Huang BH, Ju LP, Liao PC. 2018. Nuclear and chloroplast DNA phylogeography suggests an Early Miocene southward expansion of *Lithocarpus* (Fagaceae) on the Asian continent and islands. Botanical Studies **59**, 27.

Yang X, Li H, Yu H, Chai L, Xu Q, Deng X. 2017. Molecular phylogeography and population evolution analysis of *Citrus ichangensis* (Rutaceae). Tree Genetics & Genomes **13**, 29.

Yap JYS, Rossetto M, Costion C, Crayn D, Kooyman RM, Richardson J, Henry R. 2018. Filters of floristic exchange: how traits and climate shape the rain forest invasion of Sahul from Sunda. Journal of Biogeography **45**, 838–847.

Yin H, Du J, Wu J, Wei S, Xu Y, Tao S, Wu J, Zhang S. 2015. Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. Scientific Reports **5**, 17644.

Yoshioka S, Aida R, Yamamizo C, Shibata M, Ohmiya A. 2012. The *carotenoid cleavage dioxygenase 4* (*CmCCD4a*) gene family encodes a key regulator of petal color mutation in chrysanthemum. Euphytica **184**, 377–387.

Yu Y, Fu J, Xu Y, *et al.* 2018. Genome re-sequencing reveals the evolutionary history of peach fruit edibility. Nature Communications **9**, 5404.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017*a*. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data (G McInerny, Ed.). Methods in Ecology and Evolution **8**, 28–36.

Yu G, Wang LG, Han Y, He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: A Journal of Integrative Biology **16**, 284–287.

Yu H, Yang X, Guo F, Jiang X, Deng X, Xu Q. 2017*b*. Genetic diversity and population structure of pummelo (*Citrus maxima*) germplasm in China. Tree Genetics & Genomes **13**, 58.

Yuste-Lisbona FJ, Fernández-Lozano A, Pineda B, *et al.* 2020. ENO regulates tomato fruit size through the floral meristem development network. Proceedings of the National Academy of Sciences **117**, 8187–8195.

Zhang QJ, Gao LZ. 2017. Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA-Genome *Oryza* species. Genes|Genomes|Genetics **7**, 1875–1885.

Zhang L, Ma G, Shirai Y, Kato M, Yamawaki K, Ikoma Y, Matsumoto H. 2012. Expression and functional analysis of two lycopene β-cyclases from citrus fruits. Planta **236**, 1315–1325.

Zhang Z, Mao L, Chen H, *et al.* 2015. Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. The Plant Cell **27**, 1595–1604.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics **19**, 153.

Zhang C, Sayyari E, Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. Comparative Genomics. Springer, 53–75.

Zhang J, Tao N, Xu Q, Zhou W, Cao H, Xu J, Deng X. 2009. Functional characterization of *Citrus PSY* gene in Hongkong kumquat (*Fortunella hindsii* Swingle). Plant Cell Reports **28**, 1737.

Zhang X, Zhao M, McCarty DR, Lisch D. 2020. Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. Nucleic Acids Research **48**, 6685–6698.

Zhao Z, He S, Hu Y, Yang Y, Jiao B, Fang Q, Zhou Z. 2017. Fruit flavonoid variation between and within four cultivated *Citrus* species evaluated by UPLC-PDA system. Scientia Horticulturae **224**, 93–101.

Zhao D, Jiang N. 2014. Nested insertions and accumulation of indels are negatively correlated with abundance of mutator-like transposable elements in maize and rice. PLOS ONE **9**, e87069.

Zhao C, Liu X, Gong Q, *et al.* 2020. Three AP2/ERF family members modulate flavonoid synthesis by regulating type IV chalcone isomerase in citrus. Plant Biotechnology Journal **19**, 671–688.

Zhao B, Qi K, Yi X, Chen G, Liu X, Qi X, Zhang S. 2019. Identification of hexokinase family members in pear (*Pyrus* × *bretschneideri*) and functional exploration of *PbHXK1* in modulating sugar content and plant growth. Gene **711**, 143932.

Zheng X, Xie Z, Zhu K, Xu Q, Deng X, Pan Z. 2015. Isolation and characterization of *carotenoid cleavage dioxygenase 4* genes from different citrus species. Molecular Genetics and Genomics **290**, 1589–1603.

Zheng X, Zhu K, Sun Q, *et al.* 2019. Natural variation in *CCD4* promoter underpins species-specific evolution of red coloration in citrus peel. Molecular Plant **12**, 1294–1307.

Zhou G, Chen Y, Yao W, Zhang C, Xie W, Hua J, Xing Y, Xiao J, Zhang Q. 2012. Genetic composition of yield heterosis in an elite rice hybrid. Proceedings of the National Academy of Sciences **109**, 15847–15852.

Zhou Y, Duvaux L, Ren G, Zhang L, Savolainen O, Liu J. 2017. Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. Heredity **118**, 211–220.

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS. 2019. The population genetics of structural variants in grapevine domestication. Nature Plants **5**, 965–979.

Zhu A, Ibrahim JG, Love MI. 2019*a*. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics **35**, 2084–2092.

Zhu K, Wu Q, Huang Y, Ye J, Xu Q, Deng X. 2020. Genome-wide characterization of cis-acting elements in the promoters of key carotenoid pathway genes from the main species of genus *Citrus*. Horticultural Plant Journal **6**, 385–395.

Zhu C, Zheng X, Huang Y, *et al.* 2019*b*. Genome sequencing and CRISPR/Cas9 gene editing of an early flowering Mini-Citrus (*Fortunella hindsii*). Plant Biotechnology Journal **17**, 2199–2210.

Zielezinski A, Girgis HZ, Bernard G, *et al.* 2019. Benchmarking of alignment-free sequence comparison methods. Genome Biology **20**, 144.