https://doi.org/10.35662/unine-thesis-3080

Université de Neuchâtel

Characterization of domesticated clove (*Syzygium aromaticum*) and comparison with related species

A dissertation submitted to the
University of Neuchâtel
For the degree of
Docteur és Sciences

Presented by

Sonia Ouadi
Philip Morris International Research & Development, Neuchâtel
Laboratory of Plant Physiology, Institute of Biology, University of Neuchâtel

Thesis committee

Dr. Nicolas Sierro (Thesis supervisor) – Philip Morris International Research & Development
Dr. Nikolai V. Ivanov (Thesis co-director) – Philip Morris International Research & Development, University of Neuchâtel
Prof. Felix Kessler (Thesis co-director) – University of Neuchâtel
Dr. Gaétan Glauser (Internal expert) – University of Neuchâtel
Prof. Lukas Mueller (External expert) –Boyce Thompson Institute, USA

November 2023

unine
Université de Neuchâtel
Faculté des sciences

Rue Emile-Argand 11
CH-2000 Neuchâtel
doctorat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Madame Sonia OUADI

Titre :

"Characterization of domesticated clove (*Syzygium aromaticum*) and comparison with related species"

sur le rapport des membres du jury composé comme suit:

- Prof. Felix Kessler, co-directeur de thèse, Université de Neuchâtel, Suisse
- Dr Nikolai Ivanov, co-directeur de thèse, Université de Neuchâtel et Philip Morris International, Suisse
- Prof. Lukas Mueller, Cornell University, Ithaca, USA
- Dr Gaétan Glauser, Université de Neuchâtel, Suisse
- Dr Nicolas Sierro, Philip Morris International, Suisse

Neuchâtel, le 23 novembre 2023

Le Doyen, Prof. R. Bshary

Abstract

Clove, *Syzygium aromaticum*, is a valuable spice crop tree belonging to the Myrtaceae family. Its dried, unopened flower buds, essential oil, and oleoresin-rich in eugenol, are used commercially in a broad range of industries. Despite its economic importance and the promising therapeutic potential of eugenol, little -omics information was available for fundamental and applied research on clove. A reference genome is an important tool for the advancement of plant biology research and crop improvement programs. In this thesis, the clove genome was sequenced, assembled at the chromosome level and subsequently analysed using comparative genomics and multi-omics approaches to investigate clove genome evolution and the genetic basis of eugenol biosynthesis. In addition, the genome assemblies of *Syzygium malaccense* (430 Mb), *Syzygium aqueum* (392 Mb), *Syzygium jambos* (426 Mb), and *Syzygium syzygioides* (431 Mb) were constructed and compared to the genome assemblies of clove (370 Mb), *Syzygium grande* (405 Mb) and their Myrtaceae relative *Eucalyptus grandis* (690 Mb) to provide further genomic insights into the evolution of *Syzygium* genomes.

By analysing the clove genome assembly and sets of transcriptomic and metabolomic data from leaves and buds, the key gene families involved in the biosynthesis of the precursor of lignin and phenylpropene volatiles (eugenol, isoeugenol, chavicol) were identified as well as genes possibly involved in the biosynthesis of eugenol. On the basis of this multi -omics study, we hypothesized that eugenol acetate may play a key role in the high accumulation of eugenol in clove leaves and buds. Fewer genes were found in the newly assembled *Syzygium* genomes (1 to 3) than in clove (15) to encode for putative eugenol synthase—a key enzyme in the eugenol biosynthetic pathway—suggesting that this copy number variation may be involved in the high content in eugenol of clove. Synteny analyses revealed a high conservation of the chromosomal organization of the 6 *Syzygium* species studied and intrachromosomal rearrangements on 7 of the 11 chromosomes, between the *Syzygium* species and *E. grandis*.

This new set of *Syzygium* genomes is a valuable contribution to genomics resources for the Myrtaceae family, and especially to the *Syzygium* genus, the world's most species-rich tree genus. Moreover, the availability of the clove reference genome opens new perspectives for applied research aiming to optimize clove production.

Keywords: *Syzygium aromaticum*, *Syzygium malaccense*, *Syzygium jambos*, *Syzygium aqueum*, *Syzygium syzygioides*, Myrtaceae, sequencing, *de novo* assembly, synteny, eugenol.

Résumé

Le giroflier, *Syzygium aromaticum*, est un arbre à épices appartenant à la famille des Myrtacées. Les boutons floraux encore fermés et séchés (clous de girofle), l'huile essentielle et l'oléorésine produits à partir du giroflier sont riche en eugénol et utilisés commercialement dans un large éventail d'industries. Malgré son importance économique et le potentiel thérapeutique prometteur de l'eugénol, peu de données omiques étaient disponibles pour la recherche fondamentale et appliquée sur le giroflier. Un génome de référence est un outil important pour l'avancement de la recherche en biologie végétale et des programmes d'amélioration des cultures. Dans le cadre de cette thèse, le génome du giroflier a été séquencé, et assemblé à l'échelle chromosomique, puis analysé en utilisant des approches de génomique comparative et multi-omiques dans le but d'étudier l'évolution du génome du giroflier et la base génétique de la biosynthèse de l'eugénol. Les assemblages des génomes de *Syzygium malaccense* (430 Mb), *Syzygium aqueum* (392 Mb), *Syzygium jambos* (426 Mb), et *Syzygium syzygioides* (431 Mb) ont également été construits pour être comparés aux génomes du giroflier (370 Mb), de *Syzygium grande* (405 Mb) ainsi qu'au génome d'une autre espèce de Myrtacées, *Eucalyptus grandis* (690 Mb), afin d'acquérir de nouvelles connaissances sur l'évolution des génomes des espèces de *Syzygium*.

En analysant l'assemblage du génome du giroflier et en intégrant les données transcriptomiques et métabolomiques produites à partir des feuilles et des bourgeons floraux, les principales familles de gènes impliquées dans la biosynthèse des précurseurs de la lignine et des phénylpropènes volatils (eugénol, isoeugénol, chavicol) ont été identifiées ainsi que les gènes potentiellement impliqués dans la biosynthèse de l'eugénol. Sur la base de cette étude multi-omiques, nous avons émis l'hypothèse que l'acétate d'eugénol pourrait jouer un rôle clé dans la forte accumulation d'eugénol dans les feuilles et les bourgeons floraux du giroflier. En outre, un nombre inférieur de gènes codant pour une possible eugénol synthase — une enzyme clé dans la biosynthèse de l'eugénol — ont été identifiés dans les génomes des 4 espèces de *Syzygium* nouvellement assemblés (1 à 3) que dans le génome du giroflier (15), suggérant que cette variation du nombre de copie peut être impliquée dans la teneur élevée en eugénol des organes de ce dernier. Les analyses de synténie ont révélé une bonne conservation de l'organisation chromosomique des 6 espèces de *Syzygium* étudiées et des réarrangements intrachromosomiques sur 7 des 11 chromosomes entre les espèces de *Syzygium* et *E. grandis*.

Ces nouveaux génomes de *Syzygium* sont une contribution précieuse aux ressources génomiques de la famille des Myrtacées, et en particulier pour le genre *Syzygium*, le plus riche en espèces d'arbre au monde. De plus, la disponibilité d'un génome de référence pour le giroflier ouvre de nouvelles perspectives pour la recherche appliquée visant à optimiser la productivité des girofliers.

Mots clés : *Syzygium aromaticum*, *Syzygium malaccense*, *Syzygium jambos*, *Syzygium aqueum*, *Syzygium syzygioides*, Myrtacées, séquençage, assemblage *de novo*, synténie, eugénol.

Contents

Abbreviations

4CL	4-hydroxycinnamate:CoA ligase
AAT	alcohol acyltransferase
ABC	ATP-binding cassette transporter
AIMT	trans-anol O-methyltransferase
AIS	anol/isoeugenol synthase
APS	allyl-phenylpropene synthase
BAHD	benzyl alcohol-acetyl-, anthocyanin-O-hydroxy-cinnamoyl-, anthranilate-N-hydroxycinnamoyl/benzoyl-, deacetyl-vindoline
C3′H	4-coumaroyl shikimate/quinate 3′-hydroxylase
C3H	4-coumarate 3-hydroxylase
C4H	cinnamate 4-hydroxylase
CAD	cinnamyl alcohol dehydrogenase
CCoAOMT	caffeoyl CoA 3-O-methyltransferase
CCR	cinnamoyl CoA reductase
CFAT	coniferyl alcohol acyltransferase
CCV	*Corymbia citriodora* subsp. *Variegata*
COMT	caffeate/5-hydroxyferulate 3-O-methyltransferase
CSE	caffeoyl shikimate esterase
CT	*Corymbia torelliana*
CVOMT	chavicol o-methyl transferase
EGS	eugenol synthase
EL	expanded pale green leaves
EO	essential oil
EOMT	eugenol o-methyl transferase
ETS	external transcribed spacer
F5H	ferulate 5-hydroxylase/coniferaldehyde 5-hydroxylase
FB	buds in fruiting stage
FBP	peduncles of buds in the fruiting stage
FW	fresh weight
GB	green buds

GBP	peduncles of green buds
HCT	4-hydroxycinnamoyl CoA:shikimate/quinate hydroxycinnamoyltransferase
IEMT	(iso) eugenol o-methyl transferase
IFB	buds in the initial fruiting stage
IGS	isoeugenol synthase
ITS	internal transcribed spacer
LPV	Laboratory of Plant Physiology
LTR	long terminal repeats
ML	mature leaves
MLL	mature leaf laminas
PAL	phenylalanine ammonia-lyase
PCBER	phenylcoumaran benzylic ether reductases
PB	pink buds
PBP	peduncles of pink buds
PIP	pinoresinol–lariciresinol reductase, isoflavone reductase, phenylcoumaran benzylic ether reductase
PMI	Philip Morris International
PPS	propenyl-phenylpropene synthase
SAQU	*Syzygium aqueum*
SARO	*Syzygium aromaticum*
SD	standard deviation
SGRA	*Syzygium grande*
SJAM	*Syzygium jambos*
SMAL	*Syzygium malaccense*
SSYZ	*Syzygium syzygioides*
ST	stems
TE	transposable elements
VST	variance stabilizing transformation
YB	young buds
YBP	peduncles of young buds
YLI	young leaves
YLII	expanded young leaves

General introduction

Syzygium aromaticum (L.) Merr. & L. M. Perry, commonly known as clove, is a tropical evergreen tree native to Molucca Island of Indonesia, and the species with the major commercial importance of the *Syzygium* genus, the world's most species-rich tree genus with about 1200 species [1-3]. It is cultivated to produce clove (the dried unopened flower bud), oleoresins and essential oil that are commercially used in a wide range of applications within the pharmaceutical, food, cosmetic, tobacco, and agricultural industries [4, 5]. Eugenol is the most abundant and main bioactive compound found in clove products [5, 6]. Besides contributing to its flavour, eugenol is a versatile phenolic compound with multiple pharmaceutical activities (e.g., analgesic, antioxidant, anti-inflammatory, antibacterial, antifungal, and antiviral activities) and is considered a promising alternative drug for human health (e.g., cancer and pathogenic microorganism resistance, diabetes, obesity, and autoimmune diseases) and solutions for sustainable agriculture [4, 7, 8].

Despite its economic importance, no genetic map or reference genome were publicly available for basic and applied research on clove. Although clove is considered the richest plant source of eugenol, the molecular basis of the biosynthesis of eugenol has not been described and compared with other species.

Genome structure evolution

Comparative genomics enables the identification of similarities and differences between the genome structure (e.g., chromosome number and organisation) and features (e.g., genes and repeats content) of different species, and thus provides important information on the genome evolution of these species after they diverged from a common ancestor. For instance, block of genes located on the same chromosome and in same in the same order (syntenic block) in different species provide information on conserved region between these species. Interspecific genomic differences, in contrast, provide information on the genome evolution of each species, and can bring important insights into lineage-specific whole genome duplication, genome size variation, and the genetic basis of plant adaptation to their environment and agronomical traits[9-12].

Clove is placed within the Myrtales order and is a member of the Myrtaceae family which comprises several species of significant ecological and economic relevance worldwide such as *Eucalyptus* species, the most planted hardwood species worldwide and guava (*Psidium Guajava*) [13]. The genome of *Eucalyptus grandis*, published in 2014, was the first reference genome for the Myrtales order and the Myrtaceae family [9]. Anchored on 11 chromosomes, the most conserved chromosome number within the Myrtaceae, the genome of *E. grandis* is a valuable tool for comparative genomic analyses and has been compared to the chromosome-scale assemblies of the Myrtaceae species published later (Table 1).

Table 1 List of published chromosome-scale genome assemblies within the Myrtaceae family.

	Bases in chromosomes (n = 11)	**Reference**
Eucalyptus grandis	605 Mb 612.6 Mb	Myburg *et al.* 2014 [9], Bartholomé et al. 2015 [14]
Leptospermum scoparium	248.8Mb	Thrimawithana *et al.* 2019 [15]
Metrosideros polymorpha [a]	284 Mb	Izuno *et al.* 2016, 2019 [16, 17]
Corymbia citriodora	412.62 Mb	Healey *et al.* 2021 [18]
Psidium guajava	441.27 Mb	Feng *et al.* 2021 [19]
Syzygium aromaticum [b]	368 Mb	Ouadi *et al.* 2022 [20]
Syzygium grande	387.6 Mb	Low *et al.* 2022 [21]
Melaleuca alternifolia	331.1 Mb	Zheng *et al.* 2022 [22]
Rhodomyrtus tomentosa	470.35 Mb	Li *et al.* 2023 [23]
E. urophylla × *E. grandis*	536.43 Mb	Shen *et al.* 2023 [24]
Syzygium malaccense [b]	429 Mb	Ouadi *et al.* 2023[25]
Syzygium aqueum [b]	387 Mb	Ouadi *et al.* 2023[25]
Syzygium jambos [b]	416 Mb	Ouadi *et al.* 2023[25]
Syzygium syzygioides [b]	425 Mb	Ouadi *et al.* 2023[25]

[a] 11 chromosomes constructed based on the *E. grandis* assembly.

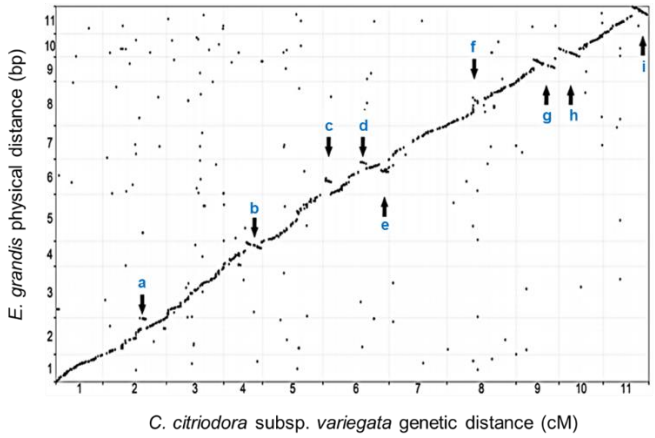
[b] Chromosome-scale genome assemblies generated during the project.

Comparative genetic mapping and genomic analyses revealed high levels of synteny and collinearity among the *Eucalyptus* genus [24, 26, 27]. In 2023, the comparison of the *de novo* genome assembly of the hybrid *E. urophylla* × *E. grandis* (EUC) with 30 *Eucalyptus* species revealed that the genome structure of EUC, *E. grandis*, and *E. globulus* had the higher collinearity level, and the absence of large-scale structural variation between these three *Eucalyptus* species.

In addition, despite a high level of genome synteny, large intra-chromosomal rearrangements (e.g., inversions, translocations) among different chromosomes (Chr 2, 3, 4, 6, 8, and 9) of the EUC and other *Eucalyptus* species were detected [24].

Intergeneric comparative analyses of *E. grandis* and its eucalypts relative *Corymbia citriodora subsp. variegata* (CCV) revealed that the genome structure of the two eucalypts was highly conserved after they diverged approximately 60 million years ago [18, 28, 29]. Nevertheless, the comparison of the *E. grandis* genome assembly with both the high-density linkage map of the CCV and later the *de novo* genome assembly revealed few large intrachromosomal rearrangements located on seven of eleven chromosomes (chromosomes 2, 4, 6, 8, 9, 10, and 11) [18, 29] (Figure 1).

a



Chromosome	Type of rearrangement
2 (a)	Inversion / translocation
4 (b)	Inversion
6 (c)	Inversion / translocation
6 (d)	Inversion / translocation
6 (e)	Inversion / translocation
8 (f)	Inversion
9 (g)	Inversion
10 (h)	Inversion
11 (i)	Inversion

b

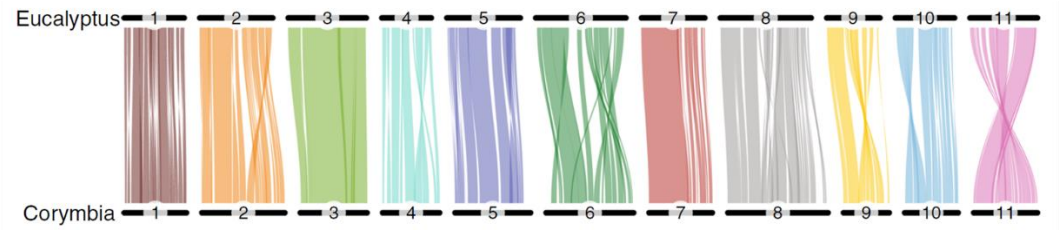


Figure 1 Comparative genomics of *Eucalyptus grandis* and *Corymbia citriodora* subsp. *Variegata* (CCV). (a) Comparative analyses performed using the markers on CCV comprehensive linkage map and *E. grandis* genome assembly and the list of intrachromosomal rearrangements detected (adapted from Butler et al., 2017 [29]). (b) Comparative analyses of the 11 chromosome sequences of CCV and *E. grandis* (copied from Healey et al. 2021 [18]).

Biosynthesis of eugenol

Eugenol (4-allyl-2-methoxyphenol) is a secondary metabolite found in clove and several other plant species belonging to the Myrtaceae (e.g., allspice, *Dioica pimenta*), Lamiaceae (e.g., basil, *Ocimum basilum*), Lauraceae (e.g., cinnamon, *Cinnamomum verum*), and Myristicaceae families (e.g., nutmeg, *Myristica fragrans*). It is classified as a phenylpropane, a large group of volatiles reported as involved in plant defense and reproduction [30, 31].

The biosynthesis of eugenol involves gene families of the general phenylpropanoid pathway which also generates the precursors of the lignin biosynthesis and several other secondary metabolites (e.g., flavonoids, coumarins, anthocyanin) [31]. Eugenol is produced via the branch of phenylpropanoid pathway producing coniferyl alcohol. The two final steps from coniferyl alcohol are catalysed by biosynthetic genes belonging to two families: BAHD (benzyl alcohol-acetyl-, anthocyanin-O-hydroxy-cinnamoyl-, anthranilate-N-hydroxycinnamoyl/benzoyl-, deacetyl-vindoline), and PIP (pinoresinol–lariciresinol reductase, isoflavone reductase, phenylcoumaran benzylic ether reductase) families [32, 33]. Coniferyl alcohol is first converted to coniferyl acetate by an alcohol acyltransferase (AAT) belonging to the BAHD family, then, coniferyl acetate is used as a substrate by eugenol synthase (EGS) belonging to the PIP family to produce eugenol [32]. Similarly, to the gene families of the general phenylpropanoid pathway, genes belonging to the BADH, and PIP families are not restricted to the biosynthesis of eugenol but are also involved in the production of other compounds such as volatile esters (BADH) or phenylpropenes (e.g., chavicol and isoeugenol) (PIP) [32, 33].

Eugenol is the most abundant compound found in clove products accounting for 72% to 96.6% of essential oil (EO) composition. The remaining components of clove EO are the phenylpropene eugenol acetate (0–21.3%) and the sesquiterpenes β-caryophyllene (1.7–19.5%) and α-humulene (0.2–2.2%) [6]. Among *Syzygium* species studied in this project, traces of eugenol were reported in EO of *S. aqueum* (0.19%) obtained by hydrodistillation of the leaves [34]. Among other Myrtaceae genera, eugenol was the main component of EO of allspice (*Pimenta dioica*) leaves and fruits (71.4 % and 65.9%) [35] and was also detected in EO of *Eucalyptus* species leaves (e.g., *E. citriodora* (3.9%), *E. camaldulensis* (0.6%), *E. crebra* (0.7%), E. *tereticornis* (1.8%), *E. melanophloia* (1.7%), and *E. microtheca* (11.8%)) [36, 37].

Analyses of the chemical composition of clove EO from leaves and buds collected at different growth phases showed that the percentages of eugenol and its ester, eugenol acetate, varied greatly during the growth of the two organs [38, 39]. Eugenol acetate was the major constituent of EO produced from young leaves (up to 65.52%) and buds (up to 57%) and only traces were found in EO from mature leaves (0.36 % - 1.64 %) and full fruiting bud (2.01% ± 0.54%). In contrary, the percentage of eugenol in EO increased from 25.43% - 30.38% in young leaves to 88.32% - 90.22% in mature leaves, and from 39.66% ± 3.45 to 94.89 ± 0.66 % in young buds and full fruiting buds respectively. This variation of the percentages of both compounds in EO suggests that eugenol acetate might be an intermediary step in eugenol biosynthesis whereas it was not reported in the eugenol biosynthetic pathway from other plants.

Eugenol is a natural compound recognized for its numerous pharmacological properties including a broad spectrum of inhibitory effect against microorganisms [4, 8]. In plants, eugenol content was reported to vary in response to abiotic stress and to biotic stress as well as to be attractive to insects such as some nocturnal bee's species [40-44]. For instance, a higher amount of eugenol was observed in the leaves of strawberry plants (*Fragaria* × *ananassa* cv. 'Xiaobai') treated with dark, drought, heat, and salt stresses than in non-treated plants. Besides, the expression of two eugenol synthase (FaEGS1 and FaEGS2) and two transcription factors (FaEOBII and FaMYB10) was up-regulated in treated strawberry plants. Cadmium treatment increased the eugenol content in treated strawberry plants, and significantly promoted the expression of FaEGS2 and FaMYB10 but inhibited the expression of FaEGS1 and FaEOBII [40]. In essential oil (EO) extracted from intact mango leaves (*Mangifera indica*), and from leaves exposed to abiotic stress (leaves pierced with a paper punch), eugenol and eugenol acetate were not detected. In contrast, both compounds were detected in EO extracted from leaves exposed to biotic stress (leaf infested with grasshoppers (*Tropidacris collaris*)) [41]. In addition, eugenol was found a promising priming agent to promote thermotolerance, and resistance to fungus and virus [45,46].

The composition of clove EO has been well investigated [6, 38, 39, 47-49]. Furthermore, the analysis of volatiles emitted by living buds revealed that both eugenol and eugenol acetate were emitted [50]. However, little is known about the biological and molecular mechanisms involved in the biosynthesis, accumulation, and regulation of eugenol and eugenol acetate in clove trees.

Aims of the project

The aim of this thesis work is to generate genomic, transcriptomic, and metabolomic data and to analyse these data to investigate the clove genome evolution and the genetic basis of the biosynthesis of eugenol. A reference genome is a valuable tool to understand crop evolution and biology [51-53]. Therefore, these -omics data were first used to construct a reference genome at the chromosome-scale for clove using next-generation sequencing technologies, biological databases, and bioinformatics tools. The clove genome was then used in comparative genomic and multi -omics analyses to start investigating the clove genome structure evolution and the genetic basis of biosynthesis of eugenol in clove leaves and buds.

During this PhD work, the genome of *E. grandis* was extensively used to investigate the clove genome evolution and function using comparative genomic approaches. Synteny and collinearity analysis findings reported between *E. grandis* and CCV were also compared to those obtained between clove and *E. grandis* thus providing a first insight into the genome structure conservation of clove and eucalypts [20]. In addition, to investigate the *Syzygium* genome structure evolution, intrageneric comparative genomics analyses were conducted between the genomes of clove and four *Syzygium* species sequenced during the project *(Syzygium malaccense*, *Syzygium aqueum*, *Syzygium jambos*, and *Syzygium syzygioides*) and the genome of *S. grande* published in 2022 [21] (Table 1). All the *Syzygium* species studied belong to the subgenus *Syzygium* (the largest of the five *Syzygium* subgenera) which originated from Sahul (paleocontinent comprising Australia, New Guinea, and the Aru Islands) and started to diversify approximately 9 Mya [21].

To gain a first insight into eugenol biosynthesis in clove, integrated analyses of multi-omics data were conducted to infer its genetic basis. Firstly, gene families associated with the biosynthesis of eugenol were identified in the clove genome by a homology-based approach and information available from *E. grandis* phenylpropanoid genes and sequences of characterised proteins belonging to the BADH and PIP family. Genes belonging to the ATP-Binding Cassette (ABC) superfamily involved in the transport of volatile compounds were also identified [54]. Secondly, transcriptomic and metabolomic data from clove leaves and buds at different growth phase were generated to identify the copies of the genes involved in the biosynthesis of eugenol on the basis of correlation between expression pattern and eugenol and eugenol acetate concentration.

This Ph.D. thesis is a collaborative project between the University of Neuchâtel and Philip Morris Products S.A. (PMI Research and Development). The genomic and transcriptomic data were generated at PMI in the genomics and transcriptomics laboratory. The metabolomic data acquisition was performed at the University of Neuchâtel, at the Neuchâtel Platform of Analytical Chemistry. The biological material was provided by Dr. Leyla Davis, who allowed and facilitated the sampling of the clove tree and the four others *Syzygium* species studied in the Masoala hall in Zürich Zoo (Figure 2).

Figure 2 Clove tree sampled in the Masoala hall and its buds at different stages of development. (a) young buds, (b) green buds, (c, d) yellow-pink buds, flower; buds in the initial fruiting stage, and (e) buds in fruiting stage.

Outline of the thesis

This PhD dissertation is organised into three chapters. The first chapter of this thesis aims to review clove's main uses and agronomical challenges. In addition, it provides examples on how the current knowledge about genetic resources and the availability of the clove genome could contributes to the development of genetic and genomic tools necessary to design modern improvement programs.

The second chapter titled "The clove (*Syzygium aromaticum*) genome provides insights into the eugenol biosynthesis pathway" was published in the scientific journal Communications Biology. The objectives of this study were to sequence and assemble a reference genome for *S. aromaticum*, as well as to generate combined transcriptomic and metabolomic data sets from clove leaves and buds at different growth phases to investigate the clove genome evolution and the genetic basis of the biosynthesis of eugenol. This article presents the genome features of the clove genome assembly, includes the results of the comparative genomics analyses performed between clove and *E. grandis* as well the findings of the multi-omics study designed to investigate the genetic basis of eugenol in clove leaves and buds.

The last chapter of this thesis reports the *de novo* chromosome-scale genome assemblies for *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides*, as well as comparative genomics findings. In this study, the features and architecture of four newly assembled genomes were compared to those of the clove genome as well as to those of *Syzygium grande* genome—published by Low *et al*.[21]— to provide an insight into the genome evolution of the *Syzygium* genus. This chapter is a research article published in the scientific journal Frontiers in Plant Science.

References

1. Beech, E., Rivers, M., Oldfield, S., and Smith, P., *GlobalTreeSearch: The first complete global database of tree species and country distributions.* Journal of Sustainable Forestry, 2017. 36(5): p. 454-489.

2. POWO, *(2023). "Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet; http://www.plantsoftheworldonline.org/ Retrieved 11 April 2023." https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:327906-2*.

3. Nair, K.N., *The Genus Syzygium: Syzygium Cumini and Other Underutilized Species*. 2017: CRC Press.

4. Danthu, P., Simanjuntak, R., Fawbush, F., Tsy, J.L.P., Razaϭimamonjison, G., Abdillahi, M.M., Jahiel, M., and Penot, E., *The clove tree and its products (clove bud, clove oil, eugenol): prosperous today but what of tomorrow's restrictions?* World, 2020. 161(58,172): p. 52,915.

5. Nurdjannah, N. and Bermawie, N., *Cloves*, in *Handbook of herbs and spices*, K.V. Peter, Editor. 2012, Elsevier. p. 197-215.

6. Razafimamonjison, G., Jahiel, M., Duclos, T., Ramanoelina, P., Fawbush, F., and Danthu, P., *Bud, leaf and stem essential oil composition of Syzygium aromaticum from Madagascar, Indonesia and Zanzibar*. International Journal of Basic and Applied Sciences, 2014. 3(3): p. 224.

7. Cortés-Rojas, D.F., de Souza, C.R.F., and Oliveira, W.P., *Clove (Syzygium aromaticum): a precious spice.* Asian Pac. J. Trop. Biomed., 2014. 2(4): p. 90–96.

8. Batiha, G.E.-S., Alkazmi, L.M., Wasef, L.G., Beshbishy, A.M., Nadwa, E.H., and Rashwan, E.K., *Syzygium aromaticum L.(Myrtaceae): Traditional uses, bioactive chemical constituents, pharmacological and toxicological activities.* Biomolecules, 2020. 10(2): p. 202.

9. Myburg, A.A., et al., *The genome of Eucalyptus grandis.* Nature, 2014. 510(7505): p. 356-362.

10. Ren, L., Huang, W., Cannon, E.K., Bertioli, D.J., and Cannon, S.B., *A mechanism for genome size reduction following genomic rearrangements.* Frontiers in Genetics, 2018. 9: p. 454.

11. Zhang, L., et al., *A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour.* Nature communications, 2019. 10(1): p. 1494.

12. Zhang, T., et al., *Genome of Crucihimalaya himalaica, a close relative of Arabidopsis, shows ecological adaptation to high altitude.* Proceedings of the National Academy of Sciences, 2019. 116(14): p. 7137-7146.

13. Grattapaglia, D., Vaillancourt, R.E., Shepherd, M., Thumma, B.R., Foley, W., Külheim, C., Potts, B.M., and Myburg, A.A., *Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus.* Tree Genet. Genomes, 2012. 8(3): p. 463–508.

14. Bartholomé, J., Mandrou, E., Mabiala, A., Jenkins, J., Nabihoudine, I., Klopp, C., Schmutz, J., Plomion, C., and Gion, J.M., *High-resolution genetic maps of E ucalyptus improve Eucalyptus grandis genome assembly.* New Phytologist, 2015. 206(4): p. 1283-1296.

15. Thrimawithana, A.H., et al., *A whole genome assembly of Leptospermum scoparium (Myrtaceae) for mānuka research.* N. Z. J. Crop Hortic. Sci., 2019. 47(4): p. 233–260.

16. Izuno, A., Hatakeyama, M., Nishiyama, T., Tamaki, I., Shimizu-Inatsugi, R., Sasaki, R., Shimizu, K.K., and Isagi, Y., *Genome sequencing of Metrosideros polymorpha (Myrtaceae), a dominant species in various habitats in the Hawaiian Islands with remarkable phenotypic variations.* J. Plant Res., 2016. 129(4): p. 727–736.

17. Izuno, A., Wicker, T., Hatakeyama, M., Copetti, D., and Shimizu, K.K., *Updated Genome Assembly and Annotation for Metrosideros polymorpha, an Emerging Model Tree Species of Ecological Divergence.* G3: Genes, Genomes, Genetics, 2019. 9(11): p. 3513-3520.

18. Healey, A.L., et al., *Pests, diseases, and aridity have shaped the genome of Corymbia citriodora.* Communications biology, 2021. 4(1): p. 1-13.

19. Feng, C., Feng, C., Lin, X., Liu, S., Li, Y., and Kang, M., *A chromosome-level genome assembly provides insights into ascorbic acid accumulation and fruit softening in guava (Psidium guajava).* Plant biotechnology journal, 2021. 19(4): p. 717-730.

20. Ouadi, S., Sierro, N., Goepfert, S., Bovet, L., Glauser, G., Vallat, A., Peitsch, M.C., Kessler, F., and Ivanov, N.V., *The clove (Syzygium aromaticum) genome provides insights into the eugenol biosynthesis pathway.* Communications Biology, 2022. 5(1): p. 1-13.

21. Low, Y.W., et al., *Genomic insights into rapid speciation within the world's largest tree genus Syzygium.* Nature communications, 2022. 13(1): p. 1-15.

22. Zheng, X., et al., *The chromosome-level Melaleuca alternifolia genome provides insights into the molecular mechanisms underlying terpenoids biosynthesis.* Industrial Crops and Products, 2022. 189: p. 115819.

23. Li, F., et al., *Gap-free genome assembly and comparative analysis reveal the evolution and anthocyanin accumulation mechanism of Rhodomyrtus tomentosa.* Horticulture Research, 2023.

24. Shen, C., Li, L., Ouyang, L., Su, M., and Guo, K., *E. urophylla× E. grandis high-quality genome and comparative genomics provide insights on evolution and diversification of eucalyptus.* BMC genomics, 2023. 24(1): p. 1-10.

25. Ouadi, S., Sierro, N., Kessler, F., and Ivanov, N.V., *Chromosome-scale assemblies of S. malaccense, S. aqueum, S. jambos, and S. syzygioides provide insights into the evolution of Syzygium genomes.* Frontiers in Plant Science, 2023. 14.

26. Hudson, C.J., Kullan, A.R., Freeman, J.S., Faria, D.A., Grattapaglia, D., Kilian, A., Myburg, A.A., Potts, B.M., and Vaillancourt, R.E., *High synteny and colinearity among Eucalyptus genomes revealed by high-density comparative genetic mapping.* Tree Genetics & Genomes, 2012. 8(2): p. 339-352.

27. Li, F., Zhou, C., Weng, Q., Li, M., Yu, X., Guo, Y., Wang, Y., Zhang, X., and Gan, S., *Comparative genomics analyses reveal extensive chromosome colinearity and novel quantitative trait loci in Eucalyptus.* PloS one, 2015. 10(12): p. e0145144.

28. Thornhill, A.H., Crisp, M.D., Külheim, C., Lam, K.E., Nelson, L.A., Yeates, D.K., and Miller, J.T., *A dated molecular perspective of eucalypt taxonomy, evolution and diversification.* Aust. Syst. Bot., 2019. 32(1): p. 29–48.

29. Butler, J., Vaillancourt, R., Potts, B., Lee, D., King, G.J., Baten, A., Shepherd, M., and Freeman, J., *Comparative genomics of Eucalyptus and Corymbia reveals low rates of genome structural rearrangement.* BMC Genom., 2017. 18(1): p. 397.

30. Atkinson, R.G., *Phenylpropenes: occurrence, distribution, and biosynthesis in fruit.* J. Agric. Food Chem., 2016. 66(10): p. 2259–2272.

31. Vogt, T., *Phenylpropanoid biosynthesis.* Mol. Plant, 2010. 3(1): p. 2–20.

32. Koeduka, T., *Functional evolution of biosynthetic enzymes that produce plant volatiles.* Biosci. Biotechnol. Biochem., 2018. 82(2): p. 192–199.

33. D'Auria, J.C., *Acyltransferases in plants: a good time to be BAHD.* Curr. Opin. Plant Biol., 2006. 9(3): p. 331–340.

34. Sobeh, M., Braun, M.S., Krstin, S., Youssef, F.S., Ashour, M.L., and Wink, M., *Chemical profiling of the essential oils of Syzygium aqueum, Syzygium samarangense and Eugenia uniflora and their discrimination using chemometric analysis.* Chemistry & biodiversity, 2016. 13(11): p. 1537-1550.

35. Mérida-Reyes, M.S., Muñoz-Wug, M.A., Oliva-Hernández, B.E., Gaitán-Fernández, I.C., Simas, D.L.R., Ribeiro da Silva, A.J., and Pérez-Sabino, J.F., *Composition and antibacterial activity of the essential oil from Pimenta dioica (L.) Merr. from Guatemala.* Medicines, 2020. 7(10): p. 59.

36. Ghaffar, A., et al., *Chemical composition and in-vitro evaluation of the antimicrobial and antioxidant activities of essential oils extracted from seven Eucalyptus species.* Molecules, 2015. 20(11): p. 20487-20498.

37. Salehi, B., et al., *Insights into Eucalyptus genus chemical constituents, biological activities and health-promoting effects.* Trends in Food Science & Technology, 2019. 91: p. 609-624.

38. Razafimamonjison, G., Boulanger, R., Jahiel, M., Ramanoelina, P., Fawbush, F., Lebrun, M., and Danthu, P., *Variations in yield and composition of leaf essential oil from Syzygium aromaticum at various phases of development.* Int. J. Basic Appl. Sci., 2016. 5(1): p. 90.

39. Razafimamonjison, G., Jahiel, M., Ramanoelina, P., Fawbush, F., and Danthu, P., *Effects of phenological stages on yield and composition of essential oil of Syzygium aromaticum buds from Madagascar.* Int. J. Basic Appl. Sci., 2013. 2(4): p. 312–318.

40. Wang, S., Zhou, D., Shi, M., Feng, H., Xie, X., Sun, P., Xue, H., Fang, C., and Zhao, J., *Expression patterns of four key genes involved in strawberry eugenol synthesis under abiotic stresses.* Acta Ecologica Sinica, 2022. 42(1): p. 68-75.

41. Silva, R.R.d., da Câmara, C.A., Almeida, A.V., and Ramos, C.S., *Biotic and abiotic stress-induced phenylpropanoids in leaves of the mango (Mangifera indica L., Anacardiaceae).* Journal of the Brazilian Chemical Society, 2012. 23: p. 206-211.

42. Rezaie, R., Abdollahi Mandoulakani, B., and Fattahi, M., *Cold stress changes antioxidant defense system, phenylpropanoid contents and expression of genes involved in their biosynthesis in Ocimum basilicum L.* Scientific reports, 2020. 10(1): p. 5290.

43. Reddy, V.A., Li, C., Nadimuthu, K., Tjhang, J.G., Jang, I.-C., and Rajani, S., *Sweet basil has distinct synthases for eugenol biosynthesis in glandular trichomes and roots with different regulatory mechanisms.* International Journal of Molecular Sciences, 2021. 22(2): p. 681.

44. Martínez-Martínez, C.A., et al., *Floral Volatiles: A Promising Method to Access the Rare Nocturnal and Crepuscular Bees.* Frontiers in Ecology and Evolution, 2021. 9: p. 676743.

45. Tsai, W.-A., Weng, S.-H., Chen, M.-C., Lin, J.-S., and Tsai, W.-S., *Priming of plant resistance to heat stress and tomato yellow leaf curl Thailand virus with plant-derived materials.* Frontiers in Plant Science, 2019. 10: p. 906.

46. Jiang, N., Wang, L., Jiang, D., Wang, M., Yu, H., and Yao, W., *Combined metabolome and transcriptome analysis reveal the mechanism of eugenol inhibition of Aspergillus carbonarius growth in table grapes (Vitis vinifera L.).* Food Research International, 2023. 170: p. 112934.

47. Hastuti, L., Saepudin, E., Cahyana, A., Rahayu, D., Murni, V., and Haib, J. *The influence of sun drying process and prolonged storage on composition of essential oil from clove buds (Syzygium aromaticum)*. in *International Symposium on Current Progress in Mathematics and Sciences 2016, ISCPMS 2016: Proceedings of the 2nd International Symposium on Current Progress in Mathematics and Sciences 2016*. AIP Publishing.

48. Alfikri, F.N., Pujiarti, R., Wibisono, M.G., and Hardiyanto, E.B., *Yield, Quality, and Antioxidant Activity of Clove (Syzygium aromaticum L.) Bud Oil at the Different Phenological Stages in Young and Mature Trees.* Scientifica, 2020. 2020: p. 9701701.

49. Murni, V., Saepudin, E., Cahyana, A., Rahayu, D., Hastuti, L., and Haib, J. *Effect of oven drying and storage on essential oil composition of clove (Syzygium aromaticum) from Toli-Toli*. in *International Symposium on Current Progress in Mathematics and Sciences 2016, ISCPMS 2016: Proceedings of the 2nd International Symposium on Current Progress in Mathematics and Sciences 2016*. AIP Publishing.

50. Kasai, H., Shirao, M., and Ikegami-Kawai, M., *Analysis of volatile compounds of clove (Syzygium aromaticum) buds as influenced by growth phase and investigation of antioxidant activity of clove extracts.* Flavour and fragrance journal, 2016. 31(2): p. 178-184.

51. Wendel, J.F., Jackson, S.A., Meyers, B.C., and Wing, R.A., *Evolution of plant genome architecture.* Genome biology, 2016. 17: p. 1-14.

52. Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., and Mayer, K.F., *Plant genome sequencing—applications for crop improvement.* Current opinion in biotechnology, 2014. 26: p. 31-37.

53. Cheng, Q.-Q., Ouyang, Y., Tang, Z.-Y., Lao, C.-C., Zhang, Y.-Y., Cheng, C.-S., and Zhou, H., *Review on the development and applications of medicinal plant genomes.* Frontiers in Plant Science, 2021. 12: p. 791219.

54. Adebesin, F., et al., *Emission of volatile organic compounds from petunia flowers is facilitated by an ABC transporter.* Science, 2017. 356(6345): p. 1386-1388.

Chapter 1 – Preface

This first chapter is an unpublished manuscript reviewing clove's main uses, the agronomical challenges and the opportunities that could be provided by the development of -omics resources for clove. It was prepared with the aim of being included as a chapter in a book on genetics on Genetics, Genomics and Breeding in Seed Spices for the book series titled "Advances in Agri Genomics" from Taylor & Francis group.

Chapter 1 - Clove spice – applications, challenges and opportunities

Sonia Ouadi[1,2], Nicolas Sierro[2], Simon Goepfert[2], Felix Kessler[1], Nikolai V. Ivanov[2].

[1] Faculty of Sciences, Laboratory of Plant Physiology, University of Neuchâtel, 2000 Neuchâtel, Switzerland.

[2] PMI R&D, Philip Morris Products S. A, Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland.

Abstract

Clove (*Syzygium aromaticum*) is a valuable crop, native to the Maluku Islands in Indonesia, belonging to the Myrtaceae family. Its dried, unopened flower buds (clove), essential oil and oleoresins, rich in eugenol, are used commercially in a wide range of industries. Despite its economic importance, popular medicinal properties and the promising therapeutic potential of eugenol, scarce genetic and genomic information was available before the publication of the clove genome in 2022. The availability of a high-quality genome assembly for clove and its small size (370 Mb) opens new perspectives for clove fundamental and applied research.

In this chapter, we describe clove's main uses, the agronomical challenges and the opportunities that could be provided by the development of -omics resources for clove. We also outline the genetic resources currently available for this species and how—combined with the availability of the clove genome—they could contribute to the development of modern characterization and selection techniques required to design crop improvement programs.

1.1 The clove *S. aromaticum* (L. Merril & Perry): a multipurpose tropical crop

Clove, *S. aromaticum*, is a medium-sized (8-12 m) tropical evergreen tree belonging to the Myrtaceae family, which also includes myrtle, eucalyptus, and guava. Clove is the species with the most commercial importance of the large and diverse *Syzygium* genus [1, 2]. It is cultivated to produce clove bud (the dried unopened flower bud), essential oil (EO), and oleoresins rich in eugenol (4-allyl-2-methoxyphenol).

Clove products are commercially used in a wide range of applications within the pharmaceutical, cosmetic, food, tobacco, and agricultural industries [3-5]. Cloves are used as an aromatic spice in gastronomy, as a flavoring agent in the food industry and in the tobacco industry to manufacture Kreteks, an Indonesian cigarette made of tobacco and minced, dried clove buds. Clove buds and stems can be used to produce oleoresin (by solvent extraction) that is used as a flavoring agent [4]. The clove EO is extracted by distillation from buds, leaves, and stems. The major components of clove EO are the phenylpropenes eugenol (72–96.6%), eugenol acetate (0–21.3%), the sesquiterpenes β-caryophyllene (1.7–19.5%), and α-humulene (0.2–2.2%) [6]. Clove EO is mainly extracted for the production of caryophyllene and eugenol and is commercially used in the fragrance and pharmaceutical industries, for instance, in perfumery, for the production of soaps and toothpastes and in medical and dental practice for its analgesic and general antiseptic properties [4, 5, 7, 8].

The medicinal properties of clove are principally attributed to its high amount of eugenol, a versatile natural compound with analgesic, anesthetic, antioxidant, anti-inflammatory, antibacterial, antifungal, and antiviral activities. These properties are exploited by several industries, and numerous scientific studies have investigated eugenol's mechanisms of action and assessed its potential as a promising natural alternative solution for human health issues (e.g., cancer and pathogenic microorganism resistance, diabetes, obesity, autoimmune diseases) and sustainable agriculture (e.g., decreasing the use of synthetic preservatives and pesticides) [4, 5, 7-10].

1.2 Clove production

1.2.1 Clove trade history

Clove is native to the Maluku Islands in east Indonesia, also known as the "spice islands." A clove dating back to 900–1000 AD was found in the ancient port of Mantai in Sri Lanka, and it constitutes an archaeobotanical evidence of an early long-distance trade of the spice [11]. The trade history of clove is marked by the establishment of the route to the "spice islands" early in the 16th century by Portuguese explorers in search of valuable spices. The Portuguese were dispossessed by the Dutch in 1605. To ensure the monopoly of the trade, the Dutch instituted a policy of extirpation, eliminating trees on the Maluku Islands to introduce and to concentrate the production on one island, Ambon [12]. Clove was traded from the "spice islands" to Europe under the control of the Dutch East India company, until a few specimens were stolen from Seram island in 1770 during an expedition ordered by Pierre Poivre, the French administrator on Île de France and Bourbon (Mauritius and Reunion, respectively), and introduced in Mauritius [12, 13]. Clove trees were then introduced to different parts of the world (e.g., Tanzania, Madagascar, India), where they adapted well to ecological conditions and gained economic importance [3, 13]. In the early 20th century, trees from Zanzibar (Zanzibar type) were reintroduced to Indonesia where they exhibited a higher yield than Indonesian traditional varieties and widely cultivated [12].

1.2.2 Clove production

Clove—together with pepper, capsicum, nutmeg, cardamom, allspice, vanilla, ginger, cinnamon, cassia, and turmeric—is one of the most important tropical spice crops in global trade, with an estimated production of 186,969 tons in 2021, a production level that lies between those of cinnamon and nutmeg (at 226,753 and 146,952 tons, respectively) [14, 15]. The current main producers are Indonesia, Madagascar, and Tanzania, but clove is also produced in other countries such as in Sri Lanka, Brazil, Comoros, Kenya, China, and India. Indonesia is currently the largest producer and consumer of clove in the world, with the majority of the Indonesian clove production being used to manufacture Kreteks [4, 5].

Clove, however, is a crop with limitations affecting farm revenues and price. Clove trees are propagated by seed and only start to bear flowers after 4 to 10 years, depending on their geographical location [3, 13]. Clusters of buds are hand-picked (before flowering), and buds are hand-separated from the cluster. The collected buds are then sun-dried or dried in a desiccator. The optimal stage for harvesting the buds lasts only a few days, and the cost of manual harvesting exceeds 40% of the total production costs in Indonesia [16]. To address the problem of manual harvesting and reduce production cost, hormone-like chemicals were applied to clove trees to induce bud fall, but this approach has not been successful [16].

Clove is also susceptible to diseases and pests such as Sumatra disease (caused by *Pseudomonas solanacearum*), Cacar daun (caused by fungal pathogen *Phyllostiticta* sp.) in Indonesia, the "sudden death" (caused by a fungus *Valsa eugeniae*) in Zanzibar. In Madagascar, the main pest is andretra, the larvae of the leptidoptera *Chrysotypus mabilianum* [12, 13].

Most importantly, clove is an irregular-bearing spice crop. Its yield fluctuates largely from year to year, and little is known about the external (environment and cultivation practices) and internal (genetic and molecular mechanism) factors causing the irregular bearing pattern [17-19]. To overcome yield fluctuations and stabilize annual production, the effect of the growth retardant paclobutrazol (PBZ) was assessed on clove trees from West Java. Application by watering with PBZ the stems and roots around the clove tree base resulted in better shoots and clove production per tree within 3 consecutive years [20]. PBZ application was also used to produce shorter seedlings of forest clove, a wild type species characterized by rapid growth and a height up to 20 m making harvesting difficult [21].

Depending on the region, clove productivity can vary by a factor of 6. For example in 2020, the yield reported for Indonesia was 241.7 kg/ha while for Tanzania it was 1261.8 kg/ha [15]. Aside from possible differences in yield calculation, productivity is likely affected by the quality of seeds used by farmers, damage to and maintenance of the trees, as well as aging of trees [22]. Depending on the climatic conditions, production of cloves can also vary. For instance, in 2018, unusually moist conditions in Madagascar resulted in an annual production of around 10 to 20% in an average year [5, 23].

Major challenges for clove improvement are to increase productivity while conserving the quality of the product, and at the same time reducing the cost of harvesting and the yield losses caused by the irregular bearing and diseases. Possible objectives of clove improvement programs could be the development of plant material more resistant to biotic and abiotic stresses with for example, rapid juvenile growth, early and synchronized bearing, low branching and producing a high yield of large buds and dense large foliage that can be used to produce large amounts of eugenol-rich EO.

1.3 Genetic resources

1.3.1 Diversity among clove trees

The clove trade history has impacted its actual geographical distribution, and the eradication of indigenous populations by the Dutch in the 17th century could have also drastically affected the biodiversity of the species, principally in the Northern islands that is referred as the species' center of origin. Nevertheless, a survey of 34 wild and cultivated clove populations based on 35 morphological characters was carried out in 1983 in the Maluku islands and it revealed considerable diversity [24]. The populations investigated comprised: popularly recognized cultivated varieties (Zanzibar, Siputih and Sikotok), cultivated types believed to be growing only on the Maluku islands (Tibobo, Tauro, Sibela, Indari, Dokiri, AFO 1, AFO 2, and Air mata), a wild type (Cengkeh hutan), hybrids (e.g., Cengkeh raja 1 to 4 that local farmers consider hybrids between Siputih and Cengkeh hutan trees) and two distinct individuals (Daun buntal and a dwarf mutant). More variation was observed among wild cloves compared to cultivated populations; nevertheless, several of the cultivated populations were found to be morphologically distinct. Based on the survey results, it was suggested that the conservation and pattern of observed variation could be due to a combination of two factors that possibly resulted in genetic isolation: the scattered and insular distribution of many of the cultivated populations on the Maluku islands and seed utilization from local mother trees.

Clove was reported as being a cross-pollinating species [22, 25, 26]. Several subsequent surveys were conducted in Indonesia to map and report morphological variations among clove trees for the purposes of crop improvement and conservation [25-30]. Some also assessed EO characteristics including the eugenol content [22, 31-34].

Among the populations grown by farmers, five superior clove varieties were selected based on their productivity and the quality of the clove: Zanzibar Karo, Zanzibar Gorontalo, Tuni Bursel, AFO and Siantan Agribun [30].

Clove trees cultivated outside the Maluku Islands are descendants of a few specimens originating from the Maluku Islands and possess low levels of genetic diversity [12]. However, variants were reported in India (dwarf and bushy trees, small and narrow leaves variants and a variant bearing long and bold cloves with long, thick, broad leaves) [35-37].

1.3.2 Example of genetic resources valuable for clove improvement

Wild relatives of crops have been very valuable genetic resources for crop improvement [38-40]. For this reason, several research groups evaluated the genetic diversity of wild clove and cultivated local clove using morphological characters of the plants and biochemical characters of EO. Populations and accessions with interesting morpho-agronomical characters have been reported and represent valuable genetic resources for clove improvement programs.

For the purpose of increasing clove production in Indonesia, studies were successfully conducted to select mother trees with high productivity. Twenty mother trees from a highly productive population belonging to a farmer in the Anambas Islands Regency were selected with a production potential of more than 100 kg of flowers per tree instead of the population's production average of 85 kg per tree. In addition, the reported content of EO and main chemical components of the selected trees met the Indonesian National Standard of clove quality [22]. With the same goal of selecting mother trees with high productivity, another study investigated morphological variations among the progenies of seeds of 129 high-yielding mother trees collected from 13 regions in Indonesia [30]. The analysis revealed 72% similarity in morphologic characters among the clove progenies. Three distinct groups were identified in the dendrogram generated from morphological characters recorded from the high-yielding mother trees' progenies, indicating different genetic backgrounds.

Mother trees with high productivity constitute a valuable source of high-quality seeds and genetic diversity that could be exploited for clove improvement programs aiming to increase the clove production yield.

Dwarf clove trees were reported in India and Indonesia and have multiple advantages. Their small size (2 m tall and a canopy width of 5 m for Indian variants) would facilitate maintenance and harvesting tasks. It would allow farmers to grow more trees per unit area [24, 35-37].

An important commercial trait for clove products is the high content of eugenol. Commercial EO from clove trees contains ~72-96.6% eugenol [6]; in contrast, EOs from forest clove (a wild type of clove) have lower eugenol content (0.66-28%) and lower oil content [31, 32, 34]. Nevertheless, forest clove leaves are larger and thicker, and other reported advantages of forest clove could be sturdier growth, resistance to pests and diseases, yearly production with less fluctuation, heavier flower weights, and rapid growth. These are characteristics that may be useful for crop improvement [27].

1.4 The clove genome: a new research tool

1.4.1 Background

To date, the genetic diversity of the clove trees was evaluated based on morpho-agronomical characters, EO yield, and chemical compound content. It enabled the identification of plant material with desirable characters but not yet characterized using modern molecular tools due to the lack of genomic and genetic data. Before the publication of the clove genome, the main DNA sequences available for clove were the complete chloroplast genome (159,370 bp) (NCBI Reference Sequence: NC_047249) as well as nuclear ribosomal DNA and internal transcribed spacer (ITS) and external transcribed spacer (ETS) regions [41]. Few studies using DNA sequences to assess the genetic diversity of clove were reported. ITS sequences were used in phylogenetic analyses between two local varieties of cloves from North Maluku: AFO, considered as the oldest clove native from Ternate, and Sibela as the oldest clove native from Bacan [42]. Random amplification of polymorphic DNA (RAPD) analysis was performed on 6 populations growing at different altitudes at Nilgiris, India, enabling the identification of 10 RAPD markers [43]. For other *Syzygium* species, RAPD, simple sequence repeat (SSR) and amplified fragment length polymorphism (AFLP) analysis have been already conducted, and RAPD markers and polymorphic microsatellite loci were reported (e.g., *S. cumini, S. samarangense, S. paniculatum, S. sayeri*) [44-48].

The publication of the clove genome is an important milestone in clove research. Henceforth, a genome assembly as well as transcriptomic and metabolomic data are available and can be exploited to investigate clove genome evolution and identify gene families of interest [41]. The clove genome is also a valuable tool that can be used to characterize genetic resources such as those described in the section 3.3 and improve the selection of clove trees using modern molecular tools [49, 50]. Lastly, in 2022, the publication of the genome of *S. grande* along with 182 re-sequenced Syzygium species and 58 re-sequenced unidentified taxa have increase the genomic resources available for comparative genomics studies among the *Syzygium* [51].

1.4.2 Clove genome features and initial findings

Based on the results of an earlier karyotype study, clove was found to be a diploid with x = 11 chromosomes, the most conserved chromosome number in the Myrtaceae family [52, 53]. Taking advantage of advances in new sequencing technologies such as the long-read from Oxford Nanopore Technologies and Hi-C, a high-quality *de novo* chromosome-scale assembly was generated from a tree originating from Madagascar [41]. The clove genome assembly is of small size (370 Mb) and highly contiguous, with 99.3% of the total assembly length contained in the 11 chromosome sequences. The assessment of the genome assembly and annotation completeness performed using the BUSCO method indicated a high level of completeness, comparable to the completeness level of the Myrtaceae relative *Eucalyptus grandis* [41, 54]. The comparison of the clove and *E. grandis* genomes revealed good conservation of the chromosomal organization between the chromosomes of the two species, as well as the presence of intrachromosomal rearrangements (principally terminal inversions) on 7 of the 11 chromosomes. Since the main characteristic of clove products is the high eugenol content, the clove genome was also exploited to investigate the genetic basis of this important character. First, analysis of the clove genome enabled the identification of gene families known to be involved in the biosynthesis of the precursor of lignin and phenylpropene volatiles (eugenol, isoeugenol, chavicol). Secondly, combined transcriptomics and metabolomics datasets enabled the identification of genes possibly involved in eugenol biosynthesis based on Pearson correlation between the expression profile of genes belonging to the biosynthetic families and the concentrations of eugenol and eugenol acetate measured in leaves and buds at different growth phases [41].

These findings constitute the basis for further research to understand how eugenol is produced and accumulated in high amounts in clove leaves and buds. It is also a starting point toward strategies to improve the clove production yield and simultaneously conserve the quality of the clove product for which a high amount of eugenol is a requirement. Nevertheless, further analyses are required to verify the functions of the list of genes identified as possibly involved in eugenol biosynthesis. *In vitro* characterization of enzymes encoded by these selected genes and comparative -omic approaches between wild cloves (e.g., forest clove that produces lesser amounts of eugenol) and cultivated clove would greatly clarify the genetic basis and environmental factors responsible for eugenol biosynthesis in cultivated clove types.

1.4.3 Possible applications for clove improvement

Wild clove and cultivated endogenous clove trees with interesting morpho-agronomical characters constitute valuable genetic resources for basic and applied research [38-40]. Combined with the availability of the clove genome, these genetic resources could facilitate assessment of intraspecific genetic polymorphisms, the development of genome-wide markers, the identification of genes encoding agronomically important traits and lead to the development of genomics resources necessary to design improvement programs.

The clove genome possesses two advantages: its small size and its high quality. It can thus be used as reference to which genotyping and resequencing data from relevant genetic resources can be mapped for the purpose of building comprehensive genetic maps and new genome assemblies more quickly and at lower cost.

New genomics information generated by genotyping and/or resequencing projects could be first exploited to further investigate the genetic diversity among clove tree populations and allow the development of new genomic resources such as genome-wide markers (e.g., single nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs)) and the estimation of important genetic parameters (e.g., narrow-sense heritability (h^2) of desirable agronomic characters) that are important for future clove improvement programs. For example, 3176 genome-wide location-specific polymorphic SSR markers were discovered using the genome assembly of black pepper (*Piper nigrum*) as reference and genotyping data (double digest restriction-site associated DNA (ddRAD) sequence data) from 29 genotypes including released varieties, important germplasm accessions, and wild relatives [55].

The availability of genomics resources for clove could enable the implementation of genomics-assisted breeding (GAB) approaches by using different methods such as genomic selection (GS) [56, 57]. GS has been reported as a promising approach to accelerate and improve the efficacy of forest tree breeding for complex traits [58, 59]. Most of the current published experimental GS studies in forest trees were conducted for species belonging to the economically important genera *Eucalyptus*, *Picea*, and *Pinus* and performed for traits related to growth and wood quality. Fewer studies were reported for other traits such disease resistance (e.g., Loblolly pine (*Pinus taeda*), American chestnut (*Castanea dentata* (Marshall) Borkh), *Fraxinus excelsior* L.), terpene and essential oil content (*Eucalyptus polybractea*), rubber production (*Hevea brasiliensis*), flowering intensity (*Eucalyptus cladocalyx* F. Muell.) and the timing of budburst (*Picea sitchensis* (Bong.) Carr) [58, 59].

For clove, depending on availability of genome-wide markers, GS could be performed to improve economically important complex traits related to yield, harvesting costs and the biosynthesis of eugenol and other secondary compounds. It would allow clove breeders to select appropriate parents more efficiently for crossing programs, as well as develop novel commercial clove varieties with desirable traits using fewer breeding cycles. However, unless varieties with rapid juvenile growth are established rapidly, a limitation for any clove breeding program is the time necessary for clove trees to reach maturity and produce buds, a generation in clove breeding would indeed require 4 to 10 years. The time and cost required for the design, implementation and execution of GS experimental trials also must be taken into consideration for the development of improved clove trees.

The development of robust vegetative propagation methods is also important to accelerate breeding strategies and for the rapid clonal multiplication of future improved clove varieties. While *in vitro* micropropagation of clove from seedling explants has been reported [60, 61], clove trees are currently commercially propagated by seeds collected from mother trees. Some cutting, layering, grafting and budding experiments were also reported indicating that clove can be propagated vegetatively [62]. However, successful micropropagation of clove from mature shoot explants has not been reported [62, 63]. The grafting approach was also attempted using dwarf clove as the scion and ordinary clove as the rootstock and inversely [64].

The results indicated that the rootstock had a definite influence on the scion growth. Graft union between ordinary clove rootstock and dwarf clove as the scion was obtained with a 64% success rate, and graft union between dwarf clove rootstock and ordinary clove as the scion was successful 50% of the time. The availability of the clove genome could also contribute to propagation method optimization. For example, omics analysis in grafting have provided insights into the molecular aspects of the rootstock-scion interaction and grafting responses for woody fruit crops (e.g., grapevine, apple, citrus) and forest trees (e.g., *Hevea brasilensis*) [65, 66]. Furthermore, future omics studies could contribute to the development of molecular markers (e.g., transcripts, proteins, and metabolites) and epi-markers to enable the early detection of compatible/incompatible graft combinations and selection of superior grafts for woody crop improvement programs [65-67].

1.5 Conclusions

Clove is one of the most ancient spice crops. It is currently used in a wide range of industries, and the great interest in eugenol properties could lead to new applications in medicine and agriculture. Yet, clove is a crop with specific challenges. The clove genome is a valuable tool for basic and applied research. The availability of the clove genome and knowledge about genetic resources open new avenues for the development of genetic and genomic tools necessary to design modern improvement programs and deliver enhanced clove varieties.

References

1. Craven, L.A. and Biffin, E., *An infrageneric classification of Syzygium (Myrtaceae).* Blumea-Biodiversity, Evolution and Biogeography of Plants, 2010. 55(1): p. 94-99.

2. Nair, K.N., *The Genus Syzygium: Syzygium Cumini and Other Underutilized Species*. 2017: CRC Press.

3. Cortés-Rojas, D.F., de Souza, C.R.F., and Oliveira, W.P., *Clove (Syzygium aromaticum): a precious spice.* Asian Pac. J. Trop. Biomed., 2014. 2(4): p. 90–96.

4. Nurdjannah, N. and Bermawie, N., *Cloves*, in *Handbook of herbs and spices*, K.V. Peter, Editor. 2012, Elsevier. p. 197-215.

5. Danthu, P., Simanjuntak, R., Fawbush, F., Tsy, J.L.P., Razaɓimamonjison, G., Abdillahi, M.M., Jahiel, M., and Penot, E., *The clove tree and its products (clove bud, clove oil, eugenol): prosperous today but what of tomorrow's restrictions?* World, 2020. 161(58,172): p. 52,915.

6. Razafimamonjison, G., Jahiel, M., Duclos, T., Ramanoelina, P., Fawbush, F., and Danthu, P., *Bud, leaf and stem essential oil composition of Syzygium aromaticum from Madagascar, Indonesia and Zanzibar*. International Journal of Basic and Applied Sciences, 2014. 3(3): p. 224.

7. Kamatou, G.P., Vermaak, I., and Viljoen, A.M., *Eugenol—from the remote Maluku Islands to the international market place: a review of a remarkable and versatile molecule*. Molecules, 2012. 17(6): p. 6953-6981.

8. Batiha, G.E.-S., Alkazmi, L.M., Wasef, L.G., Beshbishy, A.M., Nadwa, E.H., and Rashwan, E.K., *Syzygium aromaticum L.(Myrtaceae): Traditional uses, bioactive chemical constituents, pharmacological and toxicological activities.* Biomolecules, 2020. 10(2).

9. Marchese, A., et al., *Antimicrobial activity of eugenol and essential oils containing eugenol: A mechanistic viewpoint.* Critical reviews in microbiology, 2017. 43(6): p. 668-689.

10. Otunola, G.A., *Culinary spices in food and medicine: an overview of Syzygium aromaticum (L.) Merr. and LM Perry [Myrtaceae]*. Frontiers in Pharmacology, 2022. 12: p. 3817.

11. Kingwell-Banham, E., Bohingamuwa, W., Perera, N., Adikari, G., Crowther, A., Fuller, D.Q., and Boivin, N., *Spice and rice: pepper, cloves and everyday cereal foods at the ancient port of Mantai, Sri Lanka.* antiquity, 2018. 92(366): p. 1552-1570.

12. Smith, N.J., Williams, J.T., Plucknett, D.L., and Talbot, J.P., *Tropical forests and their crops*. 1992: Cornell University Press.

13. Danthu, P., et al., *The clove tree of Madagascar: a success story with an unpredictable future.* Bois et forêts des tropiques, 2014. 320(2): p. 83-96.

14. Douglas, M., Heyes, J., and Smallfield, B., *Herbs, Spices and Essential Oils: Post-Harvest Operations in Developing Countries*. Vol. 61. 2005: UNIDO and FAO.

15. FAOSTAT. [cited 2022; Available from: https://www.fao.org/faostat/en/#data/QCL.

16. Baietto, M., *Bud Fall Induction in Clove (Syzygiumaromaticum).* Academic Research International, 2014. 5(4): p. 23.

17. Martin, P.J., Butler, D.R., and Dabek, A.J., *Causes of irregular clove production in the islands of Zanzibar and Pemba.* Experimental agriculture, 1988. 24(1): p. 105-114.

18. Razakaratrimo, J., Jahiel, M., Jeannoda, V., and Normand, F. *Flowering is related to the axis phenology in the clove tree, an irregular bearing spice crop*. 2016.

19. Razakaratrimo, J., Jahiel, M., Jeannoda, V., and Normand, F. *Phenology and flowering of the clove tree, an irregular bearing spice crop*. in *XI International Symposium on Integrating Canopy, Rootstock and Environmental Physiology in Orchard Systems 1228*. 2016.

20. Rusmin, D., Darwati, I., Nurhayati, H., and Pribadi, E. *Flowering induction of clove to reduce its yield fluctuation by using a plant growth regulator*. in *IOP Conference Series: Earth and Environmental Science*. 2022. IOP Publishing.

21. Mahulette, A., Alfian, A., Zainal, M., Nendissa, J., Wattimena, A., Tanasale, V., Makaruku, M., and Laisina, J. *Growth of forest clove seedlings at different concentrations of paclobutrazol*. in *IOP Conference Series: Earth and Environmental Science*. 2020. IOP Publishing.

22. Bermawie, N., Wahyuni, S., and Ginting, R. *Selection of highly productive clove trees in the Anambas population*. in *IOP Conference Series: Earth and Environmental Science*. 2021. IOP Publishing.

23. Chalmin, P. and Jégourel, Y., *Vanille & clous de giroőle.* Cyclope-Les Marchés Mondiaux (Paris, France: Ed. Economica), 2019: p. 423-425.

24. Pool, P., Eden-Green, S., and Muhammad, M., *Variation in clove (Syzygium aromaticum) germplasm in the moluccan islands.* Euphytica, 1986. 35(1): p. 149-159.

25. Alfian, A., Mahulette, A., Zainal, M., and Bahrun, A. *Morphological character of raja clove (Syzygium aromaticum L. Merr & Perry.) native from Ambon Island*. in *IOP Conference Series: Earth and Environmental Science*. 2019. IOP Publishing.

26. Mahulette, A.S., Alfian, A., Suyadi, S., Supriyanto, S., Situmorang, J., Matatula, A.J., Kilkoda, A.K., Nendissa, J.I., and Wattimena, A.Y., *Type and morphological character of local clove (Syzygium aromaticum) from Maluku, Indonesia.* Biodiversitas Journal of Biological Diversity, 2022. 23(3).

27. Mahulette, A.S., Hariyadi, H., Yahya, S., Wachjar, A., and Marzuki, I., *Morpho-agronomical diversity of forest clove in Moluccas, Indonesia.* Hayati Journal of Biosciences, 2019. 26(4): p. 156-156.

28. Mahulette, A.S., Hariyadi, H., Yahya, S., Wachjar, A., and Alfian, A., *Morphological Traits of Maluku Native Forest Clove (Syzygium aromaticum L. Merr & Perry).* Journal of Tropical Crop Science, 2019.

29. Cahyaningrum, H., Lala, F., Polakitan, A., and Wahab, A. *Morphological characteristic of local clove varieties in East Halmahera*. in *IOP Conference Series: Earth and Environmental Science*. 2021. IOP Publishing.

30. Susilowati, M., Wahyuni, S., Setiadi, A., and Bermawie, N. *Growth variation and relationship of clove progenies of high-yielding mother trees collected from various regions in Indonesia*. in *AIP Conference Proceedings*. 2022. AIP Publishing LLC.

31. Mahulette, A., Yahya, S., and Wachjar, A., *The physicochemical components and characteristic from essential oils of forest cloves Syzygium aromaticum (Myrtaceae) in Maluku Province, Indonesia.* Plant Archives, 2019. 192: p. 466-472.

32. Mahulette, A., Yahya, S., and Wachjar, A. *Physico-chemical properties of clove oil from three forest clove accession groups in Maluku*. in *IOP Conference Series: Earth and Environmental Science*. 2020. IOP Publishing.

33. Hariyadi, M.A., Yahya, S., and Wachjar, A., *Morphological characters and essential oil constituents extracted of two clove varieties (Syzygium aromaticum (L.) Merr. & LM Perry.) from Ambon Island, Indonesia.* Plant Archives, 2020. 20(1): p. 2208-2214.

34. Mahulette, A., Riry, J., Kesaulya, H., Kembauw, E., Lawalata, I., Wattimena, A., Makaruku, M., and Alfian, A. *Essential oil components of forest clove variants from Ambon Island, Maluku*. in *IOP Conference Series: Earth and Environmental Science*. 2021. IOP Publishing.

35. Clove, K., *Three promising morphological variants in clove (Syzygium aromaticum (L.) Merr. & Perry) from Tamil Nadu, India.* Journal of Spices & Aromatic Crops, 1995. 3(2): p. 168.

36. Menon, P.S., *ISSR scientists develop new clove varieties.* Journal of herbs, spices & medicinal plants, 2000. 7(1): p. 103-105.

37. Krishnamoorthy, B., Rema, J., and Mathew, P., *Dwarf Clove (Syzygium aromaticum (L.) Merr. & Perry) Germplasm.* Indian Journal of Plant Genetic Resources, 2005. 18(3): p. 298-299.

38. Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., and Guarino, L., *Past and future use of wild relatives in crop breeding.* Crop science, 2017. 57(3): p. 1070-1082.

39. Migicovsky, Z. and Myles, S., *Exploiting wild relatives for genomics-assisted breeding of perennial crops.* Frontiers in Plant Science, 2017. 8: p. 460.

40. Kilian, B., Dempewolf, H., Guarino, L., Werner, P., Coyne, C., and Warburton, M.L., *Crop Science special issue: Adapting agriculture to climate change: A walk on the wild side.* Crop Science, 2021. 61(1): p. 32-36.

41. Ouadi, S., Sierro, N., Goepfert, S., Bovet, L., Glauser, G., Vallat, A., Peitsch, M.C., Kessler, F., and Ivanov, N.V., *The clove (Syzygium aromaticum) genome provides insights into the eugenol biosynthesis pathway*. Communications Biology, 2022. 5(1): p. 1-13.

42. SUNDARI, S., NURHASANAH, N., MAS'UD, A., AMIN, M., ARUMINGTYAS, E.L., and AZRIANINGSIH, R., *Update phylogenetic information of the local varieties of cloves (Syzygium aromaticum) from North Maluku, Indonesia based on ITS sequences data.* Biodiversitas Journal of Biological Diversity, 2019. 20(6).

43. Saranya, D. and Ravi, R., *Genetic Differentiation of Syzygium aromaticum Among the Altitudes Of Nilgiris.* Journal of Tree Sciences, 2018. 37(1): p. 33-39.

44. Shakya, R., Siddiqui, S., Srivatawa, N., and Bajpai, A., *Molecular characterization of jamun (Syzygium cumini L. Skeels) genetic resources.* International Journal of Fruit Science, 2010. 10(1): p. 29-39.

45. Khan, S. and Sharma, V., *Inter and intra-population variability of syzygium cumini using RAPD markers.* Progressive Agriculture, 2011. 11(1): p. 17-22.

46. Lai, J., Tsai, C., Yen, C., Ko, Y., Chen, S., Weng, I., Lin, Y., and Chiang, Y., *Molecular characterization of twenty polymorphic microsatellite markers in the polyploid fruit tree species Syzygium samarangense (Myrtaceae).* Genet. Mol. Res, 2015. 14: p. 13013-13021.

47. Thurlby, K.A., Connelly, C., Wilson, P.G., and Rossetto, M., *Development of microsatellite loci for Syzygium paniculatum (Myrtaceae), a rare polyembryonic rainforest tree.* Conservation Genetics Resources, 2011. 3(2): p. 205-208.

48. Hillyer, M.J., Boulter, S.L., Kitching, R.L., and Hughes, J.M., *Isolation and characterization of eight polymorphic microsatellite loci in the rainforest canopy tree, Syzygium sayeri (Myrtaceae).* Molecular Ecology Notes, 2007. 7(6): p. 1199-1201.

49. Purugganan, M.D. and Jackson, S.A., *Advancing crop genomics from lab to field.* Nature Genetics, 2021. 53(5): p. 595-601.

50. Bevan, M.W., Uauy, C., Wulff, B.B., Zhou, J., Krasileva, K., and Clark, M.D., *Genomic innovation for crop improvement.* Nature, 2017. 543(7645): p. 346-354.

51. Low, Y.W., et al., *Genomic insights into rapid speciation within the world's largest tree genus Syzygium.* Nature communications, 2022. 13(1): p. 1-15.

52. Oginuma, K., Lum, S.K., Lee, Y., and Tobe, H., *Karyomorphology of Some Myrtaceae from Singapore*. 1992: US Government Printing Office.

53. Grattapaglia, D., Vaillancourt, R.E., Shepherd, M., Thumma, B.R., Foley, W., Külheim, C., Potts, B.M., and Myburg, A.A., *Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus.* Tree Genetics & Genomes, 2012. 8(3): p. 463-508.

54. Myburg, A.A., et al., *The genome of Eucalyptus grandis.* Nature, 2014. 510(7505): p. 356-362.

55. Negi, A., Singh, K., Jaiswal, S., Kokkat, J.G., Angadi, U.B., Iquebal, M.A., Umadevi, P., Rai, A., and Kumar, D., *Rapid Genome-Wide Location-Specific Polymorphic SSR Marker Discovery in Black Pepper by GBS Approach.* 2022.

56. Crossa, J., et al., *Genomic selection in plant breeding: methods, models, and perspectives.* Trends in plant science, 2017. 22(11): p. 961-975.

57. Varshney, R.K., Bohra, A., Yu, J., Graner, A., Zhang, Q., and Sorrells, M.E., *Designing future crops: genomics-assisted breeding comes of age.* Trends in Plant Science, 2021. 26(6): p. 631-649.

58. Grattapaglia, D., Silva-Junior, O.B., Resende, R.T., Cappa, E.P., Müller, B.S., Tan, B., Isik, F., Ratcliffe, B., and El-Kassaby, Y.A., *Quantitative genetics and genomics converge to accelerate forest tree breeding.* Frontiers in Plant Science, 2018. 9: p. 1693.

59. Lebedev, V.G., Lebedeva, T.N., Chernodubov, A.I., and Shestibratov, K.A., *Genomic selection for forest tree improvement: Methods, achievements and perspectives.* Forests, 2020. 11(11): p. 1190.

60. Mathew, M.K. and Hariharan, M., *In vitro multiple shoot regeneration in Syzygium aromaticum.* Annals of botany, 1990. 65(3): p. 277-279.

61. Suparman, U. and Blake, J., *Studies on tissue culture of clove tree plant.* Indonesian Journal of Crop Science, 1990. 5(2): p. 67-75.

62. Rema, J., Krishnamoorthy, B., and Mathew, P., *Vegetative propagation of major tree spices.* Journal of spices and aromatic crops, 1997. 6(2): p. 57-405.

63. Nirmal Babu, K., Divakaran, M., Raj, R.P., Anupama, K., Peter, K., and Sarma, Y., *Biotechnological approaches in improvement of spices: a review.* Plant biology and biotechnology, 2015: p. 487-516.

64. Mathew, P., Rema, J., and Krishnamoorthy, B., *Reversal of dwarfness in a short-statured variant of clove (Syzygium aromaticum (L.) Merr. and Perry) by approach grafting.* Journal of Spices and Aromatic Crops, 2006. 15(1): p. 57-58.

65. Habibi, F., Liu, T., Folta, K., and Sarkhosh, A., *Physiological, biochemical, and molecular aspects of grafting in fruit trees.* Horticulture Research, 2022. 9.

66. Kapazoglou, A., Tani, E., Avramidou, E.V., Abraham, E.M., Gerakari, M., Megariti, S., Doupis, G., and Doulis, A.G., *Epigenetic changes and transcriptional reprogramming upon woody plant grafting for crop sustainability in a changing environment.* Frontiers in plant science, 2021. 11: p. 613004.

67. Loupit, G. and Cookson, S.J., *Identifying molecular markers of successful graft union formation and compatibility.* Frontiers in Plant Science, 2020. 11: p. 610352.

Chapter 2 - Preface

This chapter is a research article published in Communications biology (DOI: 10.1038/s42003-022-03618-z).

It is the result of a collaborative work between the Laboratory of Plant Physiology at University of Neuchâtel, PMI Research and Development, and the Neuchâtel Platform of Analytical Chemistry.

For this study, I performed the laboratory work, contributed to the data analysis, and wrote the manuscript. Nicolas Sierro conceived and supervised the study, performed computational analysis of sequencing, data and contributed to manuscript writing. Simon Goepfert and Lucien Bovet contributed to results interpretation. Gaetan Glauser and Armelle Vallat, developed the HPLC-UV and GC-MS method respectively, and performed the metabolomics experiments. Felix Kessler and Nikolai V. Ivanov conceived and supervised the study, and Manuel C. Peitsch initiated the strategy and the study. All authors reviewed the manuscript.

Chapter 2 - The clove (*Syzygium aromaticum*) genome provides insights into the eugenol biosynthesis pathway

Sonia Ouadi[1,2], Nicolas Sierro[2], Simon Goepfert[2], Lucien Bovet[2], Gaetan Glauser[3], Armelle Vallat[3], Manuel C. Peitsch[2], Felix Kessler[1], Nikolai V. Ivanov[2]✉

[1] Faculty of Sciences, Laboratory of Plant Physiology, University of Neuchâtel, 2000 Neuchâtel, Switzerland.

[2] PMI R&D, Philip Morris Products S. A, Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland.

[3] Faculty of Sciences, Neuchâtel Platform of Analytical Chemistry, University of Neuchâtel, 2000 Neuchâtel, Switzerland.

✉ email: nikolai.ivanov@pmi.com

Abstract

The clove (*Syzygium aromaticum*) is an important tropical spice crop in global trade. Evolving environmental pressures necessitate modern characterization and selection techniques that are currently inaccessible to clove growers owing to the scarcity of genomic and genetic information. Here, we present a 370-Mb high-quality chromosome-scale genome assembly for clove. Comparative genomic analysis between *S. aromaticum* and *Eucalyptus grandis*—both species of the Myrtaceae family—reveals good genome structure conservation and intrachromosomal rearrangements on seven of the eleven chromosomes. We report genes that belong to families involved in the biosynthesis of eugenol, the major bioactive component of clove products. On the basis of our transcriptomic and metabolomic findings, we propose a hypothetical scenario in which eugenol acetate plays a key role in high eugenol accumulation in clove leaves and buds. The clove genome is a new contribution to omics resources for the Myrtaceae family and an important tool for clove research.

2.1 Introduction

Clove, *Syzygium aromaticum* (L.) Merr. & L. M. Perry is a perennial tropical tree that belongs to the Myrtaceae family (which also includes myrtle, eucalyptus, and guava). Native to the Maluku Islands in eastern Indonesia, also known as the "spice islands", clove is one of the most anciently used and traded spices[1]. Beginning in the 18th century, clove trees were introduced to different parts of the world, where they became economically important[2, 3]. Cultivated for its dried, unopened flower buds (cloves), oleoresin and essential oil (EO) rich in eugenol, clove is one of the most important tropical spice crops in global trade. Its products are commercially used in a wide range of applications within the pharmaceutical, cosmetic, food, tobacco, and agricultural industries[4-7].

The Myrtaceae family is the eighth largest family of flowering plants. It includes more than 5,650 species, distributed across approximately 130 to 140 genera. The majority of species are diploid (2n = 22), with small to intermediate genome sizes (234–1788 Mb)[8, 9]. The chromosome-level genome assembly of *Eucalyptus grandis* (640 Mb) is a reference genome for all species that belong to the Myrtaceae family and is a valuable tool for comparative genomics[10-14]. Clove is the most commercialized species of the species-rich (1200–1500 species) and highly diversified genus *Syzygium*, which comprises other interesting woody species cultivated for their edible fruits, timber, and medicinal properties (e.g., *Syzygium cumini* and *Syzygium jambos*)[15, 16]. Previous karyotype studies indicate that clove is a diploid with 2n = 22 chromosomes[17]. However, its genome size has not been recorded in the C-value database and no reference genome or genetic map is currently available for clove[9]. A reference genome for clove would be fundamental to start investigating the genetic basis of economically important traits related to yield (such as organ size, disease resistance), cost of harvesting (low branching, falling buds), and biosynthesis of secondary compounds that are important for the quality of clove products.

The major compounds found in commercial-type clove EO are eugenol (72–96.6%), eugenol acetate (0–21.3%), β-caryophyllene (1.7–19.5%), and α-humulene (0.2–2.2%)[18]. Eugenol (4-allyl-2-methoxyphenol) is the main bioactive compound of clove and is a versatile compound with analgesic, antioxidant, anti-inflammatory, antibacterial, antifungal, and antiviral activities[19-21]. Eugenol belongs to the phenylpropene group, a large group of phenylpropanoid volatiles that contribute to the aroma and flavor of spices, herbs, and fruits and play a role in plant communication (pollinator attraction) and defense (against insects, herbivores, and pathogens)[22, 23]. The biosynthesis of eugenol, from the phenylalanine produced

by the plant shikimate pathway, involves gene families of the general phenylpropanoid pathway and the branch of the phenylpropanoid pathway specific to lignin biosynthesis for producing monolignol coumaryl alcohol and coniferyl alcohol (CCR, cinnamoyl CoA reductase and CAD, cinnamyl alcohol dehydrogenase). The final steps are catalyzed by biosynthetic genes belonging to families containing some members involved in the production of volatile phenylpropenes (BAHD (benzyl alcohol-acetyl-, anthocyanin-*O*-hydroxy-cinnamoyl-, anthranilate-N-hydroxycinnamoyl/benzoyl-, deacetyl-vindoline) acyltransferase superfamily and PIP (pinoresinol–lariciresinol reductase, isoflavone reductase, phenylcoumaran benzylic ether reductase) family)[23-25]. The biosynthetic pathway leading to the production of phenylpropene volatiles has been studied in several herbs and fruits, and many biosynthetic genes have been isolated and characterized[25]. Although clove products are considered the richest source of eugenol, the biological and molecular mechanisms involved in the biosynthesis, high accumulation, and regulation of eugenol in clove organs have not been extensively investigated.

Here, we report a high-quality reference genome for clove assembled into eleven chromosomes. In this study, we compared the newly assembled clove genome with the genome of *E. grandis* to investigate genome evolution between the two genera of the Myrtaceae family. In addition, we analyzed the genome to identify key gene families involved in the biosynthesis of eugenol and performed a combined transcriptomic and metabolomic analysis to gain insights into eugenol biosynthesis and its accumulation in clove leaves and buds.

2.2 Results

2.2.1 Genome *de novo* assembly and annotation

By using Illumina paired-end (PE) raw reads and the K-mer analysis software GenomeScope 2.0, the clove genome size was estimated to be 343 Mb (k = 21 bp)[26] (Supplementary Fig. 1). The use of third-generation long-read sequencing technologies from Oxford Nanopore Technologies (ONT) combined with Hi-C technology enabled *de novo* assembly of a high-quality reference genome for clove (Supplementary Data 1). The final assembly resulted in a 370 Mb genome consisting of 24 scaffolds, with 99.3% of the assembly being anchored on the 11 chromosomes (Table 1) (Supplementary Table 1). The completeness of the *S. aromaticum* genome assembly was assessed by using BUSCO, with the eudicots_odb10 dataset[27]. Of the tested BUSCO genes, 98.2% were found in a complete form, a percentage similar to that observed in the eucalyptus genome (97.9%). The low percentage of missing BUSCO genes (1.00%) suggests that the clove genome assembly is nearly complete (Supplementary Fig. 2).

By using RNAseq data from clove and transcripts of eucalyptus, a total of 27,528 protein-coding gene were predicted, representing 31% (115 Mb) of the genome assembly (Supplementary Data 1). The BUSCO results indicated high coverage of the gene space by the gene model and protein predictions: 95% of BUSCOs at the transcript level and 93.7% of BUSCOs at the protein level were found in complete form by using the eudicots_odb10 ortholog dataset (Supplementary Fig. 2). The clove genome comprises 43.4% (161 Mb) repetitive DNA, a fraction similar to that reported in the Myrtaceae relative guava (*Psidium guajava*) 443.8 Mb genome assembly (43.55%)[14]. Among the identified transposable elements, the long terminal repeats retrotransposons (LTR-RTs) represented 21.80% of the clove assembly. The LTR-RTs superfamilies Copia and Gypsy LTR-RTs were the most abundant repeat elements, representing 8.1% (30 Mb) and 13.2% (49 Mb) of the genome assembly, respectively (Table 1).

Table 1 The statistics for genome assembly of *S. aromaticum.*

	Chromosome-level assembly
Assembly	
Number of scaffolds	24
Minimum length of scaffolds (bp)	11,734
Maximum length of scaffolds (bp)	43,763,418
Average length of scaffolds (bp)	15,427,406.3
N	36,000 (0.01%)
N50 (bp)	35,418,074
Quality value of the assembly (QV)	46.738
Number of chromosome-scale scaffolds	11
Total length of chromosome-scale scaffolds	367,665,948 (99.3%)
Total length of assembly (bp)	370,257,752
Annotation	
Number of predicted genes	27,528
Number of predicted transcripts	50,053
Average transcript length (bp)	2,215.27
Average CDS length (bp)	1,277.41
Average exon per transcript	6.28
Annotated with UniRef Malvids	22,139 (80%)
Repeat sequences (bp)	160,689,965 (43.4%)
LTR Retrotransposons (bp)	80,731,053 (21.80%)
LTR/Copia (bp)	29,874,465 (8.1%)
LTR/Gypsy (bp)	48,891,039 (13.2%)
Others (bp)	79,958,912 (21.6%)

2.2.2 Clove genome evolution

According to our comparative analysis of the clove genome with eucalyptus genome, no major interchromosomal rearrangements were observed between the two Myrtaceae species (Fig. 1). The structures of chromosomes 1, 3, 5, and 7 are highly conserved between *E. grandis* and *S. aromaticum*, and ten intrachromosomal rearrangements (a to j) were found on the other seven chromosomes 2, 4, 6, 8, 9, 10, and 11. Large terminal inversions are present on chromosomes 4, 6, 9, 10, and 11, and more complex terminal rearrangements occur on chromosomes 2, 6, and 8 (Supplementary Table 2).

To determine if the rearrangements detected in the *S. aromaticum* genome assembly might be due to our assembly pipeline, we verified the Hi-C signal among the *S. aromaticum* chromosomes on the Hi-C contact maps. The strong Hi-C intrachromosomal signal among the

S. aromaticum chromosomes and the very low interchromosomal Hi-C signal implied that there is a high number of Hi-C data supporting the proposed scaffolding of contigs into the 11 *S. aromaticum* chromosomes (Supplementary Fig. 3). To further investigate the reliability of the proposed *S. aromaticum* chromosomes scaffolding and the intrachromosomal rearrangements found between the chromosomes of *S. aromaticum* and *E. grandis*, we compared our results to those from a previous comparative study between the high-density linkage map of the *Corymbia citriodora* subsp. *variegata* (370 Mb) and *E. grandis* genome (640 Mb)[28]. Interestingly, the structures of chromosomes 1, 3, 5, and 7 are highly conserved between *E. grandis* and *S. aromaticum*, as are those between the two eucalypt species. The presence of intrachromosomal rearrangements found between *S. aromaticum* and *E. grandis* and between *C. citriodora subsp. variegata* and *E. grandis* was detected on the other seven chromosomes, chromosomes 2, 4, 6, 8, 9, 10, and 11. The same type of rearrangement—large terminal inversions—was found on chromosomes 4 (c), 9 (h), 10 (i) and 11 (j) of *C. citriodora* subsp. *variegata* and *S. aromaticum* when compared to *E. grandis*. Two additional large terminal inversions were detected on *S. aromaticum* chromosomes 4 (b) and 9 (g) but not in the *C. citriodora* subsp. *variegata* linkage map (Fig. 1).

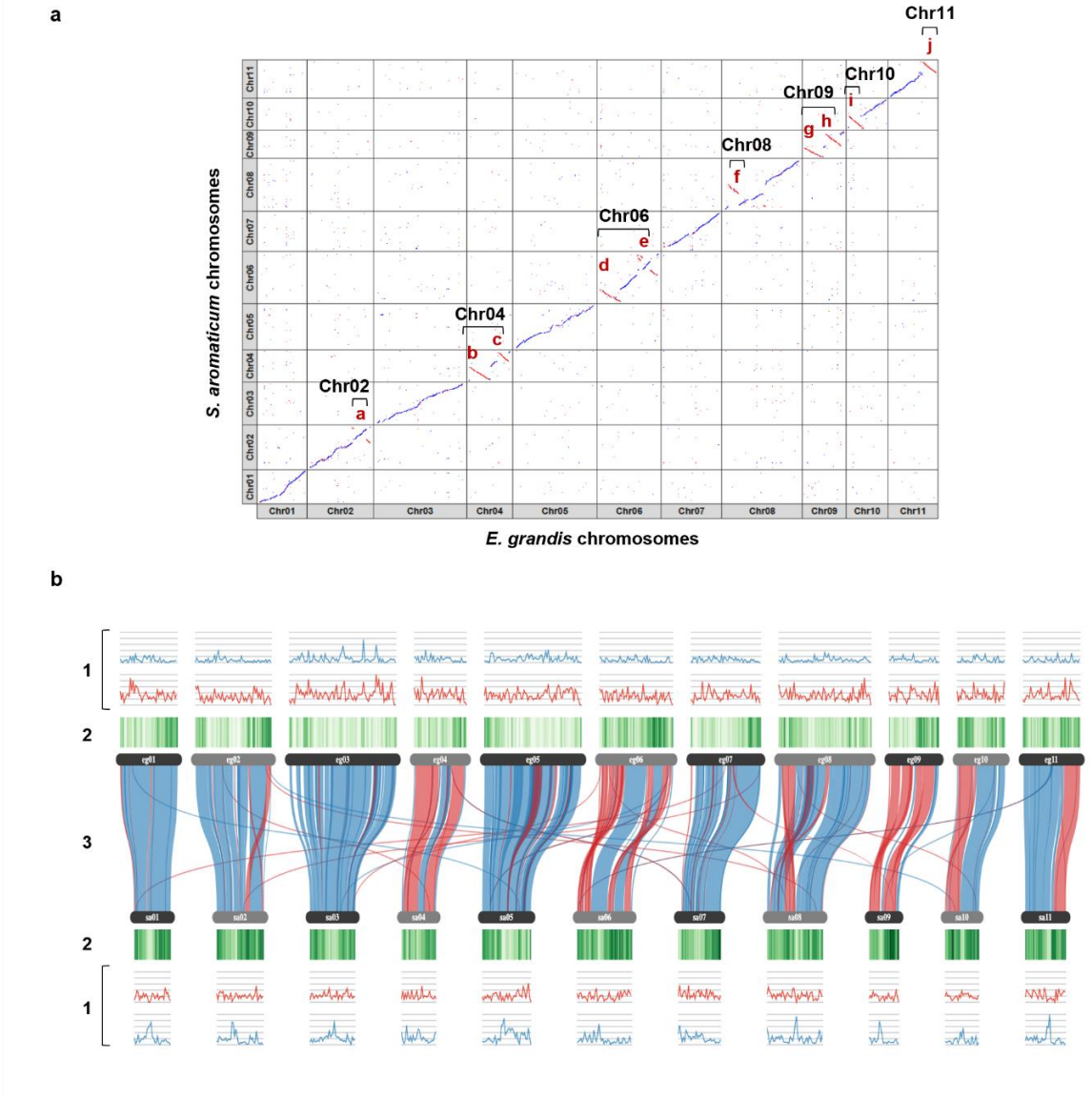


Fig. 1: Comparative analysis of the *Syzygium aromaticum* genome with *Eucalyptus grandis* genome. a DNA alignment of the S. aromaticum and E. grandis 11 chromosomes. The alignment of the DNA sequences of *E. grandis* and *S. aromaticum* chromosomes (Chr) is shown in blue. The main intrachromosomal rearrangements (a to j) are indicated in red. b SynVisio representation of *E. grandis* (eg) and *S. aromaticum* (sa) chromosomes showing the (1) distributions of Copia (red) and Gypsy (blue) LTR retrotransposons; (2) density of protein-coding genes; and (3) alignment of protein syntenic blocks between the two Myrtaceae species.

To further investigate the clove genome evolution, we classified into lineages the LTR retrotransposon (LTR-RTs) elements belonging to the superfamilies Copia and Gypsy in clove and eucalyptus genome assemblies, and estimated their insertion timeline (Fig.2, Supplementary Table 3, 4). A total of 10 Copia lineages (Ale, Bianca, Angela, Ivana, Gymco-IV, Alesia, Ikeros, SIRE, TAR and Tork) and 7 Gypsy lineages (Tekay, Athila, Galadriel, CRM, non-chromo-outgroup, Ogre and Reina) were identified in clove genome.

Among the 3,367 Gypsy elements identified, the lineages Ogre (1244 elements) and Tekay (1538) were the most expanded of the LTR-RTs lineages representing 6.31% and 5.41 % of the genome assembly, respectively. The Copia superfamily is less abundant in clove genome and comprised 2,811 elements. The most abundant Copia elements in clove genome belonged to the lineages SIRE (502 elements), Ale (609), Ikeros (486), Tork (475), Ivana (323) and TAR (241) and represented 1.65, 1.62, 1.46, 1.54, 0.78, and 0.57% of the genome assembly, respectively.

In contrast, we found a higher genomic abundance of Copia (7,736 elements) versus Gypsy (2,910) LTR-RTs in the eucalyptus genome assembly. The most abundant Copia elements identified in eucalyptus genome assembly, like for clove, belonged the lineage SIRE (2,646 elements), Ale (1,302), Ikeros (1,660), Tork (1,004), Ivana (428) and TAR (369). By estimating the time of insertion of the Copia elements, we observed a higher number of recent insertions of the Copia lineages in the eucalyptus genome (from ~2.5 Mya) when compared to clove (Fig.2).

We also noticed a distinct dynamic of Gypsy retrotransposons between the two species, especially for the Ogre elements that are more abundant in the clove genome (1,244 elements) than in the eucalyptus genome assembly (719). In the clove genome, Ogre elements were inserted into the genome between 0 and ~5Mya. In eucalyptus, we estimated that most of the Ogre elements were inserted later between 0 and ~2.5 Mya.

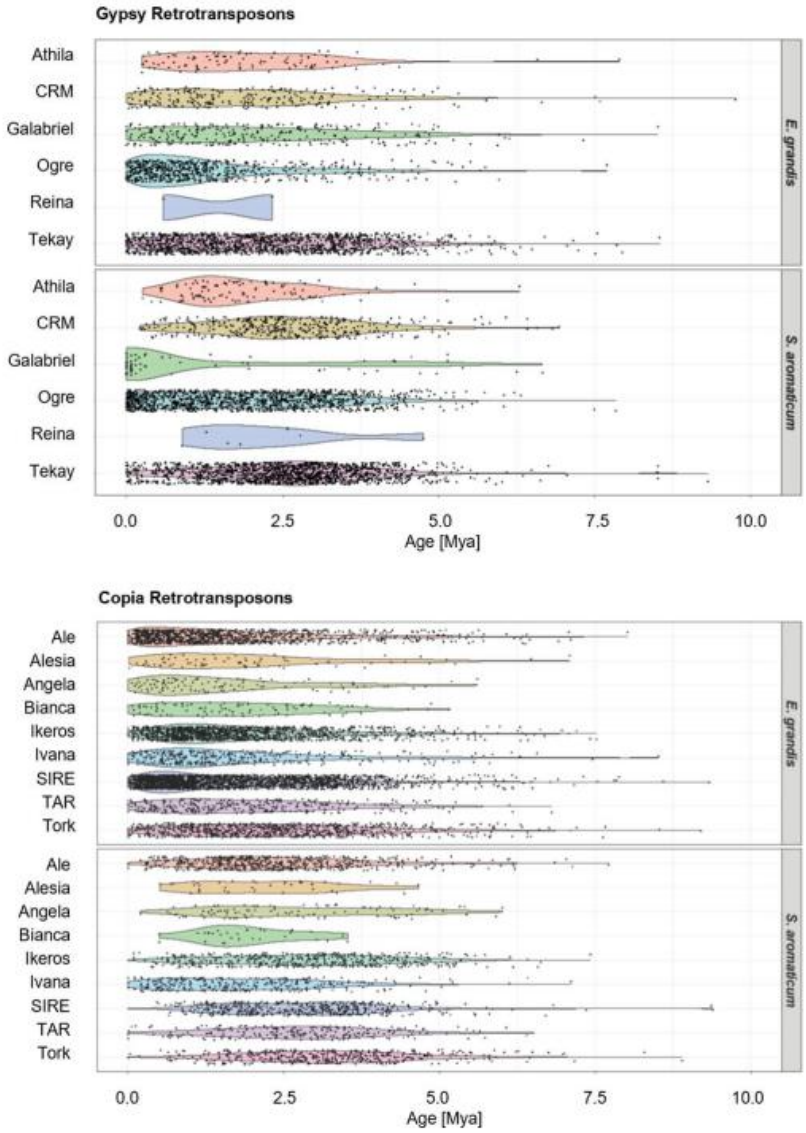


Fig. 2: Estimation of LTR-RTs element insertion time. The top panel shows the insertion time of Gypsy lineages in *E. grandis* and *S. aromaticum*, and the bottom panel shows the insertion time of Copia lineages in *E. grandis* and *S. aromaticum*. Only lineages detected in both species are shown.

2.2.3 Identification of gene families involved in the biosynthesis of eugenol

On the basis of functional annotations, clustering by ortholog best hits, and phylogenetic analysis, we identified 116 putative phenylpropanoid biosynthetic genes belonging to the 11 gene families PAL, C4H, 4CL, HCT, C3′H, CSE, COMT, CCoAOMT, F5H, CCR, and CAD in clove genome (Fig.3, Supplementary Data 2, 3). A number of genes lower than in eucalyptus genome for which 175 genes were reported for these eleven families[10, 29].

The PAL, 4CL, CCoAOMT and COMT and CAD families were significantly expanded in the eucalyptus genome when compared to the *Arabidopsis*, *Populus* and *Oryza* genomes[10]. These five families, along with HCT and C3′H families, were also more expanded in eucalyptus when compared to clove, with a notable difference for the COMT and CAD families for which 23 and 28 additional copies are present in *E. grandis* genome (Supplementary Fig. 4). For the C4H, CSE and CCR families, we selected a higher number of putative orthologs in the clove genome, with one additional copy for C4H, and CCR, and five additional copies for the CSE family.

Among genes belonging to the BAHD superfamily in clove genome, we selected 55 genes encoding for putative AATs (alcohol acyltransferase): 10 genes encoding putative AATs that clustered with selected characterised coniferyl alcohol acyltransferases and 45 putative AATs of classes III and V reported to be involved in the biosynthesis of volatiles esters (Fig.3, Supplementary Data 2, 3, Supplementary Fig. 5, and Supplementary Note)[30-33].

EGS is a key enzyme for biosynthesis of eugenol, it catalyzes eugenol production by using coniferyl acetate as a substrate. Using characterized EGS proteins belonging to the PIP family, we mined a total of 24 genes: 15 genes encoding putative EGS, and 9 encoding putative phenylcoumaran benzylic ether reductases (PCBER) (Fig.3, Supplementary Data 2, 3, and Supplementary Note)[25, 34]. A Pfam domain search allowed identification of seven additional genes encoding proteins containing an NmrA domain, which were considered possible PIP members (NmrA 1–7).

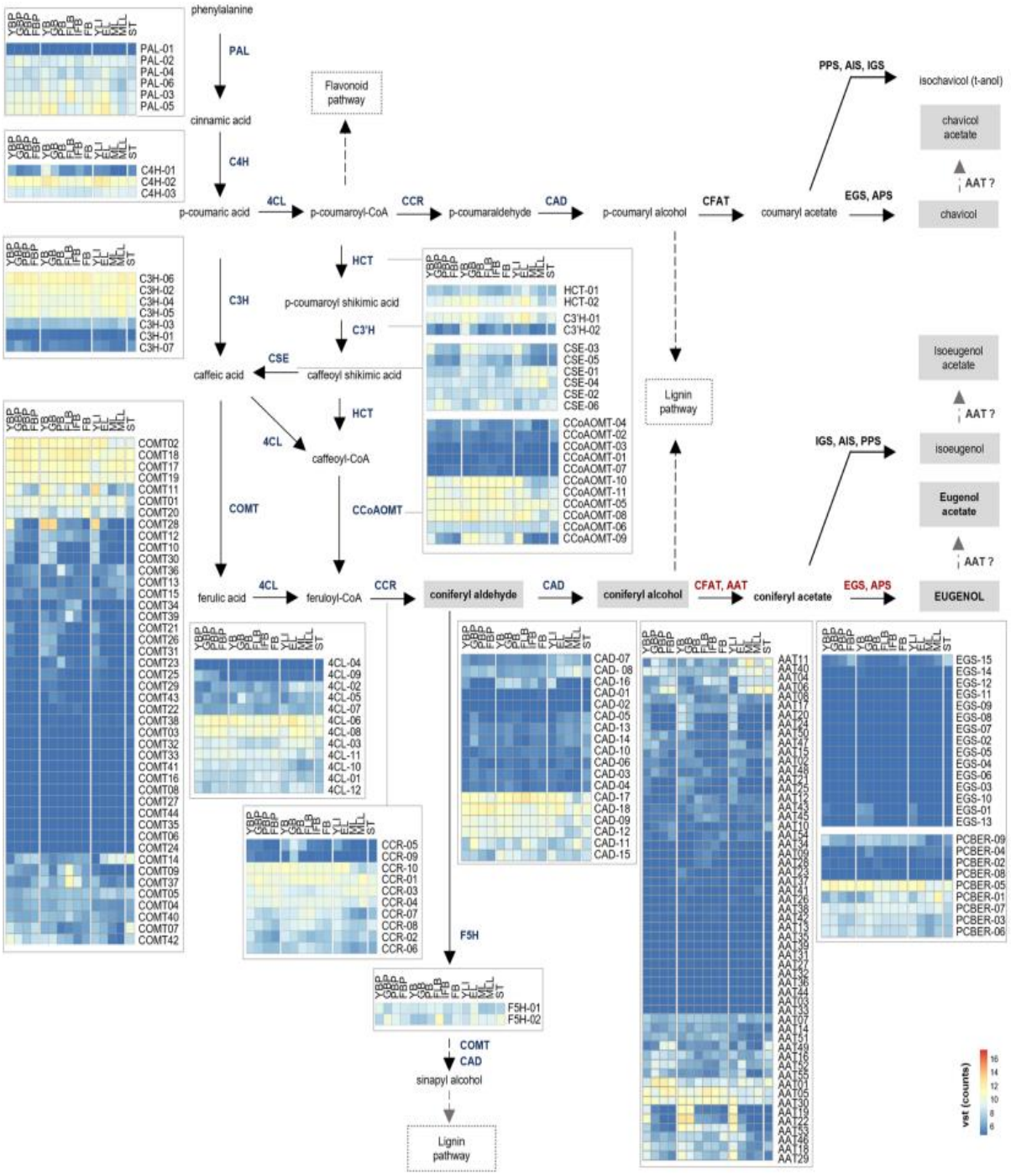


Fig. 3: Proposed biosynthetic pathways for lignin precursors and phenylpropene volatiles in *Syzygium aromaticum* (adapted from ref. 24 and ref. 25) and RNAseq evidence of expression of the identified biosynthetic genes. Gene families involved in the biosynthesis of precursors of lignin, monolignol coniferyl alcohol, coumaryl alcohol, and sinapyl alcohol are indicated in blue; gene families involved in the biosynthesis of coniferyl acetate and eugenol are indicated in red, and those involved in the biosynthesis of the phenylpropenes chavicol, isoeugenol, and t-anol are indicated in black. Compounds detected in clove leaves and buds are in gray boxes. The gray arrows indicate hypothetical reactions leading to the formation of phenylpropenes and their respective esters via the activities of alcohol acyltransferase (AAT). Heatmaps represent the relative expression profiles of the biosynthetic

genes in peduncles of young buds (YBP), peduncles of green buds (GBP), peduncles of pink buds (PBP), peduncles of buds in the fruiting stage (FBP), young buds (YB), green buds (GB), pink buds (PB), flowering bud (FLB), buds in the initial fruiting stage (IFB), buds in fruiting stage (FB), young leaves (YLI), expanded pale green leaves (EL), mature leaves (ML), mature leaf laminas (MLL), and stems (ST). n = 2 for flowering bud (FLB) and peduncles of pink buds (PBP) and n = 3 for all other groups. PAL phenylalanine ammonia-lyase, C4H cinnamate 4-hydroxylase; C3H 4-coumarate 3 hydroxylase, COMT, caffeate/5-hydroxyferulate 3-O-methyltransferase, F5H, ferulate 5-hydroxylase/coniferaldehyde 5-hydroxylase, 4CL 4-hydroxycinnamate:CoA ligase, HCT 4-hydroxycinnamoyl CoA:shikimate/quinate hydroxycinnamoyltransferase, C3′H 4-coumaroyl shikimate/quinate 3′-hydroxylase, CSE caffeoyl shikimate esterase, CCoAOMT caffeoyl CoA 3-O-methyltransferase, CCR cinnamoyl CoA reductase, CAD cinnamyl alcohol dehydrogenase, CFAT coniferyl alcohol acyltransferase, EGS eugenol synthase, APS allyl-phenylpropene synthase, IGS isoeugenol synthase, PPS propenyl-phenylpropene synthase, AIS anol/isoeugenol synthase, AAT alcohol acyltransferase, VST variance stabilizing transformation.

The fifteen genes encoding putative EGS are split into two loci. The first locus on chromosome 10 comprises fourteen putative EGS genes (EGS1 to EGS14), and a second locus on chromosome 11 is composed of a single EGS copy (EGS15). The percentage sequence similarity among the fourteen putative EGS encoded by the first cluster is high (75.5–100%). The sequence similarity decreases to 55.4% between the putative EGS encoded by the genes located on different chromosomes.

A previous phylogenetic analysis performed with EGS and other members of the PIP family of NADPH-dependent reductases suggested that phenylpropene synthases have independently evolved into two distinct lineages: a group of phenylpropene synthases composed of EGS, IGS, propenyl-phenylpropene synthase (PPS), and allyl-phenylpropene synthase (APS) and a second group of phenylpropene synthases (APS and EGS) that cluster with PCBERs and pterocarpan reductase (PTR)[25]. Our phylogenetic analysis revealed that the fifteen putative EGS clustered with the EGS–IGS protein group. Clove putative EGS showed the highest sequence similarity to the *Clarkia breweri* IGS1 and EGS1. The additional nine putative PCBERs clustered with the PCBER/PTR/EGS/IGS group and showed a closer phylogenetic relationship with *Gymnadenia conopsea* EGS1, *G. odoratissima* EGS1, *G. densiflora* IEGS1, *C. breweri* EGS2, *Fragaria. x ananassa* EGS2 (Fig. 4).

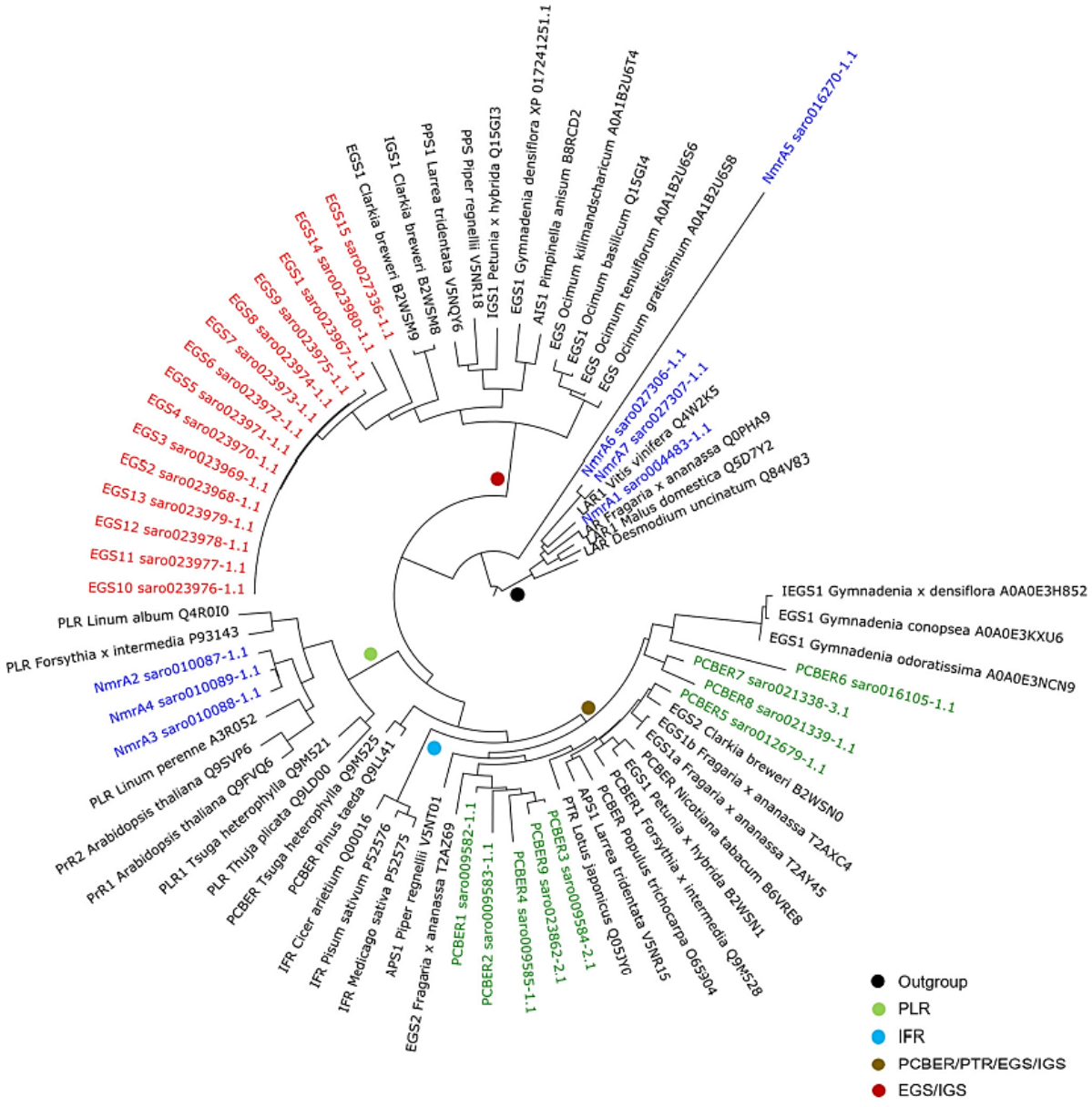


Fig. 4: Evolution of selected putative members of the PIP family in *Syzygium aromaticum*. Phylogenetic tree of protein sequences from the PIP family and *S. aromaticum* putative EGS, PCBERs, and proteins containing a NmrA domain possibly involved in the biosynthesis of eugenol. The phylogenetic trees were generated by using Clustal Omega. Groups from Koeduka[25]. EGS eugenol synthase, IGS isoeugenol synthase, AIS anol/isoeugenol synthase, PPS propenyl-phenylpropene synthase, APS allyl-phenylpropene synthase, PTR pterocarpan reductase, PCBER phenylcoumaran benzylic ether reductase, IFR isoflavone reductase, LAR leucoanthocyanidin reductase, PLR pinoresinol–lariciresinol reductase, PrR pinoresinol reductase.

2.2.4 Biosynthesis of eugenol in clove leaves and buds at different growth phases

Clove leaves and buds at different growth phases were analysed by UHPLC and GC-MS (Fig. 5). The major components found in clove leaves and buds are the major compounds found in clove EO: the phenylpropenes eugenol (8.3–29.7 mg/g fresh weight [FW]) and eugenol acetate (0.05–43.1 mg/g FW) and the sesquiterpenes β-caryophyllene (3.07–5.6 mg/g FW) and α-humulene (0.2–0.6 mg/g FW) (Supplementary Data 4). The concentration of eugenol increased during the development of clove leaves and buds to reach a maximum in mature leaves (27.9 ± 6.2 mg/g of FW) and full-budding-stage pink buds (29.7 ± 1.9 mg/g of FW), the stage at which they are collected (before the flowering stage). In contrast, the concentrations of eugenol acetate—which had accumulated in large amounts in young leaves and buds—decreased during organ growth, and only traces of the compound were detected in mature leaves and buds at the fruiting stage. In young leaves and buds, not only eugenol but also the phenylpropenes isoeugenol and chavicol (produced in smaller amounts) were found in their ester forms, while, in older leaves and buds, their alcohol forms were predominant. Coniferyl alcohol, the phenylpropanoid precursor of lignin, remained undetected in young leaves and buds. However, in older leaves and buds, the concentrations of coniferyl alcohol were significantly increased. Coumaryl acetate and coniferyl acetate (substrate of EGS) were not detected. The methyl derivative forms of phenylpropenes were not detected in either leaves or buds, despite the small percentage of methyl eugenol that was reported in EO (0.05–0.2%)[18].

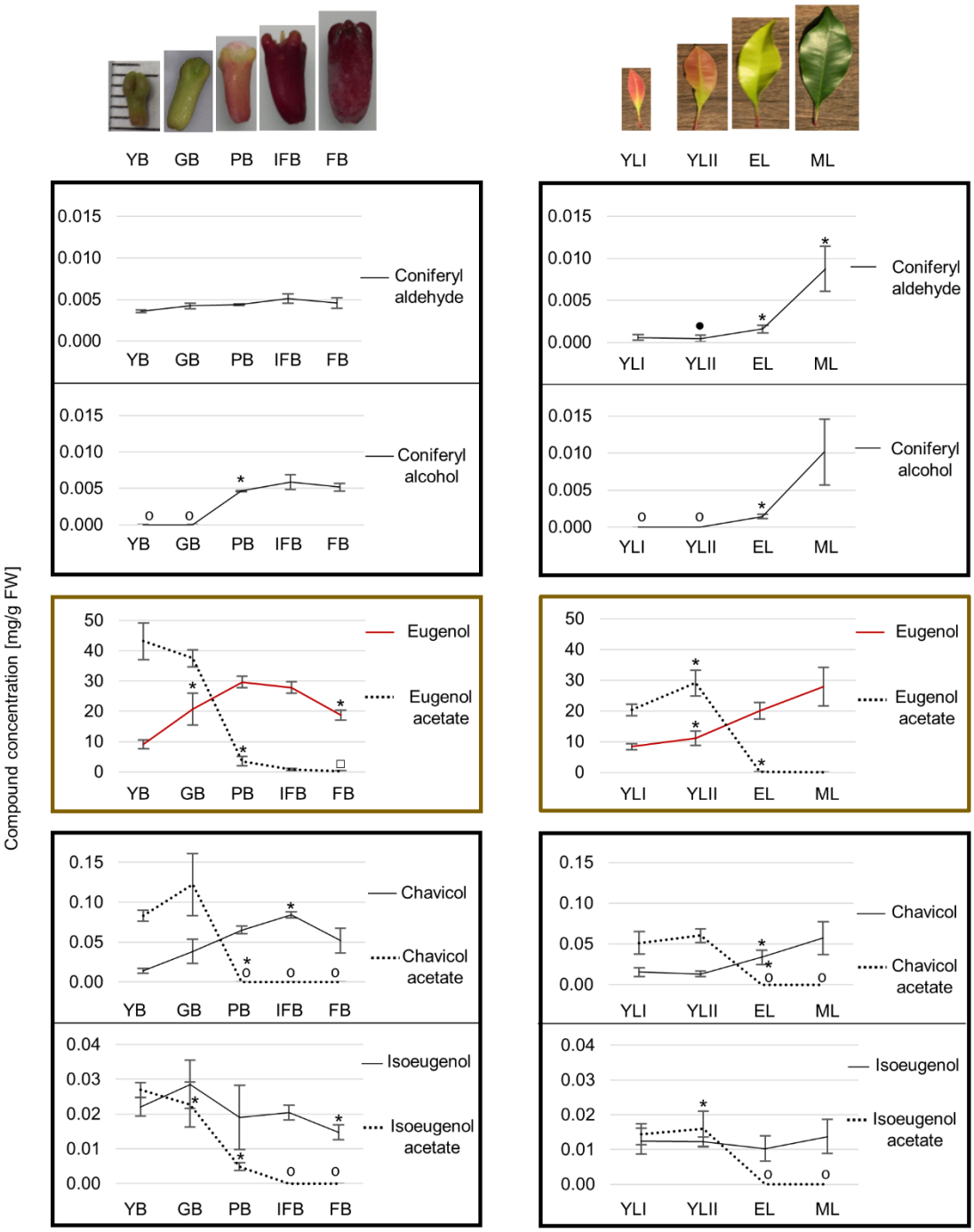


Fig. 5: Concentrations (mg/g FW) of the phenylpropenes eugenol, chavicol, and isoeugenol, their respective esters, and the precursors coniferyl aldehyde and coniferyl alcohol in *Syzygium aromaticum* leaves and buds among the growth phases. Data were represented as mean ± SD (n = 3). Asterisks indicate significant concentration differences relative to the previous growth phase. Two-tailed t-test $P < 0.05$. □ Compound not detected in two of the three organ replicates; ● compound not detected in one of the three organ replicates; O compound not detected. Fresh weight (FW). Young buds (YB), green buds (GB), pink buds (PB), buds in the initial fruiting stage (IFB), buds in fruiting stage (FB), young leaves (YLI), expanded young leaves (YLII), expanded pale green leaves (EL), and mature leaves (ML).

Twenty-four genes identified as phenylpropanoid genes from eight families (PAL, C4H, CSE, COMT, 4CL, CCoAOMT, CCR, and CAD) were found to be correlated with the concentration of the phenylpropenes isoeugenol, chavicol, and eugenol and their ester forms. These included eleven genes that were especially positively correlated with the concentration of eugenol acetate in clove leaves and buds (Fig. 6). Among the members of the PIP family, the gene expression levels of two putative EGS (EGS1 and EGS13) were positively correlated with the concentration of eugenol acetate. The expression levels of these two genes were higher in young leaves and buds than in the mature stages, where the concentration of eugenol was increased (Fig. 6). The biosynthesis of eugenol via conversion of coniferyl acetate into eugenol by EGS does seem to explain the increase in eugenol concentration in mature leaves and buds, as the expression levels of the EGS genes were lower in mature organs. The expression of nine genes encoding putative AATs (AAT-8, AAT-15, AAT-18 to AAT-22, AAT-24, and AAT-25) was positively correlated with eugenol acetate concentration and for two genes also negatively correlated with eugenol concentration (AAT-24 and AAT-25).

Similar to the results observed for the EGS protein genes, the expression levels of these nine AAT genes in young leaves and buds were higher than those in the mature stages. In this study, eugenol acetate is the main compound detected in young leaves and buds, and this large concentration of eugenol acetate is probably not completely volatilized or degraded during organ growth. However, the enzymatic hydrolysis of eugenol acetate produced in young leaves and buds to eugenol might explain the increase in eugenol levels in mature buds and leaves despite the lower expression levels of EGS genes.

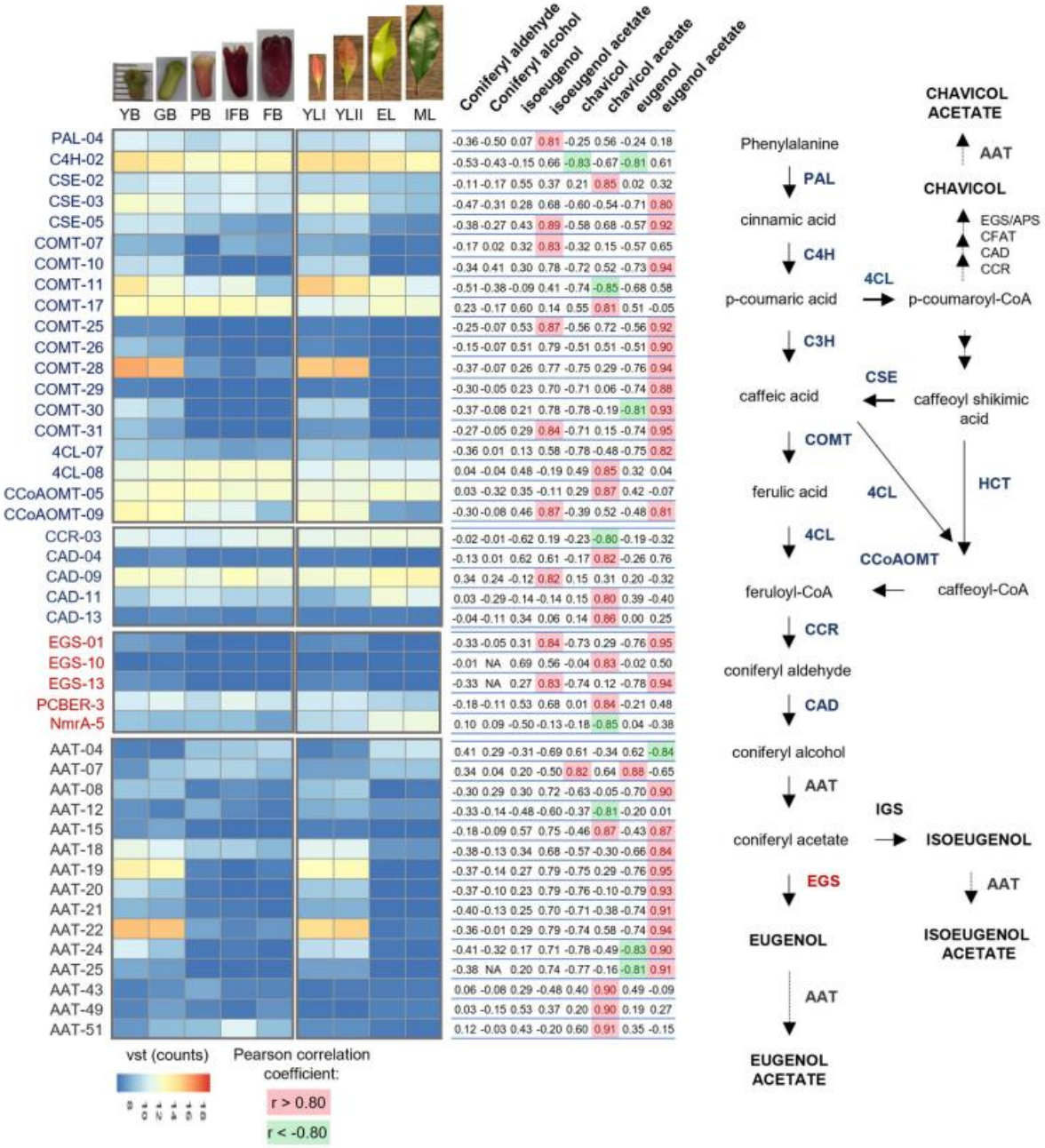


Fig. 6: List of genes possibly involved in the biosynthesis of eugenol and eugenol acetate in leaves and buds at different growth phases according to the combined transcriptomic and metabolomic data. Heatmaps indicate the expression profiles of genes correlated with the concentrations of the phenylpropenes eugenol and chavicol, their respective esters, isoeugenol acetate, and the precursor coniferyl alcohol (Pearson correlation coefficient $r > 0.8$ to $r < -0.8$, $n = 3$). Young buds (YB), green buds (GB), pink buds (PB), buds in the initial fruiting stage (IFB), buds in fruiting stage (FB), young leaves (YLI), expanded young leaves (YLII), expanded pale green leaves (EL) and mature leaves (ML). PAL phenylalanine ammonia-lyase, C4H cinnamate 4-hydroxylase, C3H 4-coumarate 3-hydroxylase, COMT caffeate/5 hydroxyferulate 3-O-methyltransferase, 4CL 4-hydroxycinnamate:CoA ligase, HCT 4-hydroxycinnamoyl CoA:shikimate/quinate hydroxycinnamoyltransferase, CSE caffeoyl shikimate esterase, CCoAOMT caffeoyl CoA 3-O-methyltransferase, CCR cinnamoyl CoA reductase, CAD cinnamyl alcohol dehydrogenase, CFAT coniferyl alcohol acyltransferase,

EGS eugenol synthase, APS allyl-phenylpropene synthase, IGS isoeugenol synthase, AAT alcohol acyltransferase, VST variance stabilizing transformation

We identified four predicted proteins annotated as hydrolase (hydrolase-149, 424 and 425) and esterase (esterase-23) that were positively correlated with the concentration of eugenol acetate and negatively correlated with the concentration of eugenol (Fig. 7). The expression levels of these four genes in young leaves and buds were higher than those in the mature stages, suggesting that these genes could be involved in the hydrolysis of eugenol acetate into eugenol during the growth of both organs. ATP-binding cassette (ABC) transporters constitute a ubiquitous superfamily of proteins involved in the transport of plant secondary metabolites[35, 36]. We found eleven genes encoding putative ABC transporters that were positively correlated with eugenol acetate. Analysis of the expression levels of these ABC genes indicated that ten of them (ABCI-1, ABCC-41, ABCG-61, ABCG-62, ABCG-64, ABC2-72, ABC2-113, ABC2-115, ABCG-131, and ABCG-143) were more highly expressed in young organs than in mature organs, suggesting that they might be involved in the transport of the eugenol acetate present in large quantities only in young leaves and buds. In contrast, one ABC transporter (ABCG-116) was positively correlated with eugenol concentration and more highly expressed in mature leaves and buds and might be involved in the transport of the eugenol (Fig. 7) (Supplementary data 3).

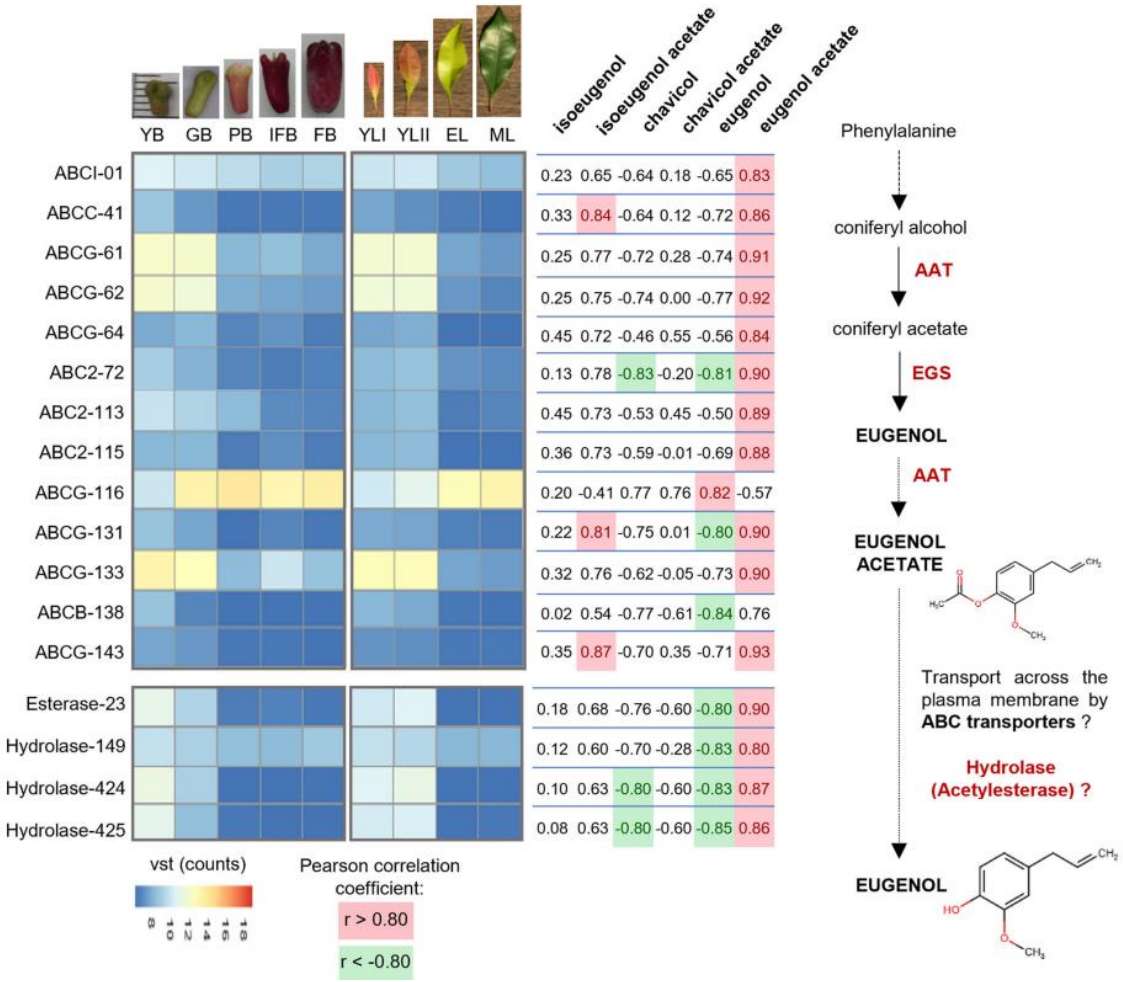


Fig. 7: Biosynthesis of eugenol in *Syzygium aromaticum* leaves and buds. Proposed hypothetical scenario in which eugenol acetate is converted to eugenol, on the basis of transcriptomic and metabolomic data analysis. Heatmaps indicate the expression profiles of genes correlated with the concentrations of the phenylpropenes eugenol, eugenol acetate, chavicol, and isoeugenol acetate (Pearson correlation coefficient $r > 0.8$ to $r < -0.8$, $n = 3$). AAT alcohol acyltransferase, EGS eugenol synthase, ABC ATP-binding cassette transporter. Young buds (YB), green buds (GB), pink buds (PB), buds in the initial fruiting stage (IFB), buds in fruiting stage (FB), young leaves (YLI), expanded young leaves (YLII), expanded pale green leaves (EL), and mature leaves (ML). VST variance stabilizing transformation.

2.3 Discussion

The current Myrtaceae infrafamilial classification recognizes seventeen tribes[37]. The *Eucalyptus* and *Corymbia* sister genera belong to tribe Eucalypteae, and clove belongs to tribe Syzygieae. The Eucalypteae tribe diverged from other Myrtaceae ~52–59 Mya[38, 39]. The estimated crown age of the Syzygieae tribe indicates that the tribe originated later, at approximately 22.5–36.4 Mya[39, 40]. An intergeneric comparison study using a high-density linkage map for the *C. citriodora* subsp. *variegata* (370 Mb) and *E. grandis* (640 Mb) genomes revealed that the overall genome structure was conserved among these eucalypt species and reported the presence of intrachromosomal rearrangements on seven of the eleven chromosomes (2, 4, 6, 8, 9, 10, and 11). However, the rearrangements located on two linkage groups, 2 and 8, of *C. citriodora* subsp. *variegata* could not be validated[28]. Our inter-tribe comparative genomic analysis between *S. aromaticum* (370 Mb) and *E. grandis* revealed that the genome architecture was also largely conserved between the two Myrtaceae species. By comparing our results with the rearrangements previously reported between eucalyptus and *C. citriodora* subsp. *variegata*, we observed that the same seven chromosomes (2, 4, 6, 8, 9, 10, and 11) were impacted by intrachromosomal rearrangements in clove and *C. citriodora* subsp. *variegat*a, suggesting that these seven chromosomes were more subjected to rearrangements and may have played a key role in diversification of the tribes and genera in this family. The presence of similar large terminal inversion on chromosomes 4 (c), 9 (h), 10 (i) and 11 (j) of clove and the corresponding linkage groups of *C. citriodora* subsp. *variegata* suggests that these rearrangements may have occurred in the *Eucalyptus* genus. In contrast, *Syzygium* and *Corymbia* seem to have retained the chromosome structure of a common ancestor to the two tribes Syzygieae and Eucalypteae, which split approximately 65–70 Mya in the late Cretaceous, during separation of the clades Syncarpieae, Eucalypteae, Leptospermeae, Chamelaucieae, and Lindsayomyrteae from the clades Backhousieae, Syzygieae, Tristanieae, Metrosidereae, Myrteae, and Kanieae[39]. Additional large terminal inversions, not reported for *C. citriodora* subsp. *variegata*, occurred in *S. aromaticum*, revealing continued evolution of the genome architecture in the Syzygieae tribe, possibly specific to genome speciation and the evolutionary process of the *Syzygium* genus.

Inversions are chromosomal rearrangements which may play a prominent role in local adaptation and speciation[41]. Nevertheless, inversions may also be the consequences of sequencing and assembly artefacts. The absence of Hi-C data to scaffold a genome and the low marker density in regions of high-density linkage maps could for instance lead to the

observations of inversions or rearrangements in comparative genomic studies. The hypothetical evolutionary scenarios of the chromosome structure presented in this study implies that the inversions observed resulted from evolutionary events. Although our scaffolding results are well supported by the Hi-C data, new comparative genomic studies with additional chromosome-scale genomes from *Syzygium* species are needed to better understand the evolutionary history of the chromosomal rearrangements that occurred in the *S. aromaticum* genome and to gain insights into the impact of the rearrangements on the evolution of the Syzygieae tribe and its actual diversity.

Transposable elements (TEs) are mobile genomic DNA that contribute to the evolution and function of genomes[42-44]. Among TEs identified in the clove genome, the long terminal repeat retrotransposons (LTR-RTs) were the most abundant representing 21.8% of the genome assembly, a fraction similar to that reported for eucalyptus (21.9%)[10]. Most of the LTR-RTs found in clove genome belonged to Gypsy superfamily as it was reported for *Populus* and the relative Myrtales species pomegranate[45, 46]. The fraction of Gypsy (13.2%) and Copia (8.1%) retrotransposons identified in clove genome were consistent with those reported for the pomegranate 274 Mb genome assembly (Gypsy (11.55%) and Copia (5.87%)). In contrast, we found a higher genomic abundance of Copia versus Gypsy LT- RTs in eucalyptus assembly (690 Mb). The differential accumulation of the Copia and the Gypsy lineages in the clove and eucalyptus genome 'assemblies indicated that a notable distinct activity of both Gypsy and Copia superfamilies occurred in the two Myrtaceae species genomes. This first analysis of clove LTR-RTs Gypsy and Copia repertories provides a valuable resource for future comparative analysis aiming to infer the impact of specific LTR-RTs lineages or element on the clove genome architecture, speciation, and biology.

The chemical composition of clove EO has been extensively investigated and found to be influenced by the geographical origin and age of the tree, the organ and its stage of development, and postharvest processing (e.g., drying and storage)[18, 47-51]. It was also reported that clove buds contain a large amount of eugenol, with an eugenol concentration of 9381.7 mg/100 g dry weight being reported for dried clove — a concentration 49 times higher than that in sweet basil (189.6 mg/100 g dry weight)[52]. In this study, we performed a combined transcriptomics and metabolomics experiment using the leaves and buds of clove at different growth phases to investigate the genomic basis of the biosynthesis of eugenol. Our results suggest that the lower concentration of eugenol found in young compared to mature buds and leaves might be the result of conversion by AATs of eugenol into eugenol acetate. This conversion might be reversed by the hydrolysis of eugenol acetate into eugenol during the

growth of both organs. On the basis of our correlation analysis, we hypothesized that eugenol acetate might be actively transported by ABC transporters across the plasma membrane of the cells occurred before being hydrolyzed into eugenol.

Genes involved in the biosynthesis of eugenol belong to large families, which are not restricted to biosynthesis of lignin precursors or production of eugenol[23, 25, 30]. The identification of genes family was performed using a homology-based approach. This approach relies on multiple factors affecting its efficacy such as the accuracy of the clove assembly sequences, of the transcript and protein predictions, the selected reference proteins, and the identity and coverage cut off used to select a number of best hits or genes of interest. The list of selected genes potentially involved in the biosynthesis of eugenol was narrowed on the basis of correlation (Pearson correlation coefficient $r \geq 0.8$ or $r \leq -0.8$) between expression pattern and eugenol and eugenol acetate concentration. In this study, we did neither validate the functional prediction of the genes selected or their role in the biosynthesis of eugenol. Nevertheless, our combined transcriptomic and metabolomic data provide a first insight into the genetic basis of eugenol biosynthesis in clove leaves and buds. In addition, the list of selected gene candidates represents a starting point for further characterization of the putative biosynthetic enzymes. Eugenol is an important compound that affects the value of clove products and a biocompound with a high potential for human health benefits[21, 53-56]. Additional studies on different organs (stems, peduncles, and roots) and specific anatomical structure (e.g., secretory cavities) will be required to further investigate the high eugenol accumulation in EO and its emission by the bud[57].

2.4 Methods

2.4.1 Biological material

The clove reference genome was generated from a single clove tree which originated in Masoala National Park in Madagascar and has been growing in Masoala Hall of the Zurich Zoo in Switzerland since 2007 under controlled climatic conditions (Collection number JA 318A). The air temperature is maintained between 20°C and 30°C, and the relative humidity is ~80% with up to 6 mm of tropical rainfall per day to replicate the climatic conditions of the Masoala Peninsula of Madagascar. Clove leaves, stems, buds, and peduncles collected from this tree were stored at -80°C until metabolite and nucleic acid extraction.

2.4.2 ONT sequencing library preparation and sequencing

High-molecular-weight DNA was isolated from frozen leaves by using the "ONT high-molecular-weight gDNA extraction from plant leaves" protocol. Sequencing libraries were prepared for sequencing on minION flow cells (FLOW-MIN106D R9 version) by using ligation sequencing (SQK-LSK109) and flow cell priming (EXP-FLP001) kits (Oxford Nanopore Technologies, Oxford, UK). Base calling was performed by using Guppy 4.0.15 in the high-accuracy mode. Raw ONT reads were cleaned by using SeqKit[58]. Reads shorter than 10000 bp or with a quality score lower than 7 were discarded.

2.4.3 DNAseq library preparation and Illumina sequencing

DNAseq libraries were generated from total DNA isolated from frozen leaves by using the "ONT high-molecular-weight gDNA extraction from plant leaves" protocol (Oxford Nanopore Technologies, Oxford, UK) and prepared using the Celero PCR workflow with enzymatic fragmentation kit from Nugen (San Carlos, CA, USA). DNAseq libraries were then loaded on an Illumina flow cell S2 and sequenced on the Illumina Novaseq6000 instrument (San Diego, CA, USA) as 2 x 151 bp PE reads. Raw reads were cleaned by using fastp to remove low complexity sequences and adapters sequences[59].

2.4.4 Hi-C library preparation and Illumina sequencing

Hi-C libraries were prepared from 0.2 g of frozen leaves by using the Proximo Hi-C Kit in accordance with the manufacturer's instructions (Phase Genomics, Seattle, WA). The libraries were then sequenced with an Illumina HiSeq 2500 instrument (San Diego, CA, USA) as 2 x 101 bp PE reads. Raw reads were cleaned by using fastp[59].

2.4.5 Estimation of genome size and percentage heterozygosity

Cleaned Illumina PE reads from DNAseq libraries were analyzed by GenomeScope 2.0 to estimate genome size and percentage heterozygosity by using a k-mer size equal to 21 bp[26].

2.4.6 *De novo* genome assembly

In order to reduce the raw error rate, cleaned ONT reads were first corrected with Ratatosk[60] by using the cleaned Illumina PE short reads from DNAseq libraries. Ratatosk corrected ONT reads were then mapped by minimap2[61] to the eucalyptus chloroplast and mitochondrial genomes to remove reads originating from plastid genomes. The filtered and corrected ONT reads were assembled by using minimap2[61] and miniasm[62]. The resulting raw assembly was first polished with minipolish[63] by using corrected ONT reads, and then with ntEdit[64] by using Illumina PE short reads from DNAseq libraries. Haplotigs were detected and removed by using Purge_dups[65].

Cleaned Illumina Hi-C reads were mapped to the haplotig-purged primary contigs by using BWA-MEM[66]. The scaffolding to a chromosome-level assembly was performed by using YAHS (https://github.com/c-zhou/yahs). Hi-C map files were generated with PretextMap (https://github.com/wtsi-hpag/PretextMap) and used to manually curate the assembly using PretextView (https://github.com/wtsi-hpag/PretextView). Finally, the curated clove genome was mapped to the *E. grandis* genome (https://eucgenie.org) by using minimap2[61], visualized by using a customized R script, and the orientation and names of the clove chromosomes set in accordance with those of *E. grandis*. The completeness of the final chromosome-level assembly was assessed by using the genome evaluation mode of BUSCO (version 5.2.2) and the eudicots_odb10 lineage dataset[27]. The quality of the final assembly was estimated using Yak[67].

2.4.7 RNAseq library preparation for Illumina sequencing

Total RNA was isolated in sextuplicate from whole leaves (young red leaves, expanded pale green leaves, and dark green mature leaves) and buds (young buds, green buds, pink buds, buds in the initial fruiting stage, and buds in the fruiting stage); in triplicate from the stem, expanded young leaves, lamina of mature leaves, peduncles (of young buds, green buds, pink buds, buds in the fruiting stage); and in duplicate from flowers and peduncles of pink and flowering buds. Total RNA was extracted from frozen powder by using Ambion PureLink Plant RNA Reagent (Ambion by Life Technologies; catalog number 12322-012) in accordance with the manufacturer's protocol (manual MAN0000243, revision 2.0). The concentration and quality of total RNA were assessed with an Agilent Bioanalyzer by using the Agilent RNA 6000 Nano Kit. mRNA libraries were prepared from 500 ng of total RNA by using the Nugen Universal Plus mRNA-Seq library preparation kit with NuQuant® (San Carlos, CA, USA). Quality control was performed by using an Agilent Bioanalyzer and Agilent DNA 1000 Kit. The libraries were sequenced on a HiSeq 4000 instrument as 2 x 150 bp PE reads.

2.4.8 Gene model prediction and functional annotation

Illumina RNAseq reads were cleaned and overlapping paired reads merged using fastp before being mapped to the assembly by using minimap2. Gene models were created by sample using Scallop[68]. Similarly, *E. grandis* transcripts were mapped to the assembly by using minimap2, and gene models were created by using bedtools bamtobed[69].

Augustus[70] was trained using Complete BUSCO genes from the eudicot_odb10 and used to predict ab initio and evidence-based genes hinted by the RNAseq and *E. grandis* gene models. The final gene models were created by merging the RNAseq and *E. grandis* gene models with evidence-based and *ab initio* Augustus predictions using taco[71] and adding CDS using Transdecoder (https://github.com/TransDecoder/TransDecoder/wiki) and gffread[72]. The Malvids portion of UniRef was used to annotate the gene models with their best hit by using DIAMOND[73]. Pfam[74] domains were annotated by using HMMER[75].

2.4.9 Prediction and annotation of repeat elements

TEnester[76] and TEsorter[77] with REXdb[78] were used to predict and annotate transposable elements and their insertion age calculated as previously reported[79]. In addition, repeats were

predicted using Red[80], MITEs using GRF[81], helitron using EAHelitron[82], tandem repeats using tantan[83].

2.4.10 Synteny between *S. aromaticum* and *E. grandis*

Identification of putative homologous chromosomal regions was performed with MCScanX[84] by using pairwise alignments of the predicted proteins, and a syntenic block size of at least five collinear protein pairs was visualized by using SynVisio (https://synvisio.github.io/#/).

2.4.11 Identification of gene families involved in eugenol biosynthesis

To identify the genes encoding putative proteins involved in the biosynthesis of eugenol, the full set of clove transcripts was compared with proteins of eucalyptus (PAL, C4H, 4CL, C3'H, HCT, CSE, COMT, CCoAOMT, and F5H) and selected C3H, AAT, and EGS sequences (characterized protein sequences from UniProt) (Supplementary Data 2). The best hits were selected, and phylogenetic analysis of the selected putative proteins and characterized proteins from UniProt was also performed to filter the selected genes.

2.4.12 Sequences and phylogenetic analysis

Multiple alignment of protein sequences was performed by using CLUSTAL OMEGA[85]. A phylogenetic tree of the PIP family protein sequences and *S. aromaticum* EGS sequences was generated by using FastTree[86], newick_utils[87], and ete3[88].

2.4.13 Combined metabolome and transcriptome analysis

Total RNA and metabolites were extracted from leaves at four different growth phases: young leaves (5.1–6.2 cm), expanded young leaves (8.7–9 cm), expanded pale green leaves (10.8–12.5 cm), and dark green mature leaves (14.7–15 cm).

Total RNA and metabolites were extracted from buds at five different growth phases: pooled young buds (<0.5 cm), green buds (1.3–1.4 cm), pink buds (full budding stage; 1.7–1.9 cm), buds at the initial fruiting stage (1.5–1.9 cm), and buds at the fruiting stage (1.8–1.9 cm).

Leaves and buds were sampled in triplicate for each growth phase and stored at -80°C before being ground in liquid nitrogen. The ground material obtained with each sample was

aliquoted into three tubes and weighed to proceed separately to RNA isolation, eugenol and eugenol acetate extraction, and volatile compound extraction.

Total RNA extraction was performed by using Ambion PureLink Plant RNA Reagent (Ambion by Life Technologies; catalog number 12322-012) in accordance with the manufacturer's protocol (manual MAN0000243, revision 2.0).

Eugenol and eugenol acetate were extracted from frozen leaf and bud powder by adding 10 volumes of methanol (w/v). The mixture of powder and methanol was then placed in a Qiagen TissueLyser II for 3 min at 30 Hz. The samples were then centrifuged for 2 min at 14800 rpm. The supernatant was diluted 1:10 with methanol for analysis by ultra-high performance liquid chromatography in ultraviolet mode (UHPLC-UV) on a Waters Acquity^{TM} UPLC instrument at 280 nm and 272 nm for eugenol and eugenol acetate, respectively. Calibration curves built from eugenol and eugenol acetate standard solutions from 2 to 200 µg/mL were used for quantification.

Volatile compounds were extracted from frozen leaf and bud powder by adding 10 volumes of dichloromethane (w/v). The mixture of powder and dichloromethane was then placed in a Qiagen TissueLyser II for 2 min at 30 Hz, with a 30-s pause every 30 s. The samples were centrifuged for 2 min at 14800 rpm, and the supernatant was filtered. The volatile compound extracts were analyzed in undiluted form and diluted (1:10 in dichloromethane) form. Gas chromatography–mass spectrometry (GC–MS) analysis was performed on an Agilent 7890B GC system equipped with a 5977A mass selective detector and an OPTIMA 5 MS Accent separation column (30 m x 0.25 mm; 0.25 µm [Macherey Nagel, Germany]). The samples were injected in the splitless mode and at an inlet temperature of 270°C. The following GC–MS conditions were used for analysis: MS source, 230°C; MS Quad, 150°C; transfer line, 280°C; oven program, 60°C for 7 min (10°C/min until 120°C, and then 5°C/min until 280°C); and run time, 45 min. Volatile compounds were identified by comparison of spectra against the database NIST MS search 2.0 and quantified by using standard peak areas (methyl chavicol, caryophyllene, caryophyllene oxide, and β-farnesene) and calibration curves.

Statistical analysis of the metabolomic results was performed by unpaired, two-tailed t-tests. $P < 0.05$ was considered significant.

2.4.14 Differential gene expression profiles between clove tissues

Gene expression analysis was performed with DESeq2[89] by using the gene-level count matrix obtained from STAR[90]. Variance-stabilizing transformation was used before generating heatmaps with pheatmap[91].

2.4.15 Reporting summary

Further information on research design is available in the Nature Research Reporting Summary. The reporting summary for this article can be found online at: https://doi.org/10.1038/s42003-022-03618-z

2.5 Data availability

Illumina and Oxford Nanopore reads are available from the National Center for Biotechnology Information Short Read Archive (SRA) under accession: PRJNA660399. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JACTMA000000000. The version described in this paper is version JACTMA0100000000. The genome annotation is available from Zenodo at DOI: 10.5281/zenodo.6579856. All additional data are available from the corresponding author (Nikolai V. Ivanov) on reasonable request.

2.6 Acknowledgements

We would like to thank Dr. Leyla Davis, curator of the Masoala Hall of the Zurich Zoo (Switzerland), who kindly authorized and facilitated the sampling of the *Syzygium aromaticum* tree for this project. We also thank Lukas Mueller (Boyce Thompson Institute) for his advice in data interpretation.

2.7 Author contributions

S.O. – performed the laboratory work, analyzed data and wrote the manuscript. N.S. – performed computational analysis of sequencing data and contributed to manuscript writing. S.G. and L.B. – contributed to results interpretation. G.G. – HPLC-UV method development and performed experiments. A.V. – GC–MS method development and performed experiments.

N.S., F.K., and N.V.I. – conceived and supervised the study. M.C.P. – Strategy and study initiation. All authors reviewed the manuscript.

2.8 Competing interests

The authors declare the following competing interests: S.O., N.S., S.G., L.B., N.V.I., and M.C.P. are employees of Philip Morris International. G.G., A.V., and F.K. are employees of the University of Neuchâtel. The remaining authors declare no competing interests.

2.9 Additional information

Supplementary information: The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03618-z.

Correspondence and requests for materials should be addressed to Nikolai V. Ivanov.

Peer review information: Communications Biology thanks Xin Liu and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Yuan Qin and Caitlin Karniski.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Kingwell-Banham, E., Bohingamuwa, W., Perera, N., Adikari, G., Crowther, A., Fuller, D.Q., and Boivin, N., *Spice and rice: pepper, cloves and everyday cereal foods at the ancient port of Mantai, Sri Lanka.* antiquity, 2018. 92(366): p. 1552-1570.

2. Smith, N.J., Plucknett, D.L., Williams, J., and Talbot, J.P., *Tropical forests and their crops*. 1992.

3. Danthu, P., et al., *The clove tree of Madagascar: a success story with an unpredictable future.* Bois et forêts des tropiques, 2014. 320(2): p. 83-96.

4. Cortés-Rojas, D.F., de Souza, C.R.F., and Oliveira, W.P., *Clove (Syzygium aromaticum): a precious spice.* Asian Pac. J. Trop. Biomed., 2014. 2(4): p. 90–96.

5. Nurdjannah, N. and Bermawie, N., *Cloves*, in *Handbook of herbs and spices*, K.V. Peter, Editor. 2012, Elsevier. p. 197-215.

6. Batiha, G.E.-S., Alkazmi, L.M., Wasef, L.G., Beshbishy, A.M., Nadwa, E.H., and Rashwan, E.K., *Syzygium aromaticum L.(Myrtaceae): Traditional uses, bioactive chemical constituents, pharmacological and toxicological activities.* Biomolecules, 2020. 10(2): p. 202.

7. Douglas, M., Heyes, J., and Smallfield, B., *Herbs, Spices and Essential Oils: Post-Harvest Operations in Developing Countries*. Vol. 61. 2005: UNIDO and FAO.

8. Grattapaglia, D., Vaillancourt, R.E., Shepherd, M., Thumma, B.R., Foley, W., Külheim, C., Potts, B.M., and Myburg, A.A., *Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus.* Tree Genet. Genomes, 2012. 8(3): p. 463–508.

9. Pellicer, J. and Leitch, I.J., *The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies.* New Phytologist, 2020. 226(2): p. 301-305.

10. Myburg, A.A., et al., *The genome of Eucalyptus grandis.* Nature, 2014. 510(7505): p. 356-62.

11. Bartholomé, J., Mandrou, E., Mabiala, A., Jenkins, J., Nabihoudine, I., Klopp, C., Schmutz, J., Plomion, C., and Gion, J.M., *High-resolution genetic maps of E ucalyptus improve Eucalyptus grandis genome assembly.* New Phytologist, 2015. 206(4): p. 1283-1296.

12. Izuno, A., Hatakeyama, M., Nishiyama, T., Tamaki, I., Shimizu-Inatsugi, R., Sasaki, R., Shimizu, K.K., and Isagi, Y., *Genome sequencing of Metrosideros polymorpha (Myrtaceae), a dominant species in various habitats in the Hawaiian Islands with remarkable phenotypic variations.* J. Plant Res., 2016. 129(4): p. 727–736.

13. Thrimawithana, A.H., et al., *A whole genome assembly of Leptospermum scoparium (Myrtaceae) for mānuka research.* N. Z. J. Crop Hortic. Sci., 2019. 47(4): p. 233–260.

14. Feng, C., Feng, C., Lin, X., Liu, S., Li, Y., and Kang, M., *A chromosome-level genome assembly provides insights into ascorbic acid accumulation and fruit softening in guava (Psidium guajava).* Plant biotechnology journal, 2021. 19(4): p. 717-730.

15. Craven, L.A. and Biffin, E., *An infrageneric classification of Syzygium (Myrtaceae).* Blumea-Biodiversity, Evolution and Biogeography of Plants, 2010. 55(1): p. 94-99.

16. Nair, K.N., *The Genus Syzygium: Syzygium Cumini and Other Underutilized Species*. 2017: CRC Press.

17. Oginuma, K., Lum, S.K., and Lee, Y., *Karyomorphology of some Myrtaceae from Singapore.* Gard. Bull., 1992. 44(2): p. 135–139.

18. Razafimamonjison, G., Jahiel, M., Duclos, T., Ramanoelina, P., Fawbush, F., and Danthu, P., *Bud, leaf and stem essential oil composition of Syzygium aromaticum from Madagascar, Indonesia and Zanzibar*. International Journal of Basic and Applied Sciences, 2014. 3(3): p. 224.

19. Kamatou, G.P., Vermaak, I., and Viljoen, A.M., *Eugenol—from the remote Maluku Islands to the international market place: a review of a remarkable and versatile molecule.* Molecules, 2012. 17(6): p. 6953-6981.

20. Marchese, A., et al., *Antimicrobial activity of eugenol and essential oils containing eugenol: A mechanistic viewpoint.* Critical reviews in microbiology, 2017. 43(6): p. 668-689.

21. Otunola, G.A., *Culinary Spices in Food and Medicine: An Overview of Syzygium aromaticum (L.) Merr. and LM Perry [Myrtaceae].* Frontiers in Pharmacology, 2021. 12: p. 793200-793200.

22. Atkinson, R.G., *Phenylpropenes: occurrence, distribution, and biosynthesis in fruit.* J. Agric. Food Chem., 2016. 66(10): p. 2259–2272.

23. Vogt, T., *Phenylpropanoid biosynthesis.* Mol. Plant, 2010. 3(1): p. 2–20.

24. Barros, J., et al., *4-Coumarate 3-hydroxylase in the lignin biosynthesis pathway is a cytosolic ascorbate peroxidase.* Nature communications, 2019. 10(1): p. 1-11.

25. Koeduka, T., *Functional evolution of biosynthetic enzymes that produce plant volatiles.* Biosci. Biotechnol. Biochem., 2018. 82(2): p. 192–199.

26. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C., *GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.* Nature Communications, 2020. 11(1): p. 1-10.

27. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. 31(19): p. 3210-3212.

28. Butler, J., Vaillancourt, R., Potts, B., Lee, D., King, G.J., Baten, A., Shepherd, M., and Freeman, J., *Comparative genomics of Eucalyptus and Corymbia reveals low rates of genome structural rearrangement.* BMC Genom., 2017. 18(1): p. 397.

29. Carocha, V., Soler, M., Hefer, C., Cassan-Wang, H., Fevereiro, P., Myburg, A.A., Paiva, J.A., and Grima-Pettenati, J., *Genome-wide analysis of the lignin toolbox of Eucalyptus grandis.* New Phytol., 2015. 206(4): p. 1297–1313.

30. D'Auria, J.C., *Acyltransferases in plants: a good time to be BAHD.* Curr. Opin. Plant Biol., 2006. 9(3): p. 331–340.

31. Dhar, N., Sreelatha, S., Reddy, V.A., Kumar, N., Panicker, D., Jin, J., Chua, N.H., and Rajani, S., *Characterization of a sweet basil acyltransferase involved in eugenol biosynthesis.* J. Exp. Bot., 2020. 71(12): p. 3638–3652.

32. Kim, S.J., Vassão, D.G., Moinuddin, S.G., Bedgar, D.L., Davin, L.B., and Lewis, N.G., *Allyl/propenyl phenol synthases from the creosote bush and engineering production of specialty/commodity chemicals, eugenol/isoeugenol, in Escherichia coli.* Arch. Biochem. Biophys., 2014. 541: p. 37–46.

33. Dexter, R., Qualley, A., Kish, C.M., Ma, C.J., Koeduka, T., Nagegowda, D.A., Dudareva, N., Pichersky, E., and Clark, D., *Characterization of a petunia acetyltransferase involved in the biosynthesis of the floral volatile isoeugenol.* Plant J., 2007. 49(2): p. 265–275.

34. Koeduka, T., et al., *The multiple phenylpropene synthases in both Clarkia breweri and Petunia hybrida represent two distinct protein lineages.* Plant J., 2008. 54(3): p. 362–74.

35. Yazaki, K., *ABC transporters involved in the transport of plant secondary metabolites.* FEBS letters, 2006. 580(4): p. 1183-1191.

36. Adebesin, F., et al., *Emission of volatile organic compounds from petunia flowers is facilitated by an ABC transporter.* Science, 2017. 356(6345): p. 1386-1388.

37. Wilson, P.G., O'brien, M., Heslewood, M., and Quinn, C., *Relationships within Myrtaceae sensu lato based on a matK phylogeny.* Plant Syst. Evol., 2005. 251(1): p. 3–19.

38. Thornhill, A.H., Crisp, M.D., Külheim, C., Lam, K.E., Nelson, L.A., Yeates, D.K., and Miller, J.T., *A dated molecular perspective of eucalypt taxonomy, evolution and diversification.* Aust. Syst. Bot., 2019. 32(1): p. 29–48.

39. Thornhill, A.H., Ho, S.Y., Külheim, C., and Crisp, M.D., *Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny.* Mol. Phylogenetics Evol., 2015. 93: p. 29–43.

40. Berger, B.A., Kriebel, R., Spalink, D., and Sytsma, K.J., *Divergence times, historical biogeography, and shifts in speciation rates of Myrtales.* Mol. Phylogenetics Evol., 2016. 95: p. 116–136.

41. Huang, K. and Rieseberg, L.H., *Frequency, origins, and evolutionary role of chromosomal inversions in plants.* Front. Plant Sci., 2020. 11: p. 296.

42. Bennetzen, J.L. and Wang, H., *The contributions of transposable elements to the structure, function, and evolution of plant genomes.* Annual review of plant biology, 2014. 65: p. 505-530.

43. Lee, S.-I. and Kim, N.-S., *Transposable elements and genome size variations in plants.* Genomics & informatics, 2014. 12(3): p. 87.

44. Zhang, L., et al., *A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour.* Nature communications, 2019. 10(1): p. 1-13.

45. Natali, L., Cossu, R.M., Mascagni, F., Giordani, T., and Cavallini, A., *A survey of Gypsy and Copia LTR-retrotransposon superfamilies and lineages and their distinct dynamics in the Populus trichocarpa (L.) genome.* Tree Genetics & Genomes, 2015. 11(5): p. 1-13.

46. Yuan, Z., et al., *The pomegranate (Punica granatum L.) genome provides insights into fruit quality and ovule developmental biology.* Plant biotechnology journal, 2018. 16(7): p. 1363-1374.

47. Razafimamonjison, G., Boulanger, R., Jahiel, M., Ramanoelina, P., Fawbush, F., Lebrun, M., and Danthu, P., *Variations in yield and composition of leaf essential oil from Syzygium aromaticum at various phases of development.* Int. J. Basic Appl. Sci., 2016. 5(1): p. 90.

48. Razafimamonjison, G., Jahiel, M., Ramanoelina, P., Fawbush, F., and Danthu, P., *Effects of phenological stages on yield and composition of essential oil of Syzygium aromaticum buds from Madagascar.* Int. J. Basic Appl. Sci., 2013. 2(4): p. 312–318.

49. Alfikri, F.N., Pujiarti, R., Wibisono, M.G., and Hardiyanto, E.B., *Yield, Quality, and Antioxidant Activity of Clove (Syzygium aromaticum L.) Bud Oil at the Different Phenological Stages in Young and Mature Trees.* Scientifica, 2020. 2020: p. 9701701.

50. Murni, V., Saepudin, E., Cahyana, A., Rahayu, D., Hastuti, L., and Haib, J. *Effect of oven drying and storage on essential oil composition of clove (Syzygium aromaticum) from Toli-Toli*. in *International Symposium on Current Progress in Mathematics and Sciences 2016, ISCPMS 2016: Proceedings of the 2nd International Symposium on Current Progress in Mathematics and Sciences 2016*. AIP Publishing.

51. Hastuti, L., Saepudin, E., Cahyana, A., Rahayu, D., Murni, V., and Haib, J. *The influence of sun drying process and prolonged storage on composition of essential oil from clove buds (Syzygium aromaticum)*. in *International Symposium on Current Progress in Mathematics and Sciences 2016, ISCPMS 2016: Proceedings of the 2nd International Symposium on Current Progress in Mathematics and Sciences 2016*. AIP Publishing.

52. Shan, B., Cai, Y.Z., Sun, M., and Corke, H., *Antioxidant capacity of 26 spice extracts and characterization of their phenolic constituents.* J. Agric. Food Chem., 2005. 53(20): p. 7749–7759.

53. Singh, P., et al., *Potential dual role of eugenol in inhibiting advanced glycation end products in diabetes: proteomic and mechanistic insights.* Scientific reports, 2016. 6(1): p. 1-13.

54. Choudhury, P., Barua, A., Roy, A., Pattanayak, R., Bhattacharyya, M., and Saha, P., *Eugenol restricts Cancer Stem Cell population by degradation of β-catenin via N-terminal Ser37 phosphorylation-an in vivo and in vitro experimental evaluation.* Chemico-Biological Interactions, 2020. 316: p. 108938.

55. Jabbari, N., Eftekhari, Z., Roodbari, N.H., and Parivar, K., *Evaluation of Encapsulated Eugenol by Chitosan Nanoparticles on the aggressive model of rheumatoid arthritis.* International Immunopharmacology, 2020. 85: p. 106554.

56. Mohamed, E.H., Alghamdi, Y.S., Mostafa Abdel-Hafez, S., Soliman, M.M., Alotaibi, S.H., Hassan, M.Y., Hany, N.A.-D., and Amer, H.H., *Susceptibility Assessment of Multidrug Resistant Bacteria to Natural Products.* Dose-response, 2020. 18(3): p. 1559325820936189.

57. Kasai, H., Shirao, M., and Ikegami-Kawai, M., *Analysis of volatile compounds of clove (Syzygium aromaticum) buds as influenced by growth phase and investigation of antioxidant activity of clove extracts.* Flavour and fragrance journal, 2016. 31(2): p. 178-184.

58. Shen, W., Le, S., Li, Y., and Hu, F., *SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation.* PloS one, 2016. 11(10): p. e0163962.

59. Chen, S., Zhou, Y., Chen, Y., and Gu, J., *fastp: an ultra-fast all-in-one FASTQ preprocessor*. Bioinformatics, 2018. 34(17): p. i884–i890.

60. Holley, G., Beyter, D., Ingimundardottir, H., Møller, P.L., Kristmundsdottir, S., Eggertsson, H.P., and Halldorsson, B.V., *Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly.* Genome Biology, 2021. 22(1): p. 1-22.

61. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. 34(18): p. 3094-3100.

62. Li, H., *Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.* Bioinformatics, 2016. 32(14): p. 2103-2110.

63. Wick, R.R. and Holt, K.E., *Benchmarking of long-read assemblers for prokaryote whole genome sequencing.* F1000Resarch, 2019. 8: p. 2138.

64. Warren, R.L., et al., *ntEdit: scalable genome sequence polishing.* Bioinformatics, 2019. 35(21): p. 4430-4432.

65. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R., *Identifying and removing haplotypic duplication in primary genome assemblies.* Bioinformatics, 2020. 36(9): p. 2896-2898.

66. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv preprint arXiv:1303.3997, 2013.

67. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H., *Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm.* Nature methods, 2021. 18(2): p. 170-175.

68. Shao, M. and Kingsford, C., *Accurate assembly of transcripts through phase-preserving graph decomposition.* Nature biotechnology, 2017. 35(12): p. 1167-1169.

69. Quinlan, A.R. and Hall, I.M., *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. 26(6): p. 841-842.

70. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D., *Using native and syntenically mapped cDNA alignments to improve de novo gene finding.* Bioinformatics, 2008. 24(5): p. 637-644.

71. Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M., and Iyer, M.K., *TACO produces robust multisample transcriptome assemblies from RNA-seq.* Nature methods, 2017. 14(1): p. 68-70.

72. Pertea, G. and Pertea, M., *GFF utilities: GffRead and GffCompare.* F1000Research, 2020. 9.

73. Buchfink, B., Xie, C., and Huson, D.H., *Fast and sensitive protein alignment using DIAMOND.* Nat. Methods, 2015. 12(1): p. 59–60.

74. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res., 2002. 30(1): p. 276–280.

75. Finn, R.D., Clements, J., and Eddy, S.R., *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Res., 2011. 39(suppl_2): p. W29–W37.

76. Lexa, M., Jedlicka, P., Vanat, I., Cervenansky, M., and Kejnovsky, E., *TE-greedy-nester: structure-based detection of LTR retrotransposons and their nesting.* Bioinformatics, 2020. 36(20): p. 4991-4999.

77. Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., and Ma, Y., *TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes.* Horticulture Research, 2022.

78. Neumann, P., Novák, P., Hoštáková, N., and Macas, J., *Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification.* Mobile DNA, 2019. 10(1): p. 1-17.

79. Marcon, H.S., Domingues, D.S., Silva, J.C., Borges, R.J., Matioli, F.F., de Mattos Fontes, M.R., and Marino, C.L., *Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic.* BMC plant biology, 2015. 15(1): p. 1-16.

80. Girgis, H.Z., *Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale.* BMC bioinformatics, 2015. 16(1): p. 1-19.

81. Shi, J. and Liang, C., *Generic repeat finder: a high-sensitivity tool for genome-wide de novo repeat detection.* Plant physiology, 2019. 180(4): p. 1803-1815.

82. Hu, K., Xu, K., Wen, J., Yi, B., Shen, J., Ma, C., Fu, T., Ouyang, Y., and Tu, J., *Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species.* BMC bioinformatics, 2019. 20(1): p. 1-20.

83. Frith, M.C., *A new repeat-masking method enables specific detection of homologous sequences.* Nucleic acids research, 2011. 39(4): p. e23-e23.

84. Wang, Y., et al., *MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.* Nucleic Acids Res., 2012. 40(7): p. e49.

85. Madeira, F., et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res., 2019. 27(W1): p. W636–W641.

86. Price, M.N., Dehal, P.S., and Arkin, A.P., *FastTree 2–approximately maximum-likelihood trees for large alignments.* PloS one, 2010. 5(3): p. e9490.

87. Junier, T. and Zdobnov, E.M., *The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell.* Bioinformatics, 2010. 26(13): p. 1669-1670.

88. Huerta-Cepas, J., Serra, F., and Bork, P., *ETE 3: reconstruction, analysis, and visualization of phylogenomic data.* Mol. Biol. Evol., 2016. 33(6): p. 1635–1638.

89. Anders, S. and Huber, W., *Differential expression analysis for sequence count data.* Genome Biol., 2010. 11(10): p. R106.

90. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. 29(1): p. 15-21.

91. Kolde, R. and Kolde, M.R., *Package 'pheatmap'.* R Package, 2015. 1(7): p. 790.

Supplementary Information

Supplementary Table 1. Statistics of clove genome assembly process

	Polished contigs	**Primary contigs** (Haplotig-purged)	**Final assembly** (Chromosome-level)
Number of sequences	415	202	24
Minimum length (bp)	43,250	10,129	11,734
Maximum length (bp)	10,157,946	10,157,946	43,763,418
Average length (bp)	1,004,870.81	1,832,780.95	15,427,406.3
N	0	0	36,000
%N	0.00	0.00	0.01%
N50	3,352,581	3,818,138	35,418,074
I50	36	30	5
Total length (bp)	417,021,388	370,221,752	370,257,752

Supplementary Table 2. Genomics coordinates of the main rearrangements (a to j) detected between the chromosomes of *S. aromaticum* and *E. grandis.*

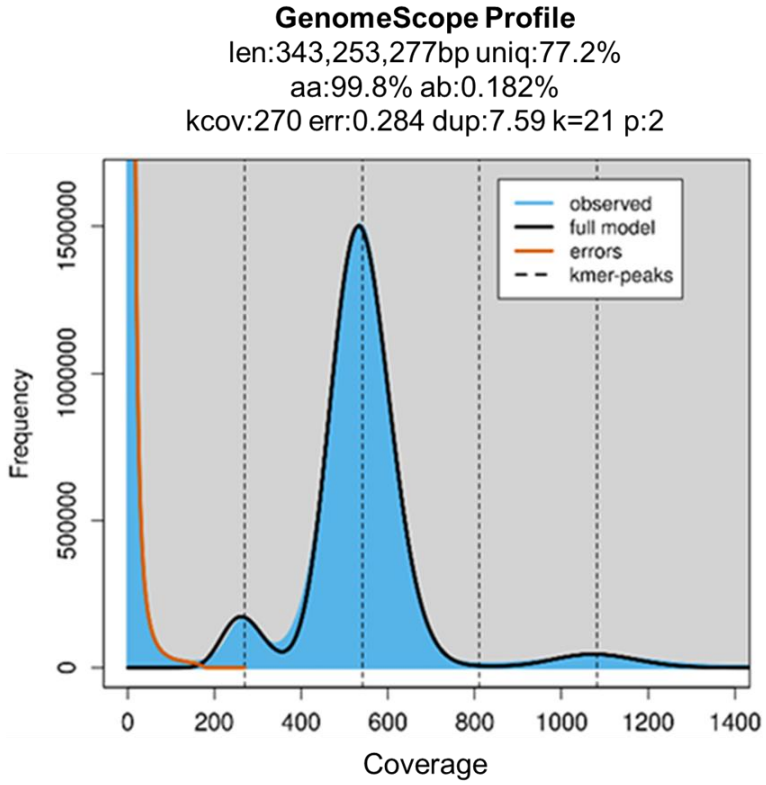
Species	Chromosome	Chr length (bp)	Rearrangement	Start (Mbp)	End (Mbp)
S. aromaticum	Chr02	36,857,792	a	22.3	25.7
	Chr02	36,857,792	a	36.1	36.5
	Chr04	27,127,432	b	1.2	8.5
	Chr04	27,127,432	c	18.2	25.6
	Chr06	42,993,409	d	0.0	9.7
	Chr06	42,993,409	e	23.7	28.3
	Chr06	42,993,409	e	36.9	41.5
	Chr08	43,763,418	f	2.4	2.5
	Chr08	43,763,418	f	14.5	21.9
	Chr09	23,380,519	g	1.1	8.3
	Chr09	23,380,519	h	11.3	20.7
	Chr10	26,565,453	i	2.4	11.3
	Chr11	31,862,996	j	22.1	31.8
E. grandis	Chr02	59,529,170	a	55.6	59.1
	Chr02	59,529,170	a	41.4	41.9
	Chr04	41,160,059	b	8.1	21.2
	Chr04	41,160,059	c	29.0	38.9
	Chr06	57,472,304	d	2.2	20.0
	Chr06	57,472,304	e	49.5	56.3
	Chr06	57,472,304	e	36.9	41.9
	Chr08	72,515,979	f	39.4	39.5
	Chr08	72,515,979	f	2.8	12.5
	Chr09	39,307,835	g	0.1	15.1
	Chr09	39,307,835	h	21.6	35.1
	Chr10	37,777,128	i	1.3	12.6
	Chr11	44,836,791	j	29.8	44.6

Supplementary Table 3. Classification of clove repeat sequences.

Class	Superfamily	Lineage	Count	Length (bp)	Percentage of assembly length
Retrotransposons			6,287	80,731,053	21.80%
	LTR Copia		2,811	29,874,465	8.07%
		Ale	609	6,008,482	1.62%
		Alesia	40	322,247	0.09%
		Angela	97	996,594	0.27%
		Bianca	30	251,488	0.07%
		Gymco-IV	1	15,732	0.00%
		Ikeros	486	5,387,472	1.46%
		Ivana	323	2,875,238	0.78%
		SIRE	502	6,105,662	1.65%
		TAR	241	2,111,029	0.57%
		Tork	475	5,715,311	1.54%
		Unknown	7	85,210	0.02%
	LTR Gypsy		3,367	48,891,039	13.20%
		Athila	105	1,095,776	0.30%
		CRM	387	3,358,133	0.91%
		Galadriel	77	704,126	0.19%
		non-chromo-outgroup	5	168,216	0.05%
		Ogre	1,244	23,362,484	6.31%
		Reina	7	79,513	0.02%
		Tekay	1,538	20,015,880	5.41%
		Unknown	4	106,911	0.03%
	LTR Unknown		109	1,965,549	0.53%
DNA transposons	MITE		2,586	499,855	0.14%
	Helitron		1,059	30,022	0.01%
Tandem repeat			134,280	3,174,970	0.86%
Repeat element (unclassified)			123,383	76,254,065	20.59%
Total			279,940	160,689,965	43.40%

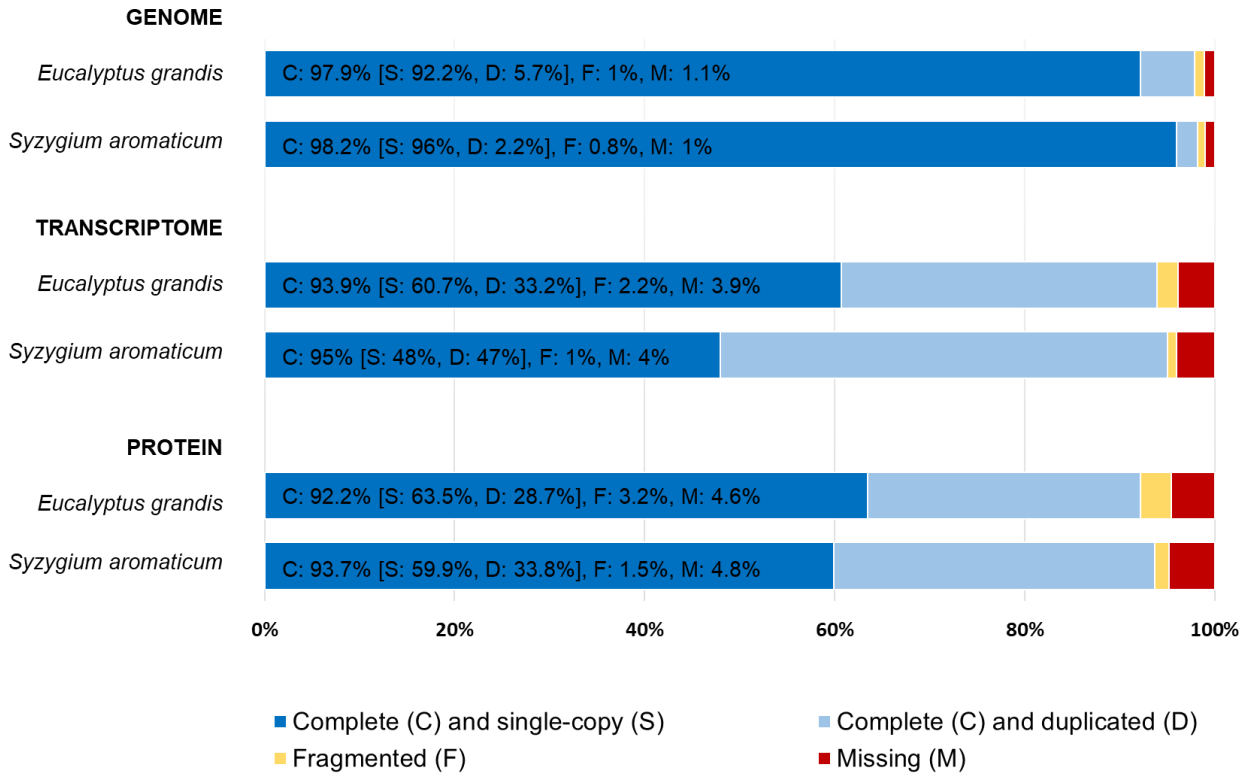
Supplementary Table 4. Classification of eucalyptus Copia and Gypsy LTR-retrotransposons.

Order	Superfamily	Lineage	Count	Length (bp)	Percentage of assembly length (690Mb)
LTR	Copia		7,736	88,014,391	12.73%
		Ale	1,302	13,907,418	2.01%
		Alesia	72	7,32,803	0.11%
		Angela	118	1,177,594	0.17%
		Bianca	102	982,202	0.14%
		Gymco-I	1	25,313	0.00%
		Ikeros	1,660	18,297,078	2.65%
		Ivana	428	4,851,223	0.70%
		SIRE	2,646	31,180,014	4.51%
		TAR	369	3,778,579	0.55%
		Tork	1,004	12,476,408	1.80%
		Unknown	34	605,759	0.09%
	Gypsy		2,910	38,170,125	5.52%
		Athila	104	1,526,850	0.22%
		Chromo-unclass	1	21,025	0.00%
		CRM	249	2,688,555	0.39%
		Galadriel	303	3,386,074	0.49%
		Non-chromo-outgroup	8	169,043	0.02%
		Ogre	719	11,205,165	1.62%
		Reina	2	19,239	0.00%
		Tekay	1,511	18,952,691	2.74%
		Unknown	13	201,483	0.03%
	Unknown		16	397,895	0.06%
	Total		10,662	126,582,411	18.31%



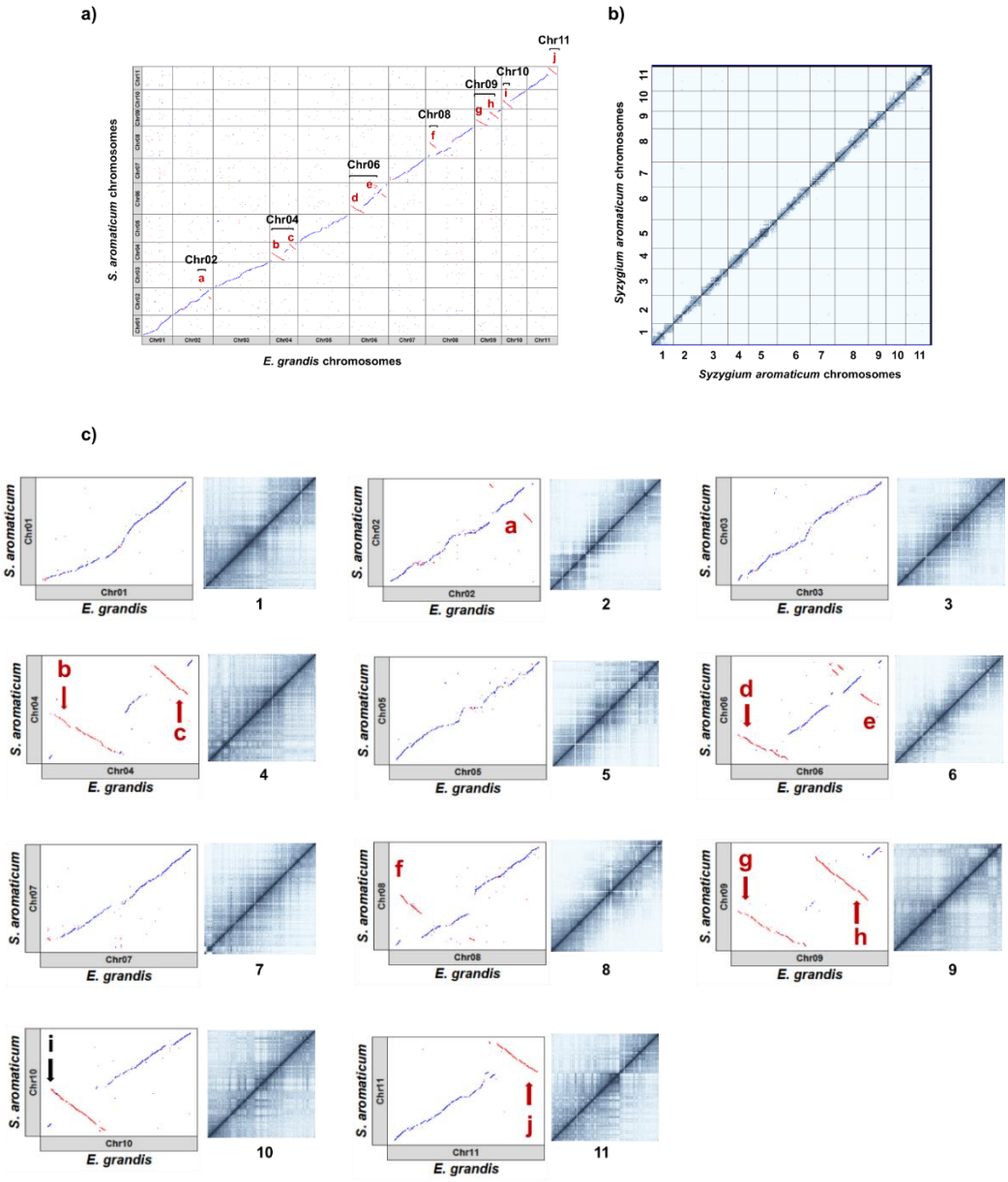
Supplementary Figure 1. GenomeScope K-mer spectra and fitted models for *Syzygium aromaticum*.

Cleaned Illumina PE reads from DNAseq libraries were analyzed by GenomeScope 2.0 to estimate genome size and percentage heterozygosity[1]. The software was run with a k-mer size equal to 21 and a ploidie level equal to 2. The resulting k-mer analysis indicated that the clove genome has a heterozygotie rate of 0.18%, and an estimated genome size of 343 Mbp.



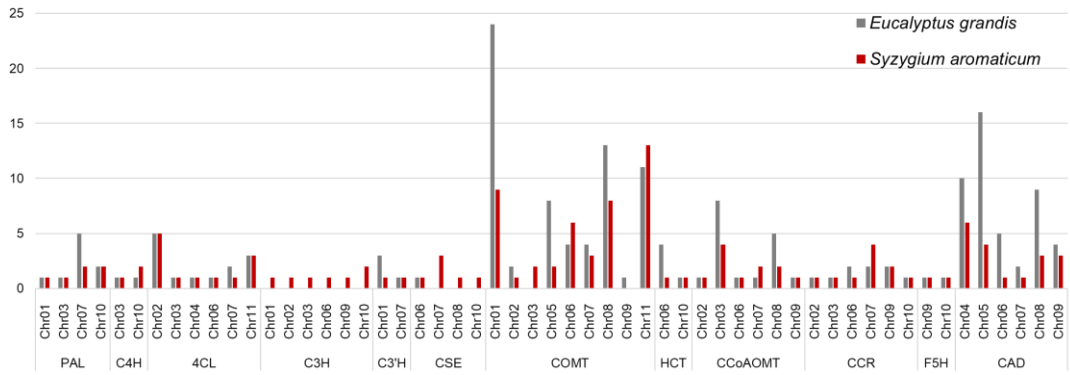
Supplementary Figure 2. BUSCO assessment of the *S. aromaticum* genome.

BUSCO results generated in the genome, transcriptome, and protein mode by using *Eucalyptus grandis* genome v.2 and the S. *aromaticum* final assembly (chromosome-level), predicted transcript and protein sets (BUSCO version 5.2.2 - dataset: eudicots_odb10, n=2326)[2].



Supplementary Figure 3. Visualization of the Hi-C data used for assembling *Syzygium aromaticum* genome into 11 chromosomes.

a) DNA alignment of *Syzygium aromaticum* and *Eucalyptus grandis* 11 chromosomes. Letters (a to j) indicate the presence of intrachromosomal rearrangements in the chromosomes 2, 4, 6, 8, 9, 10. and 11 of *S. aromaticum* when compared to the corresponding chromosome of *E. grandis*. **b)** Hi-C contact maps showing the Hi-C interactions among the 11 assembled chromosomes of *S. aromaticum.* Dark signal indicates a higher contact probability. **c)** Detailed Hi-C contact maps for individual *S. aromaticum* chromosomes and the DNA alignment plot for the corresponding chromosome of *E. grandis.* On the Hi-C contact maps, the diagonal indicates a strong Hi-C intrachromosomal signal among the *S. aromaticum* chromosomes and a very low interchromosomal Hi-C signal. This strong signal implies that there is high number of Hi-C data supporting the proposed scaffolding of contigs into the 11 *S. aromaticum* chromosomes.

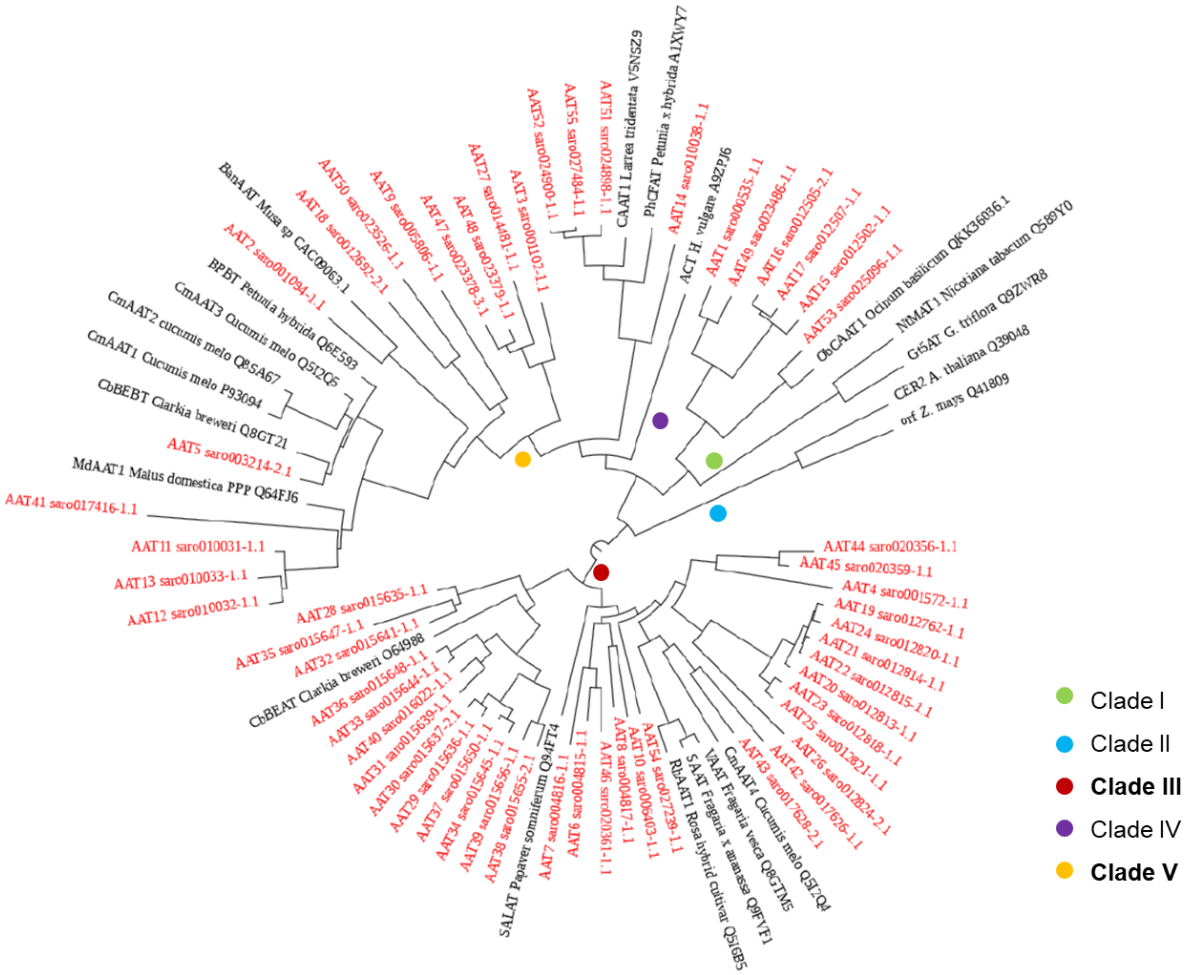


GENE FAMILY	PAL	C4H	4CL	C3H	C3'H	CSE	COMT	HCT	CCoAOMT	CCR	F5H	CAD
E. grandis	9	2	13	-	4	1	67	5	17	9	2	46
S. aromaticum	6	3	12	7	2	6	44	2	11	10	2	18

Supplementary Figure 4. Distribution and number of phenylpropanoid genes in the *S. aromaticum* genome**.**

Distribution across the 11 chromosomes (Chr) and number of phenylpropanoid genes found in *S. aromaticum*, and comparison with those reported for *E. grandis* [3, 4].

PAL, phenylalanine ammonia-lyase; **C4H**, cinnamate 4-hydroxylase; **C3H**, 4-coumarate 3-hydroxylase; **COMT**, caffeate/5-hydroxyferulate 3-*O*-methyltransferase; **F5H**, ferulate 5-hydroxylase/coniferaldehyde 5-hydroxylase; **4CL**, 4-hydroxycinnamate:CoA ligase; **HCT**, 4-hydroxycinnamoyl CoA:shikimate/quinate hydroxycinnamoyltransferase; **C3′H**, 4-coumaroyl shikimate/quinate 3′-hydroxylase; **CSE**, caffeoyl shikimate esterase; **CCoAOMT**, caffeoyl CoA 3-*O*-methyltransferase; **CCR**, cinnamoyl CoA reductase; **CAD**, cinnamyl alcohol dehydrogenase



Supplementary Figure 5. Evolution of selected *S. aromaticum* putative alcohol acyl transferases (AAT).

Phylogenetic tree of *S. aromaticum* putative AAT protein sequences. The phylogenetic trees were generated by using Clustal Omega. Uniprot entry identifiers are indicated for the reference proteins (NCBI identifiers are indicated for CAC09063.1 and QKK36036.1). The alcohol acyltransferase from *Petunia hybrida* (PhCFAT), Larrea tridentata (CAAT1) and *Ocimum basilicum* (ObCAAT1), have been shown to catalyze the formation of coniferyl acetate from coniferyl alcohol and acetyl CoA. Clade I to V from D'Auria 2006[5].

Supplementary Note. Identification of gene from the BAHD superfamily and PIP families involved in the biosynthesis of eugenol.

The two final steps of the eugenol biosynthesis are catalyzed by biosynthetic genes belonging to families containing some members involved in the production of volatile phenylpropenes (BAHD superfamily and PIP family). Coniferyl alcohol is converted to coniferyl acetate by an alcohol acyltransferase (AAT) belonging to the BAHD (benzyl alcohol-acetyl-, anthocyanin-*O*-hydroxy-cinnamoyl-, anthranilate-N-hydroxycinnamoyl/benzoyl-, or deacetyl-vindoline) acyltransferase superfamily. Then, coniferyl acetate is used as a substrate for eugenol synthesis by eugenol synthase (EGS), an NADPH-dependent reductase belonging to the pinoresinol–lariciresinol reductase, isoflavone reductase, phenylcoumaran benzylic ether reductase (PIP) family[5, 6]. Previous studies on clove EO produced from buds and leaves at different growth phases showed that eugenol acetate is the major constituent of the EO produced in young leaves and buds and that the eugenol content of EO increases during the development of the two organs, while the percentage of eugenol acetate decreases[7, 8]. In previous studies, recombinant AAT from apple (MdAAT1), wild strawberry (VAAT), and cultivated strawberry (SAAT) exhibited catalytic activity with eugenol when tested in vitro as a substrate, and MdAAT1 also exhibited catalytic activity with isoeugenol and isochavicol as substrates[9, 10]. We, therefore, evaluated AATs as good candidates not only for biosynthesis of eugenol from coniferyl acetate but also for biosynthesis of eugenol acetate from eugenol.

BADH proteins are promiscuous, making functional prediction from predicted amino acid sequences very difficult[5]. To identify the AAT genes involved in the biosynthesis of eugenol and eugenol acetate, we used reference proteins, including the characterized coniferyl alcohol acyltransferase from *Petunia hybrida* (CFAT) and *Ocimum basilicum* (ObCAAT1), and the cinnamyl alcohol acyltransferase from *Larrea tridentata* (LtCAAT1), which have been shown to catalyze the formation of coniferyl acetate from coniferyl alcohol and acetyl CoA[11-13]. We also used members of the BAHD superfamily belonging to classes III and V known to be involved in the biosynthesis of volatile esters[5].

EGS is a key enzyme for biosynthesis of eugenol and has been characterized in a few species. To identify putative EGS in clove genome we used the protein sequences of characterised EGS from eight different plant families in a phylogenetic analysis (Supplementary data set 2). For instance, EGS isolated from basil (ObEGS1) and *Clarkia breweri* (CbEGS1 and CbEGS2) catalyzes eugenol production by using coniferyl acetate as a substrate. The characterized *C. breweri* IGS1 and EGS1 have a very high percentage similarity

(96%) but catalyze the production of isoeugenol and eugenol, respectively[14]. The EGS isolated from strawberry (FaEGS2) can catalyze the production of both eugenol and isoeugenol (with a lower catalytic efficiency) from coniferyl acetate. Other members of the PIP family, such as the allyl-phenylpropene synthase (APS) isolated from creosote bush (*L. tridentata*), catalyze the conversion of *p*-coumaryl acetate and coniferyl acetate into chavicol and eugenol, respectively[6].

References

1. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C., *GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.* Nature Communications, 2020. 11(1): p. 1-10.

2. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. 31(19): p. 3210-3212.

3. Myburg, A.A., et al., *The genome of Eucalyptus grandis.* Nature, 2014. 510(7505): p. 356-62.

4. Carocha, V., Soler, M., Hefer, C., Cassan-Wang, H., Fevereiro, P., Myburg, A.A., Paiva, J.A., and Grima-Pettenati, J., *Genome-wide analysis of the lignin toolbox of Eucalyptus grandis.* New Phytol., 2015. 206(4): p. 1297–1313.

5. D'Auria, J.C., *Acyltransferases in plants: a good time to be BAHD.* Curr. Opin. Plant Biol., 2006. 9(3): p. 331–340.

6. Koeduka, T., *Functional evolution of biosynthetic enzymes that produce plant volatiles.* Biosci. Biotechnol. Biochem., 2018. 82(2): p. 192–199.

7. Razafimamonjison, G., Boulanger, R., Jahiel, M., Ramanoelina, P., Fawbush, F., Lebrun, M., and Danthu, P., *Variations in yield and composition of leaf essential oil from Syzygium aromaticum at various phases of development.* Int. J. Basic Appl. Sci., 2016. 5(1): p. 90.

8. Razafimamonjison, G., Jahiel, M., Ramanoelina, P., Fawbush, F., and Danthu, P., *Effects of phenological stages on yield and composition of essential oil of Syzygium aromaticum buds from Madagascar.* Int. J. Basic Appl. Sci., 2013. 2(4): p. 312–318.

9. Yauk, Y.K., et al., *Alcohol acyl transferase 1 links two distinct volatile pathways that produce esters and phenylpropenes in apple fruit.* Plant J., 2017. 91(2): p. 292–305.

10. Beekwilder, J., Alvarez-Huerta, M., Neef, E., Verstappen, F.W., Bouwmeester, H.J., and Aharoni, A., *Functional characterization of enzymes forming volatile esters from strawberry and banana.* Am. Soc. Plant Biol., 2004. 135(4): p. 1865–78.

11. Dhar, N., Sreelatha, S., Reddy, V.A., Kumar, N., Panicker, D., Jin, J., Chua, N.H., and Rajani, S., *Characterization of a sweet basil acyltransferase involved in eugenol biosynthesis.* J. Exp. Bot., 2020. 71(12): p. 3638–3652.

12. Kim, S.J., Vassão, D.G., Moinuddin, S.G., Bedgar, D.L., Davin, L.B., and Lewis, N.G., *Allyl/propenyl phenol synthases from the creosote bush and engineering production of specialty/commodity chemicals, eugenol/isoeugenol, in Escherichia coli.* Arch. Biochem. Biophys., 2014. 541: p. 37–46.

13. Dexter, R., Qualley, A., Kish, C.M., Ma, C.J., Koeduka, T., Nagegowda, D.A., Dudareva, N., Pichersky, E., and Clark, D., *Characterization of a petunia acetyltransferase involved in the biosynthesis of the floral volatile isoeugenol.* Plant J., 2007. 49(2): p. 265–275.

14. Koeduka, T., et al., *The multiple phenylpropene synthases in both Clarkia breweri and Petunia hybrida represent two distinct protein lineages.* Plant J., 2008. 54(3): p. 362–74.

Supplementary Data

Supplementary Data 1A. Oxford Nanopore Technologies (ONT) sequencing summary. Raw ONT reads were cleaned by using SeqKit (Shen et al., 2016)58 to discard reads shorter than 10000 bp or with quality scores lower than 7, corrected using Ratatosk, and filtered for chloroplast and mitochondria contamination.

ONT sequencing	raw reads	cleaned reads
Number of reads	1138890	1163292
Mean read length (bp)	29318.49	26644.36
Read length N50 (bp)	36902	34389
longest read length (bp)	316394	293499
Total bases (bp)	33390534296	30995175489
Coverage (based on estimated size 343-Mb)	97	90

Supplementary Data 1B. Illumina sequencing summary. Illumina raw reads were cleaned by using fastp (Chen et al., 2018)59 to discard low quality reads and adapters sequences.

Illumina sequencing	Illumina cleaned reads generated from DNAseq libraries	Illumina cleaned reads generated from Hi-C libraries
Number of reads	320000000	393890890
Total bases (bp)	47259747176	39710027944
Coverage (based on estimated size 343-Mb)	138	116

Supplementary Data 1C. RNAseq sequencing summary.

sample	tissue	definition	raw	trimmed	mapped	unique	counted
LIB01	Leaf	mature leaf	44052158	42913561	42220234	41581454	38149797
LIB02	Leaf	mature leaf	41885646	40812495	40650024	40195721	36801615
LIB03	Leaf	mature leaf	48678709	47383457	46812946	46223099	41912453
LIB04	Leaf	expanded leaf	47096030	45952450	45419732	44562911	39782604
LIB05	Bud	green bud	37204667	36274916	34994061	33049744	27784670
LIB06	Bud	green bud	55179476	53698107	52401405	50010942	43529938
LIB07	Bud	green bud	51575648	50148875	48719871	46435981	41026958
LIB08	Peduncle	green bud peduncle	31980635	30697668	29475989	28815125	25820030
LIB09	Peduncle	green bud peduncle	44736137	42918096	39292777	38460557	34240458
LIB10	Peduncle	green bud peduncle	17114088	16586844	16305487	15958086	14075331
LIB11	Bud	pink bud	34440298	33499703	32843675	30144106	27613142
LIB12	Bud	pink bud	46509387	45145570	44643262	40900238	38036812
LIB13	Bud	fruiting bud	64427777	62677464	61465962	59354866	54562800
LIB14	Bud	fruiting bud	57020894	55115649	53015446	50499225	46080164
LIB15	Peduncle	fruiting bud peduncle	10245534	9783927	5855993	5733166	5145767
LIB16	Stem	stem	16785927	15529961	10884579	10737332	9538736
LIB17	Stem	stem	48508186	46539701	41099708	40572539	36545644
LIB18	Flower	flower	76116400	73192555	70851887	64380146	58708126
LIB19	Bud	initial fruiting bud	42992441	41421034	41006029	37214600	34321941
LIB20	Flower	flower	37485994	36124542	35637610	32571410	29756759
LIB21	Leaf	expanded leaf	23596478	22783677	22536374	22245343	20593020
LIB22	Peduncle	pink and flowering bud peduncle	41627610	39422353	38907564	38078672	34808111
LIB23	Peduncle	pink and flowering bud peduncle	46225595	43633219	42864208	41834238	38326364
LIB24	Leaf	young leaf I	43053356	40771748	40158410	39244313	35910714
LIB25	Leaf	young leaf I	33777864	32077644	31783907	31102835	28587713
LIB26	Leaf	young leaf I	45668499	43616785	43201313	42234994	38712996
LIB27	Bud	young bud not pooled	37383378	35530666	34986284	33985432	31382393
LIB28	Bud	young bud not pooled	40965046	38828811	38394259	37235016	34344030
LIB29	Bud	young bud not pooled	28627612	27101578	26802618	25937460	23953139
LIB30	Peduncle	young bud peduncle	44221736	42639595	42363232	40790640	37156130
LIB31	Peduncle	young bud peduncle	37133649	35922909	35694234	34374963	31370899
LIB32	Bud	pink bud	52773984	51102281	50782532	47064417	43501513
LIB33	Bud	initial fruiting bud	43271240	41891579	41555247	39396954	36249618
LIB34	Bud	initial fruiting bud	50382836	48703872	48254426	45981475	42151286
LIB35	Peduncle	young bud peduncle	47801309	46170983	43947276	42189012	37552077
LIB36	Bud	fruiting bud	47385933	45802542	44580605	42667534	38827438

sample	tissue	definition	raw	trimmed	mapped	unique	counted
LIB37	Leaf	mature leaf lamina	57870925	56001780	55252154	54332501	49890353
LIB38	Stem	stem	3592129	3107162	3021967	2986216	2707539
LIB39	Leaf	expanded leaf	56808280	55154806	54668901	53972101	49636410
LIB40	Peduncle	fruiting bud peduncle	40429950	38485065	37203108	36436370	32907814
LIB41	Peduncle	fruiting bud peduncle	59526452	56846266	54827344	53610407	48770704
LIB42	Leaf	mature leaf lamina	43870806	42321596	41352406	40059304	36141118
LIB43	Leaf	mature leaf lamina	44700718	43125314	42407110	41043635	37444794
LIB44	Leaf	mature leaf	33486364	32590134	32129561	31738676	29187249
LIB45	Leaf	mature leaf	25047380	24277094	23730375	23417178	21582167
LIB46	Leaf	mature leaf	23701713	23022929	22901997	22582869	20717859
LIB47	Leaf	expanded leaf	27655362	26876867	26353427	26009645	24084663
LIB48	Leaf	expanded leaf	27288090	26476770	26379654	25984673	23949248
LIB49	Leaf	expanded leaf	28545192	27721832	27609292	27250246	25082005
LIB50	Leaf	young leaf II	29449068	28483762	27609156	27016314	24025829
LIB51	Leaf	young leaf II	27499774	26709966	26522789	26042575	23933649
LIB52	Leaf	young leaf II	21192164	20594114	20465975	20102946	18438988
LIB53	Leaf	young leaf I	24318806	23570138	23431977	23013486	21257861
LIB54	Leaf	young leaf I	25370388	24642986	24502245	23984657	21752714
LIB55	Leaf	young leaf I	23179271	22519435	22377379	21914341	20098768
LIB56	Bud	green bud	35943216	35363755	34617093	32842602	29697043
LIB57	Bud	green bud	27855525	27350195	27065937	25883281	23082524
LIB58	Bud	green bud	33145806	32359340	31869589	30558536	27141365
LIB59	Bud	pink bud	30665103	29930490	29686359	28054319	25618380
LIB60	Bud	pink bud	24567586	24016459	23814709	22283790	20319855
LIB61	Bud	pink bud	36428697	35827892	35597537	32518261	30404712
LIB62	Bud	initial fruiting bud	37401885	36516285	36108653	33563565	30921443
LIB63	Bud	initial fruiting bud	24716759	24296721	23961122	22743240	20972642
LIB64	Bud	initial fruiting bud	39790507	39081584	38397316	36255255	33417418
LIB65	Bud	fruiting bud	36989498	36144086	35650223	34305187	31706150
LIB66	Bud	fruiting bud	35739165	34979837	34354004	32749223	30099200
LIB67	Bud	fruiting bud	32344361	31668356	30990618	29576413	27325417
LIB68	Bud	young bud pooled	21637644	21228021	21038609	20409183	19000276
LIB69	Bud	young bud pooled	25433401	25032316	24806907	24061632	22275721
LIB70	Bud	young bud pooled	126898348	124061012	122509908	118695714	110421569

Supplementary Data 2A. List of proteins used to identify *S. aromaticum* phenylpropanoid biosynthetic genes.

Protein	species	Family	gene ID
PAL1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01144*
PAL2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C03570*
PAL3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02848*
PAL4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02849*
PAL5	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02850*
PAL6	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02851*
PAL7	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02852*
PAL8	*Eucalyptus grandis*	Myrtaceae	*Eucgr.J00907*
PAL9	*Eucalyptus grandis*	Myrtaceae	*Eucgr.J01079*
C4H1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C00065*
C4H2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.J01844*
4CL1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C02284*
4CL10	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02758*
4CL11	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02879*
4CL12	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K02927*
4CL13	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K02929*
4CL2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00087*
4CL3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B00135*
4CL4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B03468*
4CL5	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B03502*
4CL6	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B03942*
4CL7	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B03943*
4CL8	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D02624*
4CL9	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03543*
HCT1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03972*
HCT2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03973*
HCT3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03974*
HCT4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03978*
HCT5	*Eucalyptus grandis*	Myrtaceae	*Eucgr.J03126*
C3'H1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A02185*
C3'H2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A02188*
C3'H3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A02190*
C3'H4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G03199*
CCoAOMT1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I01134*
CCoAOMT10	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C03684*
CCoAOMT11	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C03939*
CCoAOMT12	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F04260*
CCoAOMT13	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H04643*
CCoAOMT14	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H04644*
CCoAOMT15	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H04646*
CCoAOMT16	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H04648*

Protein	species	Family	gene ID
CCoAOMT17	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H04650*
CCoAOMT2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G01417*
CCoAOMT3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B02687*
CCoAOMT4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C00924*
CCoAOMT5	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C00925*
CCoAOMT6	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C03667*
CCoAOMT7	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C03668*
CCoAOMT8	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C03674*
CCoAOMT9	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C03680*
COMT1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01397*
COMT10	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01797*
COMT11	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01846*
COMT12	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01863*
COMT13	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01865*
COMT14	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01867*
COMT15	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01873*
COMT16	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01874*
COMT17	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01875*
COMT18	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01876*
COMT19	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01877*
COMT2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A00759*
COMT20	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01878*
COMT21	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01880*
COMT22	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01881*
COMT23	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01884*
COMT24	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A02870*
COMT25	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B01744*
COMT26	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B01747*
COMT27	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01092*
COMT28	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E03146*
COMT29	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E03148*
COMT3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01389*
COMT30	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E03339*
COMT31	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E03341*
COMT32	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E03874*
COMT33	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E03875*
COMT34	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E03877*
COMT35	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F02623*
COMT36	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F02624*
COMT37	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F02625*
COMT38	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03794*
COMT39	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G00017*
COMT4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01392*
COMT40	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G00020*
COMT41	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G01808*
COMT42	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G01810*

Protein	species	Family	gene ID
COMT43	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00347*
COMT44	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00348*
COMT45	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00349*
COMT46	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00350*
COMT47	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00351*
COMT48	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00352*
COMT49	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00353*
COMT5	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01394*
COMT50	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00354*
COMT51	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H00356*
COMT52	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H03920*
COMT53	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H03922*
COMT54	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H03924*
COMT55	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H03926*
COMT56	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I02810*
COMT57	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00041*
COMT58	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00449*
COMT59	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00949*
COMT6	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01395*
COMT60	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00950*
COMT61	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00951*
COMT62	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00953*
COMT63	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00954*
COMT64	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00955*
COMT65	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00956*
COMT66	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K00957*
COMT67	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K01696*
COMT7	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01600*
COMT8	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01795*
COMT9	*Eucalyptus grandis*	Myrtaceae	*Eucgr.A01796*
CCR1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.J03114*
CCR2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03954*
CCR3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.B02222*
CCR4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.C01240*
CCR5	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F03605*
CCR6	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G00052*
CCR7	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02325*
CCR8	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I01552*
CCR9	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I01783*
CAD10	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D01088*
CAD11	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D01089*
CAD12	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D01090*
CAD13	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D01091*
CAD14	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01103*
CAD15	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01104*
CAD16	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01105*

Protein	species	Family	gene ID
CAD17	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01107*
CAD18	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01108*
CAD19	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01110*
CAD2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G01350*
CAD20	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01115*
CAD21	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01117*
CAD22	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E01119*
CAD23	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E02204*
CAD24	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E02310*
CAD25	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E02319*
CAD26	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E02559*
CAD27	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E02570*
CAD28	*Eucalyptus grandis*	Myrtaceae	*Eucgr.E02580*
CAD29	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F01676*
CAD3	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H03208*
CAD30	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F01677*
CAD31	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F01678*
CAD32	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F01679*
CAD33	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F01680*
CAD34	*Eucalyptus grandis*	Myrtaceae	*Eucgr.G02223*
CAD35	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H02411*
CAD36	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H02412*
CAD37	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H02414*
CAD38	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H02415*
CAD39	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H02431*
CAD4	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D00468*
CAD40	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H02433*
CAD41	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H02434*
CAD42	*Eucalyptus grandis*	Myrtaceae	*Eucgr.H04903*
CAD43	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I00570*
CAD44	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I00571*
CAD45	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I00572*
CAD46	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I00573*
CAD47	*Eucalyptus grandis*	Myrtaceae	*Eucgr.K01941*
CAD5	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D00471*
CAD6	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D00472*
CAD7	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D00473*
CAD8	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D01086*
CAD9	*Eucalyptus grandis*	Myrtaceae	*Eucgr.D01087*
F5H1	*Eucalyptus grandis*	Myrtaceae	*Eucgr.I02371*
F5H2	*Eucalyptus grandis*	Myrtaceae	*Eucgr.J02393*
CSE	*Eucalyptus grandis*	Myrtaceae	*Eucgr.F02557*

Supplementary Data 2B. List of proteins used to identify *S. aromaticum* phenylpropanoid biosynthetic genes.

Protein	species	Family	Uniprot entry number	phytozome v12 entry
C3H/APX1 (AT1G07890)	*Arabidopsis thaliana*	Brassicaceae	F4HU93	NA
C3H/APX1 (AT1G07890)	*Arabidopsis thaliana*	Brassicaceae	Q05431	NA
C3H/APX1	*Populus trichocarpa v3.0*	Salicaceae	NA	Potri.009G015400
C3H/APX1	*Populus trichocarpa v3.0*	Salicaceae	NA	Potri.016G084800

Supplementary Data 2C. List of proteins used to identify *S. aromaticum* selected members of the pinoresinol–lariciresinol reductase, isoflavone reductase, and phenylcoumaran benzylic ether reductase (PIP) family.

Protein	species	Family	Uniprot entry number	NCBI entry number
EGS1	*Clarkia breweri*	Onagraceae	B2WSM9	NA
EGS2	*Clarkia breweri*	Onagraceae	B2WSN0	NA
EGS1a	*Fragaria x ananassa*	Rosaceae	T2AY45	NA
EGS1b	*Fragaria x ananassa*	Rosaceae	T2AXC4	NA
EGS2	*Fragaria x ananassa*	Rosaceae	T2AZ69	NA
EGS1	*Gymnadenia conopsea*	Orchidaceae	A0A0E3KXU6	NA
EGS1	*Gymnadenia odoratissima*	Orchidaceae	A0A0E3NCN9	NA
IEGS1	*Gymnadenia x densiflora*	Orchidaceae	A0A0E3H852	NA
APS1	*Larrea tridentata*	Zygophyllaceae	V5NR15	NA
EGS	*Ocimum gratissimum*	Lamiaceae	A0A1B2U6S8	NA
EGS	*Ocimum kilimandscharicum*	Lamiaceae	A0A1B2U6T4	NA
EGS	*Ocimum tenuiflorum*	Lamiaceae	A0A1B2U6S6	NA
EGS1	*Ocinum basilicum*	Lamiaceae	Q15GI4	NA
EGS1	*Petunia x hybrida*	Solanaceae	B2WSN1	NA
APS1	*Piper regnellii*	Piperaceae	V5NT01	NA
IGS1	*Clarkia breweri*	Onagraceae	B2WSM8	NA
PPS	*Larrea tridentata*	Zygophyllaceae	V5NQY6	NA
IGS1	*Petunia x hybrida*	Solanaceae	Q15GI3	NA
AIS1	*Pimpinella anisum*	Apiaceae	B8RCD2	NA
PPS	*Piper regnellii*	Piperaceae	V5NR18	NA
DcE(I)GS1	*Daucus carota*	Apiaceae	NA	XP_017241251.1
RcEGS1	*Rosa chinensis*	Rosaceae	M1FGX5	NA

Supplementary Data 2D. List of proteins used to identify *S. aromaticum* alcohol acyl transferase (AAT).

Protein	species	Family	Uniprot entry number	NCBI entry number
CbBEAT	*Clarkia breweri*	Onagraceae	O64988	NA
CbBEBT	*Clarkia breweri*	Onagraceae	Q8GT21	NA
CmAAT1	*Cucumis melo*	Cucurbitaceae	P93094	NA
CmAAT2	*cucumis melo*	Cucurbitaceae	Q8SA67	NA
CmAAT3	*Cucumis melo*	Cucurbitaceae	Q5I2Q5	NA
CmAAT4	*Cucumis melo*	Cucurbitaceae	Q5I2Q4	NA
VAAT	*Fragaria vesca*	Rosaceae	Q8GTM5	NA
SAAT	*Fragaria x ananassa*	Rosaceae	Q9FVF1	NA
CAAT1	*Larrea tridentata*	Zygophyllaceae	V5NSZ9	NA
MdAAT1	*Malus domestica PPP*	Rosaceae	Q64FJ6	NA
BanAAT	*Musa sp*	Musaceae	NA	CAC09063.1
SALAT	*Papaver somniferum*	Papaveraceae	Q94FT4	NA
BPBT	*Petunia hybrida*	Solanaceae	Q6E593	NA
PhCFAT	*Petunia x hybrida*	Solanaceae	A1XWY7	NA
RhAAT1	*Rosa hybrid cultivar*	Rosaceae	Q5I6B5	NA
ObCAAT1	*Ocinum basilicum*	Lamiaceae	NA	QKK36036.1
FCAAT1	*Fragaria chiloensis*	Rosaceae	D0QJ93	NA
MdAAT1-GS	*Malus domestica PPP*	Rosaceae	V9P9T8	NA
FaAAT2	*Fragaria ananassa*	Rosaceae	G1EFQ3	NA
SpAAT1	*Solanum pennellii*	Solanaceae	A0A0A0RSG8	NA
SlAAT1	*Solanum lycopersicum*	Solanaceae	Q6QLX4	NA
VpAAT1	*Vasconcellea cundinamarcensis*	Caricaceae	D0QJ94	NA

Supplementary Data 3. List of selected genes in *S. aromaticum*.

PROTEIN ID	GENE ID	PROTEIN ID	GENE ID	PROTEIN ID	GENE ID
4CL1	*saro002210*	AAT37	*saro015650*	C3H3	*saro006040*
4CL10	*saro027091*	AAT38	*saro015655*	C3H4	*saro013498*
4CL11	*saro027249*	AAT39	*saro015656*	C3H5	*saro022087*
4CL12	*saro027250*	AAT4	*saro001572*	C3H6	*saro022916*
4CL2	*saro002211*	AAT40	*saro016022*	C3H7	*saro024461*
4CL3	*saro003720*	AAT41	*saro017416*	C4H1	*saro005105*
4CL4	*saro004831*	AAT42	*saro017626*	C4H2	*saro024000*
4CL5	*saro004858*	AAT43	*saro017628*	C4H3	*saro024007*
4CL6	*saro006451*	AAT44	*saro020356*	CAD1	*saro008030*
4CL7	*saro009285*	AAT45	*saro020359*	CAD10	*saro010523*
4CL8	*saro014422*	AAT46	*saro020361*	CAD11	*saro012757*
4CL9	*saro016984*	AAT47	*saro023378*	CAD12	*saro015897*
AAT1	*saro000535*	AAT48	*saro023379*	CAD13	*saro018287*
AAT10	*saro006403*	AAT49	*saro023486*	CAD14	*saro018288*
AAT11	*saro010031*	AAT5	*saro003214*	CAD15	*saro020520*
AAT12	*saro010032*	AAT50	*saro023526*	CAD16	*saro020856*
AAT13	*saro010033*	AAT51	*saro024898*	CAD17	*saro020857*
AAT14	*saro010038*	AAT52	*saro024900*	CAD18	*saro022325*
AAT15	*saro012502*	AAT53	*saro025096*	CAD2	*saro008031*
AAT16	*saro012505*	AAT54	*saro027239*	CAD3	*saro008032*
AAT17	*saro012507*	AAT55	*saro027484*	CAD4	*saro008033*
AAT18	*saro012692*	AAT6	*saro004815*	CAD5	*saro008034*
AAT19	*saro012762*	AAT7	*saro004816*	CAD6	*saro008376*
AAT2	*saro001094*	AAT8	*saro004817*	CAD7	*saro010003*
AAT20	*saro012813*	AAT9	*saro005806*	CAD8	*saro010004*
AAT21	*saro012814*	ABC2-072	*saro016948*	CAD9	*saro010006*
AAT22	*saro012815*	ABC2-113	*saro023078*	CCoAOMT1	*saro004326*
AAT23	*saro012818*	ABC2-115	*saro023111*	CCoAOMT10	*saro020324*
AAT24	*saro012820*	ABCB-138	*saro025277*	CCoAOMT11	*saro022299*
AAT25	*saro012821*	ABCC-041	*saro009861*	CCoAOMT2	*saro005659*
AAT26	*saro012824*	ABCG-061	*saro013997*	CCoAOMT3	*saro005660*
AAT27	*saro014481*	ABCG-062	*saro013999*	CCoAOMT4	*saro005662*
AAT28	*saro015635*	ABCG-064	*saro014171*	CCoAOMT5	*saro007134*
AAT29	*saro015636*	ABCG-116	*saro023302*	CCoAOMT6	*saro013577*
AAT3	*saro001102*	ABCG-131	*saro024701*	CCoAOMT7	*saro015962*
AAT30	*saro015637*	ABCG-133	*saro024787*	CCoAOMT8	*saro015964*
AAT31	*saro015639*	ABCG-143	*saro026067*	CCoAOMT9	*saro020321*
AAT32	*saro015641*	ABCI-001	*saro000396*	CCR1	*saro005060*
AAT33	*saro015644*	C3H1	*saro000721*	CCR10	*saro024838*
AAT34	*saro015645*	C3'H1	*saro001561*	CCR2	*saro006537*
AAT35	*saro015647*	C3H2	*saro004114*	CCR3	*saro014466*
AAT36	*saro015648*	C3'H2	*saro017357*	CCR4	*saro015340*
CCR5	*saro016645*	COMT44	*saro025510*	PAL2	*saro007067*

PROTEIN ID	GENE ID	PROTEIN ID	GENE ID	PROTEIN ID	GENE ID
CCR6	*saro016646*	COMT5	*saro001316*	PAL3	*saro017014*
CCR7	*saro016659*	COMT6	*saro001317*	PAL4	*saro017015*
CCR8	*saro021805*	COMT7	*saro001318*	PAL5	*saro022827*
CCR9	*saro021970*	COMT8	*saro001321*	PAL6	*saro022964*
COMT1	*saro000861*	COMT9	*saro001328*	PCBER1	*saro009582*
COMT10	*saro003543*	CSE1	*saro015111*	PCBER2	*saro009583*
COMT11	*saro006088*	CSE2	*saro016627*	PCBER3	*saro009584*
COMT12	*saro006980*	CSE3	*saro017029*	PCBER4	*saro009585*
COMT13	*saro009991*	CSE4	*saro017086*	PCBER5	*saro012679*
COMT14	*saro009992*	CSE5	*saro018482*	PCBER6	*saro016105*
COMT15	*saro014639*	CSE6	*saro022751*	PCBER7	*saro021338*
COMT16	*saro014861*	EGS1	*saro023967*	PCBER8	*saro021339*
COMT17	*saro014862*	EGS10	*saro023976*	PCBER9	*saro023862*
COMT18	*saro014863*	EGS11	*saro023977*		
COMT19	*saro014864*	EGS12	*saro023978*		
COMT2	*saro001049*	EGS13	*saro023979*		
COMT20	*saro014932*	EGS14	*saro023980*		
COMT21	*saro015355*	EGS15	*saro027336*		
COMT22	*saro016094*	EGS2	*saro023968*		
COMT23	*saro016249*	EGS3	*saro023969*		
COMT24	*saro017578*	EGS4	*saro023970*		
COMT25	*saro017579*	EGS5	*saro023971*		
COMT26	*saro017582*	EGS6	*saro023972*		
COMT27	*saro017583*	EGS7	*saro023973*		
COMT28	*saro017584*	EGS8	*saro023974*		
COMT29	*saro017586*	EGS9	*saro023975*		
COMT3	*saro001050*	esterase23	*saro007992*		
COMT30	*saro017955*	F5H1	*saro021382*		
COMT31	*saro017959*	F5H2	*saro024413*		
COMT32	*saro025216*	HCT1	*saro013789*		
COMT33	*saro025217*	HCT2	*saro024848*		
COMT34	*saro025218*	hydrolase149	*saro008781*		
COMT35	*saro025493*	hydrolase424	*saro026641*		
COMT36	*saro025494*	hydrolase425	*saro026642*		
COMT37	*saro025495*	NmrA1	*saro004483*		
COMT38	*saro025497*	NmrA2	*saro010087*		
COMT39	*saro025498*	NmrA3	*saro010088*		
COMT4	*saro001262*	NmrA4	*saro010089*		
COMT40	*saro025499*	NmrA5	*saro016270*		
COMT41	*saro025503*	NmrA6	*saro027306*		
COMT42	*saro025508*	NmrA7	*saro027307*		
COMT43	*saro025509*	PAL1	*saro000784*		

Supplementary Data 4. Additional HPLC-UV and GC-MS analysis results. Concentration (mg/g of FW) of metabolites in Syzygium aromaticum leaves and buds at different growth phases. Data represented are mean ± SD (n = 3). Eugenol and eugenol acetate concentrations were determined by HPLC-UV; the concentrations of the other compounds were determined by GC–MS. **Two values not detected; *one value not detected; -not detected. Young leaf (YLI), expanded young leaf (YLII), expanded pale green leaf (EL), mature leaf (ML), young bud (YB), green bud (GB), pink bud (PB), bud in the initial fruiting stage (IFB), bud in the fruiting stage (FB).

			Bud				
	Growth phase		**YB**	**GB**	**PB**	**IFB**	**FB**
Compound class	Compound	RT (min)					
Phenylpropenes	Allyl-phenylpropenes						
	Eugenol	2.88	8.607 ± 1.390	20.692 ± 5.189	29.660 ± 1.876	27.863 ± 1.867	18.687 ± 1.680
	eugenol acetate	3.35	43.117 ± 6.050	37.503 ± 2.749	3.548 ± 1.586	0.787 ± 0.314	0.163 ± 0.283**
	chavicol	15.9	0.014 ± 0.003	0.038 ± 0.015	0.065 ± 0.005	0.084 ± 0.004	0.052 ± 0.015
	Chavicol acetate	17.8	0.083 ± 0.007	0.122 ± 0.039	-	-	-
	Safrole	16.8	-	-	-	-	-
	Methyl eugenol	-	-	-	-	-	-
	Methyl chavicol (Estragole)	-	-	-	-	-	-
	Propenyl-phenylpropenes						
	Isoeugenol	20.2	0.022 ± 0.003	0.029 ± 0.007	0.019 ± 0.009	0.020 ± 0.002	0.015 ± 0.002
	isoeugenol acetate	23.7	0.027 ± 0.002	0.023 ± 0.007	0.005 ± 0.001	-	-
	t-Anol/anol	-	-	-	-	-	-
	Methyl isoeugenol	-	-	-	-	-	-
	Anethole	-	-	-	-	-	-
	Methyl isochavicol (anethole)	-	-	-	-	-	-
Terpenes	Sesquiterpenes						
	Caryophyllene	19.67	3.839 ± 1.323	3.070 ± 0.875	3.145 ± 0.470	3.983 ± 0.229	4.372 ± 3.243
	Caryophyllene oxide	23.31	-	0.040 ± 0.003	0.049 ± 0.003	0.090 ± 0.023	0.094 ± 0.036
	Humulene	20.49	0.655 ± 0.073	0.594 ± 0.165	0.357 ± 0.032	0.393 ± 0.019	0.232 ± 0.062
	Humulene oxide	23.93	-	-	0.0296 ± 0.0002	0.033 ± 0.002	0.032 ± 0.001
	alpha-Farnesene	21.37	0.286 ± 0.017	0.256 ± 0.094	0.130 ± 0.010	0.156 ± 0.015	0.074 ± 0.034
	beta-Cubebene	21	0.120 ± 0.020	0.110 ± 0.043	0.040 ± 0.035*	0.063 ± 0.007	0.024 ± 0.013
	beta-Elemene	18.97	0.065 ± 0.013	0.042 ± 0.016	0.013 ± 0.004	0.009 ± 0.001	0.001 ± 0.002**
	Copaene	18.7	0.109 ± 0.013	0.091 ± 0.031	0.041 ± 0.002	0.049 ± 0.002	0.032 ± 0.012
	Nerolidol	22.69	-	-	-	-	-
Other phenolic compound	2-Methoxy-4-propyl-phenol	20.37	0.009 ± 0.002	0.007 ± 0.001	0.005 ± 0.001	0.0056 ± 0.0004	0.004 ± 0.001
	Coniferyl acetate	-	-	-	-	-	-
	Coumaryl acetate	-	-	-	-	-	-
Benzofuran	6-Methoxy-3-methyl-benzofuran	20.87	-	0.012 ± 0.002	0.016 ± 0.003	0.015 ± 0.005	0.014 ± 0.006
Monolignol	Coniferyl alcohol	21.9	-	-	0.0046 ± 0.0001	0.006 ± 0.001	0.005 ± 0.001
	Coumaryl alcohol	na	-	-	-	-	-
Cinnamaldehyde	Coniferyl aldehyde	26.4	0.0036 ± 0.0002	0.0042 ± 0.0004	0.0044 ± 0.0001	0.005 ± 0.001	0.0046 ± 0.0006
	Coumaraldehyde	-	-	-	-	-	-
Benzoate ester	Methyl salicylate	14.8	0.074 ± 0.009	0.093 ± 0.041	0.104 ± 0.018	0.170 ± 0.043	0.084 ± 0.019

			Leaf			
	Growth phase		**YLI**	**YLII**	**EL**	**ML**
Compound class	Compound	RT (min)				
Phenylpropenes	Allyl-phenylpropenes					
	Eugenol	2.88	8.328 ± 1.059	11.076 ± 2.378	20.112 ± 2.675	27.866 ± 6.209
	eugenol acetate	3.35	20.307 ± 1.925	29.117 ± 4.184	0.107 ± 0.038	0.053 ± 0.013
	chavicol	15.9	0.015 ± 0.006	0.013 ± 0.004	0.033 ± 0.009	0.057 ± 0.020
	Chavicol acetate	17.8	0.051 ± 0.014	0.060 ± 0.008	-	-
	Safrole	16.8	-	-	0.0003 ± 0.0005*	-
	Methyl eugenol	-	-	-	-	-
	Methyl chavicol (Estragole)	-	-	-	-	-
	Propenyl-phenylpropenes					
	Isoeugenol	20.2	0.012 ± 0.004	0.012 ± 0.001	0.010 ± 0.004	0.014 ± 0.005
	isoeugenol acetate	23.7	0.014 ± 0.003	0.016 ± 0.005	-	-
	t-Anol/anol	-	-	-	-	-
	Methyl isoeugenol	-	-	-	-	-
	Anethole	-	-	-	-	-
	Methyl isochavicol (anethole)	-	-	-	-	-
Terpenes	Sesquiterpenes					
	Caryophyllene	19.67	4.107 ± 0.387	5.646 ± 1.223	3.125 ± 0.572	5.359 ± 1.354
	Caryophyllene oxide	23.31	0.010 ± 0.001	0.012 ± 0.011*	0.143 ± 0.035	0.961 ± 0.356
	Humulene	20.49	0.642 ± 0.191	0.576 ± 0.112	0.384 ± 0.071	0.649 ± 0.203
	Humulene oxide	23.93	-	0.001 ± 0.001**	0.011 ± 0.003	0.081 ± 0.033
	alpha-Farnesene	21.37	0.183 ± 0.061	0.132 ± 0.020	0.091 ± 0.022	0.026 ± 0.010
	beta-Cubebene	21	0.032 ± 0.014	0.019 ± 0.003	0.014 ± 0.004	0.011 ± 0.006
	beta-Elemene	18.97	0.029 ± 0.015	0.022 ± 0.003	0.008 ± 0.001	0.001 ± 0.001*
	Copaene	18.7	0.061 ± 0.023	0.050 ± 0.008	0.030 ± 0.005	0.063 ± 0.030
	Nerolidol	22.69	0.005 ± 0.004	0.006 ± 0.006*	0.010 ± 0.001	0.016 ± 0.005
Other phenolic compound	2-Methoxy-4-propyl-phenol	20.37	-	-	-	0.008 ± 0.005
	Coniferyl acetate	-	-	-	-	-
	Coumaryl acetate	-	-	-	-	-
Benzofuran	6-Methoxy-3-methyl-benzofuran	20.87	0.013 ± 0.004	0.011 ± 0.001	0.032 ± 0.014	0.022 ± 0.008
Monolignol	Coniferyl alcohol	21.9	-	-	0.0014 ± 0.0003	0.010 ± 0.004
	Coumaryl alcohol	na	-	-	-	-
Cinnamaldehyde	Coniferyl aldehyde	26.4	0.0006 ± 0.0003	0.0005 ± 0.0004*	0.0016 ± 0.0005	0.009 ± 0.003
	Coumaraldehyde	-	-	-	-	-
Benzoate ester	Methyl salicylate	14.8	0.007 ± 0.004	0.004 ± 0.001	0.0020 ± 0.0002	0.006 ± 0.003

Chapter 3 - Preface

This chapter is a research article published in Frontiers in Plant Science (DOI: 10.3389/fpls.2023.1248780). The final peer reviewed version of this article can be found in annex.

This article on the evolution of *Syzygium* genomes is a collaborative project between the University of Neuchâtel and PMI Research and Development.

For this study, I performed the laboratory work, analyzed data, and wrote the manuscript. Nicolas Sierro performed computational analysis of sequencing data, conceived and supervised the study, and contributed to manuscript writing. Felix Kessler and Nikolai V. Ivanov conceived and supervised the study and contributed to manuscript writing.

Chapter 3 - Chromosome-scale assemblies of *S. malaccense, S. aqueum, S. jambos*, and *S. syzygioides* provide insights into the evolution of *Syzygium* genomes

Sonia Ouadi[1, 2], Nicolas Sierro[2], Felix Kessler[1], Nikolai V. Ivanov[2*]

[1]Faculty of Sciences, Laboratory of Plant Physiology, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland

[2]PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland

*CORRESPONDING AUTHOR

Nikolai V. Ivanov

E-mail: nikolai.ivanov@pmi.com

Running Title: Evolution of *Syzygium* genomes

Abstract

Syzygium is a large and diverse tree genus in the Myrtaceae family. Genome assemblies for clove (*Syzygium aromaticum*, 370 Mb) and sea apple (*Syzygium grande*, 405 Mb) provided the first insights into the genomic features and evolution of the *Syzygium* genus. Here, we present additional de novo chromosome-scale genome assemblies for *Syzygium malaccense*, *Syzygium aqueum*, *Syzygium jambos*, and *Syzygium syzygioides*. Genome profiling analyses show that *S. malaccense*, like *S. aromaticum* and *S. grande*, is diploid (2n = 2x = 22), while the *S. aqueum*, *S. jambos*, and *S. syzygioides* specimens are autotetraploid (2n = 4x = 44). The genome assemblies of *S. malaccense* (430 Mb), *S. aqueum* (392 Mb), *S. jambos* (426 Mb), and *S. syzygioides* (431 Mb) are highly complete (BUSCO scores of 98%). Comparative genomics analyses showed conserved organization of the 11 chromosomes with *S. aromaticum* and *S. grande*, and revealed species-specific evolutionary dynamics of the long terminal repeat retrotransposon elements belonging to the Gypsy and Copia lineages. This set of *Syzygium* genomes is a valuable resource for future structural and functional comparative genomic studies on Myrtaceae species.

Keywords: *Syzygium*, Myrtaceae, *de novo* assembly, comparative genomics, synteny, long terminal repeat retrotransposons.

3.1 Introduction

Syzygium is the largest tree genus with about 1,200 species naturally occurring from the Old World tropics and subtropics to the Pacific [1-3]. In addition to their ecological importance, the genus includes several species grown for their edible fruit, medicinal properties, timber, and for the horticulture industry (e.g., *S. malaccense, S. aqueum, S. jambos,* and *S. cumini*), the most important economically being the clove tree (*S. aromaticum)* [4-7].

The *Syzygium* genus belongs to the Myrtaceae family—the eighth largest family of flowering plants—and includes economically important species such as eucalyptus, myrtle, and guava [8-10]. Although the majority of species of the Myrtaceae family are diploids (2n = 22) with small to intermediate genome sizes (234–1785 Mb), occasional polyploids derived from the most conserved chromosome number x = 11 were also reported (e.g., within the *Eugenia, Syzygium,* and *Psidium* genera) [8, 11-14]. The *Eucalyptus grandis* genome was released in 2014 as the first reference genome for the Myrtales order and the Myrtaceae family [15]. New chromosome-scale assemblies were subsequently published, enabling comparative genomics analyses within the family. Published chromosome-scale genome assemblies for the Myrtaceae currently represent major tribes of the family: Eucalypteae (*Eucalyptus grandis, Corymbia citriodora, Eucalyptus urophylla* × *Eucalyptus grandis*), Leptospermeae (*Leptospermum scoparium*), Myrteae (*Psidium guajava, Rhodomyrtus tomentosa*), Metrosidereae (*Metrosideros polymorpha*), Melaleuceae (*Melaleuca alternifolia*) and Syzygieae (*S. aromaticum, Syzygium grande*). These assemblies were generated from diploid specimens, and their size ranged from 297 Mb to 690 Mb [15-24].

The clove (*S. aromaticum* (L.) Merr. & L.M. Perry) and sea apple (*S. grande*) genomes were constructed using a combination of Oxford Nanopore Technologies long-reads and Illumina short-reads and anchored on 11 chromosomes using Hi-C technologies [20, 21]. The sea apple genome assembly (405 Mb), 182 re-sequenced *Syzygium* species and 58 re-sequenced unidentified taxa were used to generate whole genome-level phylogenies of the *Syzygium* genus, thus providing new insights into the infrageneric classification of *Syzygium*, as well as into the genus diversification patterns and their drivers.

The clove genome assembly (370 Mb) was exploited to investigate the genetic basis of the biosynthesis of eugenol, the major biocompound of clove products [25, 26]. To provide insights into the clove genome evolution, comparative genomics analyses were also performed between *S. aromaticum* and *E. grandis*. The synteny analysis performed between these two Myrtaceae species' genomes assemblies revealed good genome structure conservation. The structures of chromosomes 1, 3, 5, and 7 were found to be highly conserved between *E. grandis* and *S. aromaticum*, and 10 intrachromosomal rearrangements occurring on the 7 other chromosomes were observed (chromosomes 2, 4, 6, 8, 9, 10, and 11). Interestingly, the intrachromosomal rearrangements detected between the two eucalypt species, *E. grandis* and *C. citriodora*, were located on the same seven chromosomes [16, 27]. Long terminal repeat retrotransposons (LTR-RTs) are transposable elements (TEs) that move through the genome via a copy-and-paste mechanism using an RNA intermediate. They are considered the most abundant TE component in plant genomes and important drivers of genome size variation and diversification [28, 29]. Comparing the LTR-RTs repertoires of *S. aromaticum* and *E. grandis* revealed a differential accumulation of the LTR-RTs belonging to the superfamilies Copia and Gypsy between the two species. In *S. aromaticum* genome assembly, the LTR-RTs belonging to the Gypsy superfamily were more abundant than those belonging to the Copia superfamily. In contrast, a higher number of LTR-RTs Copia versus Gypsy was found in the *E. grandis* genome assembly.

No infrageneric comparison of chromosome-scale assemblies has been performed for the *Syzygium* genus. To further investigate the evolution of the genome architecture of *Syzygium* species and verify whether the rearrangements found between *S. aromaticum* and *E. grandis* chromosomes were the consequences of evolutionary events or due to sequencing and assembly artifacts, we generated additional chromosome-scale genome assemblies for *Syzygium malaccense* (L.) Merr. & L.M. Perry, *Syzygium syzygioides* (Miq.) Merr. & L.M. Perry, *Syzygium aqueum* (Burm.f.) Alston, and *Syzygium jambos* L. (Alston). Like *S. aromaticum* and *S. grande*, the four species belong to the subgenus *Syzygium*, the largest of the five *Syzygium* subgenera for which the crown age was estimated at 9.4 Mya by Low *et al.* [21]. Previous karyotype studies indicated that *S. malaccense* is a diploid with 2n = 22 chromosomes [30] and that *S. jambos* is a tetraploid (2n = 44); however, different chromosome numbers were also reported in the literature for the species (2n = 28, 33, 46, ~54, 66) [31, 32]. The chromosomal numbers reported in the literature indicate that *S. aqueum* is also a tetraploid (2n = 44) [33].

Here, we describe the *de novo* assembly and annotation for *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides*. To enable subsequent comparative genomic analyses, the four genomes consisting of monoploid consensus (11 chromosomes and unplaced sequences) were generated to achieve the same level of quality for the four species' genome assemblies and comparable to those of published chromosome-scale assemblies of their Myrtaceae relatives. Then, we compared the genome architecture of the four newly *Syzygium* assembled genomes with those of *S. aromaticum* and *S. grande* and their genome features (gene sets and LTR-RTs repertoires) with those of *S. aromaticum* to investigate genomic evolution from their common ancestors.

3.2 Results

3.2.1 Genome Profiling

Smudgeplot and GenomeScope 2.0 were used to perform a genome profiling step using Illumina PE short-reads from DNAseq libraries as input and a K-mer length of 21 bp [34] (Table 1, Supplemental Table 1, Supplemental Figure 1**)**. The ploidy level predicted by Smudgeplot was in accordance with previous karyotype studies for the studied *S. malaccense* and *S. jambos* specimens [30, 31]. *S. malaccense* was predicted to be a diploid specimen (2n = 2x = 22) like *S. aromaticum* and *S. grande*. The *S. aqueum*, *S. syzygioides*, and *S. jambos* specimens were predicted as being autotetraploid (2n = 4x = 44). The estimated monoploid genome sizes were similar among the four *Syzygium* species (343–372 Mb), a size range consistent with the small genome assembly sizes of *S. aromaticum* (370 Mb) and *S. grande* (405 Mb) [20, 21]. The heterozygosity rate estimated by the GenomeScope 2.0 ranged from 2.3% for the diploid specimen *S. malaccense* to 4.3% for the autotetraploid specimen *S. aqueum.* These heterozygosity rates appeared to be higher than for *S. aromaticum* (0.18%) [20] and the average reported by Ellestad et al., who performed a literature review of the genome-wide heterozygosity values estimated using the software GenomeScope and GenomeScope 2.0 [35]. They found that the average value inferred for all plant species assessed was 1.59% (1.10% for diploid plants only) noting that over half of the plant species considered were cultivated for human usage, which could affect the average value accuracy.

Table 1 Genome profiling summary.

	S. aromaticum [20]	***S. malaccense***	***S. aqueum***	***S. jambos***	***S. syzygioides***
Predicted ploidy	2n = 2x = 22	2n = 2x = 22	2n = 4x = 44	2n = 4x = 44	2n = 4x = 44
Estimated genome (1x) size	343 Mb	372 Mb	345 Mb	361 Mb	372 Mb
Estimated heterozygosity rate	0.18%	2.30%	4.30%	3.60%	4.10%

3.2.2 Genome *De Novo* Assembly

The four *de novo* chromosome-scale assemblies were constructed using long-reads from Oxford Nanopore Technologies (ONT), short paired-end reads from Illumina DNAseq libraries, and Hi-C libraries generated for each *Syzygium* species (Supplemental Tables 1–2).

ONT long-reads were first corrected using Illumina short-paired end reads from DNAseq libraries to reduce the ONT raw reads error rate. After this correction step, ONT long-reads were assembled to produce primary contigs, then polished using Illumina short-paired end reads from DNAseq libraries. To prevent assembly artifacts possibly caused by heterozygosity and polyploidy of the *Syzygium* specimens, haplotigs were detected and removed from the polished contigs. The effect of the haplotig removal step was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) in genome mode [36]. After the haplotig removal step, the number of complete and duplicated BUSCOs genes was considerably reduced in the haplotig-purged contigs (3.3% to 6.1%) when compared to the polished contigs (93.6% to 97.1%) (Figure 1A). Lastly, Hi-C data enabled the scaffolding of contigs into 11 chromosomes. On the Hi-C contact matrices, a strong intra-chromosomal signal indicates efficient scaffolding, with the 11 chromosomes of each *Syzygium* assembly supported by a high number of their respective Hi-C reads. (Figure 1B).

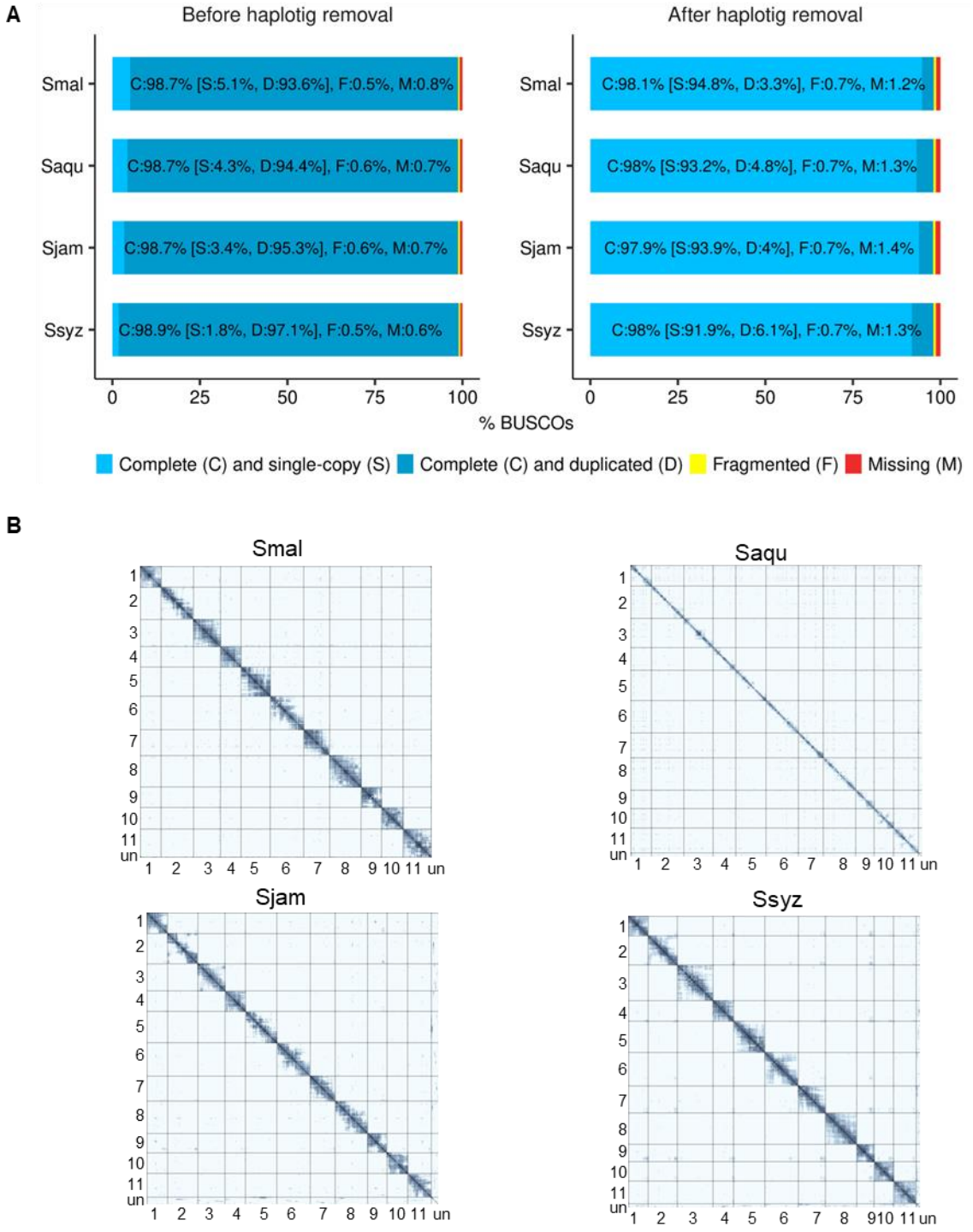


Figure 1 Assessment of the efficiency of the haplotig removal step and Hi-C scaffolding. **(A)** BUSCO completeness score comparison of the polished contigs before and after the haplotig removal step for *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz) (BUSCO version 5.4.4 - dataset: eudicots_odb10 (n = 2326)).
(B) Hi-C contact maps showing the Hi-C interactions among the 11 assembled chromosomes and unplaced scaffolds (un) for each species.

The final chromosome-scale assemblies for *S. malaccense* (430 Mb), *S. aqueum* (392 Mb), *S. jambos* (426 Mb), and *S. syzygioides* (431 Mb) consisted of monoploid consensus (11 chromosomes and unplaced sequences) with comparable quality metrics. A high level of quality at the base-scale (quality value [QV] between 44.006 and 45.114), of contiguity (97.5% to 99.8% of the assemblies length anchored on 11 chromosomes) and completeness (BUSCO complete genes scores of 98%) was reached for the four new assembled *Syzygium* genomes (Table 2, Figure 2, Supplemental Tables 3–4).

Table 2 Assembly and annotation statistics.

	S. malaccense	***S. aqueum***	***S. jambos***	***S. syzygioides***
Assembly				
Number of scaffolds	23	54	117	101
Number of chromosome-scale scaffolds	11	11	11	11
Proportion of undetermined bases (N)	0.01%	0.01%	0.02%	0.02%
QV[a] of the assembly	45.114	44.006	44.028	44.292
Length of assembly (bp)	429,836,287	391,897,832	426,159,599	431,079,378
Length of chromosome-scale scaffolds (bp)	429,008,219	386,536,673	415,622,982	424,827,227
Gene annotation				
Number of predicted genes	30,842	29,879	31,611	32,142
Number of predicted transcripts	57,144	55,010	57,897	59,495
Average transcript length (bp)	2010.89	2008.09	1991.19	2007.31
Average CDS[b] length (bp)	1122.42	1124.09	1116.93	1100.23
Average exon per transcript	5.62	5.67	5.59	5.66
Repeat annotation				
Repeat sequences (bp)	184,916,857 (43.02%)	162,020,435 (41.34%)	180,563,593 (42.37%)	184,003,101 (42.68%)
LTR[c] retrotransposons (bp)	96,086,564 (22.35%)	74,914,968 (19.12%)	77,407,268 (18.16%)	73,171,928 (16.97%)
LTR Gypsy (bp)	62,668,430 (14.58%)	48,141,612 (12.28%)	45,090,450 (10.58%)	40,369,474 (9.36%)
LTR Copia (bp)	31,769,467 (7.39%)	25,148,956 (6.42%)	30,040,740 (7.05%)	30,532,813 (7.08%)

a QV – Quality value
b CDS – Coding sequence
c LTR – Long Terminal Repeat

Despite their high heterozygosity rate, the quality metrics for the genome assemblies of the diploid specimen *S. malaccense* and the tetraploids *S. aqueum*, *S. jambos*, and *S. syzygioides* were comparable to those reported for *S. aromaticum* assembly (370 Mb) [20]. Nevertheless, BUSCO scores revealed a higher percentage of complete and duplicated BUSCOs in the four new assemblies compared to *S. aromaticum* (2.2%), principally in the genome assembly of the three autotetraploid specimens (3.3% to 5.5%) (Figure 2).

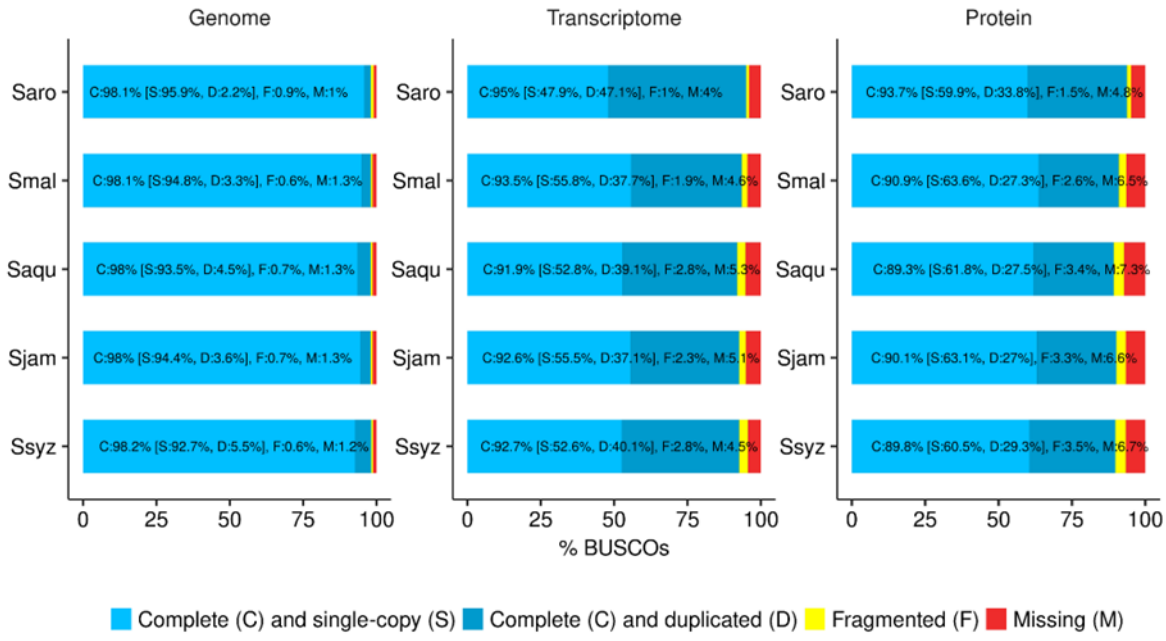


Figure 2 BUSCO completeness assessment.
Assessment of the final genome assembly, transcript set, and protein set of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz) (BUSCO version 5.4.4 - dataset: eudicots_odb10 (n = 2326)).

3.2.3 Genome Annotation

The gene model prediction was performed using sets of RNAseq data generated for each of the four newly assembled *Syzygium* species, the transcripts of *S. aromaticum,* and *E. grandis*. The average number of protein-coding genes predicted for the four newly assembled genomes is 31,119, representing 26.52% of the genome assemblies' size (Table 2).

The annotation completeness was assessed using the BUSCO method in transcriptome and protein modes and by selecting the whole set of predicted transcripts and proteins for each gene as inputs, respectively (Figure 2, Supplemental Table 3). BUSCO results indicated that the annotation completeness is comparable among the four newly assembled *Syzygium* species, with complete BUSCO scores ranging from 91.9% in *S. aqueum* assembly to 93.5% in *S. malaccense* assembly in transcript mode and from 89.3% in *S. aqueum* assembly to 90.9% in *S. malaccense* assembly in protein mode. BUSCO scores obtained for *S. aromaticum* by using the same assessment methods (95% in transcriptome mode and 93.7% in protein mode) were slightly superior to those of newly assembled genomes but still comparable. The loss of complete BUSCOs between the genome and protein mode assessments ranged from 7.2% in *S. malaccense* assembly to 8.7% in *S. aqueum* assembly, indicating acceptable quality of the predicted gene models and protein sets.

The genome assembly of *S. aromaticum* comprised multiple copies of a gene encoding for putative eugenol synthase (EGS), the enzyme that catalyzes the synthesis of eugenol from coniferyl acetate. In total, 15 copies split into 2 loci were reported: a first locus on chromosome 10 comprising 14 copies and a second locus on chromosome 11 with 1 copy [20]. The functional annotation of the four newly assembled *Syzygium* species genomes revealed fewer genes encoding for putative EGS. One gene encoding for putative EGS was identified in the genome assembly of *S. malaccense*, two in the genome assembly of *S. aqueum,* and three copies were found in the genome assemblies of *S. jambos* and *S. syzygioides*. All putative EGS genes were located on chromosome 10 except for one of the three copies of *S. syzygioides* located on chromosome 11 (Figure 3, Supplemental Table 5).

Effective lengths of repeat elements, which are different from their genomic length, were calculated by removing the length of the nested elements they contained. The proportions of genome assembly length occupied by predicted genes (25.97% to 27.37%) and repeat sequences (41.34% to 43.02%) appear to be conserved among the four newly sequenced *Syzygium* genomes (Table 2). Using the same method, repeat elements in *Syzygium aromaticum* genome assembly represents 39.98%. The most abundant repeat elements identified in the four newly sequenced *Syzygium* genomes were the LTR-RTs spanning 16.97% of the assembly length for *S. syzygioides* to 22.35% for *S. malaccense*. As reported for *S. aromaticum* and *S. grande*, LTR-RTs belonging to the Gypsy superfamily were more abundant than elements belonging to the Copia superfamily in the four newly sequenced genomes (Table 2, Supplemental Tables 6–9) [20, 21].

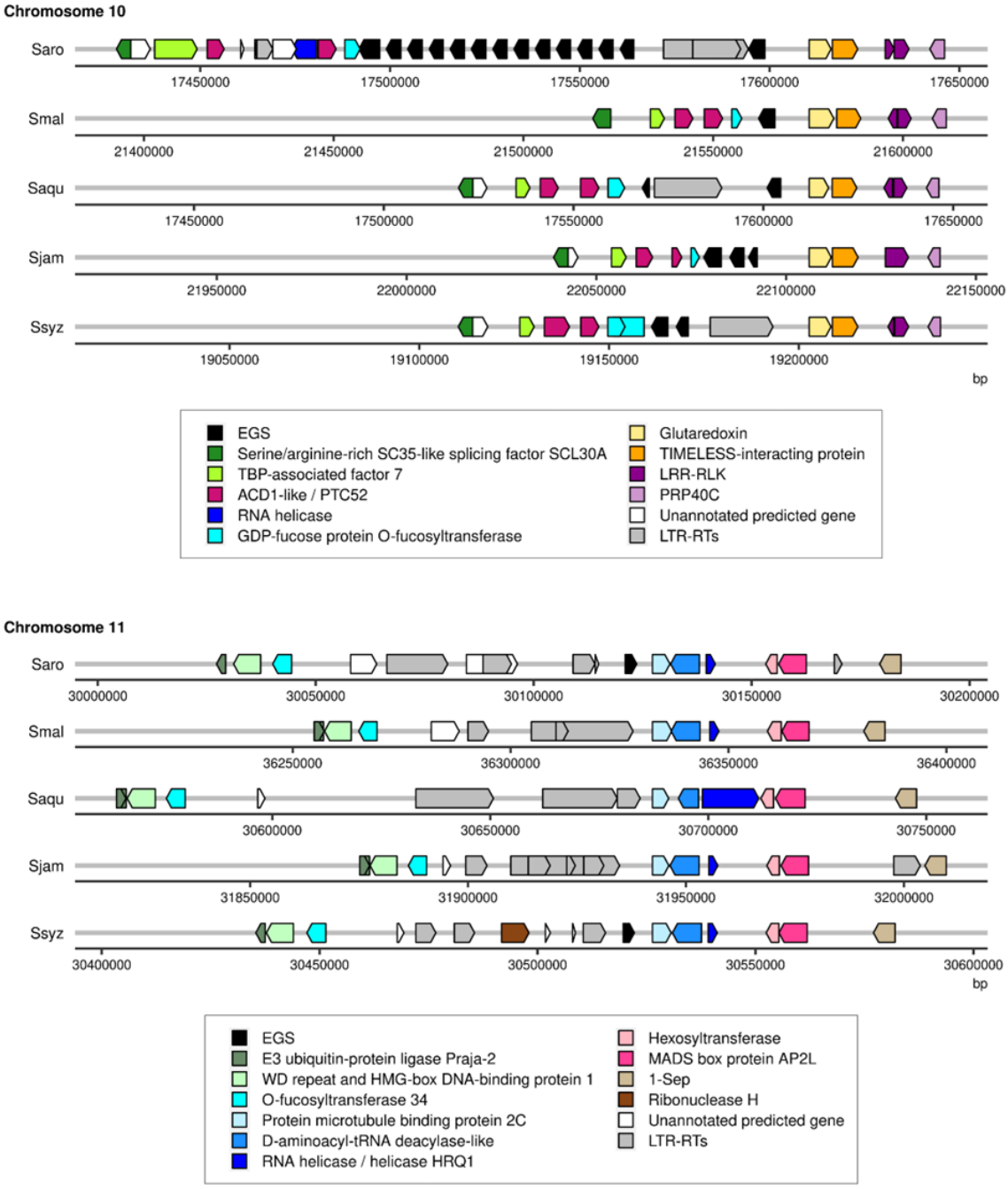


Figure 3 Illustration of the regions of chromosomes 10 and 11 of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz) where genes encoding for EGS were predicted. The position (bp) and orientation of the predicted genes on the chromosomes are indicated by arrows colored according to the functional annotation. EGS, accelerated cell death (ACD1), Protochlorophyllide-dependent translocon component Tic52 (PTC52), leucine-rich repeat receptor-like protein kinase (LRR-RLK), Pre-mRNA-processing protein 40C-like (PRP40C), TATA-binding protein-associated factor 7 (TBP-associated factor 7), LTR-RTs.

3.2.4 Synteny Analyses

To identify evolutionary structural changes among the *Syzygium* species chromosomes, we performed a synteny analysis on the four newly assembled genomes, *S. aromaticum* and *S. grande*. The alignment of the 11 chromosomes' DNA sequences of the 6 *Syzygium* species revealed a high conservation of the chromosomal organization (Figure 4A).

No large interchromosomal rearrangements were detected between the chromosomes of the six *Syzygium* species. Using a minimum region length of 20 kb, a high percentage of the five species' chromosome lengths were syntenic with *S. aromaticum*, ranging from 68.45% between *S. aromaticum* and *S. jambos* to 73.02% between *S. aromaticum* and *S. aqueum*. Intrachromosomal rearrangements such as inversions, translocations, and duplications between the chromosomes of *S. aromaticum* and those of the other five *Syzygium* species represented 5% of their 11 chromosomes length on average. In terms of number, the most frequent rearrangements observed between *S. aromaticum* and the five other species were duplications and translocations with average numbers of 1348 and 1325, respectively, spanning an average of 0.85% to 1.43% of the 11 chromosome lengths. Inversions were found less frequently for all species but occupied a larger fraction of the genome assemblies' length than duplications and translocations except for *S. syzygioides*. The percentage of assembly lengths comprising inversions between *S. aromaticum* and the five other *Syzygium* species ranged from 0.68% between *S. aromaticum* and *S. syzygioides* to 4.83% between *S. aromaticum* and *S. grande*. Overall, the size of the inversions was relatively small. For instance, 11 inversions were detected, between chromosome 5 of *S. aromaticum* and *S. grande*, representing 17.32% of the chromosome length of *S. grande* (41,797,999 bp) and 1.87% of its 11 chromosomes (387,620,547 bp) (Figure 4B, Supplemental Table 10).

In contrast, the synteny analysis performed between *S. aromaticum* and *E. grandis* revealed 10 intrachromosomal rearrangements on chromosomes 2, 4, 6, 8, 9, and 10 that included large terminal inversions representing up to 40% of the chromosome length of *S. aromaticum*. The other four chromosomes (1, 3, 5, and 7) of the two Myrtaceae species were highly syntenic [20].

To further investigate the chromosomal architecture evolution of the *Syzygium* species and verify that these rearrangements were due to biological events rather than assembly artifacts, we also performed DNA alignment of the chromosome sequences of *E. grandis* with the those of *S. malaccense*, *S. jambos*, *S. aqueum*, and *S. syzygioides*.

Chromosomes 1, 3, 5, and 7 of *E. grandis* and those of the four newly assembled species were also highly syntenic, and we observed the same 10 rearrangements on chromosomes 2, 4, 6, 8, 9, and 10 (Figure 4A, Supplemental Figure 3).

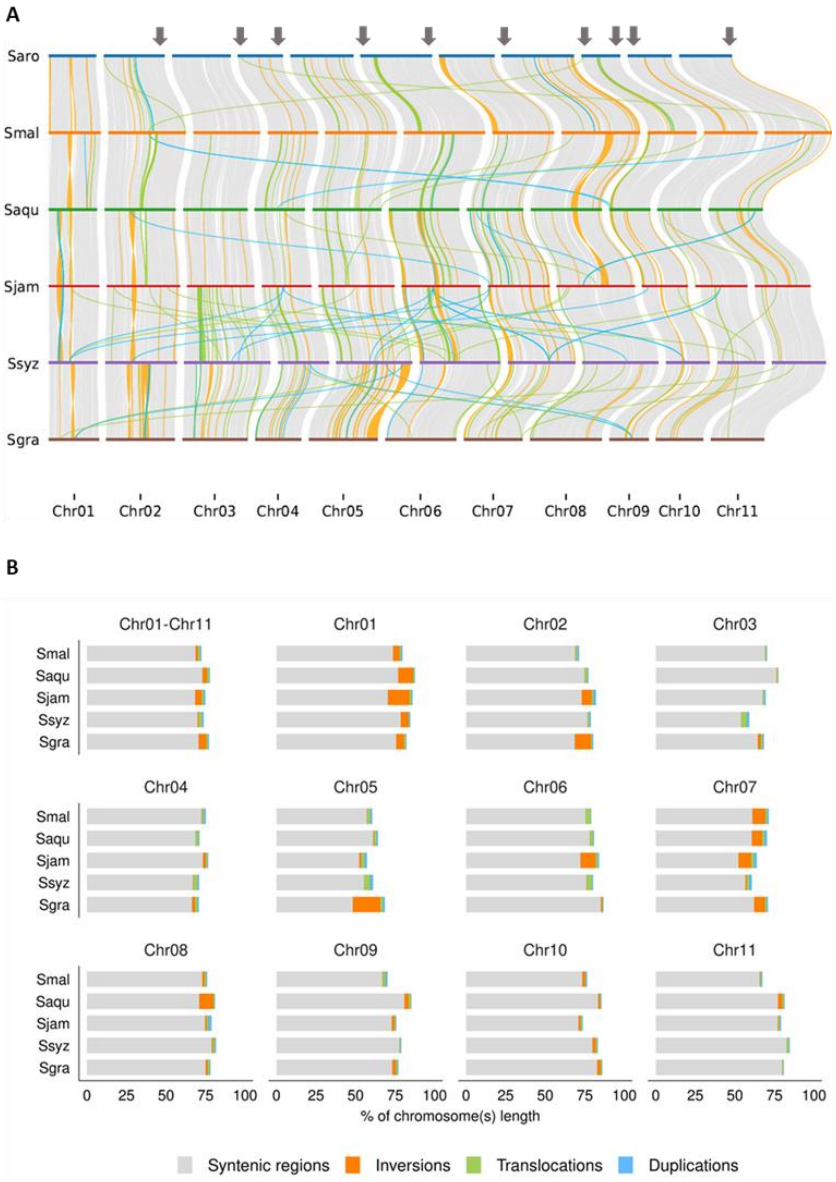


Figure 4 Identification of syntenic and rearranged regions between the 11 chromosomes of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), *S. syzygioides* (Ssyz), and *S. grande* (Sgra).
(A) Representation of the alignment of the chromosomal DNA sequences showing syntenic regions, interchromosomal, and intrachromosomal rearrangements larger than 20 kb (inversions, translocations, and duplications). Grey arrows indicate regions where rearrangements were reported between chromosomes of *E. grandis* and *S. aromaticum*.
(B) Pairwise comparison of the percentage of chromosome length occupied by syntenic regions and rearrangements between the chromosome-scale assembly (Chr01-Chr11) and 11 chromosomes (Chr01 to Chr11) of *S. aromaticum* with those of *S. malaccense*, *S. aqueum*, *S. jambos*, *S. syzygioides*, and *S. grande*.

3.2.5 Gene Orthology

To investigate the phylogenetic relationships among gene sequences of *S. aromaticum*, *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioide*s, the sets of predicted protein sequences from the five species assemblies were analyzed using OrthoFinder [37].

A total of 49,269 hierarchical orthogroups (HOGs) were identified, including 93.7 to 95.2% of each species gene set (Figure 5A). Of these, 18,963 (38.5%) HOGs contained genes from all five species, and 4,928 (10%) were species specific. In more detail, 789 were specific to *S. aromaticum*, 950 were specific to *S. malaccense*, 940 HOGs were specific to *S. aqueum*, 1009 HOGs were specific to *S. jambos*, and 1240 HOGs were specific to *S. syzygioides*. Pairwise, *S. aromaticum* and *S. aqueum* appear to share the lowest number of orthogroups (625). The highest number of shared HOGs inferred between each pair of studied species was found between *S. aqueum* and *S. syzygioides* (1218), followed by *S. aqueum* and *S. malaccense* (1152), and *S. jambos* and *S. malaccense* (1027). The species tree resulting from the analysis of the HOGs divided the *Syzygium* species studied into two groups based on closer relationships: the first group comprising *S. aromaticum* and *S. aqueum* and a second group comprising *S. jambos*, *S. malaccense*, *and S. syzygioides* (Figure 5B).

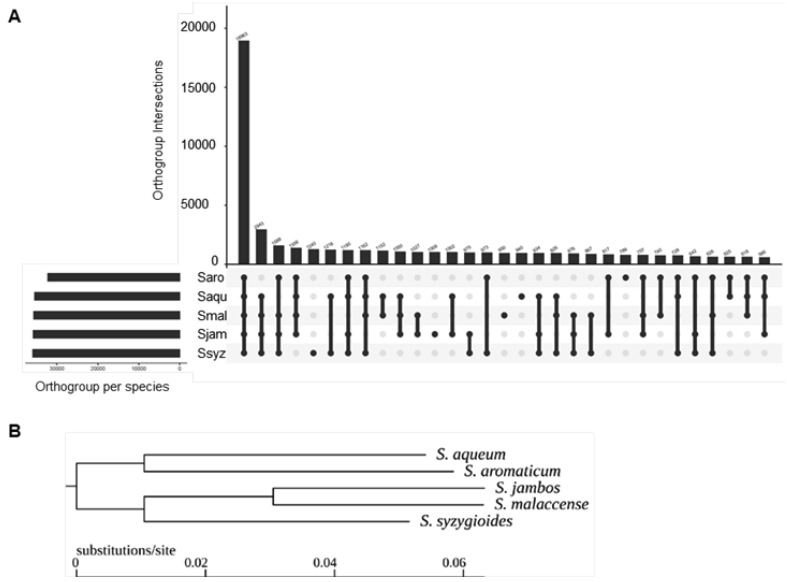


Figure 5 Hierarchical orthogroups (HOGs) inferred by OrthoFinder between S. *aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz). (A) Number of HOGs inferred by OrthoFinder using the set of predicted proteins for the five *Syzygium* species. **(B)** Rooted species tree inferred by OrthoFinder.

3.2.6 Annotation and Comparison of LTR-RTs Gypsy and Copia Repertoires

To clarify the dynamic activity of full-length LTR-RTs belonging to the superfamilies Gypsy and Copia within the *Syzygium* genus, we identified the lineages belonging to each superfamily located on the chromosomes of *S. malaccense* (429 Mbp), *S. aqueum* (387 Mbp), *S. jambos* (416 Mbp), and *S. syzygioides* (425 Mbp) and estimated their insertion time. Then, we compared the repertoires' compositions and repeat element insertion times of the four species with those of *S. aromaticum* (368 Mbp).

We found that *S. malaccense* and *S. aromaticum*, the largest and smallest chromosome-scale assemblies of this study, contained the highest (8427) and lowest number (6167) of LTR-RTs in Gypsy and Copia, respectively (Figure 6A, Supplemental Tables 6–9). In the five *Syzygium* species' chromosomes, we identified a higher number of LTR-RTs for Gypsy than Copia, with a ratio of Gypsy to Copia content ranging from 1.09 for *S. syzygioides* to 1.45 for *S. malaccense*. The Gypsy superfamily comprised a higher proportion of nested elements (17.37% to 24.47%) compared to the Copia superfamily (7.01% to 9.44%), suggesting distinct accumulation and mobile activity of both superfamilies in all five species. Our results revealed little variation in the number of Copia elements on the chromosomes of *S. aqueum* (2705 elements) and *S. aromaticum* (2809), the two smallest chromosome-scale assemblies, and on the chromosomes of *S. syzygioides* (3290), *S. jambos* (3324), and *S. malaccense* (3433). In contrast, we found a notably higher accumulation of Gypsy elements (4994) in the chromosomes of *S. malaccense* compared to the four other species. The ratio of Gypsy content varied from 1.35 when comparing *S. malaccense* with *S. jambos* to 1.49 when comparing *S. malaccense* with *S. aromaticum*. It represented a difference in Gypsy effective length of 19,402,234 bp to 21,766,176bp, respectively. In the five *Syzygium* chromosome-scale assemblies, the most abundant lineage belonged to the Gypsy superfamily, but it varied according to the species. The Gypsy lineage Tekay was the most represented for *S. aromaticum* (1534 elements), *S. jambos* (1674 elements), and *S. syzygioides* (2090 elements). At the same time, for S. *malaccense* and *S. aqueum*, we found a higher abundance of the gypsy lineage Ogre (2382 and 1799 elements, respectively). Among the Gypsy superfamily, the most abundant lineages, Tekay and Ogre, were those with the highest proportion of nested elements (19.10% to 28.55% and 16.69% to 27.92%, respectively) in all five species. For *S. aromaticum*, *S. malaccense*, and *S. syzygioides*, the proportion of nested elements belonging to the Athila lineage was also among the highest identified (Figure 6B, Supplemental Figure 4–5).

Regarding the Copia superfamily, the most represented lineages on the chromosomes of the five *Syzygium* species were Ale (608 to 873 elements), followed by the lineage Tork (456 to 762 elements) for *S. malaccense*, *S. aqueum, S. jambos* and *S. Syzygoides*, and the lineage SIRE (502 elements) for *S. aromaticum*.

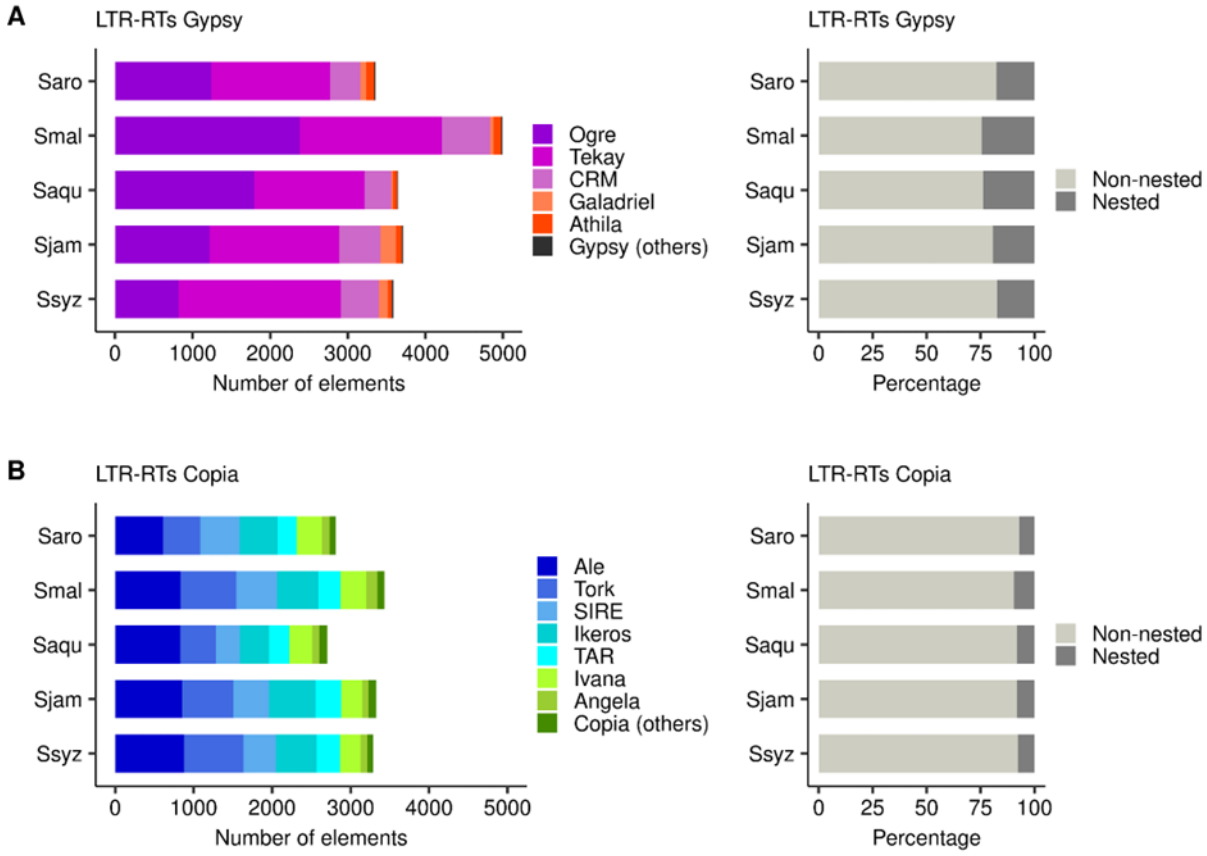


Figure 6 Composition of the full-length LTR-RTs Gypsy and Copia repertoires.
(A) Number of elements belonging to the Gypsy and Copia lineages identified on the 11 chromosomes of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz).
(B) Proportion of nested and non-nested elements. Gypsy (others) group comprises the lineages non-chromo-outgroup, Reina, Retand, tatIII, and elements Gypsy to which no lineages were assigned. Copia (others) group comprises the lineages Alesia, Bianca, Gymco-I, Gymco-IV, Gymco-II, and Osser.

The insertion times of 97.13% of the full-length Gypsy and Copia elements identified in the five *Syzygium* species were estimated (33,861elements). Nearly all elements (97.33%) were inserted in the last 5 million years (32,958 elements) (Figure 7). During this time period, distinct insertion activities of the two superfamilies occurred in the five *Syzygium* species.

Compared to the other four *Syzygium* species, the chromosomes of *S. aromaticum* underwent a more ancient wave of Gypsy insertions (peak at ~2.5 million years ago [Mya]), principally attributed to the Tekay elements, the most abundant lineage in this species (Figure 7A). We also found that a few recent insertions (18.02% of insertions) occurred in *S. aromaticum* chromosomes within the last one million years. In contrast, a recent burst of Gypsy insertions (~0–1 Mya) occurred in four other species chromosomes: most insertions of Gypsy in *S. malaccense* (44.53%), *S. aqueum* (44.43%), *S. jambos* (52.55%), and *S. syzygioides* (36.45%) were less than one million years old. We inferred that the high number of Gypsy LTR-RTs found in *S. malaccense* may be attributable to two successive waves of insertions: a peak of Tekay insertions at ~2 Mya and a more recent peak of Ogre at ~1 Mya.

Similar to what we observed for the Gypsy superfamily, the insertion of Copia elements occurred earlier in *S. aromaticum* compared to the four other species, with fewer recent insertions (Figure 7B). Compared to the Gypsy elements, a smaller proportion of recent Copia insertions (less than one million years old) were detected in *S. aromaticum* (10.66%), *S. malaccense* (24.16%), *S. aqueum* (26.49%), and *S. jambos* (36.90%) suggesting a distinct recent insertion pattern of the two superfamilies in the four species. However, we found a comparable proportion of Gypsy (36.45%) and Copia (32.22%) elements that were less than one million years old in *S. syzygioides*, the species for which we found the lowest ratio of Gypsy to Copia content (1.09).

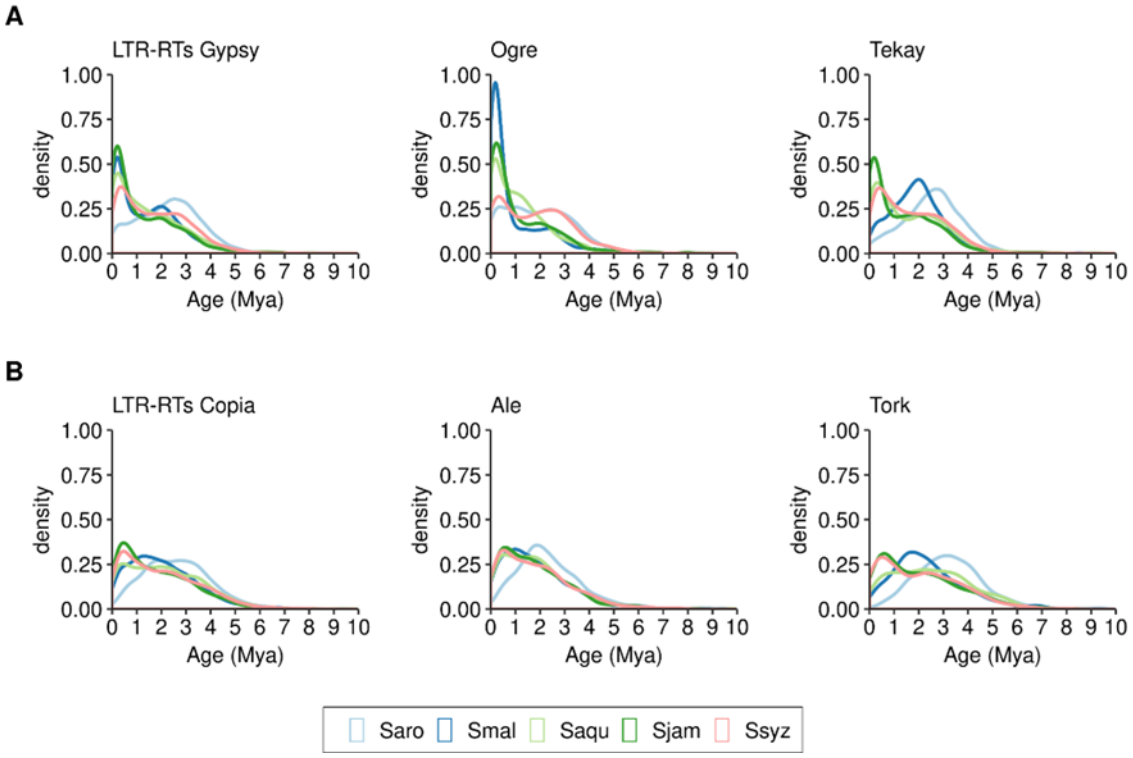


Figure 7 Distribution of insertion times of full-length LTR-RTs of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz).
(A) LTR-RTs Gypsy.
(B) LTR-RTs Copia.

3.3 Discussion

Plant genome size, ploidy level, and heterozygosity rates are challenges for genome assembly and annotation. However, lower sequencing costs and recent advances in long-read sequencing technologies, Hi-C technologies, and bioinformatics tools have facilitated the generation of assemblies with high contiguity up to the chromosome-scale also for non-model plants or non-major plant crops [38, 39]. Newly assembled and annotated genomes from related species can then be used to perform comparative genomics analyses to investigate plant genome evolution and function. Third-generation long-reads from Oxford Nanopore Technologies and Illumina short-reads combined with the Hi-C technology enabled the *de novo* assembly of the chromosome-scale genome for *S. malaccense, S. aqueum, S. jambos,* and *S. syzygioides*. A high level of quality at the base level, contiguity, and completeness was reached for the four newly sequenced genomes. The quality of the newly assembled *Syzygium* species genomes were comparable to that of the *S. aromaticum* genome. The slight differences found between the species assemblies' quality metrics may be linked to the combined impact of the ploidy level and high heterozygosity rates of the four newly sequenced species on the assembly process.

Previous infrageneric comparative genetic mapping analyses revealed high levels of synteny and collinearity among the *Eucalyptus* genus [40, 41]. In addition, genomic synteny analyses conducted between the *de novo* assembly of *E. urophylla* × *E. grandis* (EUC) and 30 *Eucalyptus* species revealed that the genome structure of EUC, *E. grandis,* and *E. globulus* showed the higher collinearity, and the absence of large-scale structural variation. Nevertheless, large structural variations among the different chromosomes of the EUC and other *Eucalyptus* species were also detected [23]. We found that the six *Syzygium* genomes studied were highly syntenic. The intrachromosomal rearrangements (duplications, translocations, and inversions) observed between *S. aromaticum* and the five other *Syzygium* species represent a small percentage (~5% on average) of the 11 chromosomes' length. These intrachromosomal rearrangements could result from contigs that were not well placed because of Hi-C signals that were not strong enough to correctly determine their position and orientation; however, they may also result from the six species' distinct genome evolutions.

Organizational conservation of chromosomes 2, 4, 6, 8, 9, 10, and 11 among the six *Syzygium* species studied constitutes new evidence supporting the 10 intrachromosomal rearrangements previously reported on these chromosomes between *S. aromaticum* and *E. grandis* genomes [20]. These 10 rearrangements were also observed when aligning the DNA sequences of the chromosomes of *E. grandis* with those of *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides.* Among the rearrangements reported between the chromosomes of *S. aromaticum* and *E. grandis*, similar large terminal inversions on chromosomes 4, 9, 10, and 11 were also reported in the two eucalypti *E. grandis* and *C. citriodora* suggesting that these terminal inversions occurred on *E. grandis* chromosomes [27]. Two other large terminal inversions were detected between chromosomes 4 and 9 of *S. aromaticum* and *E. grandis* but not between *C. citriodora* and *E. grandis*. These inversions were also observed when comparing the chromosome sequences of *E. grandis* with those of the four newly assembled genomes, suggesting that these inversions resulted from an evolution of the chromosome organization rather than from sequencing and assembly artifacts. Further comparative genomics analyses will be needed with additional *Syzygium* and Myrtaceous species to determine if these inversions are specific to the *Syzygium* genus or subgenus, for which the crown ages were estimated at 51.2 Mya and 9 Mya, respectively [21].

The analyses of the phylogenetic relationships between gene sequences of *S. aromaticum*, *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides* and comparisons of their full-length LTR-RTs repertoires provided insights into the distinct genome evolution of each species following the divergence of the *Syzygium* subg. *Syzygium* species 9 Mya [21]. The species tree inferred by OrthoFinder indicated that pairwise *S. aromaticum* and *S. aqueum* and *S. malaccense* and *S. jambos* were closely related, which is consistent with the genome-level phylogenetic trees generated by Low et al. [21]. We observed older waves of LTR-RTs Gypsy and Copia insertions in *S. aromaticum* and fewer insertions less than 1 million years old in the *S. aromaticum* chromosomes compared to those of the four other species studied. In plants, the RNA Directed DNA Methylation (RdDM) pathway, a *de novo* DNA methylation mechanism involving small interfering RNA, plays an important role in TE repression [42]. Further detailed analysis such as DNA methylation studies will be valuable to clarify the molecular causes of the recent low insertion number of LTR-RTs elements observed in *S. aromaticum*.

S. aromaticum is cultivated to produce clove bud (the dried, unopened flower bud), essential oil (EO), and oleoresins rich in eugenol [5]. The EO of *S. aromaticum* contains ~72 to 96.6% of eugenol, while the EO of *S. aqueum* has 0.19% eugenol [43, 44]. Eugenol is a phenylpropane with multiple pharmaceutical activities and is considered a promising alternative drug for human health (e.g., cancer and pathogenic microorganism resistance, diabetes, obesity, and autoimmune diseases) [25, 26, 45]. The genome assembly of *S. aromaticum* was exploited to investigate the genetic basis of this important characteristic. The identification of gene families involved in eugenol biosynthesis revealed the presence of multiple copies of genes encoding EGS, which catalyzes the synthesis of eugenol from coniferyl acetate. A cluster of 14 copies was reported on chromosome 10, and additional copies were located on chromosome 11 of *S. aromaticum*. In the genome assembly of the four newly sequenced species, we found fewer gene copies on chromosome 10 (1 to 3 copies) and no copies on chromosome 11 of *S. malaccense, S. aqueum,* and *S. jambos*. The presence of this structural variation suggested that a gene-dosage effect may be associated with the high amount of eugenol. Further studies are needed to elucidate the biological functions of the EGS gene copies in *S. aromaticum* and the four other species (e.g., *in vitro* characterization).

S. malaccense, S. aqueum, and *S. jambos* are grown for their edible fruit. Like *S. aromaticum* and other *Syzygium* species, they are also used in traditional medicine. Research on their numerous pharmaceutical properties has been undertaken (e.g., analgesic, anti-inflammatory, antioxidant, hepatoprotective, antidiabetic, antifungal, antibacterial, antiviral, and anticancer activities) [6, 7]. For instance, *S. jambos* is traditionally used to treat hemorrhages, wounds, and ulcers; *S. malaccense* is used to treat mouth ulcers and diabetes; and *S. aqueum* to treat diabetes and childbirth pain [46]. The chromosome-scale assemblies for these species are new valuable resources for the Myrtaceae family. Combined with other comparative genomics and multi-omics studies, they can be used to further investigate the genomic evolution of the Myrtaceous species and to study the genetic basis of important agronomical traits and biosynthesis of secondary metabolites.

3.4 Materials and Methods

3.4.1 Biological Materials

The *S. malaccense*, *S. aqueum*, *S. jambos,* and *S. syzygioides* genome assemblies were generated from trees growing in the Masoala Hall of the Zurich Zoo in Switzerland. Voucher specimens were deposited in the Zürich herbarium (*S. malaccense* (ZT-00170996), *S. aqueum* (ZT-00170994), *S. jambos* (ZT-00170999), and *S. syzygioides* (ZT-00170991)). Samples collected from the trees were stored at -80°C until nucleic acid extraction.

3.4.2 DNA and RNA Isolation

High-molecular-weight genomic DNA was isolated from frozen leaves using the "ONT high-molecular-weight gDNA extraction from plant leaves" protocol (Oxford Nanopore Technologies, Oxford, UK). Following the extraction, we performed a size selection step using the Circulomics Nanobind Plant Nuclei Big DNA Kit from PacBio (Menlo Park, CA, USA). (NB-900-801-001).

Total RNA from *S. malaccense*, *S. aqueum*, *S. jambos,* and *S. syzygioides* were isolated in triplicate from whole leaves (young and mature), lamina of mature leaves, and stems. Total RNA was also isolated in triplicate from *S. syzygioides*' buds (in the fruiting stage) and *S. jambos*' buds (before and after flowering) and flowers.
Total RNA was extracted from frozen powder using Ambion PureLink Plant RNA Reagent (Ambion by Life Technologies, Carlsbad, CA, USA). The concentration and quality of the total RNA were assessed with an Agilent Bioanalyzer using the Agilent RNA 6000 Nano Kit (Agilent, Santa Clara, CA, USA).

3.4.3 Illumina Sequencing Library Preparation and Sequencing

DNAseq libraries were prepared from total gDNA using the Celero PCR workflow with an enzymatic fragmentation kit from Tecan (Männedorf, Switzerland). DNAseq libraries were loaded on an Illumina S2 flow cell and sequenced on the Illumina Novaseq 6000 instrument (Illumina, San Diego, CA, USA) as 2 x 151 bp paired-end reads.

Hi-C libraries were prepared from 0.2 g of frozen leaves using the Proximo Hi-C Kit following the manufacturer's instructions (Phase Genomics, Seattle, WA, USA) and sequenced on an Illumina HiSeq 4000 instrument (Illumina) as 2 x 151 bp paired-end reads.

mRNA stranded libraries were prepared from 500 ng of total RNA using the Tecan Universal Plus mRNA-Seq library preparation kit with NuQuant® and sequenced on an Illumina HiSeq 4000 instrument as 2 x 151 bp paired-end reads.

Illumina raw reads generated from DNAseq libraries, Hi-C libraries, and RNAseq libraries were cleaned using fastp [47].

3.4.4 ONT Sequencing Library Preparation and Sequencing

Sequencing libraries were generated from high-molecular-weight gDNA and prepared for sequencing on PromethION flow cells (FLO-R0002) by using the ligation sequencing (SQK-LSK109) and flow cell priming (EXP-FLP002) kits (Oxford Nanopore Technologies, Oxford, UK). The base calling was performed by using Guppy 6.1.1 and the super accuracy plant model. Raw ONT reads were cleaned using seqKit [48] to discard reads shorter than 5,000 bp or with quality scores lower than 9.

3.4.5 Genome Profiling

Cleaned Illumina paired-end reads from DNAseq libraries were analyzed by GenomeScope 2.0 and smudgeplot to estimate the genome size, percentage of heterozygosity, and the ploidy level using a k-mer size equal to 21 [34].

3.4.6 *De Novo* Genome Assembly

ONT cleaned reads were corrected with fmlrc2 [49] using cleaned Illumina paired-end short-reads from DNAseq libraries. The corrected ONT reads were then assembled using flye [50] and polished with ntedit [51] using Illumina paired-end short reads from DNAseq libraries. Haplotigs were detected and removed from the polished contigs using purge_dups [52].

Cleaned Illumina read pairs generated from Hi-C libraries were mapped to the genomes to remove reads with low mapping scores, duplicated reads, and paired-end reads. Illumina Hi-C read pairs were mapped to the haplotig-purged contigs using minimap2 [53]. The scaffolding to a chromosome-scale assembly was performed using yahs [54]. Hi-C map files were generated with PretextMap (https://github.com/wtsi-hpag/PretextMap) and used to manually curate the assemblies using PretextView (https://github.com/wtsi-hpag/PretextView).

The curated genome assemblies were mapped to the *S. aromaticum* genome [20] using minimap2, visualized using a custom R script, and the orientation and names of the chromosomes were set in accordance with those of *S. aromaticum*. Chromosome-scale assembly completeness was assessed by using the genome evaluation mode of BUSCO 5.4.4 and the eudicots_odb10 lineage dataset [36]. The QVs of the final assemblies were estimated using yak [55].

3.4.7 Gene Annotation

To create gene models for the four *Syzygium* genomes, the Illumina RNAseq reads used for the clove genome annotation were cleaned, and overlapping paired-reads were merged using fastp [47] before being mapped to the assemblies using minimap2 [53]. Gene models were then created by sample using scallop [56]. *S. aromaticum* and *E. grandis* transcripts were also mapped to the assembly using minimap2 [53], and gene models were created by using bedtools bamtobed [57]. The final gene models were created by merging the RNAseq, *S. aromaticum*, and *E. grandis* gene models using taco [58] and adding coding sequences using Transdecoder (https://github.com/TransDecoder/TransDecoder/wiki) and gffread [59]. The eudicotyledons portion of UniProt filtered to remove proteins with poor descriptions was used to annotate the gene models with their best hit using diamond [60]. Figure 3 was generated using gggenes (https://github.com/wilkox/gggenes).

3.4.8 Repeat Annotation

Annotation of transposable elements was carried out using TE-greedy-nester [61] and TEsorter [62] with REXdb [63]. The insertion age of the predicted transposable elements was then calculated as previously reported [64]. In addition, Red [65], GRF [66], EAHelitron [67], and tantan [68] were used to predict repeats, Miniature Inverted-repeat Transposable Elements (MITEs), helitron, and tandem repeats, respectively.

3.4.9 Synteny Analyses

Synteny between the *Syzygium* species was done by pairwise mapping whole genomes using minimap2 [53], identifying structural variants using syri [69], and plotting syntenic blocks larger than 20 kb using plotsr [70].

3.4.10 Orthologue Analyses

Orthologous genes were clustered into HOGs with OrthoFinder [37] using the set of predicted protein sequences from the five species assemblies.

3.5 Data Availability

Illumina and Oxford Nanopore reads are available from the National Center for Biotechnology Information Short Read Archive (SRA) under accessions: PRJNA962868 *(S. malaccense*), PRJNA962711 (*S. aqueum*), PRJNA962713 (*S. jambos*) and PRJNA962712 (*S. syzygioides*).

Genome assemblies have been deposited at DDBJ/ENA/GenBank under accessions JASUUE000000000, JASUUB000000000, JASUUC000000000, and JASUUD000000000. The versions described in this paper are version JASUUE010000000 (*S. malaccense*), JASUUB010000000 (*S. aqueum*), JASUUC010000000 (*S. jambos*), and JASUUD010000000 (*S. syzygioides*).

Genome annotations are available from Zenodo at https: //doi.org/10.5281/zenodo.7870328 (*S. malaccense*), https://doi.org/10.5281/zenodo.7870326 (*S. aqueum*), https://doi.org/10.5281/zenodo.7870330 (*S. jambos*), and https://doi.org/10.5281/zenodo.7870334 (*S. syzygioides*).

All additional data are available from the corresponding author (Nikolai V. Ivanov) upon reasonable request.

3.6 Funding

Philip Morris International is the sole source of funding and sponsor of this research.

3.7 Author Contributions

S.O. – performed the laboratory work, analyzed data, and wrote the manuscript.
N.S. – performed computational analysis of sequencing data, conceived and supervised the study, and contributed to manuscript writing.
F.K., and **N.V.I.** – conceived and supervised the study and contributed to manuscript writing.
All authors reviewed the final manuscript.

3.8 Acknowledgments

We would like to thank Dr. Leyla Davis, curator of the Masoala Hall in the Zürich Zoo (Switzerland), for authorizing the sampling of the *Syzygium* trees for this project. We would like to also thank Remi Dulize for his technical contributions, and Lindsay Reese and Rebecca Higgins for manuscript revision.

3.9 Conflict of Interest Statement

S.O., **N.S.**, and **N.V.I.** are employees of Philip Morris International. **F.K.** is employee of the University of Neuchâtel.

References

1. Craven, L.A. and Biffin, E., *An infrageneric classification of Syzygium (Myrtaceae).* Blumea-Biodiversity, Evolution and Biogeography of Plants, 2010. 55(1): p. 94-99.

2. Beech, E., Rivers, M., Oldfield, S., and Smith, P., *GlobalTreeSearch: The first complete global database of tree species and country distributions.* Journal of Sustainable Forestry, 2017. 36(5): p. 454-489.

3. POWO, *(2023). "Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet; http://www.plantsoftheworldonline.org/ Retrieved 11 April 2023." https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:327906-2*.

4. Parnell, J.A., Craven, L.A., and Biffin, E., *Matters of scale: dealing with one of the largest genera of angiosperms*, in *Reconstructing the tree of life: taxonomy and systematics of species rich taxa*. 2007, CRC Press LLC.

5. Nurdjannah, N. and Bermawie, N., *Cloves*, in *Handbook of herbs and spices*, K.V. Peter, Editor. 2012, Elsevier. p. 197-215.

6. Nair, K.N., *The Genus Syzygium: Syzygium Cumini and Other Underutilized Species*. 2017: CRC Press.

7. Cock, I.E. and Cheesman, M., *Plants of the genus Syzygium (Myrtaceae): A review on ethnobotany, medicinal properties and phytochemistry.* Bioactive Compounds of Medicinal Plants: Properties and Potential for Human Health, 2018: p. 35-84.

8. Grattapaglia, D., Vaillancourt, R.E., Shepherd, M., Thumma, B.R., Foley, W., Külheim, C., Potts, B.M., and Myburg, A.A., *Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus.* Tree Genetics & Genomes, 2012. 8(3): p. 463-508.

9. Christenhusz, M.J. and Byng, J.W., *The number of known plants species in the world and its annual increase.* Phytotaxa, 2016. 261(3): p. 201–217-201–217.

10. Saber, F.R., Munekata, P.E., Rizwan, K., El-Nashar, H.A., Fahmy, N.M., Aly, S.H., El-Shazly, M., Bouyahya, A., and Lorenzo, J.M., *Family Myrtaceae: The treasure hidden in the complex/diverse composition.* Critical Reviews in Food Science and Nutrition, 2023: p. 1-19.

11. Pellicer, J. and Leitch, I.J., *The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies.* New Phytologist, 2020. 226(2): p. 301-305.

12. Tuler, A.C., Carrijo, T.T., Peixoto, A.L., Garbin, M.L., da Silva Ferreira, M.F., Carvalho, C.R., Spadeto, M.S., and Clarindo, W.R., *Diversification and geographical distribution of Psidium (Myrtaceae) species with distinct ploidy levels.* Trees, 2019. 33(4): p. 1101-1110.

13. Wilson, P.G., *Myrtaceae,* in *Flowering Plants. Eudicots*. 2010, Springer. p. 212-271.

14. Machado, R.M. and Forni-Martins, E.R., *Psidium cattleyanum Sabine (Myrtaceae), a neotropical polyploid complex with wide geographic distribution: insights from cytogenetic and DNA content analysis.* Brazilian Journal of Botany, 2022. 45(3): p. 943-955.

15. Myburg, A.A., et al., *The genome of Eucalyptus grandis.* Nature, 2014. 510(7505): p. 356-362.

16. Healey, A.L., et al., *Pests, diseases, and aridity have shaped the genome of Corymbia citriodora.* Communications biology, 2021. 4(1): p. 1-13.

17. Thrimawithana, A.H., et al., *A whole genome assembly of Leptospermum scoparium (Myrtaceae) for mānuka research.* New Zealand Journal of Crop and Horticultural Science, 2019. 47(4): p. 233-260.

18. Feng, C., Feng, C., Lin, X., Liu, S., Li, Y., and Kang, M., *A chromosome-level genome assembly provides insights into ascorbic acid accumulation and fruit softening in guava (Psidium guajava).* Plant biotechnology journal, 2021. 19(4): p. 717-730.

19. Izuno, A., Wicker, T., Hatakeyama, M., Copetti, D., and Shimizu, K.K., *Updated Genome Assembly and Annotation for Metrosideros polymorpha, an Emerging Model Tree Species of Ecological Divergence.* G3: Genes, Genomes, Genetics, 2019. 9(11): p. 3513-3520.

20. Ouadi, S., Sierro, N., Goepfert, S., Bovet, L., Glauser, G., Vallat, A., Peitsch, M.C., Kessler, F., and Ivanov, N.V., *The clove (Syzygium aromaticum) genome provides insights into the eugenol biosynthesis pathway.* Communications Biology, 2022. 5(1): p. 1-13.

21. Low, Y.W., et al., *Genomic insights into rapid speciation within the world's largest tree genus Syzygium.* Nature communications, 2022. 13(1): p. 1-15.

22. Li, F., et al., *Gap-free genome assembly and comparative analysis reveal the evolution and anthocyanin accumulation mechanism of Rhodomyrtus tomentosa.* Horticulture Research, 2023.

23. Shen, C., Li, L., Ouyang, L., Su, M., and Guo, K., *E. urophylla× E. grandis high-quality genome and comparative genomics provide insights on evolution and diversification of eucalyptus.* BMC genomics, 2023. 24(1): p. 1-10.

24. Zheng, X., et al., *The chromosome-level Melaleuca alternifolia genome provides insights into the molecular mechanisms underlying terpenoids biosynthesis.* Industrial Crops and Products, 2022. 189: p. 115819.

25. Kamatou, G.P., Vermaak, I., and Viljoen, A.M., *Eugenol—from the remote Maluku Islands to the international market place: a review of a remarkable and versatile molecule.* Molecules, 2012. 17(6): p. 6953-6981.

26. Otunola, G.A., *Culinary spices in food and medicine: an overview of Syzygium aromaticum (L.) Merr. and LM Perry [Myrtaceae].* Frontiers in Pharmacology, 2022. 12: p. 3817.

27. Butler, J., Vaillancourt, R., Potts, B., Lee, D., King, G.J., Baten, A., Shepherd, M., and Freeman, J., *Comparative genomics of Eucalyptus and Corymbia reveals low rates of genome structural rearrangement.* BMC Genom., 2017. 18(1): p. 397.

28. Wicker, T., et al., *A unified classification system for eukaryotic transposable elements.* Nature reviews genetics, 2007. 8(12): p. 973-982.

29. Zhou, S.-S., et al., *A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes.* Scientific Data, 2021. 8(1): p. 174.

30. Pedrosa, A., Gitaí, J., Silva, A.E.B., Felix, L.P., and Guerra, M., *Cytogenetics of Angiosperms collected in the State of Pernambuco: V.* Acta Botanica Brasilica, 1999. 13(1): p. 49-60.

31. Oginuma, K., Kato, A., Tobe, H., Mathenge, S., and Juma, F., *Chromosomes of some woody plants in Kenya.* Acta Phytotaxonomica et Geobotanica, 1993. 44(1): p. 53-58.

32. Van Lingen, T., *Syzygium jambos*, in *Edible fruits and nuts*. 1991. p. 296-298.

33. Panggabean, G., *Syzygium aqueum (Burm. f.) Alst., Syzygium malaccense (L.) M. & P, and Syzygium samarangense (Blume) M. & P.* Plant Resources of South-East Asia 2. Edible Fruits and Nuts., 1991: p. 292-294.

34. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C., *GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.* Nature Communications, 2020. 11(1): p. 1-10.

35. Ellestad, P., Pérez-Farrera, M.A., and Buerki, S., *Genomic Insights into Cultivated Mexican Vanilla planifolia Reveal High Levels of Heterozygosity Stemming from Hybridization.* Plants, 2022. 11(16): p. 2090.

36. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. 31(19): p. 3210-3212.

37. Emms, D.M. and Kelly, S., *OrthoFinder: phylogenetic orthology inference for comparative genomics.* Genome biology, 2019. 20(1): p. 1-14.

38. Pucker, B., Irisarri, I., de Vries, J., and Xu, B., *Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions.* Quantitative Plant Biology, 2022. 3.

39. Kyriakidou, M., Tai, H.H., Anglin, N.L., Ellis, D., and Strömvik, M.V., *Current strategies of polyploid plant genome sequence assembly.* Frontiers in plant science, 2018. 9: p. 1660.

40. Hudson, C.J., Kullan, A.R., Freeman, J.S., Faria, D.A., Grattapaglia, D., Kilian, A., Myburg, A.A., Potts, B.M., and Vaillancourt, R.E., *High synteny and colinearity among Eucalyptus genomes revealed by high-density comparative genetic mapping.* Tree Genetics & Genomes, 2012. 8(2): p. 339-352.

41. Li, F., Zhou, C., Weng, Q., Li, M., Yu, X., Guo, Y., Wang, Y., Zhang, X., and Gan, S., *Comparative genomics analyses reveal extensive chromosome colinearity and novel quantitative trait loci in Eucalyptus.* PloS one, 2015. 10(12): p. e0145144.

42. Wambui Mbichi, R., Wang, Q.-F., and Wan, T., *RNA directed DNA methylation and seed plant genome evolution.* Plant Cell Reports, 2020. 39: p. 983-996.

43. Razafimamonjison, G., Jahiel, M., Duclos, T., Ramanoelina, P., Fawbush, F., and Danthu, P., *Bud, leaf and stem essential oil composition of Syzygium aromaticum from Madagascar, Indonesia and Zanzibar*. International Journal of Basic and Applied Sciences, 2014. 3(3): p. 224.

44. Sobeh, M., Braun, M.S., Krstin, S., Youssef, F.S., Ashour, M.L., and Wink, M., *Chemical profiling of the essential oils of Syzygium aqueum, Syzygium samarangense and Eugenia uniflora and their discrimination using chemometric analysis.* Chemistry & biodiversity, 2016. 13(11): p. 1537-1550.

45. Batiha, G.E.-S., Alkazmi, L.M., Wasef, L.G., Beshbishy, A.M., Nadwa, E.H., and Rashwan, E.K., *Syzygium aromaticum L.(Myrtaceae): Traditional uses, bioactive chemical constituents, pharmacological and toxicological activities.* Biomolecules, 2020. 10(2): p. 202.

46. Uddin, A.N., Hossain, F., Reza, A.A., Nasrin, M.S., and Alam, A.K., *Traditional uses, pharmacological activities, and phytochemical constituents of the genus Syzygium: A review*. Food Science & Nutrition, 2022. 10(6): p. 1789-1819.

47. Chen, S., Zhou, Y., Chen, Y., and Gu, J., *fastp: an ultra-fast all-in-one FASTQ preprocessor*. Bioinformatics, 2018. 34(17): p. i884–i890.

48. Shen, W., Le, S., Li, Y., and Hu, F., *SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation.* PloS one, 2016. 11(10): p. e0163962.

49. Mak, Q.C., Wick, R.R., Holt, J.M., and Wang, J.R., *Polishing de novo nanopore assemblies of bacteria and eukaryotes with FMLRC2*. Molecular Biology and Evolution, 2023. 40(3): p. msad048.

50. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A., *Assembly of long, error-prone reads using repeat graphs.* Nature biotechnology, 2019. 37(5): p. 540-546.

51. Warren, R.L., et al., *ntEdit: scalable genome sequence polishing*. Bioinformatics, 2019. 35(21): p. 4430-4432.

52. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R., *Identifying and removing haplotypic duplication in primary genome assemblies.* Bioinformatics, 2020. 36(9): p. 2896-2898.

53. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. 34(18): p. 3094-3100.

54. Zhou, C., McCarthy, S.A., and Durbin, R., *YaHS: yet another Hi-C scaffolding tool.* Bioinformatics, 2022. 39(1).

55. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H., *Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm.* Nature methods, 2021. 18(2): p. 170-175.

56. Shao, M. and Kingsford, C., *Accurate assembly of transcripts through phase-preserving graph decomposition.* Nature biotechnology, 2017. 35(12): p. 1167-1169.

57. Quinlan, A.R. and Hall, I.M., *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. 26(6): p. 841-842.

58. Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M., and Iyer, M.K., *TACO produces robust multisample transcriptome assemblies from RNA-seq.* Nature methods, 2017. 14(1): p. 68-70.

59. Pertea, G. and Pertea, M., *GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved].* F1000Research, 2020. 9(304).

60. Buchfink, B., Xie, C., and Huson, D.H., *Fast and sensitive protein alignment using DIAMOND.* Nature methods, 2015. 12(1): p. 59-60.

61. Lexa, M., Jedlicka, P., Vanat, I., Cervenansky, M., and Kejnovsky, E., *TE-greedy-nester: structure-based detection of LTR retrotransposons and their nesting.* Bioinformatics, 2020. 36(20): p. 4991-4999.

62. Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., and Ma, Y., *TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes.* Horticulture Research, 2022. 9.

63. Neumann, P., Novák, P., Hoštáková, N., and Macas, J., *Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification.* Mobile DNA, 2019. 10: p. 1-17.

64. Marcon, H.S., Domingues, D.S., Silva, J.C., Borges, R.J., Matioli, F.F., de Mattos Fontes, M.R., and Marino, C.L., *Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic.* BMC plant biology, 2015. 15(1): p. 1-16.

65. Girgis, H.Z., *Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale.* BMC bioinformatics, 2015. 16(1): p. 1-19.

66. Shi, J. and Liang, C., *Generic repeat finder: a high-sensitivity tool for genome-wide de novo repeat detection.* Plant physiology, 2019. 180(4): p. 1803-1815.

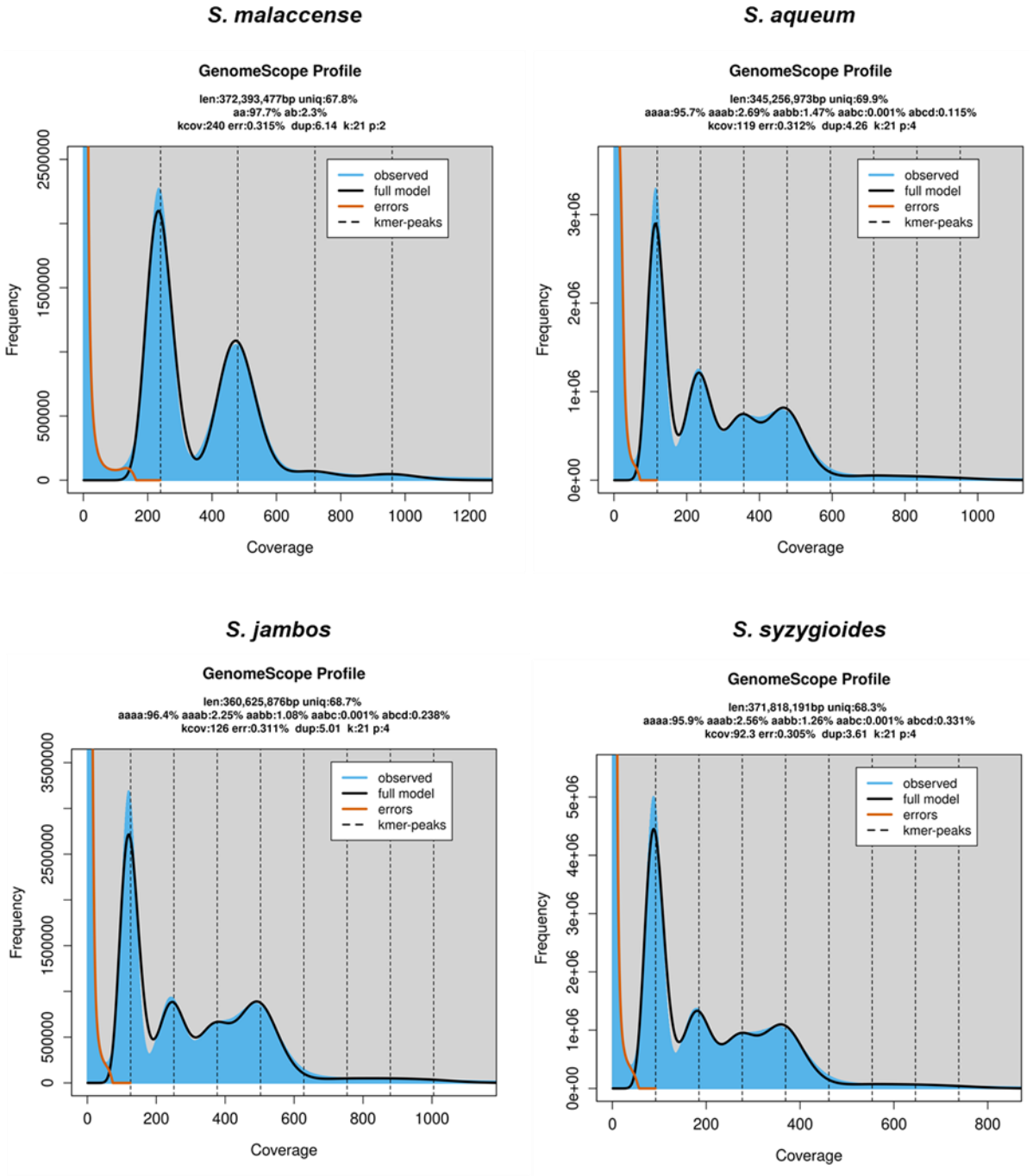
67. Hu, K., Xu, K., Wen, J., Yi, B., Shen, J., Ma, C., Fu, T., Ouyang, Y., and Tu, J., *Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species.* BMC bioinformatics, 2019. 20(1): p. 1-20.

68. Frith, M.C., *A new repeat-masking method enables specific detection of homologous sequences.* Nucleic acids research, 2011. 39(4): p. e23-e23.

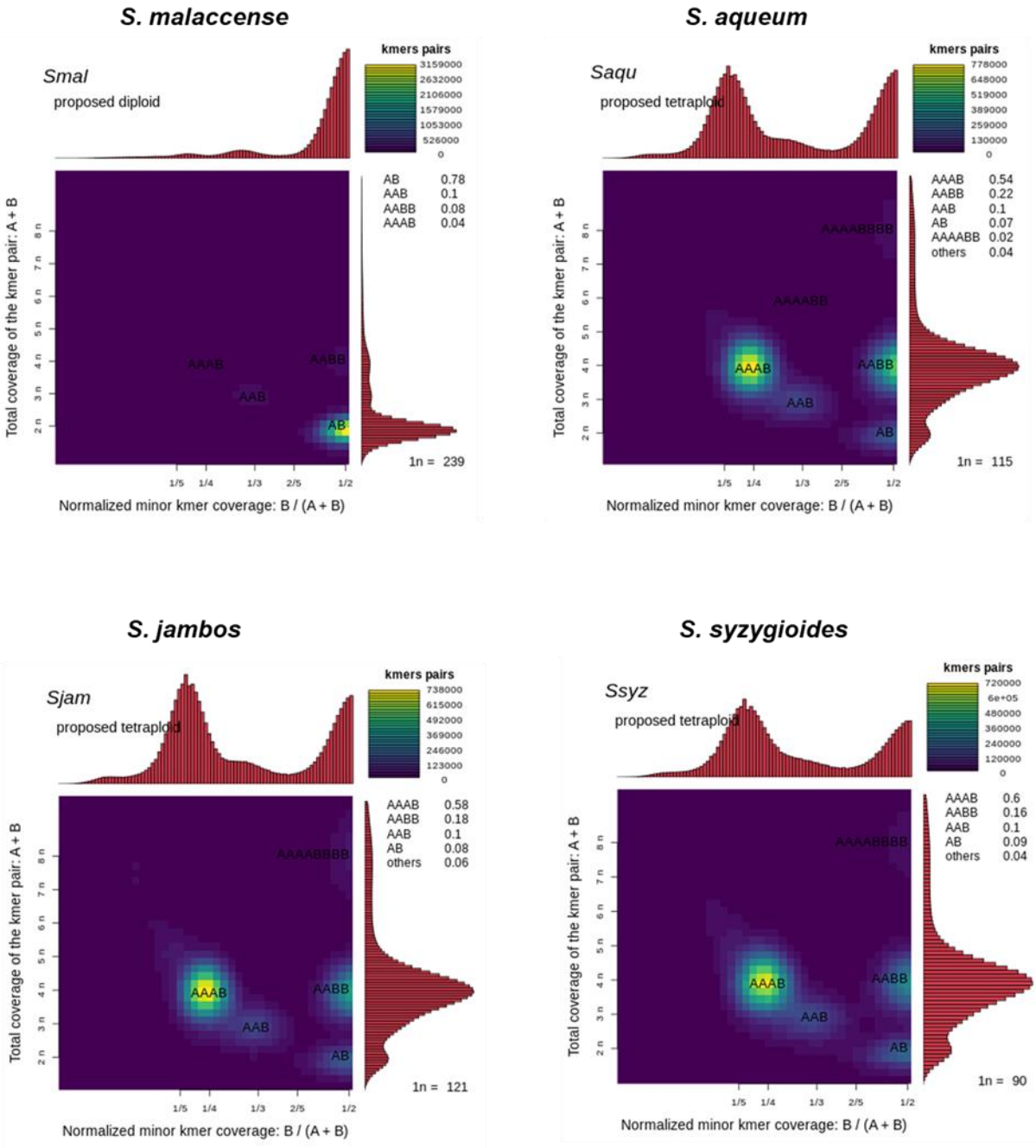
69. Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K., *SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies.* Genome biology, 2019. 20(1): p. 1-13.

70. Goel, M. and Schneeberger, K., *plotsr: visualizing structural similarities and rearrangements between multiple genomes.* Bioinformatics, 2022. 38(10): p. 2922-2926.

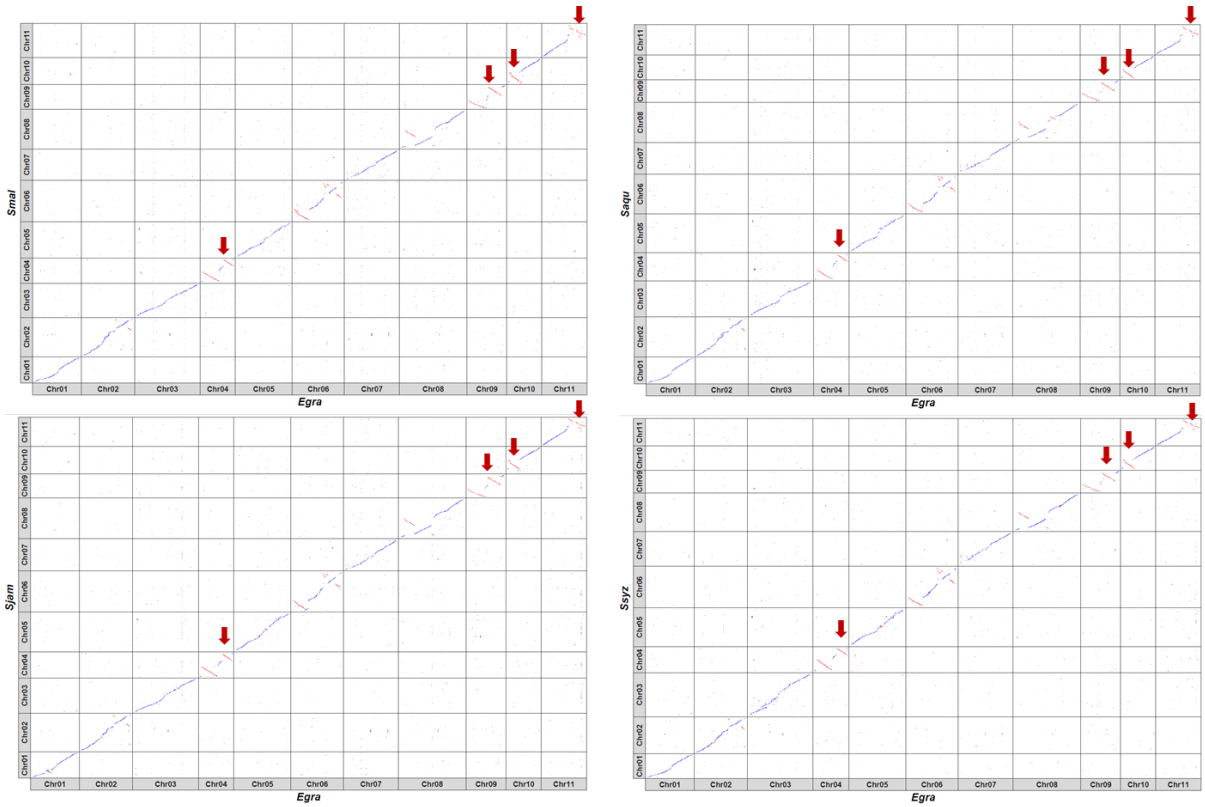
Supplemental Information



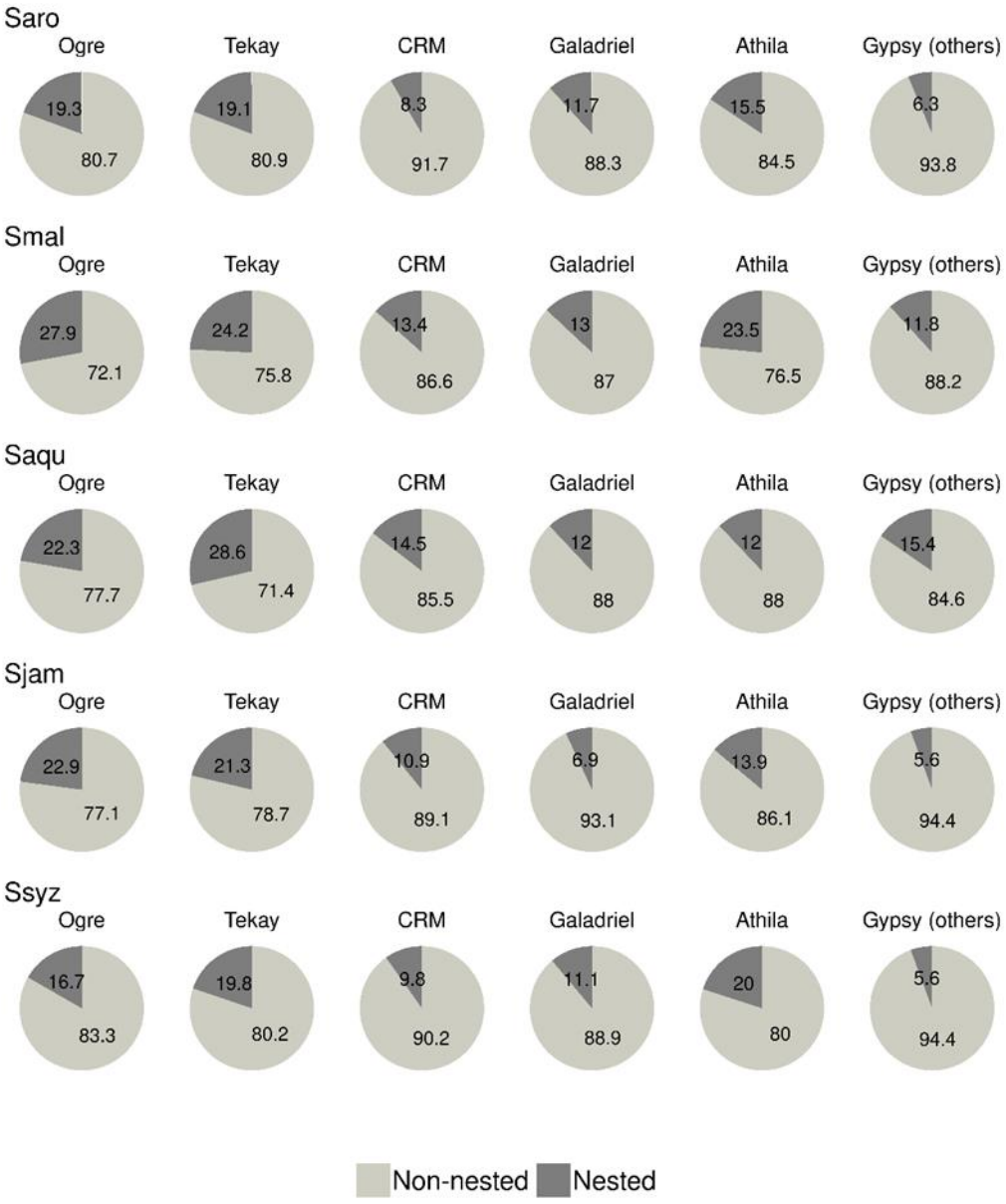
Supplemental Figure 1. GenomeScope 2.0 profile (k-mer = 21 bp) for the *S. malaccense*, *S. jambos*, *S. aqueum*, and *S. syzygioides* genomes.



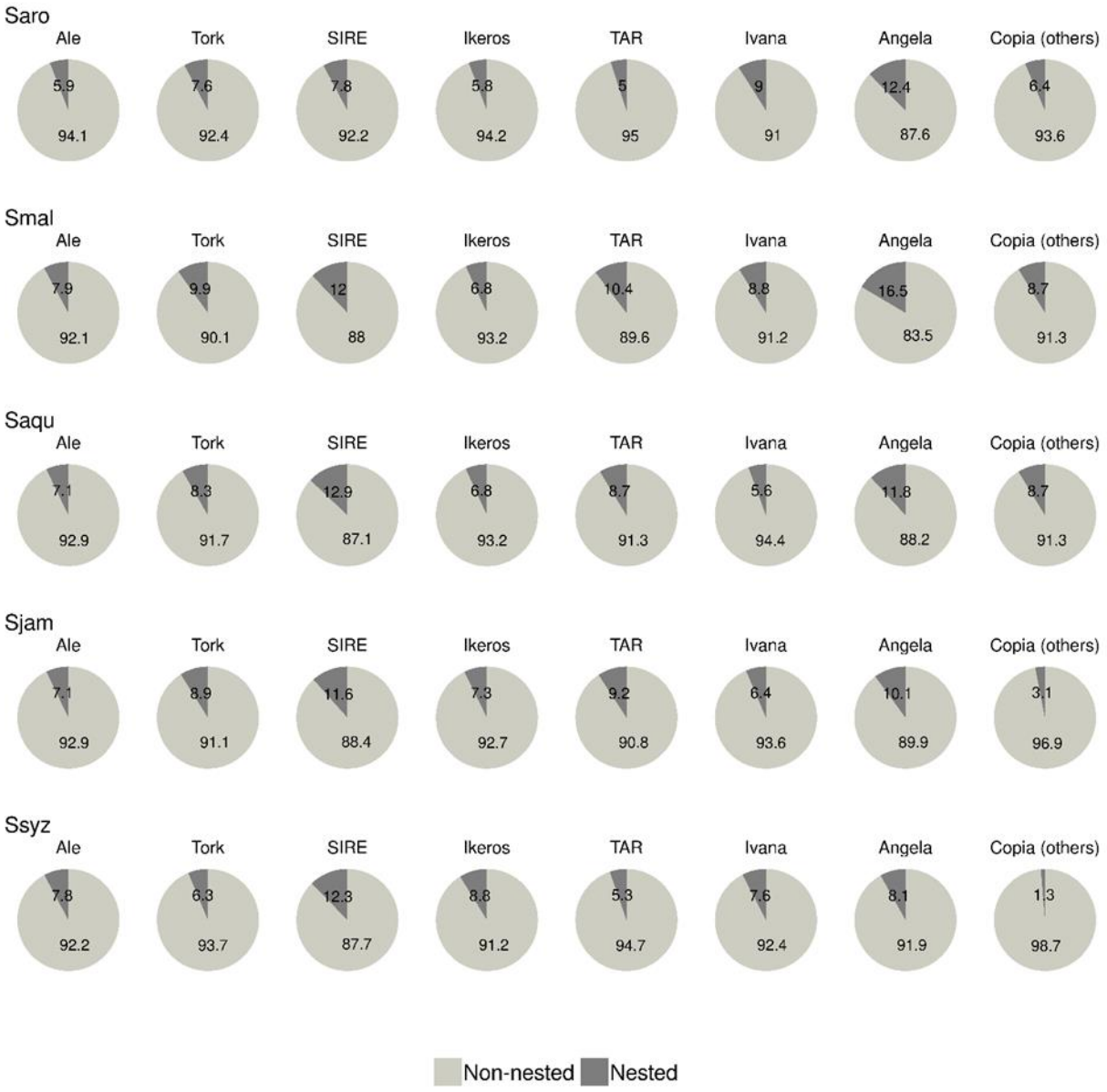
Supplemental Figure 2. Smudgeplot results for the *S. malaccense*, *S. jambos*, *S. aqueum*, and *S. syzygioides* genomes.



Supplemental Figure 3. DNA alignment of the 11 chromosomes of *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz) versus *E. grandis* (Egra) chromosomes. Aligned DNA sequences and intrachromosomal rearrangements are shown in blue and red, respectively. Red arrows indicate regions where similar large inversions were detected between *S. aromaticum* and *E. grandis* and *E. grandis* and *Corymbia citriodora*.



Supplemental Figure 4. The proportion of non-nested and nested elements in the Gypsy lineages of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz). The Gypsy (others) group comprises the lineages non-chromo-outgroup, Reina, Retand, tatIII, and elements Gypsy for which no lineages were assigned to.



Supplemental Figure 5. The proportion of non-nested and nested elements in the Copia lineages of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides*. Copia (others) group comprises the lineages Alesia, Bianca, Gymco-I, Gymco-IV, Gymco-II, Osser.

Supplemental Table 1. Illumina sequencing summary. Number of Illumina short PE reads 12 used as input for the genome profiling (using GenomeScope 2.0 and Smudgeplot), the ONT 13 reads error correction, and scaffolding steps.

	Species name	Number of reads	Total bases (bp)
Genome profiling and ONT[a] reads error correction			
	S. malaccense	1,820,074,174	268,411,454,544
	S. aqueum	1,596,353,896	1,596,353,896
	S. jambos	1,787,425,820	1,787,425,820
	S. syzygioides	1,252,027,700	1,252,027,700
Scaffolding (Hi-C reads)			
	S. malaccense	1,091,335,212	1,091,335,212
	S. aqueum	975,059,968	975,059,968
	S. jambos	1,197,724,004	1,197,724,004
	S. syzygioides	1,373,567,222	1,373,567,222

a ONT = Oxford Nanopore Technologies

Supplemental Table 2. ONT sequencing summary. ONT raw reads were first cleaned to 17 discard reads shorter than 5,000 bp or with quality scores lower than 9.

	Species name	Number of reads	Total bases (bp)
	S. malaccense	8,252,824	8,252,824
	S. aqueum	9,493,220	9,493,220
	S. jambos	9,648,395	9,648,395
	S. syzygioides	6,944,019	6,944,019

Supplemental Table 3. Genome assemblies' quality assessment using BUSCO version 5.4.4 - dataset: eudicots_odb10 (n = 2326) in mode genome, transcriptome, and Protein. Complete BUSCOs (C), Complete and single copy BUSCOs (S), Complete and duplicated BUSCOs (D), Fragmented BUSCOs (F), Missing BUSCOs (M).

	% C	% S	% D	% F	% M
BUSCO genome (Contigs before haplotig removal)					
S. malaccense	98.7	5.1	93.6	0.5	0.8
S. aqueum	98.7	4.3	94.4	0.6	0.7
S. jambos	98.7	3.4	95.3	0.6	0.7
S. syzygioides	98.9	1.8	97.1	0.5	0.6
BUSCO genome (Contigs after haplotig removal)					
S. malaccense	98.1	94.8	3.3	0.7	1.2
S. aqueum	98	93.2	4.8	0.7	1.3
S. jambos	97.9	93.9	4	0.7	1.4
S. syzygioides	98	91.9	6.1	0.7	1.3
BUSCO genome (Final assembly)					
S. aromaticum	98.1	95.9	2.2	0.9	1
S. malaccense	98.1	94.8	3.3	0.6	1.3
S. aqueum	98	93.5	4.5	0.7	1.3
S. jambos	98	94.4	3.6	0.7	1.3
S. syzygioides	98.2	92.7	5.5	0.6	1.2
BUSCO transcriptome					
S. aromaticum	95	47.9	47.1	1	4
S. malaccense	93.5	55.8	37.7	1.9	4.6
S. aqueum	91.9	52.8	39.1	2.8	5.3
S. jambos	92.6	55.5	37.1	2.3	5.1
S. syzygioides	92.7	52.6	40.1	2.8	4.5
BUSCO protein					
S. aromaticum	93.7	59.9	33.8	1.5	4.8
S. malaccense	90.9	63.6	27.3	2.6	6.5
S. aqueum	89.3	61.8	27.5	3.4	7.3
S. jambos	90.1	63.1	27	3.3	6.6
S. syzygioides	89.8	60.5	29.3	3.5	6.7

Supplemental Table 4. Chromosome length (bp) of *S. aromaticum*, *S. malaccense*, *S. aqueum*, *S. jambos*, *S. syzygioides*, and *S. grande*.

Chromosome	*S. aromaticum*	*S. malaccense*	*S. aqueum*	*S. jambos*	*S. syzygioides*	*S. grande*
Chr01	28,385,693	3,119,1675	28,960,203	30,275,079	29,286,511	30,130,455
Chr02	36,857,792	46,997,391	43,113,907	44,223,308	43,522,896	41,423,945
Chr03	35,418,074	40,588,123	38,810,027	40,716,907	52,151,277	39,560,356
Chr04	27,127,432	30,283,374	30,431,105	29,830,967	30,781,539	27,788,763
Chr05	38,056,112	43,125,459	41,576,880	45,836,902	46,132,905	41,797,999
Chr06	42,993,409	49,125,077	42,880,593	47,627,361	48,951,463	43,109,847
Chr07	33,360,290	37,421,253	33,918,699	37,171,488	40,604,280	35,655,372
Chr08	43,763,418	47,424,843	43,361,234	47,089,386	45,885,160	43,366,331
Chr09	23,380,519	29,684,569	24,382,885	27,657,921	26,538,911	23,649,235
Chr10	26,565,453	32,139,982	26,467,591	31,499,097	28,595,216	28,611,675
Chr11	31,862,996	41,026,473	32,633,549	33,694,566	32,377,069	32,526,569
Total	367,771,188	429,008,219	386,536,673	415,622,982	424,827,227	387,620,547

Supplemental Table 5. List of genes encoding for putative eugenol synthase (EGS) and their position on the chromosome (Chr) of *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides*.

Species	Gene	Protein	Annotation	Chr	Start (bp)	End (bp)
S. malaccense	Smal007380	EGS1	eugenol synthase 1-like	Chr10	21,561,954	21,566,251
S. aqueum	Saqu028928	EGS1	eugenol synthase 1-like	Chr10	17,568,090	17,569,992
S. aqueum	Saqu028929	EGS2	eugenol synthase 1-like	Chr10	17,600,976	17,604,578
S. jambos	Sjam030623	EGS1	eugenol synthase 1-like	Chr10	22,078,314	22,082,922
S. jambos	Sjam030624	EGS2	eugenol synthase 1-like	Chr10	22,085,199	22,088,927
S. jambos	Sjam030625	EGS3	eugenol synthase 1-like	Chr10	22,090,133	22,092,406
S. syzygioides	Ssyz031161	EGS1	eugenol synthase 1-like	Chr10	19,161,196	19,165,512
S. syzygioides	Ssyz031162	EGS2	eugenol synthase 1-like	Chr10	19,167,723	19,170,887
S. syzygioides	Ssyz002596	EGS3	eugenol synthase 1-like	Chr11	30,519,613	30,522,208

Supplemental Table 6. Summary of repeats elements identified in the genome assembly of *S. malaccense*.

Class	Superfamily	Lineage	Count	Length (bp)	Percentage of assembly length
Retrotransposons			8515	96,086,564	22.35%
	LTR[a] Copia		3435	31,769,467	7.39%
		Ale	835	7,018,745	1.63%
		Alesia	39	373,075	0.09%
		Angela	139	1,273,090	0.30%
		Bianca	34	264,888	0.06%
		Gymco-II	1	10,679	0.00%
		Ikeros	529	4,877,541	1.13%
		Ivana	329	2,369,071	0.55%
		SIRE	523	5,747,540	1.34%
		TAR	280	2,466,390	0.57%
		Tork	708	7,002,611	1.63%
		Unknown	18	365,837	0.09%
	LTR Gypsy		4995	62,668,430	14.58%
		Athila	98	1,025,135	0.24%
		CRM	619	4,537,097	1.06%
		Galadriel	46	459,408	0.11%
		non-chromo-outgroup	1	6945	0.002%
		Ogre	2382	37,168,658	8.65%
		Reina	8	58,597	0.01%
		Tekay	1833	19,326,726	4.50%
		Unknown	8	85,864	0.02%
	LTR Unknown		85	1,648,667	0.38%
DNA transposons	MITE[b]		5710	594,020	0.14%
	Helitron		1384	39,200	0.01%
Tandem repeats			164,955	4,058,239	0.94%
Repeats elements (unclassified)			149,758	84,138,834	19.57%
Total			3303,22	184,916,857	43.02%

a LTR = Long Terminal Repeat
b MITE = Miniature Inverted-repeat Transposable Elements

Supplemental Table 7. Summary of repeats elements identified in the genome assembly of *S. aqueum*

Class	Superfamily	Lineage	Count	Length (bp)	Percentage of assembly length
Retrotransposons			6541	74,914,968	19.12%
	LTR[a] Copia		2743	25,148,956	6.42%
		Ale	839	7,286,400	1.86%
		Alesia	38	264,985	0.07%
		Angela	97	850,005	0.22%
		Bianca	54	400,622	0.10%
		Gymco-I	1	9997	0.00%
		Ikeros	374	3,550,808	0.91%
		Ivana	290	1,997,761	0.51%
		SIRE	305	3,478,327	0.89%
		TAR	272	2,452,073	0.63%
		Tork	460	4,666,215	1.19%
		Unknown	13	191,763	0.05%
	LTR Gypsy		3710	48,141,612	12.28%
		Athila	52	577,550	0.15%
		CRM	350	2,961,734	0.76%
		Galadriel	25	305,966	0.08%
		non-chromo-outgroup	1	8977	0.002%
		Ogre	1829	27,773,931	7.09%
		Reina	10	81,745	0.02%
		Tekay	1441	16,386,086	4.18%
		Unknown	2	45,623	0.01%
	LTR Unknown		88	1,624,400	0.41%
DNA transposons	MITE[b]		4,034	580,151	0.15%
	Helitron		1198	33,858	0.01%
Tandem repeats			147,952	3,621,605	0.92%
Repeats elements (unclassified)			139,832	82,869,853	21.15%
Total			299,557	162,020,435	41.34%

a LTR = Long Terminal Repeat
b MITE = Miniature Inverted-repeat Transposable Elements

Supplemental Table 8. Summary of repeats elements identified in the genome assembly of *S. jambos.*

Class	Superfamily	Lineage	Count	Length (bp)	Percentage of assembly length
Retrotransposons			7393	77,407,268	18.16%
	LTR[a] Copia		3402	30,040,740	7.05%
		Ale	876	7,176,078	1.68%
		Alesia	42	378,260	0.09%
		Angela	82	822,474	0.19%
		Bianca	45	330,379	0.08%
		Ikeros	607	5,527,257	1.30%
		Ivana	273	1,910,486	0.45%
		SIRE	468	5,013,322	1.18%
		TAR	336	2,815,655	0.66%
		Tork	661	5,871,859	1.38%
		Unknown	12	194,970	0.05%
	LTR Gypsy		3858	45,090,450	10.58%
		Athila	73	803,315	0.19%
		CRM	544	3,657,979	0.86%
		Galadriel	209	1,213,756	0.28%
		non-chromo-outgroup	4	72,530	0.02%
		Ogre	1258	20,157,864	4.73%
		Reina	9	55,716	0.01%
		Retand	1	10,992	0.00%
		TatIII	1	3450	0.00%
		Tekay	1754	19,067,733	4.47%
		Unknown	5	47,115	0.01%
	LTR Unknown		133	2,276,078	0.53%
DNA transposons	MITE[b]		10,323	1,131,284	0.27%
	Helitron		1652	46,555	0.01%
Tandem repeats			165,419	4,055,420	0.95%
Repeats elements (unclassified)			160,407	97,923,066	22.98%
Total			345,194	180,563,593	42.37%

a LTR = Long Terminal Repeat

b MITE = Miniature Inverted-repeat Transposable Elements

Supplemental Table 9. Summary of repeats elements identified in the genome assembly of *S. syzygioides.*

Class	Superfamily	Lineage	Count	Length (bp)	Percentage of assembly length
Retrotransposons			7144	73,171,928	16.97%
	LTR[a] Copia		3348	30,532,813	7.08%
		Ale	885	7,387,426	1.71%
		Alesia	31	229,884	0.05%
		Angela	88	788,496	0.18%
		Bianca	29	267,998	0.06%
		Ikeros	520	5,028,520	1.17%
		Ivana	267	1,975,430	0.46%
		Osser	1	11,413	0.00%
		SIRE	429	4,953,376	1.15%
		TAR	307	2,767,941	0.64%
		Tork	774	6,746,234	1.56%
		Unknown	17	376,095	0.09%
	LTR Gypsy		3659	40,369,474	9.36%
		Athila	52	535,287	0.12%
		CRM	497	3,683,608	0.85%
		Galadriel	117	693,929	0.16%
		non-chromo-outgroup	3	23,449	0.01%
		Ogre	844	13,492,111	3.13%
		Reina	13	93,044	0.02%
		Tekay	2128	21,795,136	5.06%
		Unknown	5	52,910	0.01%
	LTR Unknown		137	2,269,641	0.53%
DNA transposons	MITE[b]		6525	667,396	0.15%
	Helitron		1579	44,821	0.01%
Tandem repeats			157,919	3,834,074	0.89%
Repeats elements (unclassified)			177,217	106,284,882	24.66%
Total			350,384	184,003,101	42.68%

a LTR = Long Terminal Repeat

b MITE = Miniature Inverted-repeat Transposable Elements

Supplemental Table 10. Number and length (bp) of genomic variations identified between the 11 chromosomes of *S. aromaticum* and *S. malaccense*, *S. aqueum*, *S. jambos*, *S. Syzygioides*, and *S. grande.*

Species (query)	Variation type	Count	Length (bp) on query 11 chromosomes	Percentage of query 11 chromosomes length
S. malaccense	Syntenic regions	1482	294,735,586	68.70%
	Inversions	71	6,169,822	1.44%
	Translocations	1439	5,842,278	1.36%
	Duplications	1460	3,703,684	0.86%
S. aqueum	Syntenic regions	1096	282,260,585	73.02%
	Inversions	53	10,883,986	2.82%
	Translocations	965	5,298,328	1.37%
	Duplications	922	2,222,360	0.57%
S. jambos	Syntenic regions	1493	284,500,441	68.45%
	Inversions	76	17,027,630	4.10%
	Translocations	1402	5,121,967	1.23%
	Duplications	1658	4,730,513	1.14%
S. syzygioides	Syntenic regions	1543	297,194,038	69.96%
	Inversions	48	2,906,286	0.68%
	Translocations	1538	9,027,953	2.13%
	Duplications	1680	4,265,382	1.00%
S. grande	Syntenic regions	1358	273,729,390	70.62%
	Inversions	89	18,723,392	4.83%
	Translocations	1279	4,042,562	1.04%
	Duplications	1019	2,649,137	0.68%

General conclusions

Despite the importance of clove as a spice crop and source of eugenol, scarce genomic and genetic information was available for clove research. The availability of the clove genome therefore opens new perspective for future fundamental and applied research on clove. It could be especially useful for the characterisation of genetic resources and the development of molecular breeding tools.

The characterization of clove genome, carried out using omics-based approaches, provided new insights into clove genome evolution and the genetic basis of eugenol and a lot more research will be necessary to elucidate the molecular mechanism involved in the high eugenol accumulation in clove.

In addition, high-quality genome assemblies were also generated for four *Syzygium* species, *S. malaccense*, *S. jambos*, *S. aqueum*, and *S. syzygioides*, and provided new additional evolutionary genomics insights. Comparative genomics analyses showed conserved organization of the 11 chromosomes between *S. aromaticum*, *S. grande* and the four *Syzygium* species sequenced and the occurrence of few intrachromosomal rearrangements since the five species divergence about 9 Mya. Interestingly, we found that the clove genome comprises a higher number of genes (15) encoding for putative eugenol synthase compared to the other *Syzygium* species studied (1 to 3) suggesting that this copy number variation is involved in the high content of eugenol of *S. aromaticum*. Nevertheless, the functional annotation of putative eugenol synthase and its role in the biosynthesis of eugenol was not validated by *in vitro* characterisation during my PhD work. The characterization of the selected genes encoding for putative alcohol acyltransferase (AAT), eugenol synthase (EGS), and hydrolase could bring important insights into the biosynthesis of eugenol in clove, especially on the substrate specificity of AAT (coniferyl alcohol and/or eugenol) and hydrolase (eugenol acetate) and the product that they can generate.

To further investigate the eugenol accumulation in clove, additional multi-omics studies on different organs (stems, peduncles, and roots) and specific anatomical structure (e.g., secretory cavities) will be valuable. Comparative genomic analyses between Myrtaceae species producing eugenol and clove, and between forest clove and cultivated clove populations would also be beneficial to better understand biosynthesis of eugenol (e.g., how many copies of EGS genes are present in the genome of forest clove) and the clove genome evolution (e.g., impact of the domestication).

These studies would require the sequencing/resequencing of the genome of relevant species or populations, however, the omic datasets generated for clove can also be further analysed to identify genes involved in the regulation of eugenol biosynthesis (e.g., MYB, ODORANT transcription factors) as well as other gene families encoding for other economically important traits related to the clove yield and to the biosynthesis of other secondary compounds important for the clove products value (e.g., terpene synthase gene families).

To conclude, the genome of *Syzygium aromaticum* and four additional *Syzygium* species were sequenced and assembled *de novo* at the chromosome level. This new set of *Syzygium* genomes is a valuable contribution to genomics resources for the *Syzygium* genus, the world's most species-rich tree genus, and overall, for the Myrtaceae family that comprises several economically and ecologically important species as well as many medicinal species. Combined with other comparative genomics and/or multi-omics studies, they can be used to further investigate the genomic evolution of the Myrtaceous species and to study the genetic basis of important agronomical traits and biosynthesis of secondary metabolites.

Acknowledgments

Firstly, I would like to thank my thesis directors Professor Felix Kessler and Dr. Nikolai Ivanov for giving me the opportunity to do this research work. I am deeply grateful for their insightful guidance, and continuous support throughout my thesis work.

I am very grateful to Dr. Nicolas Sierro for his tremendous help and his scientific contributions to my thesis. Without our regular meetings and his expertise in bioinformatics and plant genomes research, I could not have undertaken this project.

I would like to thank Dr. Gaetan Glauser (Neuchâtel Platform of Analytical Chemistry) and Professor Lukas Mueller (Boyce Thompson Institute) for accepting to be members of my thesis committee, as well as Dr. Simon Goepfert, Dr. Lucien Bovet, and Professor Manuel C. Peitsch for their collaboration. Thank you for sharing your knowledge and useful feedbacks.

I would like to acknowledge everyone from the genomics and transcriptomics lab at PMI and members of the LPV team at the university of Neuchatel, for their help and support. I wish to also acknowledge Dr. Leyla Davis for authorizing and facilitating the sampling of the *Syzygium* trees at the Masoala Hall in the Zürich Zoo, Dr. Armelle Vallat from the from Neuchâtel Platform of Analytical Chemistry for her help with the metabolomics experiments as well as Dr. Lindsay Reese and Dr. Rebecca Higgins from PMI for the manuscript and thesis revisions. I am also very thankful to Philip Morris Products S.A. for funding this research project.

Finally, I wish to thank my family and all my friends who were always there for encouraging me. A special thanks to my partner Dominik for his unconditional support and patience throughout all the steps of my PhD thesis

Annex

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 October 2023
DOI 10.3389/fpls.2023.1248780

OPEN ACCESS

EDITED BY
Agnieszka Zmienko,
Polish Academy of Sciences, Poland

REVIEWED BY
Pawel Wojciechowski,
Poznań University of Technology, Poland
Xiaojun Nie,
Northwest A&F University, China
Jian-Feng Mao,
Beijing Forestry University, China

*CORRESPONDENCE
Nikolai V. Ivanov
nikolai.ivanov@unine.ch

RECEIVED 27 June 2023
ACCEPTED 28 August 2023
PUBLISHED 06 October 2023

CITATION
Ouadi S, Sierro N, Kessler F and Ivanov NV (2023) Chromosome-scale assemblies of *S. malaccense, S. aqueum, S. jambos*, and *S. syzygioides* provide insights into the evolution of *Syzygium* genomes. *Front. Plant Sci.* 14:1248780. doi: 10.3389/fpls.2023.1248780

Chromosome-scale assemblies of *S. malaccense, S. aqueum, S. jambos*, and *S. syzygioides* provide insights into the evolution of *Syzygium* genomes

Sonia Ouadi[1,2], Nicolas Sierro[2], Felix Kessler[1] and Nikolai V. Ivanov[1,2]*

[1]Faculty of Sciences, Laboratory of Plant Physiology, University of Neuchâtel, Neuchâtel, Switzerland, [2]Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

Syzygium is a large and diverse tree genus in the Myrtaceae family. Genome assemblies for clove (*Syzygium aromaticum*, 370 Mb) and sea apple (*Syzygium grande*, 405 Mb) provided the first insights into the genomic features and evolution of the *Syzygium* genus. Here, we present additional *de novo* chromosome-scale genome assemblies for *Syzygium malaccense, Syzygium aqueum, Syzygium jambos*, and *Syzygium syzygioides*. Genome profiling analyses show that *S. malaccense*, like *S. aromaticum* and *S. grande*, is diploid (2n = 2x = 22), while the *S. aqueum*, *S. jambos*, and *S. syzygioides* specimens are autotetraploid (2n = 4x = 44). The genome assemblies of *S. malaccense* (430 Mb), *S. aqueum* (392 Mb), *S. jambos* (426 Mb), and *S. syzygioides* (431 Mb) are highly complete (BUSCO scores of 98%). Comparative genomics analyses showed conserved organization of the 11 chromosomes with *S. aromaticum* and *S. grande*, and revealed species-specific evolutionary dynamics of the long terminal repeat retrotransposon elements belonging to the Gypsy and Copia lineages. This set of *Syzygium* genomes is a valuable resource for future structural and functional comparative genomic studies on Myrtaceae species.

KEYWORDS

***Syzygium*, Myrtaceae, *de novo* assembly, comparative genomics, synteny, long terminal repeat retrotransposons**

1 Introduction

Syzygium is the largest tree genus with about 1,200 species naturally occurring from the Old World tropics and subtropics to the Pacific (POWO, 2023; Craven and Biffin, 2010; Beech et al., 2017). In addition to their ecological importance, the genus includes several species grown for their edible fruit, medicinal properties, timber, and for the horticulture industry (e.g., *S. malaccense, S. aqueum, S. jambos*, and *S. cumini*), the most important

economically being the clove tree (*S. aromaticum)* (Parnell et al., 2007; Nurdjannah and Bermawie, 2012; Nair, 2017; Cock and Cheesman, 2018).

The *Syzygium* genus belongs to the Myrtaceae family—the eighth largest family of flowering plants—and includes economically important species such as eucalyptus, myrtle, and guava (Grattapaglia et al., 2012; Christenhusz and Byng, 2016; Saber et al., 2023). Although the majority of species of the Myrtaceae family are diploids (2n = 22) with small to intermediate genome sizes (234–1785 Mb), occasional polyploids derived from the most conserved chromosome number x = 11 were also reported (e.g., within the *Eugenia, Syzygium*, and *Psidium* genera) (Wilson, 2010; Grattapaglia et al., 2012; Tuler et al., 2019; Pellicer and Leitch, 2020; Machado and Forni-Martins, 2022). The *Eucalyptus grandis* genome was released in 2014 as the first reference genome for the Myrtales order and the Myrtaceae family (Myburg et al., 2014). New chromosome-scale assemblies were subsequently published, enabling comparative genomics analyses within the family. Published chromosome-scale genome assemblies for the Myrtaceae currently represent major tribes of the family: Eucalypteae (*Eucalyptus grandis, Corymbia citriodora*, *Eucalyptus urophylla* × *Eucalyptus grandis*), Leptospermeae (*Leptospermum scoparium*), Myrteae (*Psidium guajava*, *Rhodomyrtus tomentosa*), Metrosidereae (*Metrosideros polymorpha*), Melaleuceae (*Melaleuca alternifolia*), and Syzygieae (*S. aromaticum, Syzygium grande*). These assemblies were generated from diploid specimens, and their size ranged from 297 Mb to 690 Mb (Myburg et al., 2014; Izuno et al., 2019; Thrimawithana et al., 2019; Feng et al., 2021; Healey et al., 2021; Low et al., 2022; Ouadi et al., 2022; Zheng et al., 2022; Li et al., 2023; Shen et al., 2023).

The clove (*S. aromaticum* (L.) Merr. & L.M. Perry) and sea apple (*S. grande*) genomes were constructed using a combination of Oxford Nanopore Technologies long-reads and Illumina short-reads and anchored on 11 chromosomes using Hi-C technologies (Low et al., 2022; Ouadi et al., 2022). The sea apple genome assembly (405 Mb), 182 re-sequenced *Syzygium* species and 58 re-sequenced unidentified taxa were used to generate whole genome-level phylogenies of the *Syzygium* genus, thus providing new insights into the infrageneric classification of *Syzygium*, as well as into the genus diversification patterns and their drivers. The clove genome assembly (370 Mb) was exploited to investigate the genetic basis of the biosynthesis of eugenol, the major biocompound of clove products (Kamatou et al., 2012; Otunola, 2022). To provide insights into the clove genome evolution, comparative genomics analyses were also performed between *S. aromaticum* and *E. grandis*. The synteny analysis performed between these two Myrtaceae species' genomes assemblies revealed good genome structure conservation. The structures of chromosomes 1, 3, 5, and 7 were found to be highly conserved between *E. grandis* and *S. aromaticum*, and 10 intrachromosomal rearrangements occurring on the 7 other chromosomes were observed (chromosomes 2, 4, 6, 8, 9, 10, and 11). Interestingly, the intrachromosomal rearrangements detected between the two eucalypt species, *E. grandis* and *C. citriodora*, were located on the same seven chromosomes (Butler et al., 2017; Healey et al., 2021). Long terminal repeat retrotransposons (LTR-RTs) are transposable elements (TEs) that move through the genome via a copy-and-paste mechanism using an RNA intermediate. They are considered the most abundant TE component in plant genomes and important drivers of genome size variation and diversification (Wicker et al., 2007; Zhou et al., 2021). Comparing the LTR-RTs repertoires of *S. aromaticum* and *E. grandis* revealed a differential accumulation of the LTR-RTs belonging to the superfamilies Copia and Gypsy between the two species. In *S. aromaticum* genome assembly, the LTR-RTs belonging to the Gypsy superfamily were more abundant than those belonging to the Copia superfamily. In contrast, a higher number of LTR-RTs Copia versus Gypsy was found in the *E. grandis* genome assembly.

No infrageneric comparison of chromosome-scale assemblies has been performed for the *Syzygium* genus. To further investigate the evolution of the genome architecture of *Syzygium* species and verify whether the rearrangements found between *S. aromaticum* and *E. grandis* chromosomes were the consequences of evolutionary events or due to sequencing and assembly artifacts, we generated additional chromosome-scale genome assemblies for *Syzygium malaccense* (L.) Merr. & L.M. Perry, *Syzygium aqueum* (Burm.f.) Alston, *Syzygium jambos* (L.) Alston, and *Syzygium syzygioides* (Miq.) Merr. & L.M. Perry. Like *S. aromaticum* and *S. grande*, the four species belong to the subgenus *Syzygium*, the largest of the five *Syzygium* subgenera for which the crown age was estimated at 9.4 Mya by Low et al. (Low et al., 2022). Previous karyotype studies indicated that *S. malaccense* is a diploid with 2n = 22 chromosomes (Pedrosa et al., 1999) and that *S. jambos* is a tetraploid (2n = 44); however, different chromosome numbers were also reported in the literature for the species (2n = 28, 33, 46, ~54, 66) (Van Lingen, 1991; Oginuma et al., 1993). The chromosomal numbers reported in the literature indicate that *S. aqueum* is also a tetraploid (2n = 44) (Panggabean, 1991).

Here, we describe the *de novo* assembly and annotation for *S. malaccense, S. aqueum*, *S. jambos*, and *S. syzygioides*. To enable subsequent comparative genomic analyses, the four genomes consisting of monoploid consensus (11 chromosomes and unplaced sequences) were generated to achieve the same level of quality for the four species' genome assemblies and comparable to those of published chromosome-scale assemblies of their Myrtaceae relatives. Then, we compared the genome architecture of the four newly *Syzygium* assembled genomes with those of *S. aromaticum* and *S. grande* and their genome features (gene sets and LTR-RTs repertoires) with those of *S. aromaticum* to investigate genomic evolution from their common ancestors.

2 Materials and methods

2.1 Biological materials

The *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides* genome assemblies were generated from trees growing in the Masoala Hall of the Zurich Zoo in Switzerland. Voucher specimens were deposited in the Zürich herbarium (*S. malaccense* (ZT-00170996), *S. aqueum* (ZT-00170994), *S. jambos* (ZT-00170999), and *S. syzygioides* (ZT-00170991)). Samples collected from the trees were stored at -80°C until nucleic acid extraction.

2.2 DNA and RNA isolation

High-molecular-weight genomic DNA was isolated from frozen leaves using the "ONT high-molecular-weight gDNA extraction from plant leaves" protocol (Oxford Nanopore Technologies, Oxford, UK). Following the extraction, we performed a size selection step using the Circulomics Nanobind Plant Nuclei Big DNA Kit from PacBio (Menlo Park, CA, USA). (NB-900-801-001).

Total RNA from *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides* were isolated in triplicate from whole leaves (young and mature), lamina of mature leaves, and stems. Total RNA was also isolated in triplicate from *S. syzygioides*' buds (in the fruiting stage) and *S. jambos*' buds (before and after flowering) and flowers.

Total RNA was extracted from frozen powder using Ambion PureLink Plant RNA Reagent (Ambion by Life Technologies, Carlsbad, CA, USA). The concentration and quality of the total RNA were assessed with an Agilent Bioanalyzer using the Agilent RNA 6000 Nano Kit (Agilent, Santa Clara, CA, USA).

2.3 Illumina sequencing library preparation and sequencing

DNAseq libraries were prepared from total gDNA using the Celero PCR workflow with an enzymatic fragmentation kit from Tecan (Männedorf, Switzerland). DNAseq libraries were loaded on an Illumina S2 flow cell and sequenced on the Illumina Novaseq 6000 instrument (Illumina, San Diego, CA, USA) as 2 x 151 bp paired-end reads.

Hi-C libraries were prepared from 0.2 g of frozen leaves using the Proximo Hi-C Kit following the manufacturer's instructions (Phase Genomics, Seattle, WA, USA) and sequenced on an Illumina HiSeq 4000 instrument (Illumina) as 2 x 151 bp paired-end reads.

mRNA stranded libraries were prepared from 500 ng of total RNA using the Tecan Universal Plus mRNA-Seq library preparation kit with NuQuant® and sequenced on an Illumina HiSeq 4000 instrument as 2 x 151 bp paired-end reads.

Illumina raw reads generated from DNAseq libraries and Hi-C libraries were cleaned using fastp 0.23.2 (--length_required 75 --low_complexity_filter) (Chen et al., 2018).10.1038/s41597-021-00968-x

2.4 ONT sequencing library preparation and sequencing

Sequencing libraries were generated from high-molecular-weight gDNA and prepared for sequencing on PromethION flow cells (FLO-R0002) by using the ligation sequencing (SQK-LSK109) and flow cell priming (EXP-FLP002) kits (Oxford Nanopore Technologies, Oxford, UK). The base calling was performed by using Guppy 6.1.1 and the super accuracy plant model. Raw ONT reads were cleaned using seqkit 2.2.0 (--min-qual 9 --min-len 5000) (Shen et al., 2016) to discard reads shorter than 5,000 bp or with quality scores lower than 9.

2.5 Genome profiling

Cleaned Illumina paired-end reads from DNAseq libraries were analyzed by GenomeScope 2.0 and smudgeplot 0.2.4 to estimate the genome size, percentage of heterozygosity, and the ploidy level using a k-mer size equal to 21 (Ranallo-Benavidez et al., 2020).

2.6 *De Novo* genome assembly

ONT cleaned reads were corrected with fmlrc2 0.1.7 (--cache_size 13 –K 21 59 79) (Mak et al., 2023) using cleaned Illumina paired-end short-reads from DNAseq libraries. The corrected ONT reads were then assembled using flye 2.9 (--read-error 0.01 --nano-hq) (Kolmogorov et al., 2019) and iteratively polished with ntedit 1.3.5 (-m 2 -i 3 -d 3 -X 0.5 -Y 0.5) using kmer profiles created with nthits 0.0.1 (--solid --outbloom -b 36) for kmers of lengths 60, 50, 40 and 30 (Warren et al., 2019) using Illumina paired-end short reads from DNAseq libraries. Haplotigs were detected and removed from the polished contigs using purge_dups 1.2.5 (Guan et al., 2020) using cutoff of 10, 315 and 645 for *S. malaccense*, 70, 440 and 960 for *S. aqueum*, and 60, 410, 960 for *S. jambos*, 10, 410 and 960 for *S. syzygioides*.

Cleaned Illumina read pairs generated from Hi-C libraries were mapped to the genomes to remove reads with low mapping scores, duplicated reads, and paired-end reads. Illumina Hi-C read pairs were mapped to the haplotig-purged contigs using minimap2 2.24 (Li, 2018) rather than bwa (Li, 2013) since we noticed that it results in assemblies of equivalent qualities in a shorter time. The scaffolding to a chromosome-scale assembly was performed using yahs 1.1a2 (-r 1000,2000,5000,10000,20000,50000,100000,200000,500000,1000000,2000000,5000000) (Zhou et al., 2022). Hi-C map files were generated with PretextMap 0.1.9 (https://github.com/wtsi-hpag/PretextMap) and used to manually curate the assemblies using PretextView 0.2.5 (https://github.com/wtsi-hpag/PretextView).

The curated genome assemblies were mapped to the *S. aromaticum* genome (Ouadi et al., 2022) using minimap2 2.24, visualized using a custom R script, and the orientation and names of the chromosomes were set in accordance with those of *S. aromaticum*. Chromosome-scale assembly completeness was assessed by using the genome evaluation mode of BUSCO 5.4.4 and the eudicots_odb10 lineage dataset (Simão et al., 2015). The QVs of the final assemblies were estimated using yak 0.1 (qv -K 2000000000) with kmer profiles created using yak 0.1 (count -k 31 -K 2000000000 -b37) (Cheng et al., 2021).

2.7 Gene annotation

The Illumina RNAseq reads from *S. malaccense*, *S. aqueum*, *S. jambos* and *S. syzygioides* as well as those used for the clove genome annotation were cleaned, and overlapping paired-reads were merged using fastp 0.23.2 (--length_required 75 --low_complexity_filter --merge) (Chen et al., 2018) before being mapped as single cDNA reads to the assemblies using minimap2 2.24 (-ax splice:hq -G5K

-N50) (Li, 2018). Gene models were then created for each RNASeq sample using scallop 0.10.5 (--min_transcript_coverage 5 --min_single_exon_coverage 50 --min_splice_bundary_hits 5 --min_mapping_quality 0) (Shao and Kingsford, 2017).This approach was used for the annotation of the clove genome, where it was observed to produce better gene models than by directly mapping paired-reads with a dedicated mapper.

To obtain models for genes that are not expressed in the RNAseq samples, the transcripts from *S. aromaticum* and *E. grandis* gene annotations were mapped to the assemblies using minimap2 2.24 (-ax splice:hq -I5G -G5K -N50 -uf) (Li, 2018), and gene models created using bedtools 2.30.0 (bamtobed -bed12) (Quinlan and Hall, 2010) and custom gawk scripts to convert the obtained bed file into a gtf file.

The final gene models were obtained by merging the RNAseq, *S. aromaticum*, and *E. grandis* gene models using taco 0.7.3 (--gtf-expr-attr TPM --filter-min-expr 10) (Niknafs et al., 2017) and adding coding sequences using Transdecoder 5.5.0 (LongOrfs -S -m 64; Predict --single_best_only --retain_blastp_hits dmd.tsv) (https://github.com/TransDecoder/TransDecoder/wiki), diamond 2.0.15 (blastp --query longest_orfs.pep --db uniref-malvids.dmnd --max-target-seqs 1 --outfmt 6 --evalue 1e-6) (Buchfink et al., 2015) and gffread 0.12.7 (Pertea and Pertea, 2020).

The eudicotyledons portion of UniProt filtered to remove proteins with poor descriptions was used to annotate the gene models with their best hit using diamond 2.0.15 (blastx --query tx.fa --db eudicotyledons.filtered.dmnd --top 10 --min-score 200 --ultra-sensitive --iterate). The illustration of the regions where genes encoding for putative eugenol synthase were predicted was generated using gggenes 0.4.0 (https://github.com/wilkox/gggenes).

2.8 Repeat annotation

Annotation of transposable elements was carried out using TE-greedy-nester 1.0.0 (--discovery_tool LTRharvest) (Lexa et al., 2020), genometools LTRharvest 1.6.2 (Ellinghaus et al., 2008) and TEsorter 1.3.0 (-db rexdb-plant --min-coverage 10 --max-evalue 0.01 --pass2-rule 70-30-80) (Zhang et al., 2022) with REXdb (Neumann et al., 2019). The insertion age of the predicted transposable elements was then calculated as previously reported (Marcon et al., 2015). In addition, Red 2.0 (Girgis, 2015), GRF 1.0 (Shi and Liang, 2019) and cd-hit 4.8.1 (grf-main -i genome.fa -c 1 -o genome.MITE --min_tr 10; cd-hit-est -i genome.MITE/candidate.fasta -o genome.MITE/clusteredCandidate.fasta -c 0.90 -n 5 -d 0 -aL 0.99 -s 0.8 -M 0; grf-mite-cluster -i genome.MITE/clusteredCandidate.fasta.clstr -g genome.fa -o genome.MITE) (Fu et al., 2012), EAHelitron (Hu et al., 2019), and tantan 39 (-f4) (Frith, 2011) were used to predict repeats, Miniature Inverted-repeat Transposable Elements (MITEs), helitron, and tandem repeats, respectively.

2.9 Synteny analyses

Synteny between the *Syzygium* species was done by pairwise mapping whole genomes using minimap2 2.24 (Li, 2018), identifying structural variants using syri 1.6 (Goel et al., 2019), and plotting syntenic blocks larger than 20 kb using plotsr 0.5.4 (Goel and Schneeberger, 2022).

2.10 Orthologue analyses

Orthologous genes were clustered into HOGs with OrthoFinder 2.5.4 (Emms and Kelly, 2019) using the set of predicted protein sequences from the five species assemblies.

3 Results

3.1 Genome profiling

Smudgeplot and GenomeScope 2.0 were used to perform a genome profiling step using Illumina PE short-reads from DNAseq libraries as input and a K-mer length of 21 bp (Ranallo-Benavidez et al., 2020) (Table 1; Supplementary Table 1; Supplementary Figures 1, 2). The ploidy level predicted by Smudgeplot was in accordance with previous karyotype studies for the studied *S. malaccense* and *S. jambos* specimens (Oginuma et al., 1993; Pedrosa et al., 1999). *S. malaccense* was predicted to be a diploid specimen (2n = 2x = 22) like *S. aromaticum* and *S. grande*. The *S. aqueum*, *S. jambos*, and *S. syzygioides* specimens were predicted as being autotetraploid (2n = 4x = 44). The estimated monoploid genome sizes were similar among the four *Syzygium* species (343–372 Mb), a size range consistent with the small genome assembly sizes of *S. aromaticum* (370 Mb) and *S. grande* (405 Mb) (Low et al., 2022; Ouadi et al., 2022). The heterozygosity rate estimated by the GenomeScope 2.0 ranged from 2.3% for the diploid specimen *S. malaccense* to 4.3% for the autotetraploid specimen *S. aqueum*. These heterozygosity rates appeared to be higher than for *S. aromaticum* (0.18%) (Ouadi et al., 2022) and the average reported by Ellestad et al., who performed a literature review of the genome-wide heterozygosity values estimated using the software GenomeScope and GenomeScope 2.0 (Ellestad et al., 2022). They found that the average value inferred for all plant species assessed was 1.59% (1.10% for diploid plants only) noting that over half of the plant species considered were cultivated for human usage, which could affect the average value accuracy.

3.2 Genome *De Novo* assembly

The four *de novo* chromosome-scale assemblies were constructed using long-reads from Oxford Nanopore Technologies (ONT), short paired-end reads from Illumina DNAseq libraries, and Hi-C libraries generated for each *Syzygium* species (Supplementary Tables 1, 2).

To prevent assembly artifacts possibly caused by heterozygosity and polyploidy of the *Syzygium* specimens, haplotigs were detected and removed from the polished contigs. The effect of the haplotig removal step was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) in genome mode (Simão et al., 2015). After

10.3389/fpls.2023.1248780

TABLE 1 Genome profiling summary.

	S. aromaticum (Ouadi et al., 2022)	*S. malaccense*	*S. aqueum*	*S. jambos*	*S. syzygioides*
Predicted ploidy	2n = 2x = 22	2n = 2x = 22	2n = 4x = 44	2n = 4x = 44	2n = 4x = 44
Estimated genome (1x) size	343 Mb	372 Mb	345 Mb	361 Mb	372 Mb
Estimated heterozygosity rate	0.18%	2.30%	4.30%	3.60%	4.10%

the haplotig removal step, the number of complete and duplicated BUSCOs genes was considerably reduced in the haplotig-purged contigs (3.3% to 6.1%) when compared to the polished contigs (93.6% to 97.1%) (Figure 1A). Hi-C data enabled the scaffolding of contigs into 11 chromosomes. On the Hi-C contact matrices, a strong intra-chromosomal signal indicates efficient scaffolding, with the 11 chromosomes of each *Syzygium* assembly supported by a high number of their respective Hi-C reads (Figure 1B).

The final chromosome-scale assemblies for *S. malaccense* (430 Mb), *S. aqueum* (392 Mb), *S. jambos* (426 Mb), and *S. syzygioides* (431 Mb) consisted of monoploid consensus (11 chromosomes and unplaced sequences) with comparable quality metrics. A high level of quality at the base-scale (quality value [QV] between 44.006 and 45.114), of contiguity (97.5% to 99.8% of the assemblies length anchored on 11 chromosomes) and completeness (BUSCO complete genes scores of 98%) was reached for the four new assembled *Syzygium* genomes (Table 2; Figure 2; Supplementary Tables 3, 4).

Despite their high heterozygosity rate, the quality metrics for the genome assemblies of the diploid specimen *S. malaccense* and

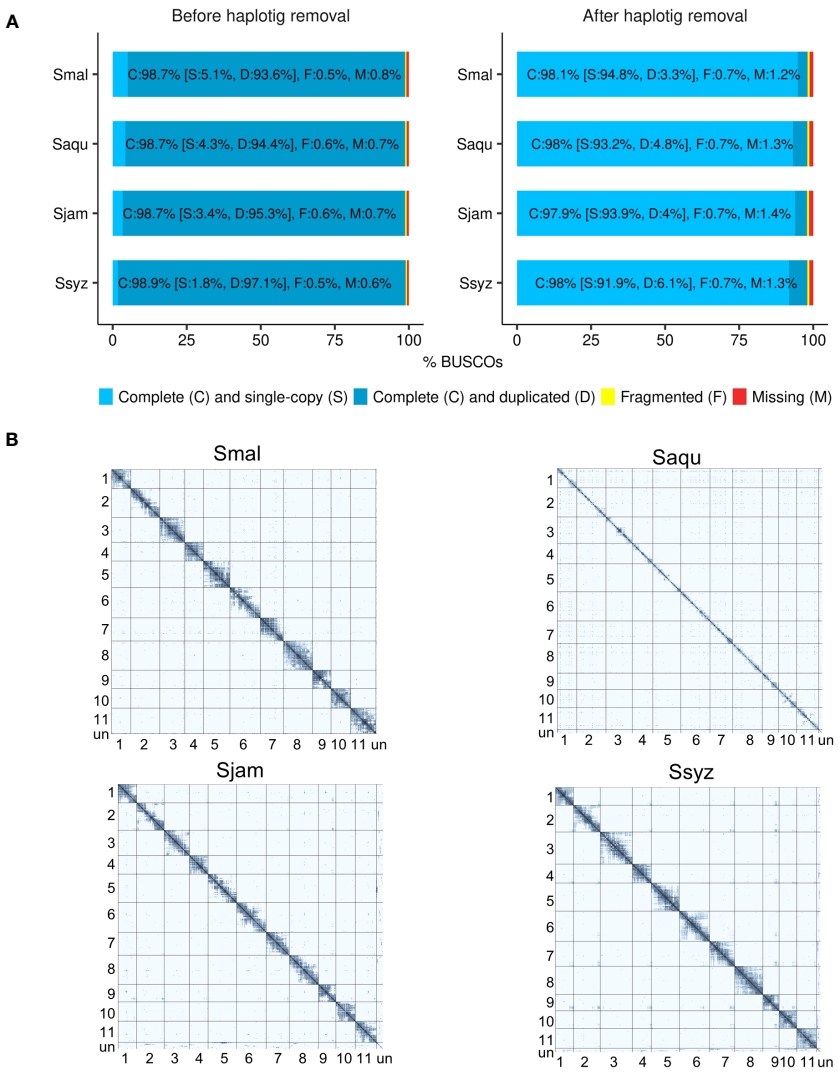


FIGURE 1
Assessment of the efficiency of the haplotig removal step and Hi-C scaffolding. **(A)** BUSCO completeness score comparison of the polished contigs before and after the haplotig removal step for *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz) (BUSCO version 5.4.4 - dataset: eudicots_odb10 (n = 2326)). **(B)** Hi-C contact maps showing the Hi-C interactions among the 11 assembled chromosomes and unplaced scaffolds (un) for each species.

10.3389/fpls.2023.1248780

TABLE 2 Assembly and annotation statistics.

	S. malaccense	*S. aqueum*	*S. jambos*	*S. syzygioides*
Assembly				
Number of scaffolds	23	54	117	101
Number of chromosome-scale scaffolds	11	11	11	11
Proportion of undetermined bases (N)	0.01%	0.01%	0.02%	0.02%
QV[1] of the assembly	45.114	44.006	44.028	44.292
Length of assembly (bp)	429,836,287	391,897,832	426,159,599	431,079,378
Length of chromosome-scale scaffolds (bp)	429,008,219	386,536,673	415,622,982	424,827,227
Gene annotation				
Number of predicted genes	30,842	29,879	31,611	32,142
Number of predicted transcripts	57,144	55,010	57,897	59,495
Average transcript length (bp)	2010.89	2008.09	1991.19	2007.31
Average CDS[2] length (bp)	1122.42	1124.09	1116.93	1100.23
Average exon per transcript	5.62	5.67	5.59	5.66
Repeat annotation				
Repeat sequences (bp)	184,916,857 (43.02%)	162,020,435 (41.34%)	180,563,593 (42.37%)	184,003,101 (42.68%)
LTR[3] retrotransposons (bp)	96,086,564 (22.35%)	74,914,968 (19.12%)	77,407,268 (18.16%)	73,171,928 (16.97%)
LTR Gypsy (bp)	62,668,430 (14.58%)	48,141,612 (12.28%)	45,090,450 (10.58%)	40,369,474 (9.36%)
LTR Copia (bp)	31,769,467 (7.39%)	25,148,956 (6.42%)	30,040,740 (7.05%)	30,532,813 (7.08%)

[1] QV, Quality value.
[2]CDS, Coding sequence.
[3]LTR, Long Terminal Repeat.

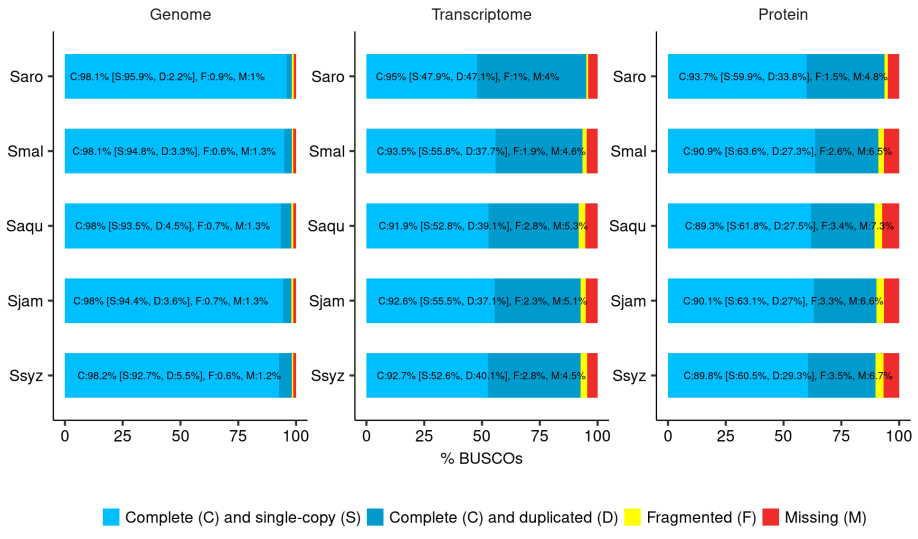


FIGURE 2
BUSCO completeness assessment. Assessment of the final genome assembly, transcript set, and protein set of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz) (BUSCO version 5.4.4 - dataset: eudicots_odb10 (n = 2326)).

the autotetraploids *S. aqueum*, *S. jambos*, and *S. syzygioides* were comparable to those reported for *S. aromaticum* assembly (370 Mb) (Ouadi et al., 2022). Nevertheless, BUSCO scores revealed a higher percentage of complete and duplicated BUSCOs in the four new assemblies compared to *S. aromaticum* (2.2%), principally in the genome assembly of the three autotetraploid specimens (3.3% to 5.5%) (Figure 2).

3.3 Genome annotation

The average number of protein-coding genes predicted for the four newly assembled genomes is 31,119, representing 26.52% of the genome assemblies' size (Table 2).

The annotation completeness was assessed using the BUSCO method in transcriptome and protein modes and by selecting the whole set of predicted transcripts and proteins for each gene as inputs, respectively (Figure 2; Supplementary Table 3). BUSCO results indicated that the annotation completeness is comparable among the four newly assembled *Syzygium* species, with complete BUSCO scores ranging from 91.9% in *S. aqueum* assembly to 93.5% in *S. malaccense* assembly in transcript mode and from 89.3% in *S. aqueum* assembly to 90.9% in *S. malaccense* assembly in protein mode. BUSCO scores obtained for *S. aromaticum* by using the same assessment methods (95% in transcriptome mode and 93.7% in protein mode) were slightly superior to those of newly assembled genomes but still comparable. The loss of complete BUSCOs between the genome and protein mode assessments ranged from 7.2% in *S. malaccense* assembly to 8.7% in *S. aqueum* assembly, indicating acceptable quality of the predicted gene models and protein sets.

The genome assembly of *S. aromaticum* comprised multiple copies of a gene encoding for putative eugenol synthase (EGS), the enzyme that catalyzes the synthesis of eugenol from coniferyl acetate. In total, 15 copies split into 2 loci were reported: a first locus on chromosome 10 comprising 14 copies and a second locus on chromosome 11 with 1 copy (Ouadi et al., 2022). The functional annotation of the four newly assembled *Syzygium* species genomes revealed fewer genes encoding for putative EGS. One gene encoding for putative EGS was identified in the genome assembly of *S. malaccense*, two in the genome assembly of *S. aqueum*, and three copies were found in the genome assemblies of *S. jambos* and *S. syzygioides*. All putative EGS genes were located on chromosome 10 except for one of the three copies of *S. syzygioides* located on chromosome 11 (Figure 3; Supplementary Table 5).

Effective lengths of repeat elements, which are different from their genomic length, were calculated by removing the length of the nested elements they contained. The proportions of genome assembly length occupied by predicted genes (25.97% to 27.37%) and repeat sequences (41.34% to 43.02%) appear to be conserved among the four newly sequenced *Syzygium* genomes (Table 2). Using the same method, repeat elements in *Syzygium aromaticum* genome assembly represents 39.98%. The most abundant repeat elements identified in the four newly sequenced *Syzygium* genomes were the LTR-RTs spanning 16.97% of the assembly length for *S. syzygioides* to 22.35% for *S. malaccense*. As reported for *S. aromaticum* and *S. grande*, LTR-RTs belonging to the Gypsy superfamily were more abundant than elements belonging to the Copia superfamily in the four newly sequenced genomes (Table 2; Supplementary Tables 6–9) (Low et al., 2022; Ouadi et al., 2022).

3.4 Synteny analyses

To identify evolutionary structural changes among the *Syzygium* species chromosomes, we performed a synteny analysis on the four newly assembled genomes, *S. aromaticum* and *S. grande*. The alignment of the 11 chromosomes' DNA sequences of the 6 *Syzygium* species revealed a high conservation of the chromosomal organization (Figure 4A).

No large interchromosomal rearrangements were detected between the chromosomes of the six *Syzygium* species. A high percentage of the five species' chromosome lengths were syntenic with *S. aromaticum*, ranging from 68.45% between *S. aromaticum* and *S. jambos* to 73.02% between *S. aromaticum* and *S. aqueum*. Intrachromosomal rearrangements such as inversions, translocations, and duplications between the chromosomes of *S. aromaticum* and those of the other five *Syzygium* species represented 5% of their 11 chromosomes length on average. In terms of number, the most frequent rearrangements observed between *S. aromaticum* and the five other species were duplications and translocations with average numbers of 1348 and 1325, respectively, spanning an average of 0.85% to 1.43% of the 11 chromosome lengths. Inversions were found less frequently for all species but occupied a larger fraction of the genome assemblies' length than duplications and translocations except for *S. syzygioides*. The percentage of assembly lengths comprising inversions between *S. aromaticum* and the five other *Syzygium* species ranged from 0.68% between *S. aromaticum* and *S. syzygioides* to 4.83% between *S. aromaticum* and *S. grande*. Overall, the size of the inversions was relatively small. For instance, 11 inversions were detected, between chromosome 5 of *S. aromaticum* and *S. grande*, representing 17.32% of the chromosome length of *S. grande* (41,797,999 bp) and 1.87% of its 11 chromosomes (387,620,547 bp) (Figure 4B; Supplementary Table 4). In contrast, the synteny analysis performed between *S. aromaticum* and *E. grandis* revealed 10 intrachromosomal rearrangements on chromosomes 2, 4, 6, 8, 9, and 10 that included large terminal inversions representing up to 40% of the chromosome length of *S. aromaticum*. The other four chromosomes (1, 3, 5, and 7) of the two Myrtaceae species were highly syntenic (Ouadi et al., 2022). To further investigate the chromosomal architecture evolution of the *Syzygium* species and verify that these rearrangements were due to biological events rather than assembly artifacts, we also performed DNA alignment of the chromosome sequences of *E. grandis* with the those of *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides*. Chromosomes 1, 3, 5, and 7 of *E. grandis* and those of the four newly assembled species were also highly syntenic, and we observed the same 10 rearrangements on chromosomes 2, 4, 6, 8, 9, 10 and 11 (Figure 4A; Supplementary Figure 3).

10.3389/fpls.2023.1248780

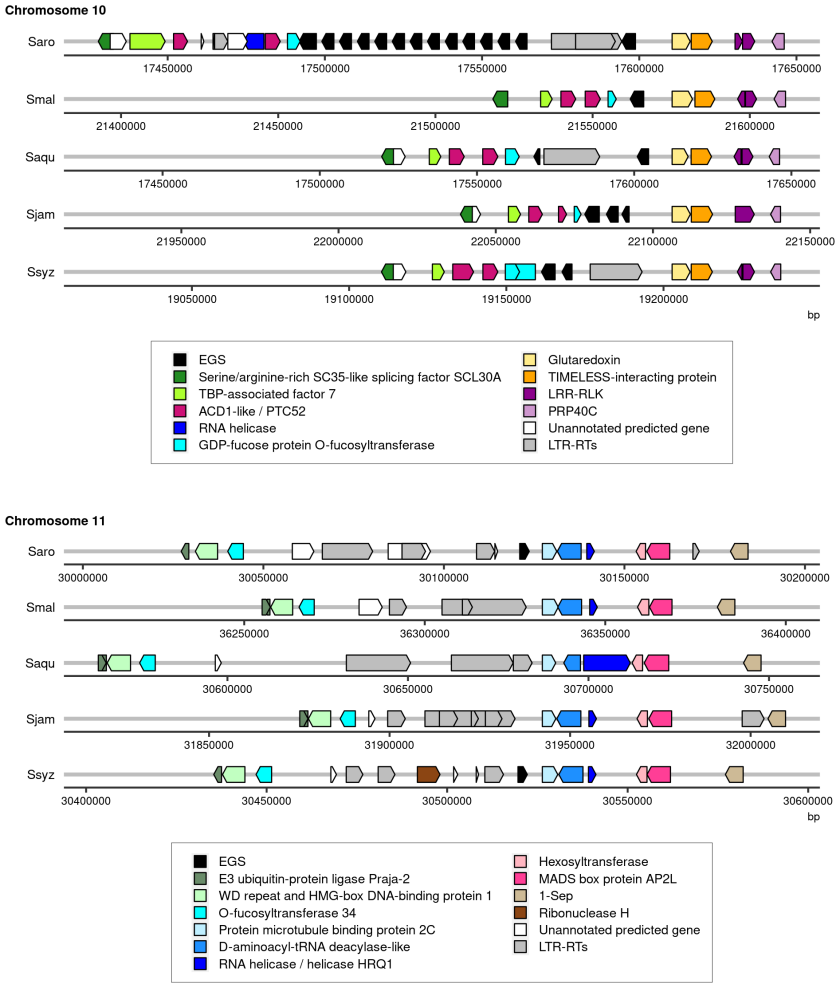


FIGURE 3
Illustration of the regions of chromosomes 10 and 11 of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz) where genes encoding for EGS were predicted. The position (bp) and orientation of the predicted genes on the chromosomes are indicated by arrows colored according to the functional annotation. EGS, accelerated cell death (ACD1), Protochlorophyllide-dependent translocon component Tic52 (PTC52), leucine-rich repeat receptor-like protein kinase (LRR-RLK), Pre-mRNA-processing protein 40C-like (PRP40C), TATA-binding protein-associated factor 7 (TBP-associated factor 7), LTR-RTs.

3.5 Gene orthology

To investigate the phylogenetic relationships among gene sequences of *S. aromaticum*, *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides*, the sets of predicted protein sequences from the five species assemblies were analyzed using OrthoFinder (Emms and Kelly, 2019).

A total of 49,269 hierarchical orthogroups (HOGs) were identified, including 93.7 to 95.2% of each species gene set (Figure 5A). Of these, 18,963 (38.5%) HOGs contained genes from all five species, and 4,928 (10%) were species specific. In more detail, 789 were specific to *S. aromaticum*, 950 were specific to *S. malaccense*, 940 HOGs were specific to *S. aqueum*, 1009 HOGs were specific to *S. jambos*, and 1240 HOGs were specific to *S. syzygioides*. Pairwise, *S. aromaticum* and *S. aqueum* appear to share the lowest number of orthogroups (625). The highest number of shared HOGs inferred between each pair of studied species was found between *S. aqueum* and *S. syzygioides* (1218), followed by *S. aqueum* and *S. malaccense* (1152), and *S. jambos* and *S. malaccense* (1027). The species tree resulting from the analysis of the HOGs divided the *Syzygium* species studied into two groups based on closer relationships: the first group comprising *S. aromaticum* and *S. aqueum* and a second group comprising *S. jambos*, *S. malaccense*, *and S. syzygioides* (Figure 5B).

3.6 Annotation and comparison of LTR-RTs Gypsy and Copia repertoires

To clarify the dynamic activity of full-length LTR-RTs belonging to the superfamilies Gypsy and Copia within the *Syzygium* genus, we identified the lineages belonging to each superfamily located on the chromosomes of *S. malaccense* (429 Mbp), *S. aqueum* (387 Mbp), *S. jambos* (416 Mbp), and *S.*

10.3389/fpls.2023.1248780

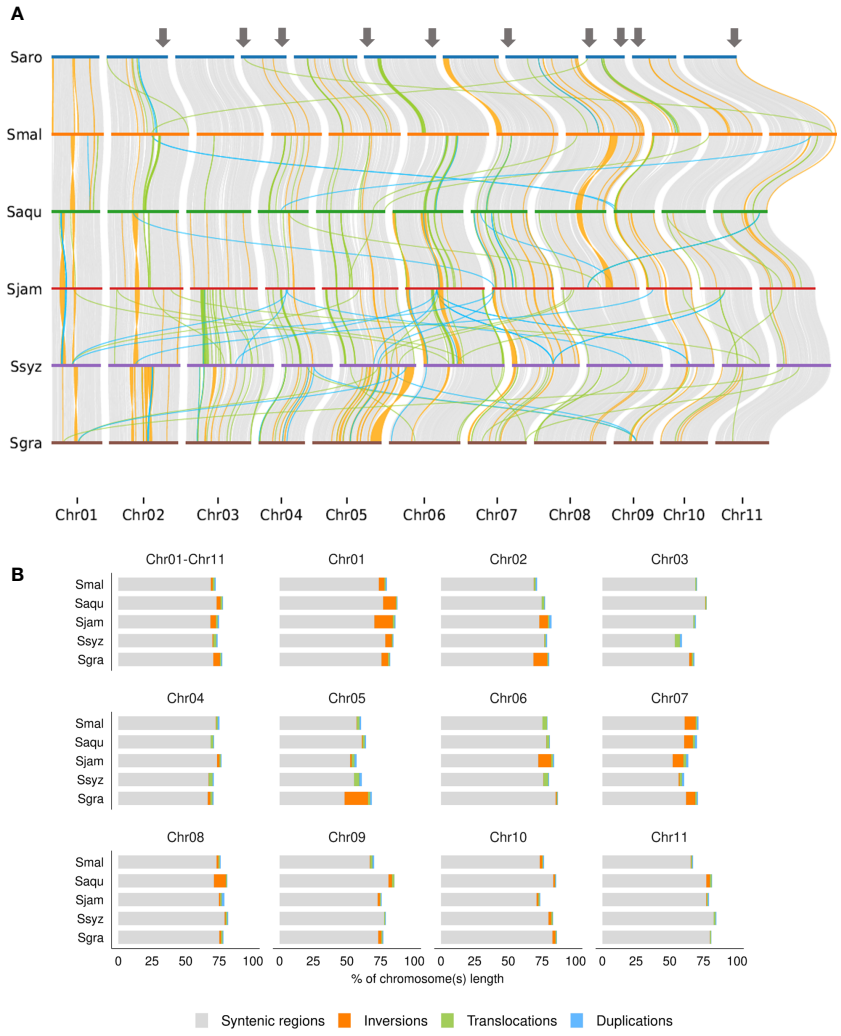


FIGURE 4
Identification of syntenic and rearranged regions between the 11 chromosomes of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), *S. syzygioides* (Ssyz), and *S. grande* (Sgra). **(A)** Representation of the alignment of the chromosomal DNA sequences showing syntenic regions, interchromosomal, and intrachromosomal rearrangements larger than 20 kb (inversions, translocations, and duplications). Grey arrows indicate regions where rearrangements were reported between chromosomes of *E grandis* and *S. aromaticum*. **(B)** Pairwise comparison of the percentage of chromosome length occupied by syntenic regions and rearrangements between the chromosome-scale assembly (Chr01-Chr11) and 11 chromosomes (Chr01 to Chr11) of *S. aromaticum* with those of *S. malaccense*, *S. aqueum*, *S. jambos*, *S. syzygioides*, and *S. grande*.

syzygioides (425 Mbp) and estimated their insertion time. Then, we compared the repertoires' compositions and repeat element insertion times of the four species with those of *S. aromaticum* (368 Mbp).

We found that *S. malaccense* and *S. aromaticum*, the largest and smallest chromosome-scale assemblies of this study, contained the highest (8427) and lowest number (6167) of LTR-RTs in Gypsy and Copia, respectively (Figure 6A; Supplementary Tables 6–9). In the five *Syzygium* species' chromosomes, we identified a higher number of LTR-RTs for Gypsy than Copia, with a ratio of Gypsy to Copia content ranging from 1.09 for *S. syzygioides* to 1.45 for *S. malaccense*. The Gypsy superfamily comprised a higher proportion of nested elements (17.37% to 24.47%) compared to the Copia superfamily (7.01% to 9.44%), suggesting distinct accumulation and mobile activity of both superfamilies in all five species. Our results revealed little variation in the number of Copia elements on the chromosomes of *S. aqueum* (2705 elements) and *S. aromaticum* (2809), the two smallest chromosome-scale assemblies, and on the chromosomes of *S. syzygioides* (3290), *S. jambos* (3324), and *S. malaccense* (3433). In contrast, we found a notably higher accumulation of Gypsy elements (4994) in the chromosomes of *S. malaccense* compared to the four other species. The ratio of Gypsy content varied from 1.35 when comparing *S. malaccense* with *S. jambos* to 1.49 when comparing *S. malaccense* with *S. aromaticum*. It represented a difference in Gypsy effective length of 19,402,234 bp to 21,766,176bp, respectively. In the five *Syzygium* chromosome-scale assemblies, the most abundant lineage belonged to the Gypsy superfamily, but it varied according to the species. The Gypsy lineage Tekay was the most represented for *S. aromaticum* (1534 elements), *S. jambos* (1674 elements), and *S. syzygioides* (2090 elements). At the same time, for S. *malaccense* and *S. aqueum*, we found a higher abundance of the gypsy lineage Ogre (2382 and 1799

10.3389/fpls.2023.1248780

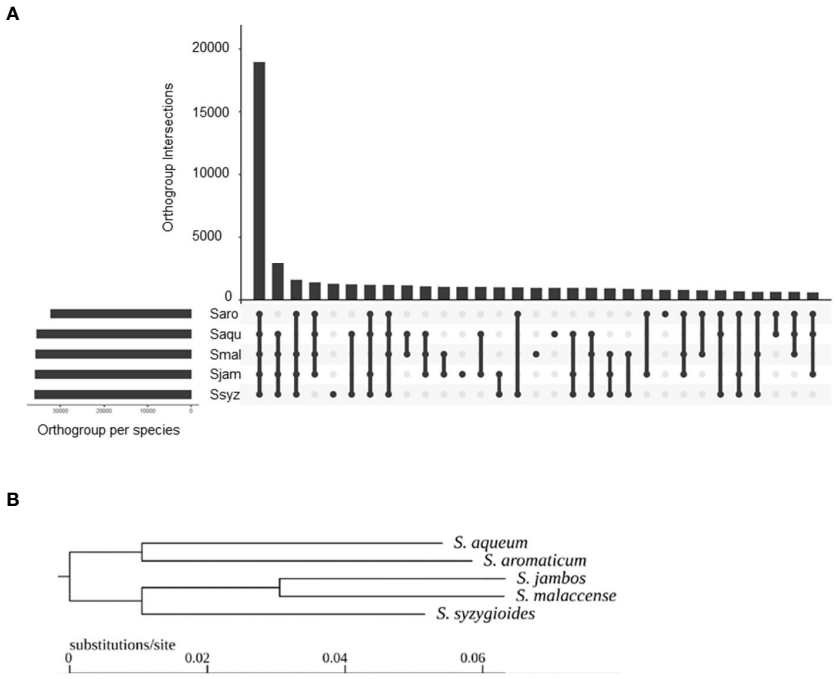


FIGURE 5
Hierarchical orthogroups (HOGs) inferred by OrthoFinder between S. *aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz). **(A)** Number of HOGs inferred by OrthoFinder using the set of predicted proteins for the five *Syzygium* species. **(B)** Rooted species tree inferred by OrthoFinder.

elements, respectively). Among the Gypsy superfamily, the most abundant lineages, Tekay and Ogre, were those with the highest proportion of nested elements (19.10% to 28.55% and 16.69% to 27.92%, respectively) in all five species. For *S. aromaticum*, *S. malaccense*, and *S. syzygioides*, the proportion of nested elements belonging to the Athila lineage was also among the highest identified (Figure 6B; Supplementary Figures 4, 5).

Regarding the Copia superfamily, the most represented lineages on the chromosomes of the five *Syzygium* species were Ale (608 to 873 elements), followed by the lineage Tork (456 to 762 elements) for *S. malaccense*, *S. aqueum, S. jambos* and *S. Syzygoides*, and the lineage SIRE (502 elements) for *S. aromaticum*.

The insertion times of 97.13% of the full-length Gypsy and Copia elements identified in the five *Syzygium* species were estimated (33,861elements). Nearly all elements (97.33%) were inserted in the last 5 million years (32,958 elements) (Figure 7). During this time period, distinct insertion activities of the two superfamilies occurred in the five *Syzygium* species.

Compared to the other four *Syzygium* species, the chromosomes of *S. aromaticum* underwent a more ancient wave of Gypsy insertions (peak at ~2.5 million years ago [Mya]), principally attributed to the Tekay elements, the most abundant lineage in this species (Figure 7A). We also found that a few recent insertions (18.02% of insertions) occurred in *S. aromaticum* chromosomes within the last one million years. In contrast, a recent burst of Gypsy insertions (~0–1 Mya) occurred in four other species chromosomes: most insertions of Gypsy in *S. malaccense* (44.53%), *S. aqueum* (44.43%), *S. jambos* (52.55%), and *S. syzygioides* (36.45%) were less than one million years old. We inferred that the high number of Gypsy LTR-RTs found in *S. malaccense* may be attributable to two successive waves of insertions: a peak of Tekay insertions at ~2 Mya and a more recent peak of Ogre at ~1 Mya.

Similar to what we observed for the Gypsy superfamily, the insertion of Copia elements occurred earlier in *S. aromaticum* compared to the four other species, with fewer recent insertions (Figure 7B). Compared to the Gypsy elements, a smaller proportion of recent Copia insertions (less than one million years old) were detected in *S. aromaticum* (10.66%), *S. malaccense* (24.16%), *S. aqueum* (26.49%), and *S. jambos* (36.90%) suggesting a distinct recent insertion pattern of the two superfamilies in the four species. However, we found a comparable proportion of Gypsy (36.45%) and Copia (32.22%) elements that were less than one million years old in *S. syzygioides*, the species for which we found the lowest ratio of Gypsy to Copia content (1.09).

4 Discussion

Plant genome size, ploidy level, and heterozygosity rates are challenges for genome assembly and annotation. However, lower sequencing costs and recent advances in long-read sequencing technologies, Hi-C technologies, and bioinformatics tools have facilitated the generation of assemblies with high contiguity up to the chromosome-scale also for non-model plants or non-major plant crops (Kyriakidou et al., 2018; Pucker et al., 2022). Newly assembled and annotated genomes from related species can then be used to perform comparative genomics analyses to investigate plant

10.3389/fpls.2023.1248780

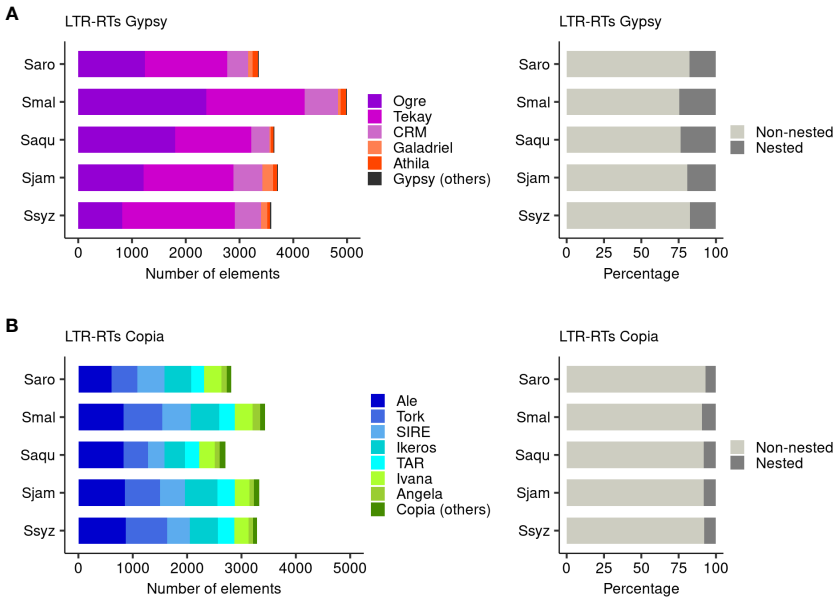


FIGURE 6
Composition of the full-length LTR-RTs Gypsy and Copia repertoires. **(A)** Number of elements belonging to the Gypsy and Copia lineages identified on the 11 chromosomes of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz). **(B)** Proportion of nested and non-nested elements. Gypsy (others) group comprises the lineages non-chromo-outgroup, Reina, Retand, tatIII, and elements Gypsy to which no lineages were assigned. Copia (others) group comprises the lineages Alesia, Bianca, Gymco-I, Gymco-IV, Gymco-II, and Osser.

genome evolution and function. Third-generation long-reads from Oxford Nanopore Technologies and Illumina short-reads combined with the Hi-C technology enabled the *de novo* assembly of the chromosome-scale genome for *S. malaccense, S. aqueum*, *S. jambos*, and *S. syzygioides*. A high level of quality at the base level, contiguity, and completeness was reached for the four newly sequenced genomes. The quality of the newly assembled *Syzygium* species genomes were comparable to that of the *S. aromaticum* genome. The slight differences found between the species assemblies' quality metrics may be linked to the combined impact of the ploidy level and high heterozygosity rates of the four newly sequenced species on the assembly process.

Previous infrageneric comparative genetic mapping analyses revealed high levels of synteny and collinearity among the

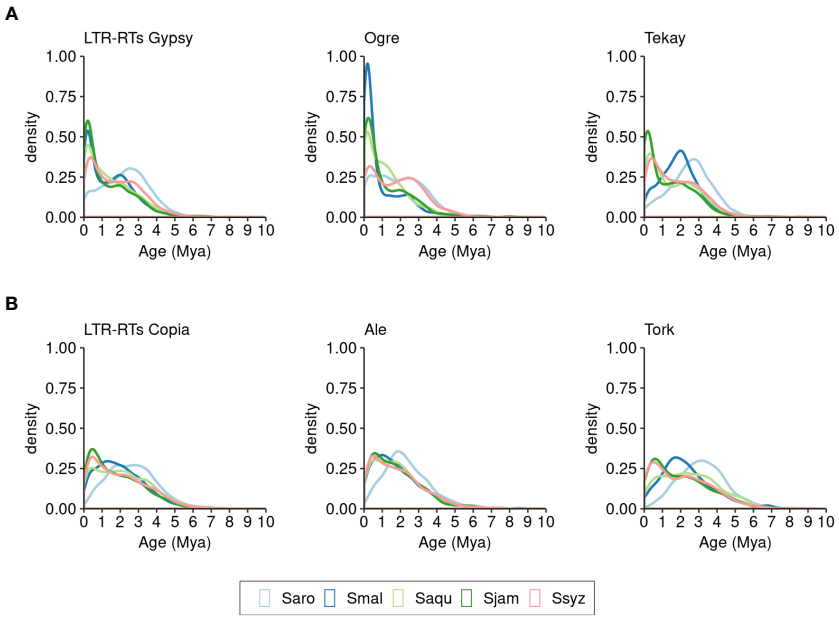


FIGURE 7
Distribution of insertion times of full-length LTR-RTs of *S. aromaticum* (Saro), *S. malaccense* (Smal), *S. aqueum* (Saqu), *S. jambos* (Sjam), and *S. syzygioides* (Ssyz). **(A)** LTR-RTs Gypsy. **(B)** LTR-RTs Copia.

Eucalyptus genus (Hudson et al., 2012; Li et al., 2015). In addition, genomic synteny analyses conducted between the *de novo* assembly of *E. urophylla* × *E. grandis* (EUC) and 30 *Eucalyptus* species revealed that the genome structure of EUC, *E. grandis*, and *E. globulus* showed the higher collinearity, and the absence of large-scale structural variation. Nevertheless, large structural variations among the different chromosomes of the EUC and other *Eucalyptus* species were also detected (Shen et al., 2023). We found that the six *Syzygium* genomes studied were highly syntenic. The intrachromosomal rearrangements (duplications, translocations, and inversions) observed between *S. aromaticum* and the five other *Syzygium* species represent a small percentage (~5% on average) of the 11 chromosomes' length. These intrachromosomal rearrangements could result from contigs that were not well placed because of Hi-C signals that were not strong enough to correctly determine their position and orientation; however, they may also result from the six species' distinct genome evolutions.

Organizational conservation of chromosomes 2, 4, 6, 8, 9, 10, and 11 among the six *Syzygium* species studied constitutes new evidence supporting the 10 intrachromosomal rearrangements previously reported on these chromosomes between *S. aromaticum* and *E. grandis* genomes (Ouadi et al., 2022). These 10 rearrangements were also observed when aligning the DNA sequences of the chromosomes of *E. grandis* with those of *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides*. Among the rearrangements reported between the chromosomes of *S. aromaticum* and *E. grandis*, similar large terminal inversions on chromosomes 4, 9, 10, and 11 were also reported in the two eucalypts *E. grandis* and *C. citriodora* suggesting that these terminal inversions occurred on *E. grandis* chromosomes (Butler et al., 2017). Two other large terminal inversions were detected between chromosomes 4 and 9 of *S. aromaticum* and *E. grandis* but not between *C. citriodora* and *E. grandis*. These inversions were also observed when comparing the chromosome sequences of *E. grandis* with those of the four newly assembled genomes, suggesting that these inversions resulted from an evolution of the chromosome organization rather than from sequencing and assembly artifacts. Further comparative genomics analyses will be needed with additional *Syzygium* and Myrtaceous species to determine if these inversions are specific to the *Syzygium* genus or subgenus, for which the crown ages were estimated at 51.2 Mya and 9.4 Mya, respectively, (Low et al., 2022).

The analyses of the phylogenetic relationships between gene sequences of *S. aromaticum*, *S. malaccense*, *S. aqueum*, *S. jambos*, and *S. syzygioides* and comparisons of their full-length LTR-RTs repertoires provided insights into the distinct genome evolution of each species following the divergence of the *Syzygium* subg. *Syzygium* species 9.4 Mya (Low et al., 2022). The species tree inferred by OrthoFinder indicated that pairwise *S. aromaticum* and *S. aqueum* and *S. malaccense* and *S. jambos* were closely related, which is consistent with the genome-level phylogenetic trees generated by Low et al. (Low et al., 2022). We observed older waves of LTR-RTs Gypsy and Copia insertions in *S. aromaticum* and fewer insertions less than 1 million years old in the *S. aromaticum* chromosomes compared to those of the four other species studied. In plants, the RNA Directed DNA Methylation (RdDM) pathway, a *de novo* DNA methylation mechanism involving small interfering RNA, plays an important role in TE repression (Wambui Mbichi et al., 2020). Further detailed analysis such as DNA methylation studies will be valuable to clarify the molecular causes of the recent low insertion number of LTR-RTs elements observed in *S. aromaticum*.

S. aromaticum is cultivated to produce clove bud (the dried, unopened flower bud), essential oil (EO), and oleoresins rich in eugenol (Nurdjannah and Bermawie, 2012). The EO of *S. aromaticum* contains ~72 to 96.6% of eugenol, while the EO of *S. aqueum* has 0.19% eugenol (Razafimamonjison et al., 2014; Sobeh et al., 2016). Eugenol is a phenylpropane with multiple pharmaceutical activities and is considered a promising alternative drug for human health (e.g., cancer and pathogenic microorganism resistance, diabetes, obesity, and autoimmune diseases) (Kamatou et al., 2012; Batiha et al., 2020; Otunola, 2022). The genome assembly of *S. aromaticum* was exploited to investigate the genetic basis of this important characteristic. The identification of gene families involved in eugenol biosynthesis revealed the presence of multiple copies of genes encoding EGS, which catalyzes the synthesis of eugenol from coniferyl acetate. A cluster of 14 copies was reported on chromosome 10, and additional copies were located on chromosome 11 of *S. aromaticum*. In the genome assembly of the four newly sequenced species, we found fewer gene copies on chromosome 10 (1 to 3 copies) and no copies on chromosome 11 of *S. malaccense*, *S. aqueum*, and *S. jambos*. The presence of this structural variation suggested that a gene-dosage effect may be associated with the high amount of eugenol. Further studies are needed to elucidate the biological functions of the EGS gene copies in *S. aromaticum* and the four other species (e.g., *in vitro* characterization).

S. malaccense, *S. aqueum*, and *S. jambos* are grown for their edible fruit. Like *S. aromaticum* and other *Syzygium* species, they are also used in traditional medicine. Research on their numerous pharmaceutical properties has been undertaken (e.g., analgesic, anti-inflammatory, antioxidant, hepatoprotective, antidiabetic, antifungal, antibacterial, antiviral, and anticancer activities) (Nair, 2017; Cock and Cheesman, 2018). For instance, *S. jambos* is traditionally used to treat hemorrhages, wounds, and ulcers; *S. malaccense* is used to treat mouth ulcers and diabetes; and *S. aqueum* to treat diabetes and childbirth pain (Uddin et al., 2022). The chromosome-scale assemblies for these species are new valuable resources for the Myrtaceae family. Combined with other comparative genomics and multi-omics studies, they can be used to further investigate the genomic evolution of the Myrtaceous species and to study the genetic basis of important agronomical traits and biosynthesis of secondary metabolites.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA962868 https://www.ncbi.nlm.nih.gov/, PRJNA962711 https://www.ncbi.nlm.nih.gov/, PRJNA962713 https://www.ncbi.nlm.nih.gov/, PRJNA962712 https://www.ncbi.nlm.nih.gov/genbank/, JASUUE000000000 https://www.ncbi.nlm.nih.gov/genbank/, JASUUB000000000 https://

www.ncbi.nlm.nih.gov/genbank/, JASUUC000000000 https://www.ncbi.nlm.nih.gov/genbank/, JASUUD000000000 https://zenodo.org/, 7870328 https://zenodo.org/, 7870326 https://zenodo.org/, 7870330 https://zenodo.org/, 7870334.

Author contributions

SO performed the laboratory work, analyzed data, and wrote the manuscript. NS performed computational analysis of sequencing data, conceived, and supervised the study, and contributed to manuscript writing. FK, and NI conceived and supervised the study and contributed to manuscript writing. All authors contributed to the article and approved the submitted version.

Funding

The authors declare that this study received funding from the company Philip Morris International. The funder had the following involvement in the study: the study design, collection, analysis, interpretation of data, the writing of this article and the decision to submit it for publication.

Acknowledgments

We would like to thank Dr. Leyla Davis, curator of the Masoala Hall in the Zürich Zoo (Switzerland), for authorizing the sampling of the *Syzygium* trees for this project. We would like to also thank Remi Dulize for his technical contributions, and Lindsay Reese and Rebecca Higgins for manuscript revision.

Conflict of interest

Authors SO, NS, and NI were employed by the company Philip Morris International.

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1248780/full#supplementary-material

References

Batiha, G. E.-S., Alkazmi, L. M., Wasef, L. G., Beshbishy, A. M., Nadwa, E. H., and Rashwan, E. K. (2020). *Syzygium aromaticum* L.(Myrtaceae): Traditional uses, bioactive chemical constituents, pharmacological and toxicological activities. *Biomolecules* 10 (2), 202. doi: 10.3390/biom10020202

Beech, E., Rivers, M., Oldfield, S., and Smith, P. (2017). GlobalTreeSearch: The first complete global database of tree species and country distributions. *J. Sustain. For.* 36 (5), 454–489. doi: 10.1080/10549811.2017.1310049

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12 (1), 59–60. doi: 10.1038/nmeth.3176

Butler, J., Vaillancourt, R., Potts, B., Lee, D., King, G. J., Baten, A., et al. (2017). Comparative genomics of *Eucalyptus* and *Corymbia* reveals low rates of genome structural rearrangement. *BMC Genom.* 18 (1), 397. doi: 10.1186/s12864-017-3782-7

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17), i884–i890. doi: 10.1093/bioinformatics/bty560

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18 (2), 170–175. doi: 10.1038/s41592-020-01056-5

Christenhusz, M. J., and Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa* 261 (3), 201–217-201–217. doi: 10.11646/phytotaxa.261.3.1

Cock, I. E., and Cheesman, M. (2018). Plants of the genus *Syzygium* (Myrtaceae): A review on ethnobotany, medicinal properties and phytochemistry. *Bioactive Compounds Medicinal Plants: Properties Potential Hum. Health* 35–84. doi: 10.1201/b22426

Craven, L. A., and Biffin, E. (2010). An infrageneric classification of *Syzygium* (Myrtaceae). *Blumea-Biodiver. Evol. Biogeogr. Plants* 55 (1), 94–99. doi: 10.3767/000651910X499303

Ellestad, P., Pérez-Farrera, M. A., and Buerki, S. (2022). Genomic Insights into Cultivated Mexican Vanilla planifolia Reveal High Levels of Heterozygosity Stemming from Hybridization. *Plants* 11 (16), 2090. doi: 10.3390/plants11162090

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinf.* 9, 1–14. doi: 10.1186/1471-2105-9-18

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20 (1), 1–14. doi: 10.1186/s13059-019-1832-y

Feng, C., Feng, C., Lin, X., Liu, S., Li, Y., and Kang, M. (2021). A chromosome-level genome assembly provides insights into ascorbic acid accumulation and fruit softening in guava (*Psidium guajava*). *Plant Biotechnol. J.* 19 (4), 717–730. doi: 10.1111/pbi.13498

Frith, M. C. (2011). A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.* 39 (4), e23–e23. doi: 10.1093/nar/gkq1212

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28 (23), 3150–3152. doi: 10.1093/bioinformatics/bts565

Girgis, H. Z. (2015). Red: an intelligent, rapid, accurate tool for detecting repeats *de-novo* on the genomic scale. *BMC Bioinf.* 16 (1), 1–19. doi: 10.1186/s12859-015-0654-5

Goel, M., and Schneeberger, K. (2022). plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* 38 (10), 2922–2926. doi: 10.1093/bioinformatics/btac196

Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 20 (1), 1–13. doi: 10.1186/s13059-019-1911-0

Grattapaglia, D., Vaillancourt, R. E., Shepherd, M., Thumma, B. R., Foley, W., Külheim, C., et al. (2012). Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus. *Tree Genet. Genomes* 8 (3), 463–508. doi: 10.1007/s11295-012-0491-x

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 36 (9), 2896–2898. doi: 10.1093/bioinformatics/btaa025

Healey, A. L., Shepherd, M., King, G. J., Butler, J. B., Freeman, J. S., Lee, D. J., et al. (2021). Pests, diseases, and aridity have shaped the genome of *Corymbia citriodora*. *Commun. Biol.* 4 (1), 1–13. doi: 10.1038/s42003-021-02009-0

Hu, K., Xu, K., Wen, J., Yi, B., Shen, J., Ma, C., et al. (2019). Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. *BMC Bioinf.* 20 (1), 1–20. doi: 10.1186/s12859-019-2945-8

Hudson, C. J., Kullan, A. R., Freeman, J. S., Faria, D. A., Grattapaglia, D., Kilian, A., et al. (2012). High synteny and colinearity among Eucalyptus genomes revealed by high-density comparative genetic mapping. *Tree Genet. Genomes* 8 (2), 339–352. doi: 10.1007/s11295-011-0444-9

Izuno, A., Wicker, T., Hatakeyama, M., Copetti, D., and Shimizu, K. K. (2019). Updated genome assembly and annotation for *metrosideros polymorpha*, an emerging model tree species of ecological divergence. *G3-Genes Genom. Genet.* 9 (11), 3513–3520. doi: 10.1534/g3.119.400643

Kamatou, G. P., Vermaak, I., and Viljoen, A. M. (2012). Eugenol—from the remote Maluku Islands to the international market place: a review of a remarkable and versatile molecule. *Molecules* 17 (6), 6953–6981. doi: 10.3390/molecules17066953

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37 (5), 540–546. doi: 10.1038/s41587-019-0072-8

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2018). Current strategies of polyploid plant genome sequence assembly. *Front. Plant Sci.* 9, 1660. doi: 10.3389/fpls.2018.01660

Lexa, M., Jedlicka, P., Vanat, I., Cervenansky, M., and Kejnovsky, E. (2020). TE-greedy-nester: structure-based detection of LTR retrotransposons and their nesting. *Bioinformatics* 36 (20), 4991–4999. doi: 10.1093/bioinformatics/btaa632

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2*. doi: 10.48550/arXiv.1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34 (18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, F., Xu, S., Xiao, Z., Wang, J., Mei, Y., Hu, H., et al. (2023). Gap-free genome assembly and comparative analysis reveal the evolution and anthocyanin accumulation mechanism of Rhodomyrtus tomentosa. *Hortic. Res* 10 (3). doi: 10.1093/hr/uhad005

Li, F., Zhou, C., Weng, Q., Li, M., Yu, X., Guo, Y., et al. (2015). Comparative genomics analyses reveal extensive chromosome colinearity and novel quantitative trait loci in Eucalyptus. *PloS One* 10 (12), e0145144. doi: 10.1371/journal.pone.0145144

Low, Y. W., Rajaraman, S., Tomlin, C. M., Ahmad, J. A., Ardi, W. H., Armstrong, K., et al. (2022). Genomic insights into rapid speciation within the world's largest tree genus Syzygium. *Nat. Commun.* 13 (1), 1–15. doi: 10.1038/s41467-022-32637-x

Machado, R. M., and Forni-Martins, E. R. (2022). Psidium cattleyanum Sabine (Myrtaceae), a neotropical polyploid complex with wide geographic distribution: insights from cytogenetic and DNA content analysis. *Braz. J. Bot.* 45 (3), 943–955. doi: 10.1007/s40415-022-00829-w

Mak, Q. C., Wick, R. R., Holt, J. M., and Wang, J. R. (2023). Polishing de novo nanopore assemblies of bacteria and eukaryotes with FMLRC2. *Mol. Biol. Evol.* 40 (3), msad048. doi: 10.1093/molbev/msad048

Marcon, H. S., Domingues, D. S., Silva, J. C., Borges, R. J., Matioli, F. F., de Mattos Fontes, M. R., et al. (2015). Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic. *BMC Plant Biol.* 15 (1), 1–16. doi: 10.1186/s12870-015-0550-1

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., et al. (2014). The genome of Eucalyptus grandis. *Nature* 510 (7505), 356–362. doi: 10.1038/nature13308

Nair, K. N. (2017). *The genus Syzygium: Syzygium Cumini and Other Underutilized Species* (United States: CRC Press).

Neumann, P., Novák, P., Hoštáková, N., and 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, 1–17. doi: 10.1186/s13100-018-0144-1

Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M., and Iyer, M. K. (2017). TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods* 14 (1), 68–70. doi: 10.1038/nmeth.4078

Nurdjannah, N., and Bermawie, N. (2012). "Cloves," in *Handbook of herbs and spices* (Amsterdam, Neatherlands: Elsevier), 197–215.

Oginuma, K., Kato, A., Tobe, H., Mathenge, S., and Juma, F. (1993). Chromosomes of some woody plants in Kenya. *Acta Phytotax. Geobot.* 44 (1), 53–58.

Otunola, G. A. (2022). Culinary spices in food and medicine: an overview of Syzygium aromaticum (L.) Merr. and LM Perry [Myrtaceae]. *Front. Pharmacol.* 12, 3817. doi: 10.3389/fphar.2021.793200

Ouadi, S., Sierro, N., Goepfert, S., Bovet, L., Glauser, G., Vallat, A., et al. (2022). The clove (Syzygium aromaticum) genome provides insights into the eugenol biosynthesis pathway. *Commun. Biol.* 5 (1), 1–13. doi: 10.1038/s42003-022-03618-z

Panggabean, G. (1991). "Syzygium aqueum (Burm. f.) Alst., Syzygium malaccense (L.) M. & P, and Syzygium samarangense (Blume) M. & P. *Plant Resources of South-East Asia 2*," in *Edible fruits and nuts* (Pudoc, Wageningen: Pudoc Scientific Publishers), 292–294.

Parnell, J. A., Craven, L. A., and Biffin, E. (2007). "Matters of scale: dealing with one of the largest genera of angiosperms," in *Reconstructing the tree of life: taxonomy and systematics of species rich taxa* (Boca Raton, FL: CRC Press LLC), 253–270.

Pedrosa, A., Gitaí, J., Silva, A. E. B., Felix, L. P., and Guerra, M. (1999). Cytogenetics of angiosperms collected in the state of Pernambuco: V. *Acta Bot. Bras.* 13 (1), 49–60. doi: 10.1590/S0102-33061999000100006

Pellicer, J., and Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* 226 (2), 301–305. doi: 10.1111/nph.16261

Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research* 9 (304). doi: 10.12688/f1000research.23297.2

POWO (2023) *Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew*. Available at: http://www.plantsoftheworldonline.org/. Retrieved 11 April 2023.

Pucker, B., Irisarri, I., de Vries, J., and Xu, B. (2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quant. Plant Biol.* 3 (5), e5. doi: 10.1017/qpb.2021.18

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6), 841–842. doi: 10.1093/bioinformatics/btq033

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11 (1), 1–10. doi: 10.1038/s41467-020-14998-3

Razafimamonjison, G., Jahiel, M., Duclos, T., Ramanoelina, P., Fawbush, F., and Danthu, P. (2014). Bud, leaf and stem essential oil composition of Syzygium aromaticum from Madagascar, Indonesia and Zanzibar. *Int. J. Basic Appl. Sci.* 3 (3), 224. doi: 10.14419/ijbas.v3i3.2473

Saber, F. R., Munekata, P. E., Rizwan, K., El-Nashar, H. A., Fahmy, N. M., Aly, S. H., et al. (2023). Family Myrtaceae: The treasure hidden in the complex/diverse composition. *Crit. Rev. Food Sci. Nutr.*, 1–19. doi: 10.1080/10408398.2023.2173720

Shao, M., and Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* 35 (12), 1167–1169. doi: 10.1038/nbt.4020

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One* 11 (10), e0163962. doi: 10.1371/journal.pone.0163962

Shen, C., Li, L., Ouyang, L., Su, M., and Guo, K. (2023). E. urophylla× E. grandis high-quality genome and comparative genomics provide insights on evolution and diversification of eucalyptus. *BMC Genom.* 24 (1), 1–10. doi: 10.1186/s12864-023-09318-0

Shi, J., and Liang, C. (2019). Generic repeat finder: a high-sensitivity tool for genome-wide de novo repeat detection. *Plant Physiol.* 180 (4), 1803–1815. doi: 10.1104/pp.19.00386

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

Sobeh, M., Braun, M. S., Krstin, S., Youssef, F. S., Ashour, M. L., and Wink, M. (2016). Chemical profiling of the essential oils of Syzygium aqueum, Syzygium samarangense and Eugenia uniflora and their discrimination using chemometric analysis. *Chem. Biodivers.* 13 (11), 1537–1550. doi: 10.1002/cbdv.201600089

Thrimawithana, A. H., Jones, D., Hilario, E., Grierson, E., Ngo, H. M., Liachko, I., et al. (2019). A whole genome assembly of Leptospermum scoparium (Myrtaceae) for mānuka research. *N. Z. J. Crop Hortic. Sci.* 47 (4), 233–260. doi: 10.1080/01140671.2019.1657911

Tuler, A. C., Carrijo, T. T., Peixoto, A. L., Garbin, M. L., da Silva Ferreira, M. F., Carvalho, C. R., et al. (2019). Diversification and geographical distribution of Psidium (Myrtaceae) species with distinct ploidy levels. *Trees* 33 (4), 1101–1110. doi: 10.1007/s00468-019-01845-2

Uddin, A. N., Hossain, F., Reza, A. A., Nasrin, M. S., and Alam, A. K. (2022). Traditional uses, pharmacological activities, and phytochemical constituents of the genus Syzygium: A review. *Food Sci. Nutr.* 10 (6), 1789–1819. doi: 10.1002/fsn3.2797

Van Lingen, T. (1991). "Syzygium jambos (L.) Alston. Plant Resources of South-East Asia 2," in Edible fruits and nuts (Pudoc, Wageningen: Pudoc Scientific Publishers), 296–298.

Wambui Mbichi, R., Wang, Q.-F., and Wan, T. (2020). RNA directed DNA methylation and seed plant genome evolution. *Plant Cell Rep.* 39, 983–996. doi: 10.1007/s00299-020-02558-4

Warren, R. L., Coombe, L., Mohamadi, H., Zhang, J., Jaquish, B., Isabel, N., et al. (2019). ntEdit: scalable genome sequence polishing. *Bioinformatics* 35 (21), 4430–4432. doi: 10.1093/bioinformatics/btz400

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8 (12), 973–982. doi: 10.1038/nrg2165

Wilson, P. G. (2010). "Myrtaceae," in *Flowering Plants. Eudicots* (Berlin, Heidelberg: Springer), 212–271.

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., et al. (2022). TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* 9. doi: 10.1093/hr/uhac017

Zheng, X., Chen, X., Lin, G., Chen, J., Li, H., Xiao, Y., et al. (2022). The chromosome-level Melaleuca alternifolia genome provides insights into the molecular mechanisms underlying terpenoids biosynthesis. *Ind. Crops Prod.* 189, 115819. doi: 10.1016/j.indcrop.2022.115819

Zhou, C., McCarthy, S. A., and Durbin, R. (2022). YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* 39 (1). doi: 10.1093/bioinformatics/btac808

Zhou, S.-S., Yan, X.-M., Zhang, K.-F., Liu, H., Xu, J., Nie, S., et al. (2021). A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. *Sci. Data* 8 (1), 174. doi: 10.1038/s41597-021-00968-x