NICOLAUS COPERNICUS UNIVERSITY IN TORUŃ

Building a Model: Developing Genomic Resources for *Ferula communis* (Apiaceae), a Traditional Medicinal Plant

Author:

Mergi Daba Dinka

Doctoral Supervisor:

Professor dr hab. Krzysztof Szpila

Co-supervisor:

Dr. Marcin Piwczynski

A dissertation submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy in the

Department of Ecology and Biogeography

Faculty of Biological and Veterinary Sciences

Nicolaus Copernicus University in Toruń

Toruń, November 2023

This study was supported by the National Science Center (NCN) grant number **2015/18/E/NZ8/00716**.

Author’s statement: I affirm that this thesis has been authored entirely by me.

(Oświadczenie autora rozprawy: oświadczam, że niniejsza rozprawa została napisana przeze mnie samodzielnie)

_____	_____
Date	The author’s signature

Doctoral Supervisor’s Statement:

_____	_____
Date	The advisor’s signature

Doctoral co-supervisor’s Statement:

_____	_____
Date	The co-advisor’s signature

Abstract

Angiosperms, representing the most diverse group of land plants, have complex genomes that are believed to underpin their evolutionary success. However, due to the absence of data for numerous angiosperm groups, our understanding of genome architecture remains limited. In this work, I address this knowledge gap by assembling and analyzing the genome of *Ferula communis*, a species from the economically important plant family Apiaceae.

The nuclear genome of *F. communis* was assembled utilizing sequences obtained from next-generation sequencing technologies, encompassing both Illumina short-read and Oxford Nanopore Technology (ONT) libraries. A wide spectrum of bioinformatic tools, including short-read, long-read, and hybrid genome assemblers, were used in this process. The most superior assembly, based on multiple metrics, was produced by the long-read assembler Flye. This resulted in a genome size of 2.8 Gb, composed of 59,178 contigs, an N50 of 0.17 Mb, and a genome completeness covering 93.2% of BUSCO genes. Notably, transposable elements constituted 86.67% of the genome, with *Gypsy* and *Copia* retrotransposons being particularly prevalent, highlighting their pivotal role in genome evolution. The functional genome annotation identified 68,318 protein-coding genes, and 1,262 tRNA genes. When compared to the genomes of *Daucus carota* (carrot), *Apium graveolens* (celery), and *Corindrum sativum* (coriander) from the Apiaceae family, *F. communis* exhibits a higher count of annotated genes and a more extensive gene family expansions. Within these expanded gene families, I identified genes tied to defense mechanisms, stress responses, and vernalization. Surprisingly, *F. communis* shares more unique genes with the distantly related coriander than with carrot and celery. A notable instance is the genes tied to terpenoid biosynthesis. The probable reason for this enhanced gene sharing with coriander over other studied Apiaceae members is genome reduction in carrot, leading to extensive gene deletion. Conversely, in celery, its primary aquatic habitat may have necessitated a specific gene set distinct from those required in terrestrial environments.

The plastid genome of *F. communis* was assembled using two assemblers: GetOrganelle and Novoplasty. Spanning 166,696 bp, it encodes 132 genes, comprising 87 protein-coding, 37 tRNA, and 8 rRNA genes. This genome reflects the standard quadripartite structure observed in other angiosperm plastid genomes, which includes paired IR, LSC, and SSC regions. The *F. communis* plastid DNA houses 217 simple sequence repeats (SSRs), with a predominant concentration in the LSC region. Several intergenic spacer (IGS) regions demonstrated pronounced nucleotide diversity and have potential as informative phylogenetic markers. Interestingly, only the *ccsA* gene, which is responsible for heme attachment to cytochrome c, displayed sites of strong positive selection within the *F. communis* plastid genome.

The mitochondrial genome was assembled using GetOrganelle and analyzed for gene content and structural variations. Interestingly, rather than forming a single master circle, the *F. communis* mtDNA genome assembly resulted in 16 scaffolds with a total genome size of 250,278 bp. This suggests a non-circular genome structure in this species. Within these scaffolds, 37 protein-coding genes, 3 rRNA genes, and 20 tRNA genes are present. Among the annotated genes, eight of the protein-coding genes and three tRNAs contain introns of varying lengths. In agreement with other assembled mtDNA genomes in the Apiaceae family, 183 SSRs were identified, 82% of which are mono- and dimeric repeats. Moreover, 385 RNA editing sites were identified within 25 protein-coding genes. Of these edits, 61.14% transition amino acids from hydrophilic to hydrophobic states, while 31.09% involve alterations between hydrophobic amino acids.

The results presented here establish a foundation and provide a valuable genomic resources for future investigations into the genomics, evolution, and medicinal properties of this important plant species.

Streszczenie

Rośliny okrytozalążkowe reprezentują najbardziej zróżnicowaną grupę roślin lądowych. Za jedną z przyczyn ich sukcesu ewolucyjnego uważa się bardzo specyficzną budowę genomów charakteryzującą się dużą plastycznością. Jednak z powodu braku danych dla większości grup okrytonasiennych nasze zrozumienie architektury genomu jest na dzień dzisiejszy ograniczone. W tej pracy podjąłem badania, których celem jest wypełnienie luki w naszej wiedzy o genomach roślin okrytozalążkowych, składając i analizując genom *Ferula communis*, gatunku z ekonomicznie ważnej rodziny baldaszkowatych (Apiaceae).

Genom jądrowy *F. communis* został złożony z wykorzystaniem sekwencji uzyskanych dzięki technologii sekwencjonowania nowej generacji, obejmującej zarówno technologię Illumina, jak i Oxford Nanopore Technology (ONT). W procesie składania genomu użyto szerokiego spektrum narzędzi bioinformatycznych, w tym programów wykorzystujących tylko krótkie fragmenty, tylko długie oraz oba typy odczytów. Najlepszy jakościowo genom został złożony przez program Flye wykorzystujący tylko długie odczyty z technologii ONT. Wielkość uzyskanego genomu wynosiła 2,8 Gb, na który składało się 59 178 kontigów. Jakość genomu oszacowana na podstawie metryki N50 wynosiła 0,17 Mb, zaś kompletności genomu bazująca na genach BUSCO została określona na 93,2%. Elementy mobilne stanowiły 86,67% genomu, przy czym retrotranspozony *Gypsy* i *Copia* były szczególnie liczne, co podkreśla ich kluczową rolę w ewolucji genomu *F. communis*. Dzięki funkcjonalnej adnotacji genomu zidentyfikowałem 68 318 genów kodujących białka i 1 262 genów tRNA. W porównaniu z genomami marchwi, selera i kolendry należących do rodziny baldaszkowatych, *F. communis* wykazuje większą liczbę unikalnych genów oraz charakteryzuje się większą ekspansją rodzin genów. Wśród tych rozszerzonych rodzin zidentyfikowałem geny związane z mechanizmami obronnymi, odpowiedziami na stres i wernalizacją. Co ciekawe, *F. communis* dzieli więcej genów z

odlegle spokrewnioną kolendrą niż z marchwią i selerem. Przykładem są geny związane z biosyntezą terpenoidów. Prawdopodobną przyczyną tego zaskakującego wyniku jest redukcja genomu u marchwi, prowadząca do redukcji liczby genów. Natomiast u selera, jego pierwotne środowisko wodne mogło wymagać specyficznego zestawu genów, który różni się od tego wymaganego w środowiskach lądowych.

Genom plastydowy *F. communis* został złożony przy użyciu dwóch programów: GetOrganelle i Novoplasty. Złożony genom obejmuje 166 696 pz i koduje 132 geny, w tym 87 kodujących białka, 37 tRNA i 8 genów rRNA. Genom *F. communis* odzwierciedla standardową strukturę czteroczęściową obserwowaną w innych genomach plastydowych okrytonasiennych, która obejmuje regiony IR, LSC i SSC. W genomie znajduje się 217 sekwencji mikrosatelitarnych (SSR), szczególnie w regionie LSC. Kilka spośród przestrzeni międzygenowych charakteryzowało się dużą różnorodnością nukleotydową co umożliwia potencjalne wykorzystanie ich jako znaczników filogenetycznych. Wśród wszystkich genów kodujących białka w genomie plastydowym *F. communis*, tylko gen *ccsA*, który odpowiada za przyłączanie hemu do cytochromu c, podlega silnej selekcji pozytywnej.

Genom mitochondrialny został złożony przy użyciu programu GetOrganelle i przeanalizowany pod kątem zawartości genów i wariantów strukturalnych. Analiza sekwencji mitochondrialnych dała wynik w postaci 16 kontigów o łącznej długości 250 278 pz co wskazuje, że genom *F. communis* nie występuje w postaci jednej kolistej cząsteczki DNA. W obrębie tych kontigów udało się scharakteryzować 37 genów kodujących białka, 3 geny rRNA i 20 genów tRNA. Spośród adnotowanych genów, osiem genów kodujących białka i trzy tRNA zawierały introny o różnych długościach. W genomie zidentyfikowano 183 sekwencji mikrosatelitarnych, z których 82% to powtórzenia mono- i dimerowe. Ponadto, określono 385 miejsc edycji RNA w 25 genach kodujących białka. 61,14% edytowanych miejsc prowadzi do zmiany aminokwasu z hydrofilowego na hydrofobowy, zaś 31,09% obejmuje zmiany między aminokwasami hydrofobowymi.

Prezentowane tutaj wyniki stanowią podstawę i dostarczają cenne informacje dla przyszłych badań nad genomiką, ewolucją i właściwościami leczniczymi tego ważnego gatunku z rodziny baldaszkowatych.

Acknowledgments

As I bring this academic chapter of my life to a close, I do so with a profound sense of accomplishment and satisfaction. This journey has been extensive and rewarding, and I extend my heartfelt gratitude to all those who have been instrumental in my pursuit of this PhD degree.

First and foremost, I want to extend my gratitude to Professor dr hab. Krzysztof Szpila for serving as my main supervisor during my PhD journey. Your presence and support, though not directly involved in the research, have provided valuable guidance and a supportive academic environment.

I would like to express my sincere gratitude to my co-supervisor, Dr. Marcin Piwczynski, for granting me the privilege to embark on my doctoral journey under his guidance. His consistent support, mentorship, astute guidance, constructive feedback, and unwavering dedication to enhancing the caliber of my research, as well as his wholehearted commitment to my academic and personal development throughout this research endeavor, have been truly invaluable. His wisdom and encouragement have left an indelible mark on my academic journey.

To my beloved wife, Chaltu Zergaw, I offer my profound thanks for your unwavering support during the extraordinary journey of pregnancy and childbirth, even when I couldn't be there physically, is a testament to your incredible strength and love. You not only bore our child but also carried me through some of life's most challenging moments. Your love and resilience are the pillars of our bond, and I'm forever grateful for your extraordinary heart. This accomplishment is as much yours as it is mine. I love You. To my cherished daughter, Hilani, whose boundless enthusiasm and endless smiles brought light to my life and served as a reminder of the importance of balance and family.

I would like to extend my profound gratitude to Dr. Paulina Trzeciak. Her continuous mentorship in the molecular laboratory, assistance with administrative tasks, and, above all, her

kindness, helpful personality, and support during critical times have been greatly appreciated. Dr. Andrzej Grzywacz, with his supportive nature, patience, and willingness to allow me to use his computer for analyzing my data (a time-consuming task), and reviewing the thesis has not gone unnoticed. To my fellow PhD students, Madalina Oana Popa, Kinga Walczak, and Drashti Parmar, I want to convey my heartfelt appreciation for your positive attitude, help, and support during my doctoral studies.

My heartfelt appreciation also goes out to my extended family, my father Daba Dinka, my mother Tsige Ormamena, my brother Beri, my sisters Genet, Tigist, and Ayantu; father and mother-in-law Zergaw Nine and Beshera Bati. I am also grateful for the support of my friends who have been my source of strength throughout this academic journey. This PhD thesis would not have been possible without the collective support and encouragement of all these wonderful individuals. Thank you from the bottom of my heart.

Last but not least, I would like to acknowledge the divine presence of Almighty God in my life. Your guidance, strength, and faith have been my source of resilience and hope.

ምስጋና

ከሁሉ አስቀድሜ በነገር ሁሉ የረዳኝን በተከበበ ከተማ አስደናቂ ምህረቱን ከኔ ያላራቀ፤ የህይዋቴ ጀማሪ እና ፈፃሚ የሆነውን እግዚአብሔርን አመሰግናለሁ።

ይህንን የህይወቴን የትምህርት ምዕራፍ ሳጠናቅቅ፣ ፒኤችዲ ዲግሪዬን ለመከታተል አስተዋፅዖ ላበረከቱት ሁሉ ልባዊ ምስጋናዬን አቀርባለሁ። በመጀመሪያ ደረጃ ለአማካሪዬ **ዶክተር ማርቺን ፒሺቺኝስኪ** ያለኝን ልባዊ አድናቆት መግለጽ እፈልጋለሁ። በሱ አማካሪነት የዶክትሬት ትምህርት ጉዞዬን እንድጀምር እድል ሰለሰጠኝ የእሱ ያልተቋረጠ ድጋፍ፣ አስተዋይ መመሪያ፣ ገንቢ አስተያየት እና የጥናቴን ጥራት ደረጃ ከፍ ለማድረግ ያላሰለሰ ቁርጠኝነት፣ እንዲሁም በዚህ ትምህርት የምርምር ጥረቴ እና ለግል ሕይወት እድገቴ ላበረከተው መልካም ተፃእኖ ሳላመሠግነው አላልፊም። የእሱ ድጋፉ እና ማበረታቻ በዚህ ጉዞዬ ላይ የማይጠፋ አሻራ ኣሳረፉአል።

ለምወዳት ባለቤቴ **ማር (ጫልቱ ዘርጋው)** ፣ ለምርምር ባደረኩት ብዙ ምሽቶች ቅዳሜና እሁዶች በእርግዝና ወራት እና በወለድሽበት ጊዜ አብሬሽ ባለመሆኔ ብፀፀትም በዛ ሁሉ ጭንቅ ውሰጥ ሆነሽ እኔን ሰታበረታችኝ ነበር ለዚህ ሁሉ ድጋፍሽ ጥልቅ ምስጋናዬን አቀርባለሁ። የአንቺ ፍቅር እና ማበረታቻ ቃላቶችሽ ሰኬታማ እንድሆን ኣድርገውኛል አፈቅርሻለሁ ። ለምወዳት ልጄ **ሂላኒ** ፣ ገደብ የለሽ ጉጉትሽ እና ማለቂያ የለሽ ፈገግታሽ በህይወቴ ላይ ብርሃን እናም ሚዛናዊ የቤተሰብን አሰፈላጊነት ምን ያህል ጠቃሚ እንደሆነ ሁልግዜ እንዳሰታውሰ ምክንያት ሰልሆንሽኝ እግዚአብሄር በዘመንሽ ሁሉ ይባርክሽ።

ለዶክተር **ፓውሊና ትሼቻክ** ታላቅ ምስጋናዬን ማቅረብ እፈልጋለሁ። ይህን ምርምር ሰከውን በሞለኪውላር ቤተ ሙከራ ውሰጥ ላበረከትሽው መልካም አስተዋፃኦ፣ በአስተዳደራዊ ተግባራት እርዳታ ከሁሉም በላይ አጋዥ ሰብዕናዋ እና ደግነቷ በአስቸጋሪ ጊዜያት ያበረከተችልን ድጋፍ አድናቆት ተችሯታል። ለዶ/ር **አንዜይ ግዢቫች** ያለው የደጋፉነት ባህሪው፣ ትዕግስት እና ኮምፒውተሩን ተጠቅሜ ረጂም ጊዜ የሚወስደውን ውሂቤን እንድመረምር ሰለፈቅደልኝ ልባዊ ምስጋናዬን አቀርባለሁ። ለዶክትሬት ተማሪዎች እና ጓደኞቼ ፣ **ማዳሊና ኦዋና ፖፓ፣ ኪንጋ ቫልቻክ** እና **ድራሸቲ ፓርማር**፣ በዶክትሬት ትምህርቴ ወቅት ላደረጋችሁት አዎንታዊ አመለካከት፣ እርዳታ እና ድጋፍ ያለኝን አድናቆት ማሳየት እፈልጋለሁ።

ልባዊ አድናቆት ለሚቸራቸው ቤተሰቦቼ ለአባቴ **ዳባ ዲንቃ፣** እናቴ **ጽጌ ኦርማመና፣** ወንድሜ **በሪ፣** እህቶቼ **ገነት፣ ትግስት** እና **አያንቱ፤** እንደ አማችም እንደ አባትም **አባቢ** (ዘርጋው ኒኔ) እና እንደ አማችም እንደ እናትም **አዬ** (በሼራ ባቲ) ከነልጆቻቸው እንዲሁም በዚህ ትምህርት ጉዞዬ ሁሉ የጥንካሬ ምንጭ ለነበሩት ጓደኞቼ ለሰጡኝ ድጋፍ አመሰግናለሁ።

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Whole genomes as a key to understand evolution of flowering plants

Angiosperms are the most diverse and species-rich group of land plants, that contribute substantially to global photosynthesis and carbon sequestration (Judd et al., 1999). With over 350,000 species (Kenneth, 2013), they occupy nearly every habitat, from forests and grasslands to sea margins, and are adapted to a wide range of environments, including extreme ones such as marine, arctic, alpine, and desert, which are largely uncolonized by other vascular plants (Folk et al., 2020). Angiosperms exhibit a wide variety of life forms, including trees, shrubs, herbs, submerged aquatics, vines and epiphytes (Yang et al., 2020). As a result, they are the most important source of food for animals, including humans. In addition, flowering plants are the most economically important group of green plants, serving as a source of pharmaceuticals, fibre products, timber, ornamentals, and other commercial products. Despite their importance, the origin and extraordinary diversification of flowering plants remain poorly understood.

In recent years, however, the development of massively parallel sequencing approaches has shed light on the evolution of angiosperms, revealing unprecedented genomic diversity that is key to understand their current domination in almost all ecosystems (Pellicer et al., 2018). One of the main discoveries in the past 20 years of plant genome studies is that all flowering plants are polyploids, having gone through multiple whole genome duplication (WGD) events. Each of these events was superimposed on earlier duplications, the earliest of which occurred approximately at the root of the seed plants. However, the direct relationship between WGD and evolutionary diversification remains debatable. For example, the analysis of recent WGD events has shown lower net diversification rates in polyploids than in their diploid relatives (Mayrose et al., 2011; Scarpino et al., 2014). Similarly, there was no significant difference in the number of diploid and polyploid species in sister clades in various

plant taxa (Wood et al., 2009). In contrast, analyses of ancient WGDs mapped on the angiosperm phylogeny showed that out of 106 events, 61 were associated with changes in the rate of species diversification (Landis et al., 2018). Interestingly, WGDs are not distributed evenly throughout time. The majority of them are associated with one of three waves of duplication events. The most ancient wave occurred around 120 Mya in the early evolutionary history of eudicots and monocots. The second wave, which includes the highest number of WGDs, occurred at the K-Pg boundary (Cretaceous-Paleogene), while the third one within the last 20 Myr. However, there are no simple mechanisms linking WGD events with any morphological innovations found in angiosperms. For example, the origin of flowering plants was not associated with any of the gene families' expansion, but rather with contraction as seen in the case of C_2H_2 transcription factor family, glycosyltransferase GT1 family, or leucine rich repeat (NBS-LRR) domains (Leebens-Mack et al., 2019). This result is consistent with the hypothesis that innovations in angiosperms may have involved the functional co-option of genes that were duplicated earlier in the evolutionary history of seed plants. Additionally, Clark and Donoghue (2017) showed, using more precise molecular dating techniques than earlier authors, that some of the main WGD events predated diversification bursts by dozens of millions of years, precluding their deterministic role in the evolution of innovations.

Compared to other land plants, such as ferns and gymnosperms, angiosperms have on average 2.5–3.5 times smaller genome size (Table 1, Benton et al., 2021). This is related with a small average length and low number of chromosomes in angiosperms, with an ancestral haploid chromosome number estimated at n = 7 (Carta et al., 2020). In contrast, gymnosperms distribute their larger genomes among larger chromosomes (Leitch & Leitch, 2012), whereas ferns retain duplicated genomes in a larger number of chromosomes (Wu et al., 2019). At first glance, the large number of WGD events and a relatively small average genome size in flowering plants compared to other taxonomic groups might seem contradictory. However, the balance between WGD and small genomes is assured by processes,

which are collectively called post-polyploidization diploidization (PPD), which includes repetitive DNA loss, chromosome rearrangements, and a complex pattern of gene loss. PPD mechanisms are now considered one of the key properties of angiosperms that allow for accelerated genome rearrangements after WGD events (Dodsworth et al., 2016). This unusual genomic plasticity may have driven diversification into ecological niches, unavailable for other plant groups, by pulse or stepwise bursts of speciation events (Dodsworth et al., 2016).

Table 1. The comparison of genome sizes among various groups of land plants

	No. of recognized species	No. of species with genome size data	Representation (%)	Min. (pg)	Max. (pg)	Mean (pg)	Std Dev	Range (max./min.)
Non-vascular plants								
Bryophyte	c. 20,000	334	1.7	0.16	20.46	0.92	1.75	128-fold
Vascular plants								
Pteridophytes	c. 13,000	303	2.3	0.08	150.61	12.11	13.84	1,883-fold
Seed plants								
Gymnosperms	c. 1,000	421	42.1	2.25	36.00	18.35	7.31	16-fold
Angiosperm	c. 352,000	10,770	3.05	0.065	152.23	5.13	8.94	2,342-fold

Source: The plant list (http://www.theplantlist.org/1.1/browse/B/) and Royal Botanical Kew garden (https://cvalues.science.kew.org/search, access date 30 March 2022)

Besides WGD events, the genome size in angiosperms is largely determined by transposable elements (TE), which often constitute the major fraction of plant DNA (up to 84%) (Oliver et al., 2013). The burst of TE movements is often triggered by polyploidization or hybridization, leading to genomic modification that includes changes in genome size. In addition, TE proliferation can increase genome size without any obvious causative event. The movements of TE within the genome have a tremendous impact on genomic and transcriptomic variation. Regions enriched with TEs are often

associated with chromatin modification that is responsible for transcriptional silencing or DNA methylation (Sahebi et al., 2018). TE also contribute to coding genes variation by forming new introns, exons or chimeric genes. Although rare, TE might be expected to form completely new genes. Recently, there have been several reports of TE-derived genes in angiosperms, but not gymnosperms, indicating that some TE characteristics are unique to flowering plants and contribute to their success (Joly-Lopez et al., 2012; Knip et al., 2012).

1.1.1. General properties of plant genomes

As mentioned in the previous section, polyploidy, or WGD, is far more important in the evolutionary history of angiosperms than was previously recognized and is now considered a main characteristic that has shaped the angiosperm genomes for millions of years (Landis et al., 2018). Polyploidy can arise through several processes, the most prevalent of which are autopolyploidy, which involves doubling of a genome within a single individual or doubling through hybridization among individuals of the same species, and allopolyploidy, which involves an increase number of genomes by interspecific or even intergeneric hybridization (del Pozo and Ramirez-Parra, 2015). The ubiquity of repeating WGD events in angiosperms has led to enormous genome size variation in this group. Among studied land plants, angiosperms (Pellicer and Leitch, 2020), have the smallest reported genome size which belongs to the carnivorous plant *Genlisea tuberosa* with a C-value of 0.0623 pg, equivalent to 61 Mbp (Fleischmann et al., 2014). The largest reported genome size belongs also to angiosperm *Paris japonica* (Pellicer et al., 2010), a monocot with a C-value of 152,000 pg (149,000 Mbp).

The process of genome doubling theoretically confers several advantages to a polyploid, including the evolution of novel gene functions through neofunctionalization. However, the recent studies have shown that the prevalent fate of duplicated genes in angiosperms is silencing by mutation or epigenetic modification (Jiao and Paterson, 2014). Interestingly, the process of genome

rearrangements after WGD is not random for all duplicated genes. According to Jiao and Paterson (2014), three categories has been defined based on 'fates' of individual gene pairs:

(1) The majority of genes are randomly lost or preserved after WGD at a rate indistinguishable from the genome-wide average.

(2) Genes representing specific categories, such as those involved in signal transduction and transcription, are preferentially preserved in duplicate.

(3) Other specific genes, especially those having broader expression domains and higher expression levels than genes retained in duplicate, tend to be restored to a single copy status. Examples of these genes include those involved in chloroplast-related functions or DNA repair. Furthermore, genes having TE in their adjacency and may have a repressive effect on expression are more likely to be silenced.

Recent studies have shown that the forces underlying the fate of individual gene copies include selection to maintain balanced protein interactions and higher-order interactions in molecular networks (Wendel et al., 2016). For example, genes encoding monomeric proteins with few interacting partners or genes that function downstream of biological networks are less constrained than genes with numerous of protein-protein interactions, which function as hubs in biological pathways (Wendel et al., 2016).

The major fraction of the angiosperm genome consists of repetitive elements, which may account up to 90% of genome size, rather than functional genes. Due to the differential evolution of these repetitive elements and their functional significance, independent of genome size, there is no simple correlation between their number and number of whole-genome duplication (WGD) events. Repetitive DNA can broadly divided into two broad categories of repeat types: dispersed mobile elements and tandem repeats (Bennetzen and Wang, 2014). Tandem repeats include the ribosomal DNA and satellite DNA. Ribosomal DNA comprises two elements 5S rDNA and 18S-5.8S-26S known also as 35S rDNA. Both elements are present in high copy numbers to meet the high cellular requirement

for ribosomes. However, they are usually clustered at different chromosomal loci, i.e. there is typically a spatial separation between them on chromosomes (Garcia et al., 2014). Satellite DNA, on the other hand, are long arrays of repeated sequence monomers found in centromeric, telomeric and interstitial chromosomal regions of the genome. These elements are heterogeneous elements which can be divided into families that differ in location, nucleotide sequence, sequence complexity, repeat unit length and abundance. The majority of these families have high rates of genomic change and are usually either species specific or shared by a particular phylogenetic lineage, although some satellite DNA can be conserved over long evolutionary times (Garrido-Ramos, 2015). For decades, the satellite DNA was considered to be 'junk DNA'. However, recent studies have shown that they are important structural elements of heterochromatic regions of centromers and likely play a role in chromosome segregation during mitosis and meiosis (Plohl et al., 2014). Additionally, their role during chromosome organization and pairing has also been suggested (Plohl et al., 2012).

Among the repetitive sequences, tandem repeats typically constitute a small portion of the genome, while TE are responsible for the majority of the repetitions in angiosperm genomes (Wicker et al., 2018). TEs are classified according to their transposition mechanism into two classes: class I, also called retrotransposons, and class II, often simply called DNA transposons (Table 2) (Bennetzen and Wang, 2014). Retrotransposons transpose through a process involving an RNA intermediate (a 'copy-and-paste' mechanism), thereby amplifying their copy count during each transposition event (Fig. 1a) (Macko-Podgorni et al., 2013). Meanwhile, most class II DNA transposons transpose via a DNA intermediate—transposons are excised from the chromosome using the transposase enzyme and subsequently integrated elsewhere in the genome by the same transposase (Fig. 1b) (Bennetzen and Wang, 2014).

Based on sequence similarity, various structural features and phylogenetic relations among reverse transcriptase, class I retrotransposons are classified into five orders: LTR retrotransposons,

DIRS-like elements, Penelope-like elements, LINEs, and SINEs (Table 2, Wicker et al., 2007).

Table 2. Classification of transposable elements (TEs) and their transpose mechanism in angiosperms according to Feschotte et al. (2002), Wicker et al. (2007), Makałowski et al. (2019), and Macko-Podgórni et al. (2021)

Class	Order	Superfamily	Transpose mechanism	TSD (bp)
Class I (retrotransposons)	LTR	*Copia*	Autonomous	4-6
		Gypsy	Autonomous	4-6
	IDRS	*DIRS*	Autonomous	0
	PLE	*Penelope*	Autonomous	variable
	LINE	*L1*	Autonomous	variable
		I	Autonomous	variable
	SINE	*tRNA*	Non-autonomous	5-15
		7SL	Non-autonomous	variable
Class II (DNA transposons) Subclass 1	TIR	*Tc1-Mariner*		
		hAT:		
		• *Ac*	Autonomous	8
		• *Ds*	Non-autonomous	
		Mutator		9-11
		P		8
		PIF-harbinger:		
		• *PIFa*	Autonomous	3
		• *mPIF*	Non-autonomous	
		CACTA:		
		• *Spm*	Autonomous	2-3
		• *CAC1*	Autonomous	
		• *dSpm*	Non-autonomous	
		• *CAC2*	Non-autonomous	
Class II (DNA transposons) Subclass 2	Helitron	*Helitron*		0

(LTR-long terminal repeats, DIRS-Dictyostelium intermediate repeat sequence, PLE-Penelope-Like elements, LINE-long interspersed nuclear element, SINE-short interspersed nuclear element, TIR-terminal inverted repeat, TSD-target site duplication, MITEs-Miniature inverted repeat transposable elements)

The LTR retrotransposons are less common in animals but are the most prevalent order in plants. They consist of two major superfamilies, *Gypsy* and *Copia,* which range in size from a few hundred base

pairs up to 25 kb. Upon integration, they generate a target site duplication (TSD) of 4-6 bp flanking the element. The most distinctive class I retrotrasposons is the DIRS order (Goodwin and Poulter, 2004). This order differs in structure from other retrotrasposons, encodes a distinct complement of proteins, has a different replication mechanism, and does not produce a TSD upon integration because it uses tyrosine recombinase instead of an integrase for transposition (Goodwin and Poulter, 2004).

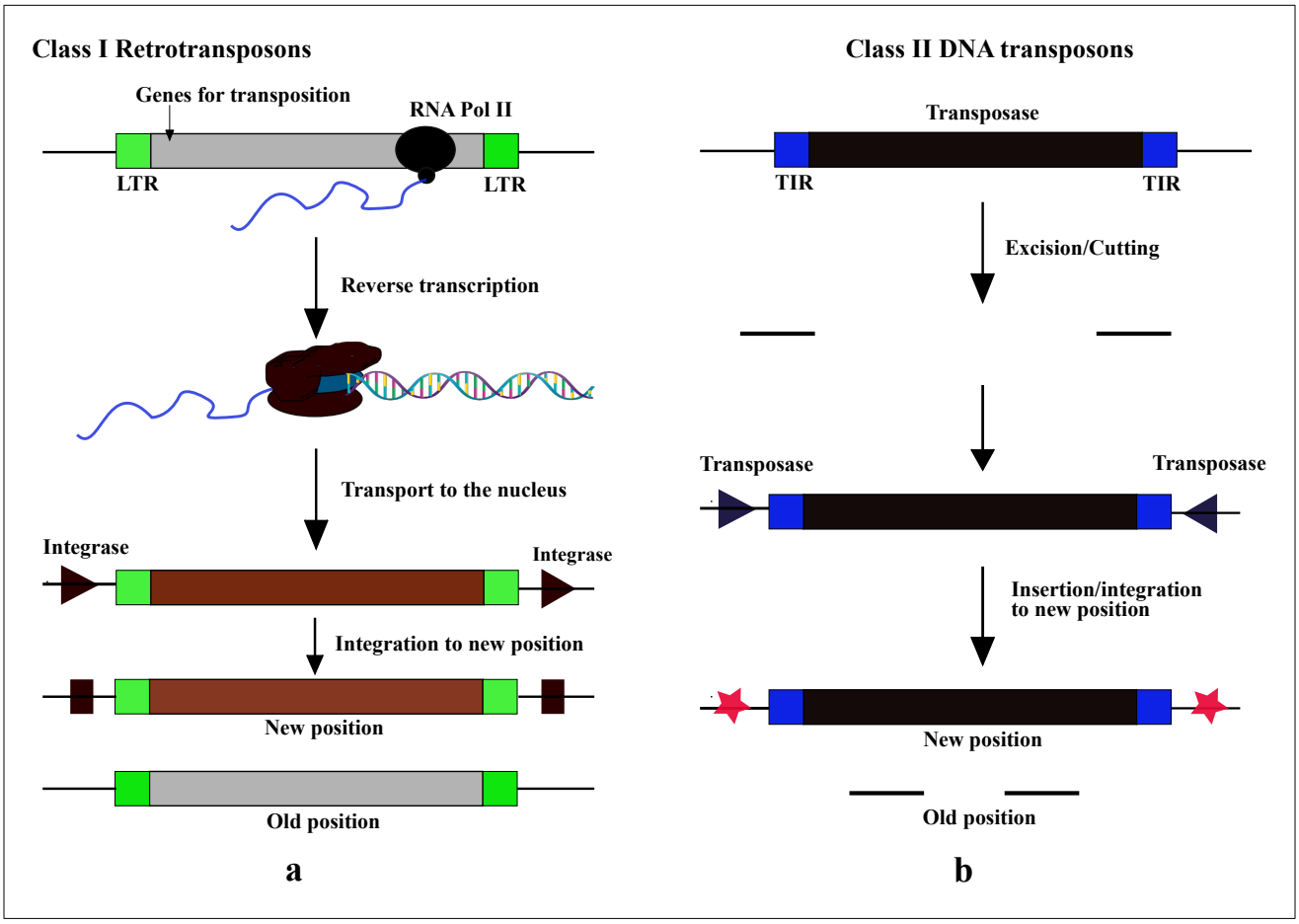

Figure 1. Types of transposable elements (TEs) and their mechanism of transposition. Adopted and modified from Lisch (2013).

The Penelope-like (PLE) order was first found in *Drosophilia virilis* (Evgen'ev et al., 1997), but it has since been found in more than 50 species, including plants (Evgen'ev and Arkhipova, 2005). PLE encodes a reverse transcriptase that is more closely related to telomerase than reverse transcriptase

from LTR retrotransposons or LINEs, as well as an endonuclease that is related to both intron-encoded endonuclease and the bacterial DNA repair protein UvrC. The LINEs, which are several kilobases in length, are found in all eukaryotic kingdoms. They are divided into five major superfamilies (Eickbush and Malik, 2007), but the most common ones in green plants are *L1, I* and *RTE* superfamilies (Županski et al., 2001; Eickbush and Malik, 2007). They form variable TSD upon insertion (Eickbush and Malik, 2007). The retrotransposon elements classified under the SINE order originate from accidental retrotranspositions of various polymerase III transcripts, unlike autonomous class I retrotransposon elements (Kramerov and Vassetzky, 2005). They are non-autonomous, non-deletion derivatives with sizes ranging between 80-500 bp and generate TSDs upon integration.

Class II DNA transposable elements contains two subclasses, distinguished by the number of DNA strands cleaved during transpositions. However, none of them moves via an RNA mediator (Wicker et al., 2007). Subclass 1 consists of the TIR order, which is characterized by variable length of terminal inverted repeats (TIRs). There are nine known superfamilies of TIR, distinguished by TIR sequences and TSD size (Wicker et al., 2007). However, only six of them are found in plant genomes. In contrast to subclass 1, subclass 2 contains TEs that undergo a transposition process that involves replication without double-stranded cleavage (Mhiri et al., 2022). This transposition mechanism involves the displacement of only one strand and does not generate TSD (Kapitonov and Jurka, 2007).

TEs affect genomes in a several ways, including genome size variation, genome rearrangement, and regulation of gene expression (Bennetzen and Wang, 2014). Genome size variation in angiosperms is mainly the result of the amplification of a small number of TE orders, particularly the LTR retrotransposons (Sanmiguel and Bennetzen, 1998). LTR elements can proliferate very rapidly, and in consequence, they can constitute a significant portion of plant genomes (Zhao et al., 2016). This differentiates plant genomes from mammal ones, where non-LTR retrotransposons dominate (Lander et al., 2001). There is a positive correlation between genome size and TEs number, especially LTR (Wang

et al., 2021). When comparing the LTR retrotransposon content among plants, such as the rice genome (400 Mb) and maize (2,400 Mb), both have approximately the same number of TE super-families. However, the maize genome is enriched in some LTR retrotransposon super-families with thousands of copies per genome, whereas in rice, copy number of each family does not exceed hundreds per nucleus (Baucom et al., 2009a; b). LTR retrotransposon amplification increases genome size primarily in bursts (Wicker et al., 2007; Baidouri and Panaud, 2013). These bursts can be fairly long-lived, lasting several millions of years or more. TE outbreaks have occurred several times throughout plant genome evolution, allowing a large number of TE families to evade silencing mechanisms (Lisch and Slotkin, 2011; Baidouri and Panaud, 2013). However, because the genome must maintain stability, most TEs are deleted or silenced, and only a few TEs may counteract these processes by inactivating systems that have evolved to recognize them (Fu et al., 2013). As a result, ancient TEs typically make up a minor portion of the genome, while contemporary TEs are primarily responsible for genome size variation (Oliver et al., 2013; Divashuk et al., 2020).

Transposable elements (TEs) have impacts on gene expression and function that go beyond their role in inducing mutations by inserting into coding sequences (Butelli et al., 2012). Nevertheless, insertion of TEs in non-coding regions can influence gene expression in various ways, including:

(1) Impact on transcript splicing/processing: TEs can affect the splicing and processing of RNA transcripts. This occurs through the presence of splice sites or other regulatory sequences within the TEs, which influence the joining of exons and introns during the mRNA formation. Consequently, alternative splicing patterns can arise, leading to the generation of different mRNA isoforms with diverse protein coding potentials or functional characteristics. For example, the insertion of *Mutator* transposons within the introns of *Adh1-S* (Luehrsen and Walbot, 1990) or brown *midrib1* (*bm1*) (Chen et al., 2012) genes in maize has been shown to generate new isoforms. In most cases, however, the

presence of splicing events between a gene and transposon leads to the production of non-functional proteins (Hirsch and Springer, 2017).

(2) Provision of new promoters: TEs may contain regulatory elements like promoters that regulate the initiation of transcription. When TEs are inserted close to genes, they have the ability to introduce their own promoters, thus impacting the expression of neighboring genes. This can lead to the emergence of new sites for transcription initiation, ultimately causing changes in the expression levels or patterns of adjacent genes. For instance, in citrus species, the natural variation in fruit pigmentation is caused by allelic variations at the *Ruby* locus, which encodes a *MYB* transcriptional activator of anthocyanins (Butelli et al., 2012). The alleles that give rise to fruit color, such as the ones observed in blood oranges, are attributed to the insertion of a retrotransposon in the regulatory regions of the *Ruby* gene.

(3) Positive and negative regulation of gene expression levels: TEs can function either as enhancers or silencers, influencing the expression levels of nearby genes. TEs may harbor binding sites for transcription factors or other regulatory elements that modulate gene expression either positively or negatively. In a study focused on class II DNA transposons in four grass species, researchers observed distinct associations between various transposable element (TE) families and gene expression levels (Han et al., 2013).

1.1.2. Structure and organization of organellar genomes in Angiosperms

In addition to their nuclear genome, plants have two small genomes inside semi-autonomous organelles such as plastids and mitochondria (Mahapatra et al., 2021). Both organelles have originated endosymbiotically with plastids originating from cynobacteria and mitochondria originating from alpha-proteobacteria (Gutman and Niyogi, 2009; Smith and Keeling, 2015). Mitochondria generate the most of the cell's supply of energy in the form of adenosine triphosphate (ATP), while plastids are mainly responsible for conducting photosynthesis (chloroplast) or for storage and synthesis of pigments

(chromoplasts) (Zhang et al., 2012). Plastid and mitochondrial genomes occur in large number in plant cells, but the quantity frequently changes among different tissue types and stages of plant development (Oldenburg and Bendich, 2015; Krupinska et al., 2020).

Plastid genome (pDNA) of angiosperms is a uniparentally inherited circular DNA molecule with a quadripartite structure containing a large single copy region (LSC), a small single copy region (SSC), and two copies of an inverted repeat (IR) region, Ira and Irb (Yang et al., 2010). The size of the genome ranges from 120 to 180 kb (Park et al., 2019). The differences in size are mainly due to the contraction and/or expansion of inverted repeat (IR) into or out of adjacent single-copy regions and/or changes in sequence complexity due to insertions or deletions of sequences (Khan et al., 2021). Genome sequencing of pDNA has unveilled notable sequence and structural variations both within and between different plant species. Plastid heteroplasmy, the presence of more than one type of pDNA genome within a cell or individual with different single copy region orientations, has been reported in different plant species, for example in *Pelargonium* (Tilney-Bassett and Birky, 1981), *Oryza sativa* L. (rice) (Moon et al., 1987), *Gossyplum hirsutum* L. (Lax et al., 1987), *Oenothera* (Chiu et al., 1988), *Medicago sativa* (Johnson and Palmer, 1989), *Actinidia deliciosa* (Kiwifruit) (Chat et al., 2002), genus *Passiflora* (Passifloraceae) (Hansen et al., 2007), and *Phoenix dactylifera* (Sabir et al., 2014). Heteroplasmy can arise from biparental inheritance, when each parent passes on its organelles to the zygote, as it has been observed in *Pelargonium* (Tilney-Bassett and Birky, 1981) or from uniparental inheritance when sorting in the parent is so incomplete that some heteroplasmic gametes are produced as it was observed in *Gossyplum hirsutum* L. (Lax et al., 1987). A recent study suggested that structural pDNA genome heteroplasmy should be extremely common across all angiosperms (Wang and Lanfear, 2019).

Most Angiosperms' pDNA contains a total of 110-130 genes, which includes up to 80 protein coding genes, approximately 30 transfer RNA (tRNA) genes, and four ribosomal RNA (rRNA) genes

(Jansen & Ruhlman, 2012). The pDNA genome is more conserved in terms of gene structure and composition in comparison with mitochondrial and nuclear genome (Asaf et al., 2016). However, a number of modifications at the genomic and gene level among the angiosperm pDNA genomes have been reported (Lei et al., 2016). These modifications include, the loss of a single copy of the IR (Palmer and Thompson, 1982), the occurrence of 50 kb and 78 kb inversions (Bruneau et al., 1990), gene loss (Wicke et al., 2013), and the variation in the size of IR and intergenic spacers (Wakasugi et al., 2001; Raubeson and Jansen, 2005). Gene and intron loss events have been observed multiple times during pDNA genome evolution (Gao et al., 2010), including genes such as *infA* (Millen et al., 2001; Mardanov et al., 2008), *rps15* (Tsuji et al., 2007), *rps16* (Doyle et al., 1995; Jansen et al., 2006; McCoy et al., 2008), *rpl22* (Jansen et al., 2006), *rpl23* (McCoy et al., 2008), *accD* (Lee et al., 2007), *ycf1* (Cai et al., 2008), *ycf15* (Mardanov et al., 2008), and introns such as those in the *rpl2*, *clpP* and *rps12* genes (Doyle et al., 1995; Jansen et al., 2008).

Angiosperm pDNA genomes consist of up to 60% protein and RNA coding genes, with the remaining 40% comprising non-coding regions such as intergenic spacers (IGS) and introns (Borsch and Quandt, 2009). These non-coding regions are the most rapidly evolving sequences and are frequently used to study evolutionary relationship between organisms (Degtjareva et al., 2012). They contain specific structural elements, including stem loops, which can be highly dynamic and AT-rich, resulting in a mosaic-like pattern of conserved and variable element (Korotkova et al., 2014). Some of these elements are specific to certain introns and IGS and are restricted to particular lineages (Borsch and Quandt, 2009; Korotkova et al., 2009).

The mitochondrial genome of flowering plants exhibits several distinctive features in comparison to both plastid and animal mitochondrial genomes (Schuster and Brennicke, 1994). These include a wide range genome sizes, ranging from 66 kb in *Viscum scurruloideum* to 11,300 kb in *Silene conica* (Palmer and Herbon, 1988), as well as a complex structure that may be linear or

multichromosomal in addition to the typical circular structure found in most mitochondrial genomes (Wu et al., 2022). The mitochondrial genome of flowering plant also contains a significant amount of non-coding DNA, low gene density, a large number of introns (Chapdelaine and Bonen, 1991), and a relatively high level of RNA editing (Hao et al., 2021).

Recent evidence has challenged the long-held view that the mitochondrial genome of plants is conserved as a single circular molecule (Sloan, 2013). Among land plants, particularly angiosperms, the whole array of various structural modifications has been observed. Some species, such as *Arabidopsis thaliana*, have a one standard circular genome (Unseld et al., 1997). In contrast, *Cucumis sativus* (cucumber) has been found to have three autonomous circular chromosomes of varying sizes (1556, 84 and 45 kb) that replicate independently of one another (Alverson et al., 2011). Similarly, *Silene noctiflora*, has numerous circular chromosomes, some of which are entirely autonomous and others that show signs of recombination (Wu et al., 2022). Except circular structures, mtDNA can occur in linear form as observed in one of the cytoplasmic-male-sterile groups of maize, where the mtDNA genome was assembled in a linear form, that likely originated from recombination of the original circular genome with a linear plasmid (Allen et al., 2007). Plasmids are known to exist autonomously in mitochondria in a variety of species and play an active role in genome rearrangements (Handa, 2008).

Even with one standard circular genome, genome size, and gene content may vary considerably among and within species of flowering plants. This huge variation in the mtDNA is mainly due to three processes: transferring of genes or sequences from the mitochondrial genome to the nuclear genome or vice versa (Wu et al., 2017), acquiring sequences from other species through horizontal gene transfer (HGT) (Rice et al., 2013), and sequence duplications (Albert et al., 1998). For example, Goremykin et al. (2012) found that approximately 20% of mtDNA in *Malus domestica* (apple) originated through transferring from the nucleus. *Amborella trichopoda*, which is one of the most basal Angiosperms,

contains DNA sequences in mtDNA from green algae, which is likely assimilated through HGT (Rice et al., 2013). However, the majority of the size and structural variation among mitochondrial genomes reflects differences in repetitive DNA content. Large repeats, typically several kb or more in length, recombine frequently and are a main cause of genome isoforms (Kozik et al., 2019). In addition, there are often other repeated sequences in the size range of 1kb and lower. However, they do not recombine so often as large repeats.

Similarly to structural variation, mitochondrial gene content is also highly variable across extant angiosperms, ranging from 19 to 64 genes and 18 to 25 introns (mainly group II intron RNAs), excluding duplicate genes and ORFs (Skippington et al., 2015). The variable gene content reflects a pattern of differential losses and functional transfers to the nucleus across angiosperms (Adams et al., 2002). This is well illustrated by dynamic evolution of tRNAs content which may have mitochondrial, plastid and possibly bacterial origin. The number of tRNAs varies widely among species, and to date, there is no sequenced angiosperm mitochondrial genome that contains a full set of native tRNAs required for translation of full mitochondrial gene set. Missing tRNAs are encoded in the nucleus and imported from the cytosol (Joyce and Gray, 1989). The entire mtDNA is a dynamic structure, and the gene order also varies substantially at various taxonomic levels from family to plant cultivars of the same species, although several syntenic blocks of genes are conserved all the way back to their bacterial progenitors (Oda et al., 1992). The variation extends not only to structural changes in gene order, but also to the products of functional genes encoded by mtDNA. Although genes in angiosperm mitochondrial genomes evolve slowly at the sequence level relative to plant nuclear and chloroplast genomes, there is a massive diversity in their products due to heavy RNA editing. RNA editing occurs in flowering plants, almost exclusively in the form of cytidine (C) to uridine (U) substitutions. Numerous C to U conversions can modify the final product of a gene in many ways, such as altering

the coding sequences of the transcripts, by inserting AUG start sites or eliminating premature stop codons, or by influencing splicing and altering the stability of RNAs (Small et al., 2020).

1.1.3. Genomic resources of Angiosperms

New sequencing technologies, i.e. next-generation and third-generation sequencing, have provided powerful tools to explore the genomic resources of organisms in an unprecedented way. These advancements have made possible to sequence genomes of 10,770 angiosperm species to date (Table 1). Although this represents only a small fraction of all described species of flowering plants, even this small sample clearly shows that plant genomes exhibit spectacular diversity in size, composition, and complexity. Nonetheless, it is still poorly understood how this variation drives changes in plant chemistry, morphology and ecology. Understanding the link between spectacular diversity of plant species and their genomes is unquestionably important for solving many issues of societal relevance, such as predicting, and mitigating plant species' responses to climate change, ensuring food security for the world's expanding population, discovering new sources of medicines, and facilitating effective conservation strategies (Soltis and Soltis, 2021). Sequencing projects, especially those focusing on economically important plant families, are first steps towards this goal. The celery family (Apiaceae), which includes economically relevant species such as *Daucus carota* (carrot), *Apium graveolens* (celery), *Corindrum sativum* (coriander), and *Petroselinum crispum* (parsley), is one of those families whose genetic resources are important, although poorly explored.

1.2. Characteristics of Scandiceae—a model system for comparative genomics in the celery family (Apiaceae)

Apiaceae (Umbelliferae) is a family of flowering plants with 466 accepted genera and approximately 3,820 species (Plunkett et al., 2018). The family is one of the largest families of seed plants and is most likely the oldest recognized and thoroughly characterized (Dalechamps, 1586; Morison, 1672). Representatives of Apiaceae are mainly distributed in the temperate regions of Eurasia, North America

and Africa, with the highest diversity found in Central Asia (Liu et al., 2014). The largest number of genera (289) as well as the most generic endemism (177) have been found in Asia, followed by Europe (126) with 17 endemic genera. Species from this group are extremely rare in the tropics, where they are restricted to high mountains, whereas the highest number of species is found in Mediterranean and arid areas (Palumbo et al., 2021). Apiaceae can be found in nearly every type of temperate habitats including such extremes as aquatic and arid environments (Reduron, 2021).

The plants of the Apiaceae family are mostly herbaceous (annual, biennial or perennial), though a minority are woody sub-shrubs, shrubs, or rarely trees (Reduron, 2021). The members of Apiaceae are easily recognized through their distinguishing morphological characteristics, such as hollow or pith-filled stems, pinnately divided leaves with sheathing bases (Downie, Katz-Downie, et al., 2000), inflorescences in the form of compound umbels (less commonly simple-umbellate, capitulate or cymose), often subtended by involucral bracts, and specialized dry fruits (schizocarp) broken up into two single seeded segments (mericarps) attached to a central stalk (carpophore) (Plunkett et al., 2018). The family is well known for their distinctive flavors due to the secretory canals consisting of schizogenous oil ducts with resin, ethereal oil, or mucilage located in fruits, stems, leaves and roots (Kljuykov et al., 2004).

Apiaceae is ranked as one of the most economically important families, best known as a source of a wide variety of crop plants (Wang et al., 2022). For instance, *Daucus carota* (carrot) is one of the world's most widely grown root crops, *Apium graveolens* (celery) is widely consumed for its stalk, *Petroselinum crispum* (parsley) is known for its entire edible plant, and *Foeniculum vulgare* (fennel) is a popular above-ground vegetable. The family includes many herbs and spices such as *Pimpinella anisum* (aniseed), *Carum carvi* (caraway), *Coriandrum sativum* (coriander), *Cuminum cyminum* (cumin), and *Anethum graveolens* (dill). In addition to edible plants, many species exhibits the poisonous properties due to toxic polyacetylenes. The best known toxic plants are *Cicuta virosa* (water

hemlock), *Conium maculatum* (hemlock), and *Aethusa cynapium* (fool's parsley). Since the family is a source of a variety of phytochemicals, it is intensively studied for its potential industrial and therapeutic applications (Sayed-Ahmad et al., 2017).

The current classification of Apiaceae is almost exclusively based on phylogenetic analyzes of molecular data. The family is now classified into four subfamilies: Mackinlayoideae, Azorelloideae, Saniculoideae and Apioideae (Chandler and Plunkett, 2004; Calviño et al., 2016). Apioideae is the largest and the most morphologically diverse subfamily, currently comprises about 380 genera and 3200 species, representing 85.2% genera and 83.8% species of Apiaceae including the majority of crop plants. The most comprehensive classification framework to date identifies 41 major clades in Apioideae, of which 21 have obtained tribal or subtribal rank.

Among distinct groups of Apioideae, tribe Scandiceae has arisen as a model clade for testing various biological hypotheses for two reasons. Firstly, this taxon is highly diversified in terms of its life history, habit, ecology, floral morphology, umbel structure, and fruit anatomy and morphology. For instance, Scandiceae, unlike other clades in Apioideae, displays all types of the fruit morphology including compressed diaspores with appendages in the form of wings regarded as anemochorous (dispersed by wind), rounded fruits with bristles and hooks regarded as epizoochorous (carried away on animal fur or feathers), and fruits without any distinct structures facilitating dispersal regarded as barochorous (gravity dispersed) or autochorous (self-dispersing). Interestingly, the evolution of epizoochory in Scandiceae was related to a unique developmental modification among Apioideae—the enlargement of secondary ribs and their transformation into bristles or hooks (Wojewódzka et al., 2019).

Secondly, Scandiceae has become a model system due to extensive studies on one of the most economically important crop plants—*Daucus carota*. *Daucus carota* represents the first species from Apiaceae for which the entire genome was sequenced, annotated and mapped (Iorizzo et al., 2016).

Moreover, *Daucus carota* is often used for studying in vitro propagation and regeneration, biosyntheisis of various compounds such as carotenoids or for studying cytoplasmic male sterility (Kalia et al., 2019; Que et al., 2019; Simon, 2021). Recent studies have shown that *Daucus carota* is a morphologically complex species with relatively low genetic variation and small genome size (Spooner et al., 2014). This makes it a perfect system to study the genetic architecture that allows rapid evolution under strong selection, low effective population size, and low standing genetic variation.

Except for *Daucus carota* and their relatives, recent years have introduced a new group from Scandiceae, the genus *Ferula*, which is a potentially promising source of bio-active compounds and an interesting model for studying various evolutionary problems. *Ferula* is one of the most species-rich genera of Apiaceae, represented by 180–185 perennial species (Plunkett et al., 2018). The geographical distribution of *Ferula* ranges from Central and West Asia and extending westward to the Mediterranean and Macaronesian regions. *Ferula*, which has dorsally flattened mericarps adapted to wind dispersal, was traditionally classified in the tribe Peucedanae along with other species characterized by this type of fruit (Drude, 1898; Pimenov and Leonov, 1993). However, this placement was first questioned by immunological studies (Shneyer et al., 1995) and later by molecular studies that firmly established the position of *Ferula* within the tribe Scandiceae (Ajani et al., 2008; Kurzyna-Młynik et al., 2008; Panahi et al., 2015; Piwczyński et al., 2018). A recent study by Piwczyński et al. (2021) established the position of subtribe Ferulinae, to which *Ferula* belongs, as a sister group to a clade consisting of Daucinae and Torilidinae, both subtribes with species having epizoochorous fruits. Furthermore, a new infrageneric classification system for *Ferula*, based on molecular data, was proposed. However, the authors indicated that this new system is far from definitive due to two problems. First, there is a low phylogenetic signal in the molecular markers used, which failed to resolve phylogeny within tribes and subtribes. This suggests that members of *Ferula* evolve much faster at the molecular level than other

studied species from other clades in Apioideae. Rough molecular dating estimates (Banasiak et al., 2013) suggest that it took no more than 5-6 Myr for evolution to form such a diversity of species.

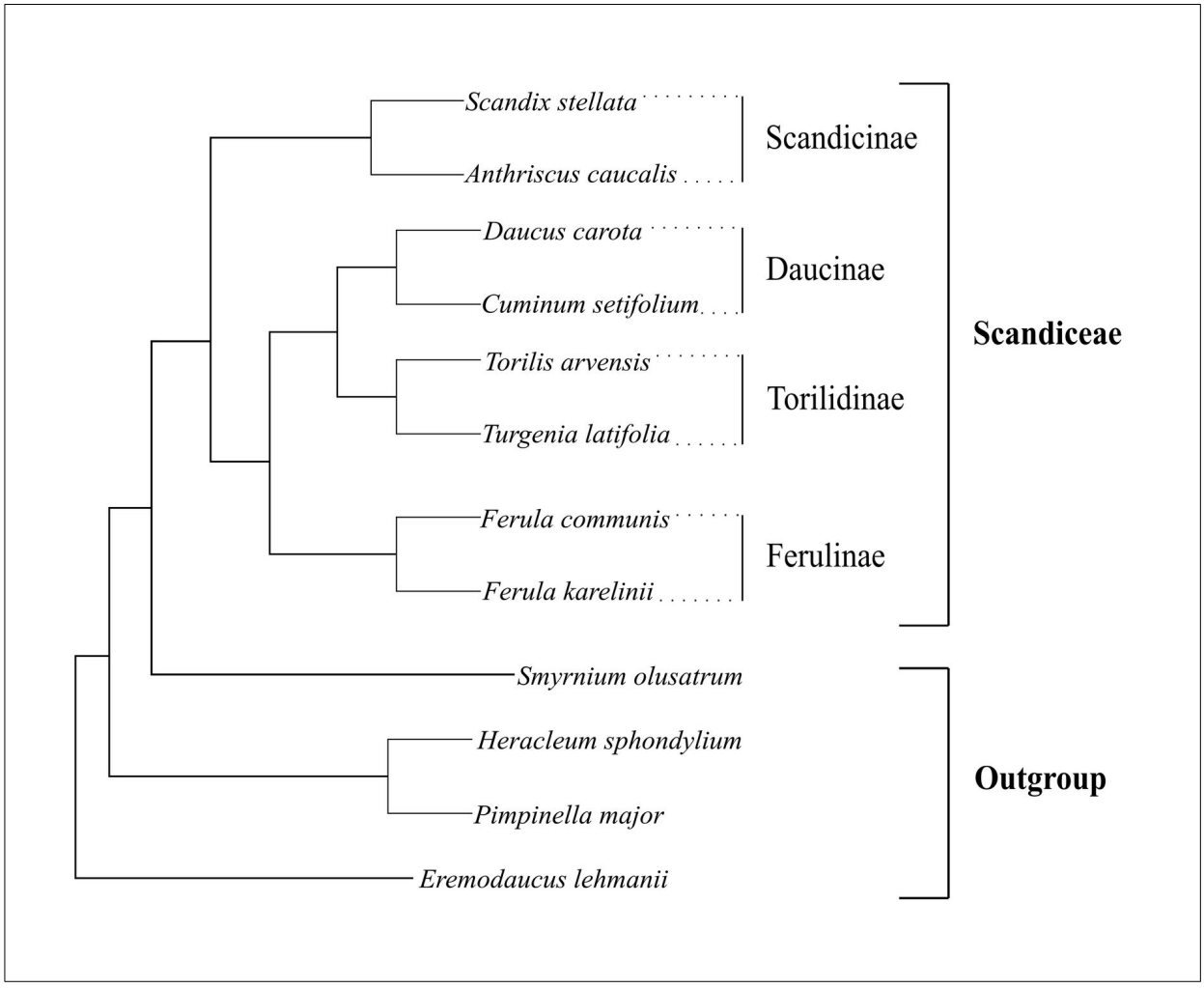

Figure 2. Phylogenetic position of Ferulinae in the tribe Scandiceae according to Piwczyński et al. (2021).

The second problem confounding the establishment of a robust taxonomic system is the incongruence between nuclear and plastid markers. This incongruence in *Ferula* can be attributed to hybridization and introgression among species. Although these processes are relatively rare in Apiaceae in comparison with other angiosperm families, several cases have been reported (Lee and Downie, 2006; Spalik et al., 2009; Zhou et al., 2009; Bone et al., 2011; Yi et al., 2015). Hybridization may

facilitate adaptive and non-adaptive radiations. This was postulated as the main process responsible for rapid evolution in *Ferula.*

An interesting example of high genetic variation of potentially of hybrid origin is the *Ferula communis* species complex. This species is found in various habitats such as pastures, mountain planes, maquis, or road sites, and is distributed mainly in the Mediterranean region, reaching as far west as the Canary Islands and as far east as Turkey and Israel. The southernmost populations occur in the Arabic Peninsula (e.g. Yemen) and the Horn of Africa, including Tanzania (Figure 3a).

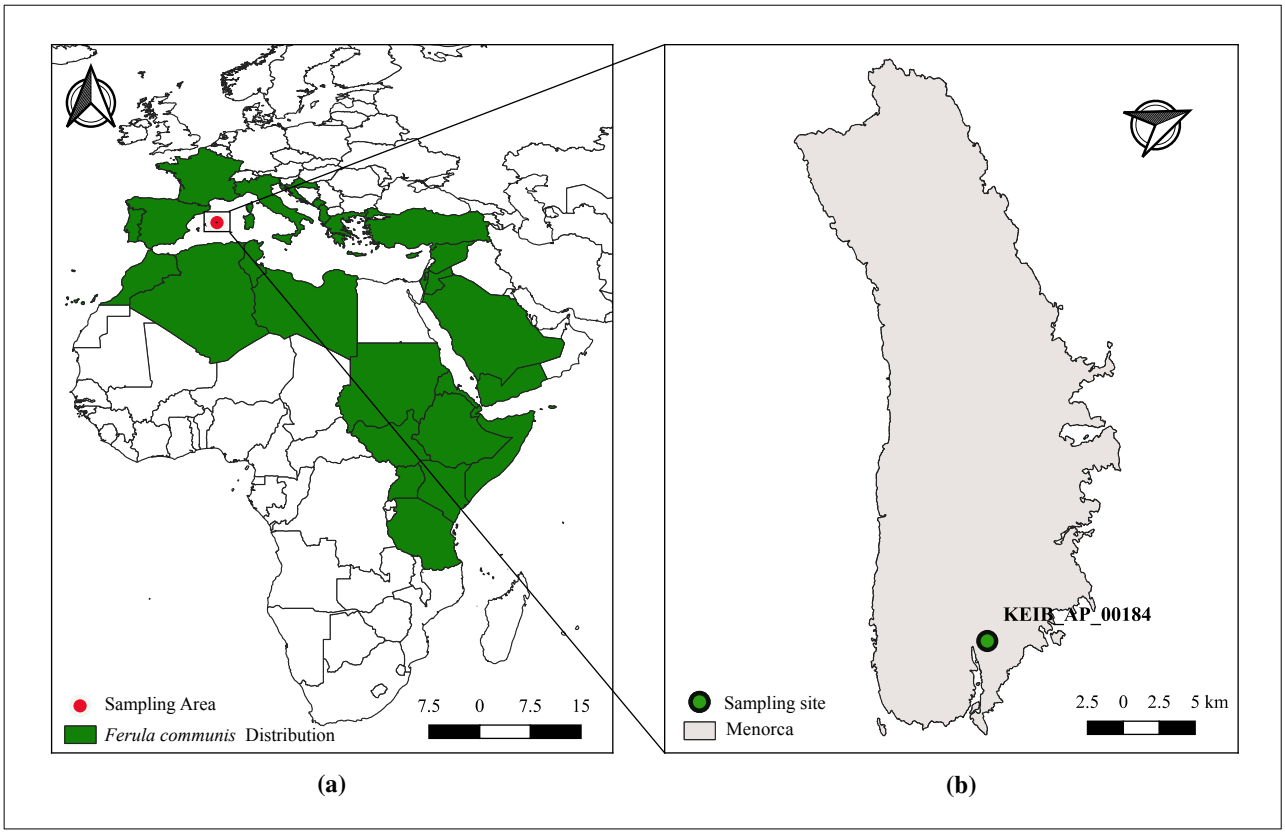

Figure 3. Distribution map of *Ferula communis* and sampling location of specimen used for genome assembly.

Several sibling species, subspecies, and varieties have been described to date within the complex, although, their delimitation is still not clear due to their morphological similarity. Furthermore, recent molecular studies have shown that there is a restricted gene flow among sympatric

populations with nearly identical morphotypes. For example, the Corso-Sardinian endemic *F. arrigonii* (Dettori et al., 2014), which was relatively recently described as a new species (Bocchieri, 1988), based mainly on phenological data, was shown to form distinct genetic clusters from *F. communis* occurring in the same area (Dettori et al., 2014, 2016). Interestingly, the same studies revealed that despite its endemicity and fragmented distribution, *F. arrigonii* is characterized by a high level of genetic diversity and a low level of differentiation among populations. A high genetic diversity and restricted gene flow were also shown for other taxa from *F. communis* complex; *F. c*. subsp. *glauca* from Malta and Tunisia formed a separate genetic cluster from *F. communis* (including *F*. *c*. subsp. *communis* and *F*. *c*. subsp. *cardonae*) originated from other Mediterranean islands and the Italian Peninsula (Dettori et al., 2016). Even within genetic clusters, there is a substructure that is often correlated with cryptic variation at a chemical level. For example, two chemotypes were described in Sardinia—one is poisonous to animals and the other one is non-poisonous. These two types are indistinguishable morphologically and karyotypically, but they are genetically distinct based on allozyme analysis (Marchi et al., 2003). These examples raise many questions concerning the origin of genetic variation in *F. communis* despite small effective population size, founder effects and operation of natural selection. To address these questions, a comparison of the whole genome of *F. communis* with other members of Scandiceae and Apioideae is an important first step.

Unlike the genus *Daucus,* which includes diploid species with a number of chromosomes ranging from 2n = 18, to 2n = 22 (e.g. *D. carota* 2n = 18, *D. littoralis* 2n = 20 or *D. crinitus* 2n = 22), as well as polyploid species (*D. glochidiatus* 2n = 44 or *D. montanus* 2n = 66), *Ferula,* with over one hundred karyotyped species to date, is exclusively diploid and has the same chromosome number of 2n = 22 (Rice et al., 2015), although B chromosomes have been observed (Sànchez-Cuxart and Mercè, 1998). On the other hand, the nuclear genome size (C-value) varies in both genera. For diploid representatives of the genus *Daucus*, the haploid genome size estimated based on flow cytometry data

ranges from approximately 460 Mb for *D. carota* to 1,476 Mb for *D. littoralis* (Nowicka et al., 2016), with estimates for polyploid species being much higher (e.g. *D. montanus* 5,390 Mb). In *Ferula*, there are also substantial differences among studied species, although chromosome number remains constant. For example, two closely related species (sometimes considered as subspecies), *F. communis* and *F. glauca* have haploid genome sizes of 1,764 and 1,470 Mb, respectively, based on estimates obtained using cytophotometric methods (Olmedilla et al., 1985). A larger genome of 2,411 Mb was discovered by flow cytometry in the Balkan endemic *F. heuffelli* (Siljak-Yakovlev et al., 2010). This size difference in the genome of *Ferula* without changes in chromosome number, high genetic variation as exemplified by *F. communis* complex, and incongruence between plastid and nuclear markers suggests that homoploid hybridization (hybridization without a change in ploidy) might be the main mode of speciation in this genus. Homoploid hybrids are often characterized by changes in karyotype, gene expression levels, and number of transposable elements in comparison with parental species (Renaut et al., 2014). However, the lack of any genomic resources for *Ferula* prevents testing this interesting hypothesis.

1.3. Genome evolution in Apiaceae – state of the art

Despite the occurrence of many herbs, vegetables, spices and medicinal plants in the Apiaceae family, the number of genomic resources in the form of high quality, annotated and mapped genomes, is restricted only to a few species. At the time of writing this thesis (April 2, 2022), three genomes have been assembled at the chromosomal level, all of them belonging to crop plants: *Daucus carota* (carrot) (Iorizzo et al., 2016), *Coriandrum sativum* (coriander) (Song et al., 2020), and *Apium graveolens* (celery) (Song et al., 2021). The *Daucus carota* genome was assembled from several paired-end, mate-paired, and BAC libraries sequenced on an Illumina platform. The genomes of two remaining species resulted from a combination of short reads from second-generation platforms (Illumina) and long reads from third-generation sequencing platforms (PacBio). In addition, incomplete and/or assembled at the

scaffold level genomes are available for the crop plant *Foeniculum vulgare* (Palumbo et al., 2018) and wild plants such as *Angelica gigas* (Gil et al., 2017), *Oenanthe javanica* (Liu et al., 2021) and *Bupleurum falcatum* (Zhu et al., 2019).

Of the three species for which almost complete genomes are available, *Daucus carota* has the smallest size estimated at 473 Mb per haploid genome organized into nine pairs of chromosomes (Iorizzo et al., 2016). The assembly itself was slightly smaller, 422 Mb, covering around 90% of the estimated genome size. In contrast, the nuclear genomes of *Corindrum sativum* (Song et al., 2020) and *Apium graveolens* (Song et al., 2021) were much larger, with the estimated genome size of 2,130 and 3,470 Mb, respectively, arranged into 11 pairs of chromosomes in both species. The assemblies covered approximately 99% of the estimated genome size (2,119 Mb) in *Corindrum sativum* and 96% (3,330 Mb) in *Apium graveolens* (Table 3).

Among the incomplete genome assemblies, the nuclear genome of *Foeniculum vulgare* (Palumbo et al., 2018) has an estimated size of 1,320 Mb per haploid genome and has an assembly coverage of 75% (1,010 Mb). The nuclear genome of *Angelica gigas* (Gil et al., 2017) and *Bupleurum falcatum* (Zhu et al., 2019) have an estimated sizes of 2,670 and 2,120 Mb, respectively. However, they have less assembly coverage, with 43.5% (920 Mb) for *Bupleurum falcatum* and 30% (804 Mb) for *Angelica gigas* (Table 3). The assembled nuclear genome of *Oenanthe javanica,* as reported by Liu et al. (2021b), has a size of 1,280 Mb.

Despite the limited number of complete genomes available for comparative studies, several interesting patterns have emerged from recent analyses. First, there is a positive relationship between genome size and the proportion of repetitive elements. In *Daucus carota,* 46% of the genome comprises repetitive sequences, while in *Corindrum sativum* and *Apium graveolens*, this number reaches 70.59% and 92.91%, respectively. Second, a noteworthy finding is that the majority of these repetitions belong to class I LTR type of transposable elements in all species. Interestingly, class I and

class II transposons are almost perfectly inversely distributed, with DNA transposons concentrated in gene-rich terminal chromosomal regions, while retrotransposons are almost absent in these areas.

All three species for which the entire genome were sequenced, *Daucus carota*, *Corindrum sativum* and *Apium graveolens*, show traces of several rounds of genome duplications (WGDs), although the estimated time frames for these events do not overlap. The WGD event unique to the *Daucus carota* linage (Dc-α and Dc-β) are thought to have occurred ~43 and ~70 Mya, respectively (Iorizzo et al., 2016). Estimated time-frame for the Dc-β WGD overlaps with the boundary between Cretaceous-Paleogene geological periods, which supports the theory that a burst of WGD events at that time, possibly suggesting a selective advantage of polyploidy (Vanneste et al., 2014). The organization of *Apium graveolens* and *Corindrum sativum* genome was also shaped by two whole-genome duplication events: Apiaceae-α and Apiaceae-ω, which dated back to ~34–38 and ~66–74 Mya (Song et al., 2021), and A-beta and A-alpha, which dated back to ~54–61 and ~45–52 Mya, respectively (Song et al., 2020).

Due to advances in sequencing technology and the relatively small sizes of plastid genomes compared to nuclear genomes, it is now possible to sequence pDNA quickly and cost-effectively (Khan et al., 2018). The pDNA genomes in the Apiaceae family range in size from 141,948 bp (*Heracleum candicans*) to 178,668 bp (*Bupleurum scorzerifolium*), with an average GC content of 37.60%, and encode between 120 to 130 genes (Li et al., 2020). Like typical angiosperm pDNA genome, pDNA genome of Apiaceae consists of a conserved quadripartite structure, with single circular DNA molecule that includes a LSC, SSC, and a set of inverted repeats (IRa and IRb) (Ruhlman et al., 2006; Daniell et al., 2016).

Restriction site mapping studies on the Apiaceae family have shown that four different IR size classes contribute to the size variation at the position of J_{LB} region (Plunkett and Downie, 1999), the junction where large single copy and inverted repeat regions intersect, usually found within or close to

Table 3. Chromosomal and scaffold level genome assembly statistics for *Daucus carota*, *Apium graveolens*, *Coriandrum sativum*, *Foeniculum vulgare*, *Angelica gigas*, *Oenanthe javanica,* and *Bupleurum falcatum* genomes

Assembly feature	***Daucus carota***	***Apium graveolens***	***Coriandrum sativum***	***Foeniculum vulgare***	***Angelica gigas***	***Oenanthe javanica***	***Bupleurum falcatum***
Assembled genome size	421 Mb	3,330 Mb	2,118.68 Mb	1,010 Mb	804 Mb	1,280 Mb	920 Mb
GC content (%)	34.8	n.s	n.s	n.s	n.s	32.97	n.s
Scaffolds	4907	4863	6186	9,443	395,007	149,941	n.s
N50 scaffold	13.4 Mb	289.78 Mb	160.99 Mb	n.s	n.s	0.0233 Mb	n.s
N50 contigs	0.0312 Mb	0.791 Mb	0.604 Mb	n.s	n.s	0.0130 Mb	0.000313 Mb
Predicted genes	32113	31326	40747	n.s	n.s	42,270	n.s
Repetitive elements (%)	46	92.91	70.59	n.s	n.s	n.s	n.s
Non-coding RNA	0.189 Mb	1.99 Mb	1.65 Mb	n.s	n.s	n.s	n.s
Estimated haploid genome size							
Flow cytometry	473 Mb	n.s	2,484 Mb	n.s	n.s	n.s	n.s
K-mer analysis	473 Mb	3,470 Mb	2,130.29 Mb	1,320 Mb	2,670 Mb	n.s	2,120 Mb
Assembly genome coverage (%)	90	96	99	75	30	n.s	43.5

(n.s: not specified)

the *rps19* gene of the ribosomal protein S10 operon and remains relatively stable in its location (Palmer, 1985; Goulding et al., 1996). Among these size variations, a difference of ~17 kb compared to the pDNA genome of tobacco J_{LB} region, characterized by one expansion and three contractions, was restricted to the Apiaceae subfamily Apioideae (Plunkett and Downie, 2000). In addition to these, there have been documented cases of expansion of IR into *rps3*, *rpl2* and contraction of IR into either the *ycf2-trnL* IGS or *ycf2* gene (Peery, 2015), as well as significant reduction of IR in *Corindrum sativum* that encompasses *trnH* and *psbA* genes in the IR region (Palmer, 1985).

However, despite the importance and species richness of the Apiaceae, genomic resources for this family are relatively limited in the NCBI database. As of March 3, 2022, there are only 120 sequenced pDNA genomes and 5 sequenced mitochondrial genomes available (https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/Apiaceae).

The Apiaceae family has five sequenced mitochondrial genomes, with the *Corindrum sativum* mtDNA genome being the smallest (82.9 kb), while the others range from 212 to 463.79 kb with an average GC content of 45.15%. The mtDNA genome encodes 30–40 protein genes, 5 ribosomal RNA genes, and 18–20 transfer RNA genes. The intergenic spacer region and repetitive sequences occupy the largest part of the mtDNA genome. For example, in the mtDNA genome of *Daucus carota*, the intergenic spacer regions accounts for 79.9% of the genome size, with repetitive elements comprising the majority of this space (49%) (Spooner et al., 2019).

Several studies have also reported the presence of mtDNA in the plastome of some Apiaceae species, suggesting that the introgression of mtDNA into the plastid genome is not unprecedented in the family (Ruhlman et al., 2006; Iorizzo et al., 2011, 2012a; Downie et al., 2014; Downie and Jansen, 2015).

1.4. Objectives of the study

The objectives of the study are as follows:

- To assemble the nuclear genome of *F. communis* using long reads from the MinION platform (Oxford Nanopore Technologies) and short reads from Illumina platform, applying various assembly methods.
- To annotate the protein-coding genes and transposable elements, and identify duplicated genes through homology-based orthologous method.
- To perform comparative analyses of the assembled genome of *F. communis* with high quality genomes of Apiaceae available in public repositories to study the history of duplication events, evolution of gene content, transposable elements and to test the hypothesis that *Ferula communis* harbors lineage-specific gene families that have rapidly evolving (significantly expanded or contracted), contributing to their successful adaptation to arid and semi-arid environments.
- To assemble, annotate, and characterize the plastid genome of *F. communis* and identify structural variations and diversity hotspot genomic regions by comparison with available plastid genomes in public repositories.
- To assemble, annotate the mitochondrial genome of *F. communis* and identify its structural organization and arrangement.

2. Materials and Methods

2.1. Sample collection of *Ferula communis*

The ripe fruits of *Ferula communis* were collected from Menorca (Figure 4b), one of the islands in the Mediterranean region, and then planted in the greenhouse at the Institute of Biodiversity and Genetic Resources Research Center (CIBIO-InBIO), Vairão, Portugal. Once the seedlings reached the first leaf stage, they were harvested and placed in tea bags, which were then labeled, and stored in sealed plastic bags filled with silica gel. To preserve the genomic DNA, the seedling bags were kept in a room at a temperature of 15 ^{0}C while drying. The dried seedlings were transported to the Department of Ecology and Biogeography at the Nicolaus Copernicus University in Toruń, Poland, where their DNA was subsequently extracted.

2.2. DNA isolation and whole genome sequencing

Total genomic DNA was isolated from four *Ferula communis* seedlings originated from the same mother, (KEIB_AP_00184), using the Dneasy® Plant Mini kit (Qiagen). To obtain a high-quality DNA yield, four individual dry seedling samples were mechanically disrupted with a pestle after adding liquid nitrogen in a 1.5 ml Eppendorf tube. To break the membranes and digest the RNA within the samples, 400 µl tissue lysis buffer (p3) and 5 µl RNase enzyme were added, vortexed, and incubated in a thermocycler for 10 minutes at 65˚C with 300 revolutions per minute (rpm). After that, 130 µl of P3 buffer was added, vortexed, and incubated on ice for 5 minutes to precipitate the proteins and polysaccharides. To remove the cell debris and precipitates, the lysate was centrifuged for 5 minutes at 15,000 rpm using an Eppendorf centrifuge 5424R device. Next, 550–600 µl of the lysate was pipetted into a QIAsheredder spin column placed in a 2 ml collection tube and centrifuged for 2 min at 15,000 rpm.

After transferring the flow-through to a 1.5 ml Eppendorf tube, 1.5 volumes of wash buffer (AW1) were added and mixed by pipetting. Following this, 650 µl of the mixture were transferred into Dneasy mini spin column and placed on the 2 ml collection tube. The tube was spun at 10,000 rpm for one minute, and the flow-through was discarded. This procedure was carried out again with the remaining mixture.

To remove impurities such as proteins and polysaccharides, the Dneasy mini spin column was placed in a new 2 ml collection tube. It was washed twice with the AW2 wash buffer: first at 10,000 rpm for one minute and then, at 15,000 rpm for 2.5 minutes. An additional cleaning step using 99.8 % EtOH was performed, centrifuging at 15,000 rpm for 2.5 minutes to remove secondary metabolites.

The spin column was transferred to a 1.5 ml Eppendorf tube. 80 µl of elution buffer (EB; Qiagen, Valencia, California, USA) was added and incubated at room temperature for 5 minutes, followed by centrifugation at 10,000 rpm for one minute. Another 30 µl of EB was added, incubated for 5 minutes at room temperature, and then centrifuged at the same speed for one minute.

The quality and quantity of DNA was verified using a 1% agarose gel electrophoresis stained with GelRed, and the Qubit 3.0 fluorometer with a dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, United States). The average DNA concentration was estimated at 38.43 ng/µl. A 60 µl sample was sent to Macrogen company for library preparation and sequencing. The library was prepared using the TruSeq DNA PCR-Free kit (550) with a 550 bp insert size. Genome-wide paired-end sequencing was conducted on the Illumina NovaSeq 6000 platform (PE 2 x 150 bp).

2.3. Library preparation and MinION sequencing

The library was prepared using Oxford Nanopore Technologies' (ONT) ligation sequencing kit (SQK-LSK110) in accordance with the manufacturer's protocol. The high molecular weight genomic DNA from *Ferula communis* (39.6 ng/µl) was diluted with 23 µl of nuclease-free water to reach a final volume of 49 µl, meeting the minimum requirement of 1 µg or 100-200 fmol of DNA for R9.4.1 flow

cells. Subsequently, the DNA repair and end-prep steps were carried out by combining 47 µl of DNA with 1 µl of DNA control strand (DNA CS), 3.5 µl of NEBNext FFPE DNA Repair buffer, 2 µl of NEBNext FFPE DNA repair mix, 3.5 µl of Ultra II End-prep reaction buffer, and 3 µl of Ultra II End-prep enzyme mix in a 0.5 ml thin-walled PCR tube, resulting in a final volume of 60 µl. The mixture was gently flicked to ensure thorough mixing, followed by centrifugation and incubation at 20°C and 60°C for 5 minutes each.

Next, the end-prep reaction mixture was transferred to a clean 1.5 ml LoBind Eppendorf tube. Then, 60 µl of resuspended AMPure XP beads were added, and the tube was gently flicked to mix the content. The mixture was incubated on a hula mixer (rotator mixer) at room temperature for 5 minutes. The end-prep reaction mixture was then subjected to centrifugation. It was placed on a magnetic rack to separate and to pellet the components until the eluate became clear and colorless. While the tube was still attached to the magnet, the supernatant was carefully pipetted out, and the pellet was washed twice with 200 µl of freshly prepared 70% ethanol, ensuring the pellet remained undisturbed. After the second wash, the tube was spun down, and placed back on the magnetic rack. Any residual ethanol was pipetted out, and the pellet was air dried for 30 seconds.

Subsequently, the tube was removed from the magnetic rack, and the pellet was resuspended in 61 µl of nuclease-free water and incubated for 2 minutes at room temperature. The resuspended mixture was then placed back on the magnetic rack until the eluate became clear and colorless. Finally, 61 µl of the eluate was carefully removed and transferred to a clean 1.5 ml Eppendorf DNA LoBind tube. One µl of elute sample were quantified for its concentration using Qubit fluorometer and the remaining 60 µl were used for the adapter ligation step.

The adapter ligation step involved combining 60 µl of DNA sample, 25 µl of ligation buffer (LNB), 10 µl of NEBNext quick T4 DNA ligase, and 5 µl of adapter mix F (AMX-F) in a tube. The mixture was gently mixed by flicking the tube and then spun down before being incubated at room

temperature for 10 minutes. To clean the ligated DNA fragments and remove unligated adapters and contaminants, 40 µl of resuspended AMPure XP beads were added to the reaction. After flicking the tube to mix the content, the reaction mixture was incubated on a hula mixer at room temperature for 5 minutes. Following incubation, the reaction sample was centrifuged, and the tube was placed on a magnet rack to pellet the components. The supernatant was carefully removed. This washing step was repeated twice, with the addition of 250 µl of Long Fragment Buffer (LFB) each time, followed by removal of the supernatant.

To remove any remaining supernatant, the tube was spun down, placed on a magnetic rack, and the residual supernatant was pipetted out. The pellet was allowed to air dry for 30 seconds. Next, the tube was taken off the magnetic rack, and the pellet was resuspended in 15 µl of EB. After resuspension, the tube was spun down and incubated at room temperature for 10 minutes. Subsequently, the tube was placed back on the magnetic rack until the eluate became clear and colorless. Finally, 15 µl of eluate containing the DNA library was transferred to a clean 1.5 ml Eppendorf DNA LoBind tube, and one µl of the eluate was quantified for concentration using a Qubit fluorometer. The final concentrations obtained for the three libraries prepared for MinION sequencing were 27.8 ng/µl, 37.4 ng/µl, and 48.4 ng/µl, respectively.

In order to avoid a detrimental effects on sequencing throughput caused by loading more than 50 fmol of DNA, each library was diluted with 10 µl of EB, resulting in a final volume of 24 µl that could be loaded onto the flow cell in two rounds.

Before loading the samples and starting the sequencing, the following steps were performed: the MinION Mk1c lid was opened, the flow cell placed on the device, and 800 µl of priming mix was pipetted into the flow cell via the priming port. The flow cell was left for 5 minutes. During this 5 minutes, the library was prepared by mixing 37.5 µl of sequencing buffer II (SBII), 25.5 µl of loading beads II (LBII), and 12 µl of the DNA library, resulting in a final volume of 75 µl. The SpotON sample

cover lid was lifted from the flow cell and 200 µl of the priming mix was loaded via the priming port. 75 µl of the DNA library was loaded into the flow cell via the SpotON sample port in a dropwise manner. Finally, the SpotON and priming ports were closed, the MinION Mk1c lid was replaced to cover the flow cell, and the sequencing was initiated to run for 72 hours.

After 48 hours, the MinION MK1c sequencing device was paused, and the flow cell was washed with a mixture of 398 µl of wash diluent (DIL) and 2 µl of wash mix for 60 minutes. The waste was then removed through the waste port, and the library was reloaded using the same protocol as explained above. The sequencing continued for an additional 24 hours. Each of the three libraries was sequenced in three different flow cells, undergoing two rounds of sequencing each.

The sequencing and generation of raw fast5 files were conducted on the MinION MK1c device using the MiniKNOW software version 4.2.8 from Oxford Nanopore Technologies. To expedite processing time, the live base-calling feature was disabled. Subsequently, the raw fast5 files underwent base-calling using Guppy version 5.0.7 software (Oxford Nanopore Technologies). The '`device cuda: 0`' was activated to leverage GPU (Graphics Processing Unit) for faster processing, and the '`super-accurate base-calling`' or '`sup`' model option was chosen to achieve super high-accuracy in base-calling.

2.4. Quality check for raw Illumina and ONT reads

The quality of raw Illumina sequence reads from whole-genome sequencing was assessed using FastQC software version 0.11.3 (Andrews, 2010). For trimming ambiguous bases, low-quality bases and Illumina adapters, Trimmomatic version 0.39 (Bolger et al., 2014) was employed with the following settings: '`ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10`', '`SLIDINGWINDOW:4:15`', and '`MINLEN:50`'.

The parameter 'ILLUMINACLIP:TrueSeq3-PE-2.fa:2:30:10' specifies removal or clipping of adapters using the TrueSeq3-PE-2.fa adapter sequence file, with 2 as the maximum mismatch count, 30 as the palindromic clip threshold, and 10 as the simple clip threshold. 'SLIDINGWINDOW:4:15' trim bases when the average quality in a 4-base window falls below a score of 15. The 'MINLEN:50' parameter ensures only reads longer than 50 bases post-trimming are retained. The clean data was then used for assembly of nuclear and organelle genomes.

The ONT raw reads from the three libraries were concatenated into a single fastq file, which was subsequently cleaned from adapters and chimeric reads using Porechop version 0.2.4 (Wick et al., 2017), with the '--discard_middle' flag. Afterwards, the adapter-trimmed ONT sequence reads were further processed using NanoFilt version 2.8.0 (De Coster et al., 2018). Only bases with a quality score (Phred+33) above 10 were retained, and only reads longer than 500 bp were kept. This step enhances data reliability and accuracy.

2.5. Assembly of the nuclear genome

The flow chart illustrating the main steps of assembly and annotation, as well as the corresponding software, is presented in Figure 4. Additionally, a brief introduction to various algorithms used for genome assembly is provided in the Supplementary Material Appendix 2.

2.5.1. Estimation of the size and hetrozygosity of the nuclear genome

The short subsequence frequency distribution, also known as the k-mer spectrum, was used method to estimate the genome size. The k-mer distribution depth was calculated from Illumina reads using the Jellyfish software version 2.3.0 (Marçais and Kingsford, 2011), with a k-mer sizes of 17, 21, and 23. The Jellyfish k-mer counts were converted to histogram files using the 'histo' command and used for further analysis. The genome size and heterozygous of the *F. communis* genome were estimated using a k-mer size of 21, as recommended by the authors of GenomeScope (Vurture et al., 2017).

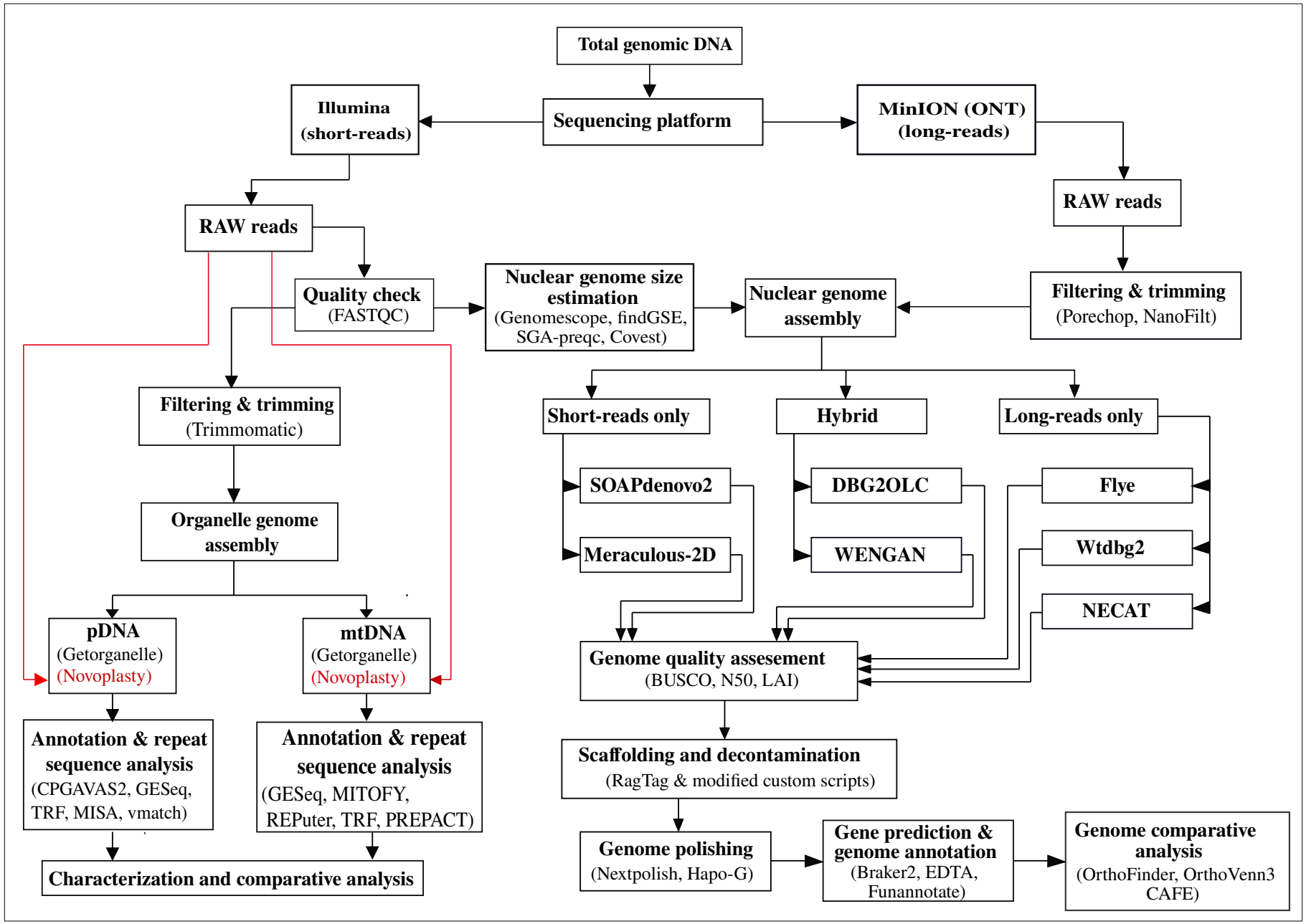

Figure 4. The flowchart for assembling the nuclear, plastid, and mitochondrial genomes of *Ferula communis*. Software used for genome assembly and analysis is listed in brackets. The red arrow indicates raw Illumina sequence data utilized by Novoplasty.

Five software: GenomeScope (Vurture et al., 2017), findGSE (Sun et al., 2018), SGA-preqc (Simpson and Durbin, 2012), Covest Basic, and Covest Repeat (Hozza et al., 2015) were used to estimate the genome size. Since Covest does not estimate hetrozygosity, Genomescope and findGSE were employed for this purpose.

2.5.2. Nuclear genome assembly and quality assessment

2.5.2.1. *De novo* genome assembly using Illumina reads

Two *de novo* genome assembly tools, SOAPdenovo2 (Luo et al., 2015) and Meraculous-2D (Goltsman et al., 2017) were used to assemble the *F. communis* genome from Illumina reads. SOAPdenovo2 and Meraculous-2D genome assemblers provide significant advantages over other short-read genome assemblers. They excel in handling short reads, such as those generated by Illumina paired-end or mate-pair sequencing, enabling better assembly of complex genomes with repetitive regions, and structural variations. These assemblers are adept at managing reads with high and low coverage, resulting in more comprehensive genome assemblies that encompass regions with low coverage or missed sequencing.

For the genome assembly of *F. communis* with SOAPdenovo2 pipeline, I utilized k-mer sizes of 31, 51, 63, and 71, while keeping other parameters at their default values. For Meraculous-2D, k-mer sizes of 41 and 51 were used, and the diploid mode flag was set to 2. All the other parameters were left at their default values. The diploid mode setting 2 is employed to preserve alternative paths from each bubble within the assembly graph and is recommended for genomes that display a high degree of heterozygosity (Goltsman et al., 2017).

2.5.2.2. *De novo* genome assembly using long ONT reads

Despite recent advancements, long reads sourced from ONT still exhibit relatively higher error rate in raw sequences compared to standard NGS devices like Illumina (Stoler and Nekrutenko, 2021). Two primary techniques are used for *de novo* assembly using single-molecule sequencing (SMS) long reads:

one involves correcting read errors before genome assembly and then using these error-corrected reads for assembly, the other involves assembling the genome with error-prone reads and then correcting them afterward (Chen et al., 2021). For the assembly of the *F. communis* genome, I employed the NECAT software, which utilizes the former correction technique, while Flye and Wdbg2 adopt the latter method.

I executed the NECAT pipeline with the following parameters: '`MIN_READ_LENGTH=1000`', '`PREP_OUTPUT_COVERAGE=10`', '`USE_GRID=true`', while leaving the remaining parameters at their default values. For Flye, I enabled a scaffolding step using the graph method ('`-scaffold`'), specified a genome size of 2.4 GB ('`-g 2.4g`'), and set a specific number of polishing iterations with the ('`-i 2`') parameter. In the wtdbg2 pipeline, I used the following settings: sequencing technology flag '`-x ont`' and estimated genome size '`-g 2.4g`', keeping the other settings at their default values.

2.5.2.3. Hybrid genome assembly

Hybrid genome assembly combines short-accurate, high-quality second-generation sequencing data (e.g. Illumina) and long, error-prone third-generation sequencing data (e.g. PacBio and Oxford Nanopore Technology) to assemble genomes and tackle complex repetitive DNA segments (English et al., 2012). A significant advantage of hybrid assembly is the utilization of error correction algorithms specifically designed for each sequencing technology. By combining precise short reads with long error-prone ones, inaccuracies in latter can be amended, yielding a more accurate final assembly. Moreover, long reads offer important information about genetic variations and haplotypes (Sedlazeck et al., 2018a), enabling the assembly of diploid or polyploid genomes and the differentiation between haplotypes.

However, hybrid assembly also comes with challenges. Integrating data from multiple sequencing technologies adds complexity to the assembly process and may require additional computational resources, longer processing times, and specialized algorithms. For a successful hybrid assembly, meticulous integration and processing of datasets are essential, considering the disparities in read lengths, error tendencies, and data formats. These factors can lead to potential complications and biases. It is important to note that each sequencing technology has its own strengths and limitations, and these can impact the accuracy of the final assembly.

Considering factors such as computational speed and the algorithm employed for genome assembly, I chose two hybrid genome assemblers, WENGAN (Di Genova et al., 2021) and DBG2OLC (Ye et al., 2016), for assembling the *F. communis* genome.

WENGAN is a hybrid genome assembler that follows a series of procedures aimed at optimizing the accuracy and contiguity of the genome assemblies. For the assembly of the *F. communis* genome, I employed this pipeline, utilizing the Abyss2 de Bruijn graph assembler to construct short-read contigs with the '`ontraw`' option for long Nanopore reads. The assembly was conducted using an estimated genome size of 2.4 GB, while the other parameters were maintained at their default settings.

The DBG2OLC pipeline (Ye et al., 2016) incorporates the SparseAssembler (Ye et al., 2012), which leverages the de Bruijn graph to generate contigs from short-reads. Following this, DBG2OLC aligns and anchors the contigs produced by SparseAssembler with the long-reads, creating an optimized overlay graph. The final assembly step involves using Sparc (Ye and Ma, 2016) to generate a consensus assembly, which results in the final assembled genome. For the assembly the default parameter set was utilized.

2.5.2.4. Genome quality assessment

To choose the best assembled genome from among the different genome assembly pipelines, Benchmarking Universal Single Copy Orthologs (BUSCO) (Simão et al., 2015) was initially employed

to assess genome completeness for single-copy orthologs. Those assembled genomes with a BUSCO score exceeding 90% were then scaffolded with RagTag version 2.1.0 (Alonge et al., 2022), using the genome of *Corindrum sativum*, *Apium graveolens*, and *Daucus carota* as references. The contig N50, which measures genome contiguity, and the BUSCO score were subsequently calculated using QUAST version 5.2 (Gurevich et al., 2013) and BUSCO version 5.4.2 (Simão et al., 2015), respectively.

In addition to these, I also evaluated the quality of assemblies using other metrics, including the long-terminal repeat (LTR) assembly index (LAI), mapping rates, base-level error rates, and structural variant error rates. The LAI quantifies the proportion of intact LTR sequences in the genome, independent of genome size (Ou et al., 2018). A higher LAI score indicates a more contiguous and complete assembly (Ou et al., 2018). The base-level and structural error rates and structural variant error rates were determined with Qualimap version 2.2.1 (Okonechnikov et al., 2015) and Sniffles version 1.0.8 (Sedlazeck et al., 2018b), respectively. This was achieved by mapping both short and long reads to the final assemblies utilizing Bowtie2 (Langmead and Salzberg, 2012) for the former and and Ngmlr (Sedlazeck et al., 2018b) for the latter.

2.5.2.5. Assembly decontamination

To eliminate any sequences with significant matches to fungal, bacterial, protozoan, or viral sequences, the non-BUSCO contigs were blasted against the local copy of the NCBI NT database using the ncbi-blast+ version 2.12.0 tool (Sayers et al., 2022). The search parameters were set to a maximum of 10 target sequences to keep in the output (‘`-max_target_seq`’), a maximum of one high-scoring segment pairs (‘`-max_hsps`’), and a maximum threshold expected value (‘`-evalue`’) of 1e-25. I manually curated the BLAST results to identify contigs containing only fungal, bacterial, protozoan, or viral sequences. These contaminant sequences were then removed from the assembly using a modified custom script from the *Drosophila* genome assembly paper workflows (Kim et al., 2021). Finally, I

characterized the draft version of the *F. communis* genome for its assembly statistics, gene prediction, and functional annotation metrics. This was done using the GenomeQC pipeline, with *Arabidopsis thaliana* TAIR10.1_chr as a gold standard reference genome (Manchanda et al., 2020).

2.5.2.6. Genome polishing

Genome polishing is a critical procedure used to rectify errors in the draft genome assembly and enhance the reliability of genome analysis (Huang et al., 2022a). A polished genome provides a more reliable foundation for downstream analyses, such as gene prediction, functional annotation, comparative genomics, and evolutionary studies. In order to improve the quality of assembled draft genome of *F. communis*, two genome polishing software tools, Nextpolish (Hu et al., 2020) and Hapo-G (Aury and Istace, 2021), were selected based on their assembly accuracy and speed.

The initial long-read genome assembly was polished using Nextpolish (Hu et al., 2020) with default parameters. The NextPolish software consists of two core modules: an error correction module and a consensus generation module. By utilizing a stepwise approach, NextPolish aims to address errors in the reference genome. The error correction module thoroughly examines the genome, identifying potential errors like base-level inaccuracies, misassemblies, and gaps. The subsequent consensus generation module utilizes the corrected regions obtained from the error correction module to construct a more precise and reliable consensus sequence, representing a polished version of the genome (Hu et al., 2020).

The Hapo-G tool (Aury and Istace, 2021), similar to other equivalent polishing tools, performs faster on simple and homozygous genomes; yet, it also improves the polishing of heterozygous genomic regions (Aury and Istace, 2021). In the homozygous genomic region, NextPolish appears to be the best performer for obtaining high-quality results, whereas Hapo-G outperforms other polishing tools in the heterozygous genomic region. It has been recommended to use NextPolish in both regions with Hapo-G to achieve good quality in homozygous and heterozygous regions (Aury and Istace,

2021). In order to achieve an improved outcome, the draft *F. communis* genome underwent polishing with Nextpolish following the genome assembly stage, and subsequently with Hapo-G after removing contaminants.

2.5.2.7. Identification and annotation of transposable elements

To identify and annotate TEs in the *F. communis* genome, I utilized a combination of homology-based and *de novo* approaches with Repeatmasker version 4.1.4 (Hubley S, 2013; Flynn et al., 2020). Two sets of libraries were employed: (1) reference repeat libraries obtained from the Repbase database (release 20181026) (Bao et al., 2015), and (2) species-specific, high-quality non-redundant TE repeat libraries created using the Extensive *de novo* TE Annotator (EDTA) pipeline (Ou et al., 2019). The *de novo* repeat libraries improve the accuracy of TEs detection and annotations.

2.5.2.8. *Ferula communis* genome annotation

The draft genome of *F. communis* was annotated using the Funannotate pipeline version 1.8.14 (Jon Palmer, 2019) and Braker pipeline version 2.1.6 (Lomsadze, 2005; Stanke et al., 2006; Iwata and Gotoh, 2012; Buchfink et al., 2015; Hoff et al., 2019; Brůna et al., 2020, 2021). First, I cleaned the assembled genome with the Funannotate '`clean`' command. Then, the cleaned genome was sorted and renamed with the Funannotate '`sort`' command using default parameters to simplify contig names for prediction Augustus software version. Then, I soft-masked the cleaned and sorted genome with Repeatmasker using the custom repeat library generated from TE identification and annotation step. Tandem repeats were identified with the Tandem Repeat Finder (TRF) version 4.0.9 software (Benson, 1999). This cleaned, sorted, and soft-masked genome was used for both gene prediction and annotation pipelines.

To generate protein homology evidence in Braker, I used the high-quality genomes of closely related species of Apiaceae (*Apium graveolens*, *Coriandrum sativum* and *Daucus carota*) and the

publicly available the OrthoDB database for Viridiplantae (https://v100.orthodb.org/download/odb10_plants_fasta.tar.gz). ProtHint (Brůna et al., 2020), a protein mapping pipeline, was used to generate protein hints from supplied protein evidence. ProtHint automatically determined protein alignments for close or distant relatives. Finally, I used the Augustus gene prediction tool (Stanke et al., 2006), trained with the ProtHint output results, to predict genes in the Braker pipeline.

For gene prediction using the Funannotate pipeline, I used the embryophyta database (https://busco-archive.ezlab.org/datasets/prerelease/embryophyta_odb10.tar.gz) and protein evidence from closely related species of Apiaceae (*Apium graveolens*, *Coriandrum sativum,* and *Daucus carota*) to train Augustus (Stanke et al., 2006), Glimmerhmm (Majoros et al., 2004), and Snap (Korf, 2004). The identification of tRNAs in the outputs generated by both prediction pipelines was carried out using the tRNAscan-SE version 2.0 software (Chan and Lowe, 2019; Chan et al., 2021).

The predicted genes from Braker2 and Funannotate were subjected to functional annotation after the protein-coding genes were checked for different protein-coding models from different databases. The protein-coding models from Pfam domains (Finn et al., 2014), CAZYme (Drula et al., 2022), proteases (MEROPS) (Rawlings et al., 2018), and BUSCO (Simão et al., 2015) groups were used for functional annotation. Additionally, I incorporated additional annotations from InterPro terms (Paysan-Lafosse et al., 2023), the Gene Ontology (GO) (Ashburner et al., 2000; Carbon et al., 2021), and Clusters of Orthologous Groups (COGs) (Tatusov et al., 2003) to enhance the functional annotation. These diverse databases were processed to generate functional annotations for *F. communis* protein-coding genes using '`funannotate annotate`' command in the Funannotate pipeline.

2.5.2.9. Analysis of gene orthology, expansion and contraction of gene families

To infer gene orthology and orthogroups, I performed a comparative genome analysis of *F. communis* with three other Apiaceae species (*Apium graveolens*, *Coriandrum sativum* and *Daucus carota*), along with three outgroup species (*Brassica carinata, Vitis vinifera,* and *Lactuca sativa*). For the analysis, I utilized the draft protein-coding sequences of *F. communis* generated from the Braker2 genome annotation pipeline mentioned earlier. The protein-coding sequences for *Apium graveolens*, *Coriandrum sativum, Daucus carota* and *Brassica carinata* were obtained from the celery genome database (CGD:http://celerydb.bio2db.com/), *Vitis vinifera* from the grape genomics database (http://www.grapegenomics.com/pages/VvCabSauv/download.php), and *Lactuca sativa* from LettuceGDB database (https://www.lettucegdb.com/genome).

The OrthoFinder version 2.5.4 (Emms and Kelly, 2015) pipeline was used to infer othologs. By default, OrthoFinder infers orthologs from orthogroup trees (a gene tree for the orthogroup) with the following steps. The protein-coding sequences are provided by the user in one FASTA file per species, that contains the amino acid sequences for protein in that species. The orthogroups were inferred among amino acid sequences using an all-versus-all blastp search with an e-value threshold set to 10e-3 (Emms and Kelly, 2015). Putative cognate gene-pairs were identified and connected in the form of orthograph generated by the software, based on normalized BLAST bit scores. The MCL graph clustering algorithm was employed to cluster the gene families into single-copy and multi-copy gene families. Then, the unrooted gene trees were inferred from each orthogroup using DendroBLAST (Kelly and Maini, 2013) or from multiple sequence alignment using mafft version 7 (Katoh and Standley, 2013). The construction of these unrooted trees was carried out using FastTree version 1.0 (Price et al., 2009), a maximum-likelihood phylogenetic tree inference software.

Once the unrooted gene trees were inferred, Orthofinder proceeds to construct a species tree based on the collection of gene trees using Species Tree Inference from All Genes (STAG) algorithm (Emms and Kelly, 2018). STAG utilizes information from the gene trees to infer a species tree. By considering the gene trees collectively, STAG can account for the complex evolutionary processes, such as gene duplication and loss, that occur in different lineages (Emms and Kelly, 2018).

To root the tree, OrthoFinder employs Species Tree Root Inference from Gene Duplication Events (STRIDE) method (Emms and Kelly, 2017). Rooting a tree involves determining the common ancestor of all species and placing it as a root of the tree.

The analysis of the expansion and contraction of gene families across lineages were conducted using the maximum likelihood approach with CAFÉ software version 5 (De Bie et al., 2006) using default parameters. CAFÉ employs a stochastic birth and death process, which is used to simulate the evolutionary dynamics of gene gain and loss across a given phylogenetic tree.

The observed family size distribution is compared to the expected distribution generated by the model. If the observed distribution significantly deviates from the expected distribution ($P < 0.01$), it indicates a significant difference in gene family size among taxa (De Bie et al., 2006).

2.5.2.10. Comparative genome analysis

The comparative genome analysis among four Apiaceae species, namely *Apium graveolens*, *Coriandrum sativum*, *Daucus carota*, and *F. communis* was conducted utilizing OrthoVenn3 (Sun et al., 2023), an online whole-genome comparative analysis tool. OrthoVenn3 employs advanced algorithms such OrthoFinder (Emms and Kelly, 2015) among others to infer orthologous gene families through a combination of sequence similarity searches, clustering methods, and graph-based algorithms (Sun et al., 2023).

The inferred orthologous gene families were then analyzed using a Venn diagram-based approach. OrthoVenn3 compares the gene families across species and identifies shared and unique gene

clusters among species. The functional annotation of the identified gene clusters in OrthoVenn3 was performed using various databases, such as Gene Onthology (GO) (Carbon et al., 2021), Kyota Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2000), and InterPro (Blum et al., 2021).

Except Venn diagrams, OrthoVenn3 generates heatmaps, and hierarchical clustering trees to illustrate the shared and unique gene clusters across species as well as the functional annotations of these clusters.

2.6. Organelle genome assembly

2.6.1. Plastid genome assembly

The plastid genome of *F. communis* was assembled *de novo* using GetOrganelle version 1.7.1 (Jin et al., 2020) and Novoplasty version 4.3.1 (Dierckxsens et al., 2016). Clean, filtered Illumina pair-end reads were used with GetOrganelle, employing k-mer lengths of 21, 45, 65, 85, and 127 under default settings (Jin et al., 2020). In the initial step, Bowtie2 (Langmead and Salzberg, 2012) was utilized to map the reads to seed sequences from the default plant seed databases in GetOrganelle. These seed-mapped reads were used as initial "baits" to recruit more target-associated reads. Subsequently, all reads were assembled using SPAdes, *de novo* assembler tool (Bankevich et al., 2012), through the GetOrganelle pipeline. SPAdes generated output for each k-mer, including an assembly graph in FASTG format. This assembly graph captured the connections between contigs, representing them as a graph that could contain allelic polymorphism and assembly uncertainty (Jin et al., 2020). The assembly was further curated using the '`slim_fastg_by_blast`' script under default settings to assess its quality and then visualized with Bandage version 0.8.1 (Wick et al., 2015).

For the Novoplasty *de novo* organelle genome assembler, raw Illumina reads were used without filtering or quality trimming, as recommended by the software developers (Dierckxsens et al., 2016). A k-mer size of 39 was employed to assemble the plastid genome of *F. communis*, with the plastid

genome of *Daucus carota* (accession number: NC_008325) adopted as a seed input and reference sequence. The other options were set to default.

2.6.1.1. Plastid genome annotation and repeat sequence analysis

The plastid genome was annotated using two web server-based annotation pipelines, namely CPGAVAS2 (Shi et al., 2019) and GESeq (Tillich et al., 2017). In CPGAVAS2, the annotation process utilized the 2,544 plastome datasets, which contain the largest number of plastome sequences among similar tools, along with RNASeq validated or corrected sequences. Blast+ version 2.8.1 (Camacho et al., 2009) was internally employed to search for homologous proteins and tRNA sequences in reference protein and tRNA databases, respectively. The set of homologous genes identified through BLAST were specifically searched for tRNA using tRNAScan-SE version 2.0.2 (Chan and Lowe, 2019) and ARAGON version 1.2.36 (Laslett and Canback, 2004). The identification of inverted repeat (IR) regions and dispersed repeats was performed using vmatch version 2.3.0 (Kurtz, 2010). Additionally, two different pipelines, MISA version 1.0 (Beier et al., 2017) and TRF version 4.0.9 (Benson, 1999), were utilized for the identification of simple sequence repeats (SSRs) ranging from 1 to 6 bp in length, and long tandem repeats (greater than 6 bp), respectively. The annotations obtained from these pipelines were then saved in GFF3 files using Maker version 2.31.10 (Campbell et al., 2014).

For GESeq, a core annotation pipeline based on a BLAT-driven best match approach was utilized, supplemented by additional profile HMM searches for protein, rRNA-coding, and *de novo* prediction of tRNA genes (Tillich et al., 2017). Two databases were employed for the annotation of protein-coding (CDS) and non-protein-coding (rRNA and tRNA) sequences. Third party tRNA annotators such as tRNAscan-SE version 2.0.7 (Chan and Lowe, 2019) and ARAGORN version 1.2.38 (Laslett and Canback, 2004), along with the NCBI reference sequence set (*Daucus carota,* accession number: NC_008325), were used in conjunction with the BLAT search, while keeping all other settings at default values.

Since GESeq does not internally incorporate repeat identification software, repeat sequences and their patterns were analyzed with the MISA with a parameters set at ('1-8 2-4 3-4 4-3 5-3 6-3') for simple sequence repeats (SSRs), and the web-based TRF for tandem repeats, with parameter set at 2, 7, and 7 for match, mismatch, and indel, respectively. The position of forward, reverse, complement, palindromic, and dispersed repeats were identified using the web-based tool REPuter (Kurtz et al., 2001), with 90% sequence identity, a Hamming distance of 3, and a minimum repeat size of 30 bp.

The circular chloroplast genome maps were drawn using Organellar Genome DRAW version 1.3.1 (Greiner et al., 2019).

2.6.1.2. Sequence divergence and selective pressure analysis

To asses the nucleotide sequence diversity within Ferulinae, the pDNA of *F. communis* and eight other *Ferula* plastid genomes, including *F. ovina* (accession number: ON324036), *F. sibirica* (accession number: ON324038), *F. transiliensis* (accession number: ON324040), *F. gigantea* (accession number: ON324042), *F. fedtschenkoana* (accession number: ON324043), *F. litwinowiana* (accession number: ON324045), *F. olivacea* (accession number: ON324046), and *F. renardii* (accession number: ON324048) were downloaded from the NCBI database. All nine pDNA genomes were aligned using mafft version 7 (Katoh and Standley, 2013). A sliding window analysis was conducted to generate the nucleotide diversity of pDNA genome using the DnaSP version 5.10 software (Librado and Rozas, 2009). The step size was set to 200 bp, with an 800 bp window length. Additionally, comparative nucleotide diversity analysis was also performed on *Daucu carota,* accession number: NC_008325) following the same procedure explained above. The gene boundaries on the inverted regions were visualized with IRscope, an online program for visualizing the junction sites of plastids (Amiryousefi et al., 2018).

To identify protein-coding genes under selection in *Ferula*, a set of 79 protein-coding genes was aligned using mafft (Katoh and Standley, 2013). Subsequently, a maximum likelihood phylogenetic tree based on the coding sequences (CDS) was constructed using RAxML version 8.2 (Stamatakis, 2014). To calculate the nucleotide substitution rates for non-synonymous (dN) and synonymous (dS) mutations, along with their ratio (ω = dN/dS), the codeml program from the PAML package version 4.10 (Yang, 2007) was employed. This program is specifically designed for analyzing and estimating selection pressure acting on protein-coding genes. The purpose of these analyses was to identify genes that exhibit evidence of positive or negative selection, as indicated by significant deviations in the dN/dS ratio from the neutral expectation of ω = 1. For this analysis, a guideline were established as follows: dN/dS < 0.6 indicates purifying selection, dN/dS ranging from 0.6 to 1 represents neutral selection and, dN/dS > 1 signifies positive selection.

However, traditional models assume a single ω ratio (the ratio of non-synonymous to synonymous substitutions) for all sites in a gene, assuming that selection operates uniformly across the gene. In reality, selection pressure can vary among different amino acid positions. To address this, site-specific models were employed using 79 protein-coding genes in *Ferula*. These models allow the estimation of ω values independently for each site, enabling the identification of sites under positive selection (ω > 1), negative selection (ω < 1), or neutral evolution (ω = 1). Specifically, the site-specific models M0, M1a, M2a, M7, and M8 were used in the analysis, allowing the ω ratio to vary among sites while maintaining a fixed ratio in all branches. Likelihood ration test (LRT) was conducted to evaluate the selection strength and the p-values of Chi square smaller than 0.05 was thought as significant, and the Bayes Emperical Bayes (BEB) analysis were used to calculate posterior probabilities for positively selected sites within genes ($p < 0.05$). It's is important to note that the BEB calculation is performed under models of positive selection only (i.e., M2a and M8) but not under the null models (i.e., M1a or M7) (Álvarez-Carretero et al., 2023).

The following parameters was used in codeml : '`seqtype = 1`', '`model = 0`', '`Nssites = 0, 1, 2, 3, 7, 8`', and codon frequency '`F3 x 4`' model. Likelihood rates were then compared between the the site-specific models M0 (one ratio) vs. M3 (discrete), M1a (neutral) vs. M2a (positive selection), and M7 (beta) vs. M8 (beta and ω) to detect positive selection at specific sites (Yang and Nielsen, 2002; Zhang et al., 2005). The M7 model in the codeml program of PAML package uses a beta distribution to describe how different parts of a gene can change over time—rapidly indicating positive selection, slowly indicating purifying selection or remaining unchanged indicating neutrality.

2.6.2. Mitochondrial genome assembly

The mitochondrial genome of *Ferula communis* was assembled using the same pipelines that were used for assembling plastid genomes: GetOrganelle and NOVOplasty.

The GetOrganelle pipeline was employed with various parameter combinations, encompassing word size ('`-w`'), k-mer size ('`-k`'), number of rounds to extend reads ('`-R`'), and the option to use a custom seed database with and without ('`-s`') (*F. sinkiagensis* accession number: OK585063.1). A total of 10 analyses were conducted with different parameter combinations, involving variation in the number of rounds to extend reads ('`-R 30`' and '`45`'), word size ('`-W 120`' and '`125`'), and k-mer size ('`-k 55,85,95,115`' and '`55,85,95,115,125`'). Additionally, two specific combinations were employed: one with k-mer size ('`-k 55,85,95,115`') word size ('`-w 69`'), and another with k-mer size ('`-k 55,85,95,115,125`') and word size ('`-w 83`'), both used in conjunction with custom seed database ('`-s Ferula sinkiagensis`') and 45 rounds of extension ('`-R 45`'). These settings facilitated the extraction and assembly of the mitochondrial genome from the trimmed and filtered Illumina sequence reads.

On the other hand, the NOVOplasty software employed the raw Illumina sequences for the assembly of the mitochondrial genome. Since NOVOplasty relies on input seed sequences to assemble the mitochondrial genome, the mitochondrial genome of *F. sinkiagensis* (accession number: OK585063.1) was used as a seed, with a k-mer size of 39.

2.6.2.1. Mitochondrial genome annotation and repeat sequence analysis

The assembled genomes were subjected to annotation using two mitochondrial genome annotation tools: GeSeq (Tillich et al., 2017) and MITOFY (Alverson et al., 2010). The mitochondrial genomes of *Daucus carota* (accession number: NC_017855)*, Nicotina tabacum* (accession number: NC_006581), and *Arabidopsis thaliana* (accession number: NC_037304) were used as references. The protein search identity threshold was set to 55% for rRNA, tRNA, and 85% for protein coding sequences.

In addition to GeSeq, the mitochondrial genome of *F. communis* was annotated using MITOFY software with default parameters. The MITOFY web server utilized tRNAscan-SE version 1.3.1 and NCBI blast version 2.11.0+ programs to annotate tRNA and protein-coding genes, respectively. The annotated mtDNA genomes were further checked and manually adjusted using ugene version 43.0 (Okonechnikov et al., 2012). The mitochondrial map was generated using Organellar Genome DRAW (OGDRAW) (Greiner et al., 2019). To identify SSRs, the microsatellite identification web server tool MISA was employed. The analysis focused on identifying SSRs with repeat units of 1, 2, 3, 4, 5, and 6 bases, and repeat numbers of 8, 4, 4, 3, 3, and 3, respectively. The inverted, palindromic and direct repeats with a minimal repeat size of 20 bp, and a Hamming distance of 3, were investigated using REPuter software (Kurtz et al., 2001). Lastly, the TRF was employed with default parameters to detect tandem repeats larger than 6 bp.

2.6.2.2. RNA editing analysis

To investigate the RNA editing sites in the mitochondrial genome of *F. communis*, the protein-coding mitochondrial genes of *Arabidopsis thaliana* were employed as reference. The analysis was carried out using the Plant RNA Editing-Prediction and Analysis Computer Tool (PREPACT) version 3.12 (Lenz et al., 2018), via a web-server (http://www.prepact.de/prepact-main.php) with default parameters.

3. Results

3.1. Genome sequencing

A total of 1,399,948,868 pair-end 150-bp reads were obtained from the Illumina platform, resulting in 211.4 Gb of raw sequence data. The raw reads exhibited a GC content of 34.49%, and 90.8% of the reads had a quality score of at least Q30 (Supplementary Table 1: S1). After trimming the raw Illumina reads to remove ambiguous bases, low-quality regions, and adapters using Trimmomatic, a set of 1,371,937,762 pair-end high-quality reads was obtained and used for nuclear genome and organelle genome assembly (Supplementary Table 1: S1).

On the ONT-MinION platform, a total of 4,880,744 raw reads were generated from three libraries, resulting in 64 Gb of sequencing data. The N50 read length was 14,193 bp. Quality assessment of the raw ONT-MinION reads revealed that 34.23 Gb of sequence data had a quality score greater than Q10 (Supplementary Table 1: S2). After trimming the low-quality bases (quality score less than 10) from both ends of the reads and discarding reads shorter than 500 bp, a total of 33.8 Gb of clean sequence data with a quality score greater than Q10 was obtained (Supplementary Table 1: S3). These filtered reads consisted of 4,552,204 reads with an average length of 7,419 bp and N50 read length of 14,266 bp, which were used for nuclear genome assembly.

3.2. Estimation of nuclear genome size and heterozygocity

The estimated genome size using Covest Basic for both raw and cleaned Illumina sequences was 1,748 Mb and 1,696 Mb, respectively (Table 4). These values closely matched the genome sizes of *F. communis* and *F. glauca,* which were estimated at 1,764 Mb and 1,470 Mb, respectively, using the cytophotometric method. However, GenomeScope estimates of the genome size were 2,364 and 2,368 Mb (Table 4), aligning more closely with the genome size of *F. haufelli* (2,411 Mb) measured by the flow cytometry method. Covest Repeat provided the highest estimated genome size (3,168 Mb), while SGA-preqc (3,120 Mb) and findGSE for filtered and unfiltered reads (2,776 and 2,820 Mb) estimated

genome sizes closer to the assembled genome of *F. communis* using DBG2OLC (2,952 Mb) and Flye (2,751Mb) pipelines (see below), respectively (Table 4).

The percentage of heterozygocity estimated by both GenomeScope and findGSE indicates a value exceeding 1% but less than 3% (Supplementary Figure F1 and Supplementary Figure F2). This suggests the presence of a moderate level of genetic diversity or the presence of moderate variant of alleles at specific loci within the genome of a diploid organism (Vurture et al., 2017).

Table 4. Estimated genome sizes and heterozygosity for the genome of *F. communis* using various software

	Genome size (Mb)					**Heterozygocity (%)**	
	GenomScope	**findGSE**	**Covest Basic**	**Covest Repeat**	**SGA-preqc**	**GenomScope**	**findGSE**
Raw Illumina sequence reads	2,364	2,776	1,748	3,763	3,168	2.22	1.83
Clean Illumina sequence reads	2,368	2,820	1,696	3,811	3,120	2.22	1.79

3.3. Nuclear genome assembly

The largest and smallest genome assemblies were obtained using short-read genome assemblers, namely the SOAPdenovo2 pipeline with a k-mer size of 31 (6,341 Mb) and Meraculous-2D pipeline with a k-mer size of 41 (814 Mb) (Table 5), respectively. However, the smallest genome size assembled with Meraculous-2D had a better N50 (0.001 Mb), a smaller number of contigs (1,115,797), and a higher percentage of complete BUSCO genes (71.3%) compared to the SOAPdenovo2 assembly (Table 5).

Among the long-read genome assemblers, the Flye genome assembler produced a genome with a size of 2,751 Mb, 91.3% complete BUSCO genes, and N50 of 0.072 Mb. In contrast, NextDenovo

assembled a genome with a size of 1,117 Mb, 44.3% complete BUSCO genes, and outperforming all other genome assemblers used in the analysis with a superior N50 of 0.137 Mb (Table 5). The Wtdbg2 pipeline assembled a genome with a size of 2,035 Mb, 82.7% of BUSCO genes, and N50 of 0.051 Mb (Table 5).

Table 5. Initial genome assembly statistics for short-read, long-read and hybrid genome assemblers

Assembly	K-mer size	Length (Mb)	Contig No.	Contig N50 (Mb)	BUSCO score (Total genes : 2326)					
					Complete genes		Duplicated genes		Fragmented genes	
Short-read assembly										
SOAPdenovo2	31	6,341	33,429,388	0.000176	774	33.3%	47	2.0%	530	22.8%
	51	4,854	18,480,974	0.00037	815	35.1%	57	2.5%	514	22.1%
	63	4,877	19,050,656	0.000355	870	37.1%	54	2.3%	538	23.1%
	71	5,093	16,429,012	0.000502	847	36.4%	63	2.7%	554	23.8%
Meraculous-2D	41	814	1,115,797	0.001	1659	71.3%	482	20.7%	319	13.7%
	51	2,062	30,088,012	0.000849	509	21.8%	15	0.6%	255	11.0%
Long-read assembly										
Flye		2,751	75,814	0.072	2,123	91.3%	312	13.4%	36	1.5%
Wtdbg2		2,035	65,099	0.051	1,924	82.7%	105	4.5%	89	3.8%
NextDenovo		1,117	8,399	0.137	1,032	44.30%	66	2.8%	42	1.8%
Hybrid assembly										
WENGAN		999	71,839	0.018	1,539	66.2%	120	5.2%	172	7.4%
DBG2OLC		2,952	152,369	0.033	2,221	95.5%	905	38.9%	38	1.6%

The genome assembled using hybrid assembler DBG2OLC had a genome size of 2,952 Mb, which is the largest among the obtained genomes. It also demonstrated the highest recovery of BUSCO genes, encompassing 2,221 out of the total 2,326 BUSCO genes, resulting in a completeness of 95.5%. However, it exhibited N50 of 0.033 Mb and the highest count of contigs (152,369) when compared to the other long-read and hybrid assemblers employed. This indicates that the DBG2OLC- assembled genome is considerably fragmented in contrast to the assemblies produced by Flye and Wtdbg2 (Table 5). On the other hand, the WENGAN hybrid genome assembler generated the second smallest genome, with a size of 999 Mb, following Meraculous-2D, which consisted of 71,839 contigs and achieved recovery of 66.2% of the BUSCO genes (Table 5).

3.4. Assessment of genome assembly quality

Among the genome assemblers utilized in the initial assembly of the *F. communis* genome, two assembly outcomes were chosen for subsequent analysis. One of the criteria employed for this selection was the BUSCO score, as genomes with the highest BUSCO scores tend to have notably better N50 values. The Flye and DBG2OLC genome assemblers yielded higher BUSCO score compared to the other short-read, long-read, and hybrid genome assemblers used in the *F. communis* genome assembly, as indicated in Table 5. The genomes from these assemblers then underwent additional scaffolding and assembly quality assessment procedures.

The results obtained after the scaffolding step of initial assemblies of Flye and DBG2OLC with RagTag revealed that the *Corindrum sativum* genome, used as a reference, yielded a higher N50 value and a small number of scaffolds compared to *Apium graveolens* and *Daucus carota* references (Supplementary Table 1: S4). The quality assessment metrics after scaffolding and polishing steps indicated that the Flye assembly outperformed the DBG2OLC assembly in several aspects. The Flye assembly showed superior performance in terms of LAI score (8.65), N50 (0.17 Mb), having the smallest number of contigs (59,178), a lower error rate (0.0225) for short read, and a higher mapping

percentage for short (96.3%) and long reads (83.01%). On the other hand, DBG2OLC exhibited better results in terms of the complete BUSCO score (96.2%) and the number of structural variants (5,391) (Table 6).

Table 6. Genome assembly assessment metrics

Assembly	Length (Mb)	# of contigs	Contig N50 (Mb)	BUSCO score (Total genes : 2326)						LAI score	Short-read mapping		Long-read mapping		Structural variants
				Complete genes		Duplicated genes		Fragmented genes			Mapping rate	Error rate	Mapping rate	Error rate	
Flye	2,772	59,178	0.17	2,167	93.2%	259	11%	17	0.7%	8.65	96.3%	0.023	83.01%	0.114	7,104
DBG2OLC	2,996	130,777	0.04	2,239	96.2%	712	30.6%	29	1.2%	7.46	95.5%	0.026	65.5%	0.113	5,391

Based on these results, the Flye assembly was considered the most accurate compared to DBG2OLC, and was further examined to identify and eliminate any potential contaminant sequences. A total of 24 contigs comprising bacterial contaminant sequences with a cumulative size of 667.0 kb, along with 20 contigs that exclusively representing virus sequences spanning 516.4 kb, were identified. Subsequently they were removed from assembled genome of *F. communis* (Supplementary Table 1: S5).

3.5. Characterization of the *Ferula communis* genome

The Flye-assembled *F. communis* genome was characterized by having a genome size of 2,772 Mb, an N50 value of 0.17 Mb, an L50 value of 548, and a total of 59,178 scaffolds. The longest scaffold in the assembly was 152.51 Mb long, and approximately 2,398.9 Mb of the assembled genome consisted of scaffolds with more than 25,000 nucleotide sequences. The genome had a GC content of ≈34% and was

found to be composed of repetitive sequences accounting for ≈87% (2,365 Mb) of the genome (Table 7).

Table 7. Genome statistics for the Flye assembled *Ferula communis* genome

Assembly feature	**Size**
Assembled genome size	2,772 Mb
# of Scaffolds	59,178
Scaffold sequences (≥ 25K nt)	2,398.9 Mb
Longest Scaffold	152.51 Mb
N50	0.17
L50	548
NG50	100,326
LG50	2,063
GC content	34.35%
Total repetitive sequence	86.67%

3.6. Gene prediction and functional gene annotation

Gene prediction using eudicots and embryophyta databases, along with phylogenetically related Apiaceae species as a source of proteins, revealed variations in the number of predicted genes. Braker2 and Funannotate predicted a total of 68,318 and 79,391 genes, respectively. The disparities in gene prediction results could be attributed to the different databases and gene prediction tools utilized by the pipelines. Additionally, the tRNAscan-SE tool predicted approximately 1,262 tRNAs.

However, the comparative structural gene annotation metrics using *Arabidopsis thaliana* as a reference identified 65,090 for Braker2 and 80,369 for Funnannotate protein-coding gene models for the *F. communis* genome (Table 8). Among these metrics, the minimum gene length was 63 bp, the

maximum gene length was 78,740 bp, and the average gene length was 2,416.3 bp. Notably, the highest number of exons (332,429) were recovered from Funannotate (Table 8). Similar results were obtained for the average number of exons (4.2 and 4.1) and average number of transcripts (1 and 1) per gene model for Braker2 and Funannotate, respectively.

Table 8. Genome-wide structural gene annotation metrics for the *Ferula communis* genome in comparison to the genome of *Arabidopsis thaliana*

Annotation metrics	***Arabidopsis thaliana* TAIR10.1_chr**	***Ferula communis***	
		Braker2	**Funannotate**
Number of gene models (bp)	33,467	65,090	80,369
Minimum gene length (bp)	3	63	64
Maximum gene length (bp)	27,265	78,740	48,358
Average gene length (bp)	2,043.2	2,416.3	2,003.5
Number of exons	324,691	270,731	332,429
Average number of exons per gene model	9.7	4.2	4.1
Average exon length (bp)	307.3	224.8	201.9
Number of transcripts	53,886	68,318	79,391
Average number of transcripts per gene model	1.6	1	1
Number of gene models less than 200 bp length	1,847	52	2,545

These metrics for genome-wide functional annotation revealed that, in *F. communis,* the number of annotated genes and the maximum gene length are approximately twice that of *Arabidopsis thaliana.* However, the average gene length, remains similar to that of *Arabidopsis thaliana*. Yet, a clear distinction emerges in terms of average exon length: the average exon length in *Arabidopsis thaliana* was longer than that of *F. communis*.

Furthermore, *Arabidopsis thaliana* exhibited approximately twice the average number of exons per gene model compared to *F. communis,* despite the fact that the number of exons is similar between *Arabidopsis thaliana* and *F. communis* genomes. Unfortunately, comprehensive information about alternative splicing and the number of isoforms for each gene in *F. communis* is not available. Therefore, the average number of exons per gene model tends to be underestimated primarily due to the absence of tissue-specific transcriptomic data within the scope of this study.

3.7. Transposable elements in the *Ferula communis* genome

I found that 86.67% of the assembled *F. communis* genome consists of repetitive sequences, which is 1.9 times that of *Daucus carota* (46%) (Iorizzo et al., 2016), 1.23 times that of *Corindrum sativum* (70.59%) (Song et al., 2020), and is closer to the estimated repetitive elements in the *Apium graveolens* genome (92.91%) (Song et al., 2021). Among the TEs identified in the *F. communis* genome, the class I retrotransposon element, the long-terminal repeat (LTR) category, accounts for 71.37% of the whole genome with a total size over 1.98 Gb. The two most frequent LTR types, *Copia* and *Gypsy*, account for 31.41% and 21.45% of the total repeat sequences, respectively (Table 9). In comparison to the *Apium graveolens* and *Corindrum sativum* genomes, the two types of LTR elements in *F. communis* (31.41% *Copia* and 21.45% *Gypsy*) were found in a lower percentages than in *Apium graveolens* and in *Corindrum sativum*. The *Copia* LTR type was found in a higher number in *Corindrum sativum* (55.85%) (Song et al., 2020), while both *Corindrum sativum* and *Apium graveolens* have almost identical amounts of *Gypsy* LTR elements (36.47% and 36.57%, respectively) (Song et al., 2020, 2021). However, 18.51% of the whole genome was identified as unknown LTR TEs in the *F. communis* genome (Table 9).

Among the class II TEs, helitron (2.91%) and CACTA (2.75%) were the most abundant DNA TEs, while polinton was the least abundant.

Tandem repeat regions account for 3.30% of the genome, and 3.53% of the repeats are categorized as unclassified repeat elements (Table 9).

Table 9. Types and percentages of various repeat elements in the *Ferula communis* genome excluding duplicated genes

Class	Count	bpMasked	%Masked
LTR			
- Copia	748,678	870,292,843	31.41%
- Gypsy	470,383	594,314,065	21.45%
- unknown	761,697	512,839,495	18.51%
TIR			
- CACTA	146,636	76,293,005	2.75%
- Mutator	77,189	34,788,077	1.26%
- PIF Harbinger	31,409	10,718,568	0.39%
- Tc1 Mariner	29,089	8,127,234	0.29%
- hAT	62,352	22,498,352	0.81%
- polinton	40	8,067	0.00%
Non-LTR			
- LINE	3,358	1,655,395	0.06%
- unknown	94	74,808	0.00%
Non-TIR			
- helitron	204,532	80,707,845	2.91%
Tandem repeat region	287,049	91,338,787	3.30%
Unclassified	382,277	97,800,965	3.53%
Total TEs	2,088,560	2,401,457,506	86.67%

3.8. Orthology analysis

The comparative genomics analysis of 400,108 protein sequences from *F. communis, Daucus carota, Apium graveolens, Coriandrum sativum, Brassica carinata, Vitis vinifera,* and *Lactuca sativa* identified

40,502 orthogroups. A total of 367,846 proteins (out of 400,108, or 91.9%) were assigned to these orthogroups (Supplementary Table 1: S6). The average size of an orthogroup was 9.1 genes per species. A total of 8,343 orthogroups were shared among all seven species, and 32 of these orthogroups were entirely comprised of single-copy genes (Supplementary Table 1: S7). The highest number of species-specific orthogroups was identified in the outgroup species such as *Brassica carinata* (6,929) and *Vitis vinifera* (5,692), while the lowest was in *Apium graveolens* (386) (Supplementary Table 1: S8).

The percentage of genes assigned to orthogroups varies among these species, with *Apium graveolens* having the highest percentage (94.4%) and *Lactuca sativa* having the lowest (88.5%) (Supplementary Table S8). The number of species-specific orthologs in each species was mostly consistent with their phylogenetic relationship. For example, the outgroup species *Brassica carinata, Vitis vinifera* and *Lactuca sativa* had the highest number of species-specific orthologs (Supplementary Table 1: S8). However, among compared species from Apiaceae family, *F. communis* had the highest number of species-specific orthologs even surpassing the outgroup species *Lactuca sativa* (Supplementary Table 1: S8). The number of shared orthologs between each pair of species (Supplementary Table 1: S9).

The results of the comparative genome analysis of four Apiaceae species, *F. communis*, *Daucus carota, Apium graveolens,* and *Coriandrum sativum*, obtained from OrthoVenn3 revealed a total of 15,575 gene cluster, encompassing approximately 92,386 protein-coding genes shared among these species (Figure 5a, b). Specifically, *F. communis* shared 1,453 gene clusters with *Coriandrum sativum*, 830 gene clusters with *Daucus carota*, and 577 gene clusters with *Apium graveolens*.

Surprisingly, the number of gene clusters shared between *F. communis* and *Coriandrum sativum* was higher (1,453 gene clusters or 6,542 protein-coding genes) compared to the gene clusters shared between *F. communis* and *Daucus carota* (830 gene clusters or 4,418 protein-coding genes), despite their distant phylogenetic relationship. Among the gene clusters shared between *F. communis*

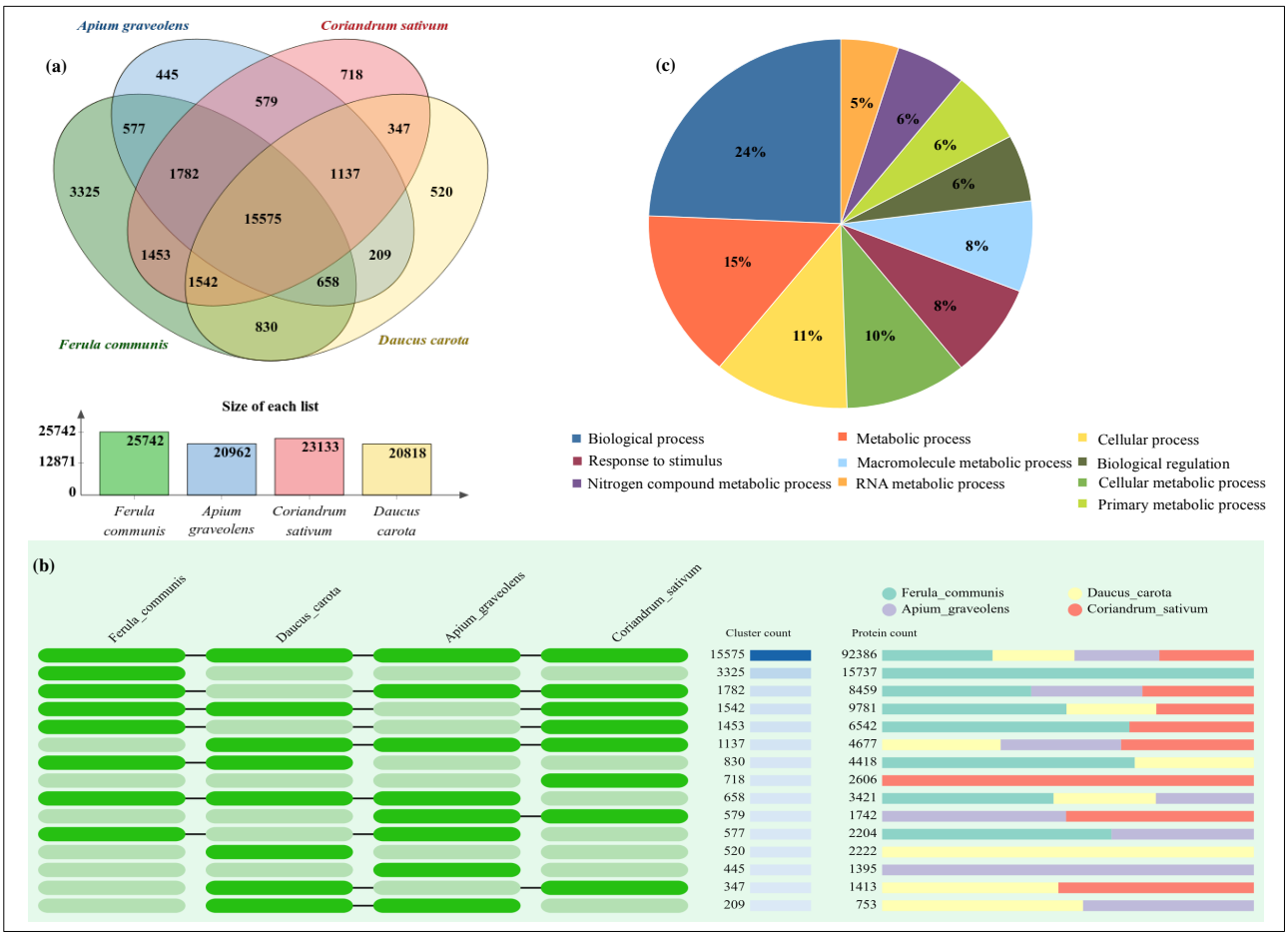

Figure 5. Comparative genomic analysis among four Apiaceae species. (a, b) Shared clusters, unique cluster and protein counts among *F. communis*, *Daucus carota, Apium graveolens,* and *Coriandrum sativum*. (c) Percentage of unique cluster genes divided according to their biological role in the *F. communis* genome.

and *Coriandrum sativum*, 2,203 genes lack annotations from GO and Swiss-prot. However, among the remaining, 82 genes responsible for transaminase activity, 64 genes for terpenoid biosynthesis process, 61 genes responsible lipid catabolic process, 56 genes negative regulation of transcription and 51 genes responsible for response to red or far red light (for the remaining refer to Supplementary Table 1: S22).

The highest number of unique gene clusters and protein-coding genes was observed in *F.*

communis (3,325 and 15,737, respectively), followed by *Coriandrum sativum* (718 and 2,606), *Daucus carota* (520 and 2,222), and *Apium graveolens* (445 and 1,395) (Figure 5a, b).

Among the unique gene clusters identified in *F. communis*, the highest number was classified under genes responsible for biological processes (24%), followed by metabolic processes (15%) and cellular processes (11%) (Figure 5c). The number of gene clusters responsible for cellular metabolic processes, stimulus response, and macromolecule metabolic processes accounted for 10%, 8%, and 8%, respectively (Figure 5c). The gene clusters identified as the RNA metabolic process in *F. communis* has the lowest percentage (4%), followed by the nitrogen compound metabolic process gene cluster (5%). The percentage of cluster genes that control biological regulation and primary metabolic activity shows comparable results (6%) in *F. communis* (Figure 5c).

The gene ontology enrichment analysis revealed that out of the 3,325 unique gene clusters identified in the *F. communis* genome, approximately 180 gene clusters, or 1,675 protein-coding genes, were assigned specific biological and molecular functions. Among these assignments, the most abundant are, 869 genes were associated with the DNA integration gene family, 279 genes with DNA recombination, 136 genes with signal transduction, 77 genes with RNA-directed DNA polymerase activity, and 75 genes with oxidoreductase activity involved in the incorporation or reduction of molecular oxygen (Table 10).

However, the remaining 3,145 gene clusters, comprising 14,062 genes were not specifically annotated in any of the databases. Instead, they have been categorized under broad ontological terms. This suggests that these gene families might represent false predictions or artifacts due to the lack of experimental evidence, such as transcriptome data or they could potentially be unique genes specific to *F. communis*, which have not yet been precisely identified as orthologs in other compared species.

Table 10. List of unique gene clusters identified through gene ontology enrichment in OrthoVenn3

SN	Ortholog ID	Gene family name	# of cluster	# of genes	P-value
1	GO:0015074	DNA integration	31	869	3.4 x 10e-27
2	GO:0016705	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	28	75	1.9 x 10e-07
3	GO:0006310	DNA recombinationg	22	279	4.6 x 10e-20
4	GO:0007165	Signal transduction	22	136	2.5 x 10e-05
5	GO:0048544	Recognition of pollen	15	43	3.1 x 10e-04
6	GO:0050832	Defense response to fungus	11	29	8.6 x 10e-04
7	GO:0006334	Nucleosome assembly	10	37	4.2 x 10e-05
8	GO:0009742	Brassinosteroid mediated signaling pathway	9	28	3.4 x 10e-06
9	GO:0016114	Terpenoid biosynthetic process	7	20	3.2 x 10e-04
10	GO:0003964	RNA-directed DNA polymerase activity	6	77	7.2 x 10e-07
11	GO:0016024	CDP-diacylglycerol biosynthetic process	6	9	1.0 x 10e-04
12	GO:0031047	Gene silencing by RNA	5	41	6.4 x 10e-04
13	GO:0009306	Protein secretion	5	13	6.4 x 10e-04
14	GO:0042659	Regulation of cell fate specification	3	19	8.5 x 10e-04
		Total	**180**	**1675**	

3.9. Gene expansion and contraction analysis

Ferula communis (27,386) and *Coriandrum sativum* (8,511) had the highest number of duplications within the Apiaceae family, followed by *Daucus carota* (4,789) and *Apium graveolens* (4,232) (Supplementary Figure F5). Notably, the most substantial gene duplications occurred at the internal node leading to the Apiaceaefamily compared to other internal nodes (Supplementary Figure F5).

Among the duplicated genes the outgroup species, *Vitis vinifera* and *Brassica carinata* exhibited the highest number of significant gene expansions, 341 and 177, respectively (Figure 6). In contrast, *Lactuca sativa* experienced a significant contraction, with 511 genes contracting instead of expanding (Figure 6).

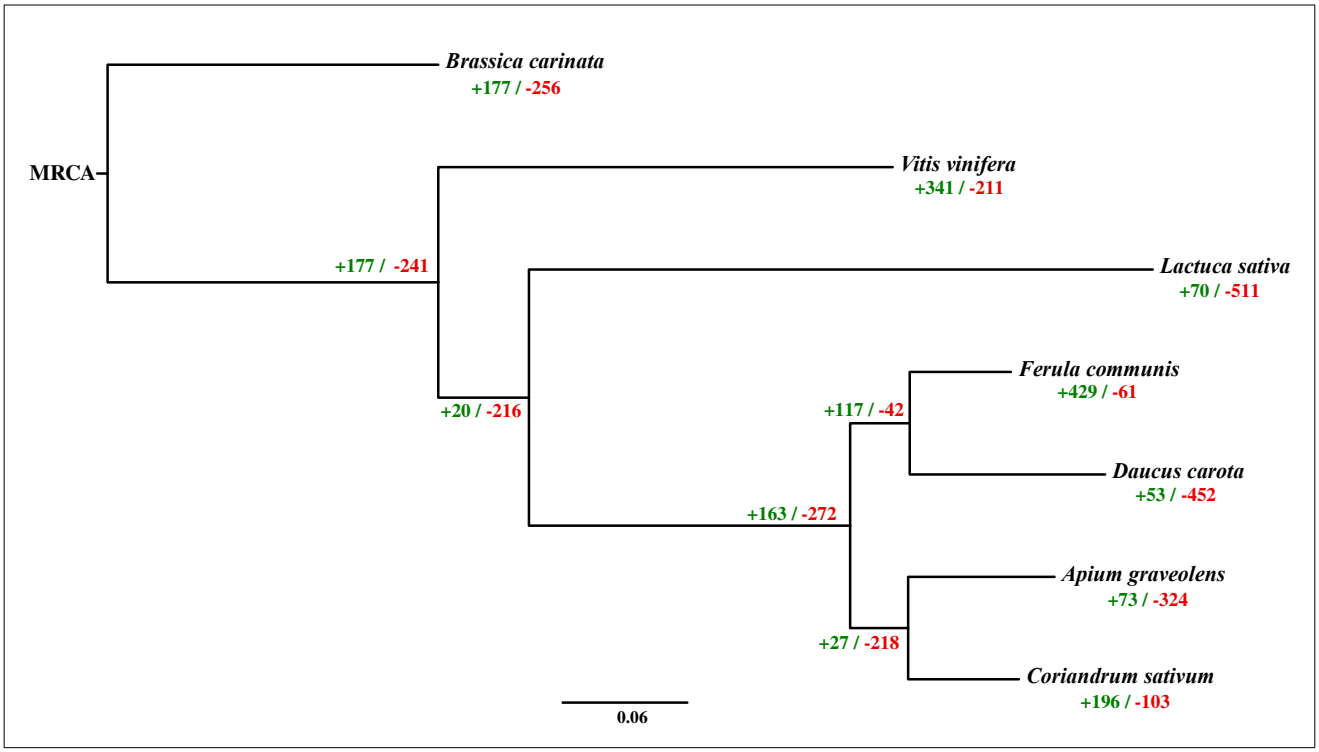

Figure 6. The expansion and contraction of gene families, along with gene duplications, modeled on the phylogenetic tree of four species from the Apiaceae family and three outgroup species are presented. The tree reconstruction was based on shared genes identified using OrthoFinder. (green = expansion, red = contraction)

When comparing the Apieaceae species, *F. communis* (429) and *Coriandrum sativum* (196) showed the highest number of significant gene expansions, while *Daucus carota* (53) and *Apium graveolens* (73) had the fewest expansions. Conversely, *Daucus carota* (452) and *Apium graveolens* (324) had the highest number of significant gene contractions, while *F. communis* (61) and *Coriandrum sativum* (103) had the fewest contractions (Figure 6).

3.10. The plastid genome of *Ferula communis*

Out of the total 1,399,948,868 whole genome Illumina sequence reads, 20,167,730 plastid reads were obtained, and 18,998,667 (94.2%) read pairs were utilized to assemble the pDNA genome of *F. communis*. The result obtained from both GetOrganelle and Novoplasty pDNA assembly resulted in a plastid genome size of 166,696 bp with a GC percentage of 38%, which is comparable to the 14 sequenced plastid genomes of *Ferula* species (Yang et al., 2022). The size of the genome is larger than those of *Daucus carota* (Ruhlman et al., 2006) and *Anthriscus sylvestris* (accession no: MT561042.1). However, when using GetOrganelle, the pDNA assembly produced two separate scaffolds (assembly paths 1 and 2) or in other words two structural haplotypes. These structures (haplotypes) have the same size but a reverse arrangement of genes in the short-single copy (SSR) region. In contrast, Novoplasty resulted in only one pDNA for *F. communis*. Since both pDNA annotation pipelines, CPGAVAS2 and GeSeq, produced identical results for *F. communis*, the CPGAVAS2-annotated pDNA genome was used for further investigation.

The pDNA of *F. communis* possessed a quadripartite structure similar to other angiosperms pDNAs, consisting of a pair of IR, SSC, and LSC regions (Figure 7). The LSC and SSC regions had sizes of 85,349 bp and 17,685 bp, respectively, and they were separated by two IR regions with a size of 31,831 bp. The pDNA genome of *F. communis* comprised 132 genes, including 87 protein-coding genes, 37 tRNA genes, and 8 rRNA genes (Figure 7 and Table 11). Among these genes, 12 protein-coding genes and 8 tRNAs contained one intron, while *ycf3* and *clpP* had two introns (Table 12).

Notably, the *ycf15* gene, which has an unknown functions, was lost in *F. communis,* instead of being reported in other *Ferula* species (Yang et al., 2022).

Based on their functions, the pDNA genes of *F. communis* can be classified into four groups: (a) genes related to photosynthesis, which include Photosystem I, Photosystem II, ATP synthase, NADH-dehydrogenase, Cytochrome b/f complex, and RUBISCO groups, (b) self-replication genes, including ribosomal and transfer RNAs, (c) non-photosynthetic genes with known functions, and (d) genes of unknown function such as *ycf1*, *ycf2*, and *ycf4* (Table 11).

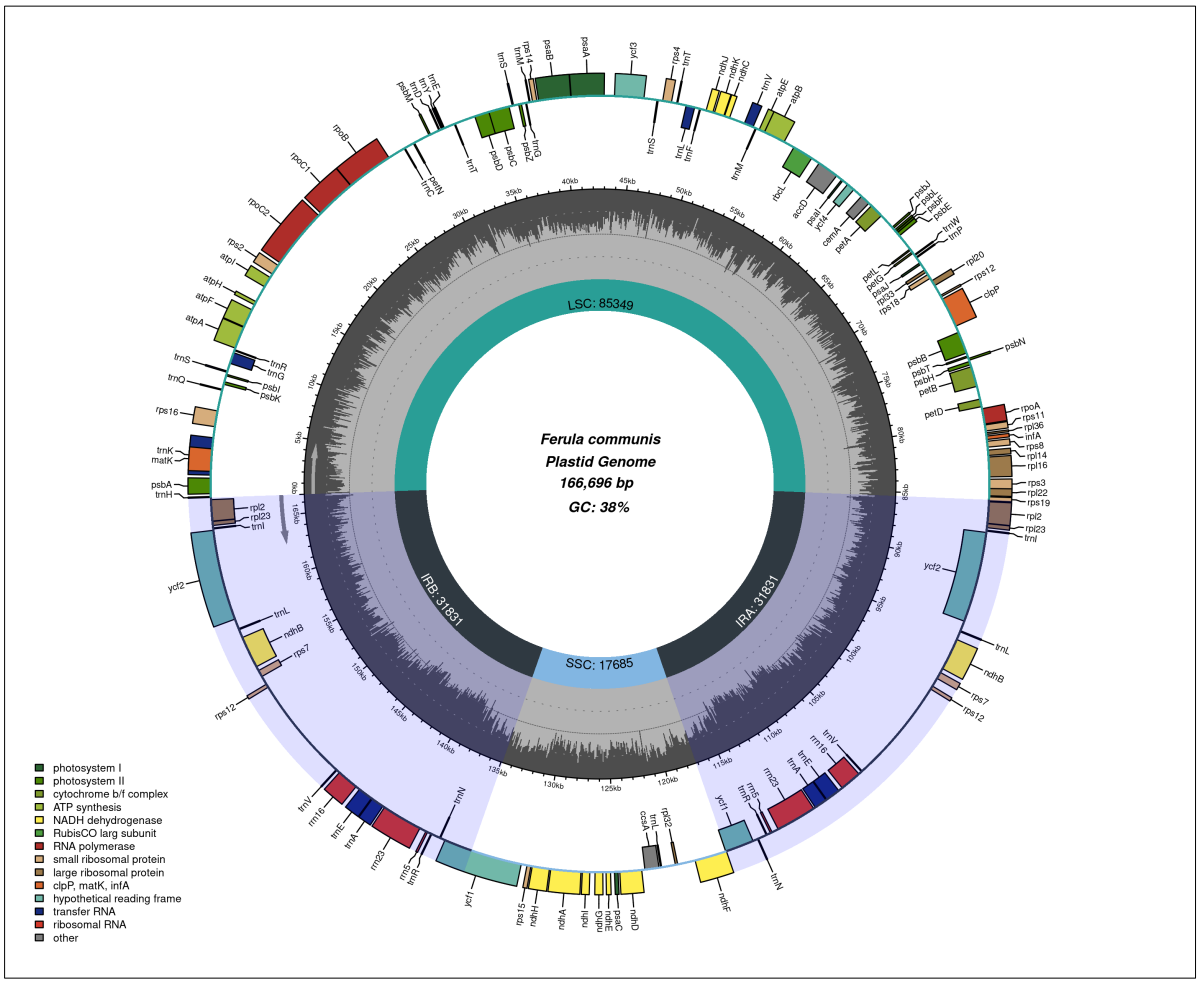

Figure 7. Map of plastid genome of *Ferula communis.* Genes inside of the circle are transcribed clockwise and those on the outside are transcribed counterclockwise. The darker gray inner circle corresponds to the GC content estimated at 5 kb window size. Different colors represent different functional genes. The inverted repeat regions are also marked (light violet).

3.10.1. Expansion and contraction of inverted repeat (IR) region

The comparison of IR boundary regions of nine *Ferula* species and *Daucus carota* showed that this region in *F. communis* was slightly different at IRb/SSC and Ira/LSC compared to *Daucus carota* and

Table 11. *Ferula communis* pDNA gene composition

Gene category	Functional groups	Gene names
Photosynthesis	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Subunit of NADH-dehydrogenase	*ndhA, ndhB, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	RUBISCO	*rbcL*
Self-replication	Large subunit of ribosome	*rpl14, rpl16, rpl2, rpl2, rpl20, rpl22, rpl23, rpl23, rpl32, rpl33, rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Small subunit of ribosome	*rps11, rps12, rps12, rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7, rps7, rps8*
Other genes	Subunit of Acetyl-CoA-carboxylase	*accD*
	C-type cytochrome synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
	Protease	*clpP*
	Translational initiation factor	*infA*
	Maturase	*matK*
Unkown	Conserved open reading frames	*ycf1, ycf1, ycf2, ycf2, ycf4*

Table 12. The lengths of introns and exons for the genes in the *Ferula communis* pDNA genome

Gene	Strand	Start	End	Exon I	Intron I	Exon II	Intron II	Exon III
rps16	-	4856	5948	40	856	197		
atpF	-	11992	13270	145	733	401		
rpoC1	-	21394	24167	432	737	1605		
ycf3	-	43589	45601	124	727	230	779	153
clpP	-	71103	73165	71	846	291	626	229
petB	+	76133	77529	6	737	654		
rpl16	-	82429	83790	9	954	399		
rpl2	-	85491	86966	391	651	434		
ycf2	+	87594	93857	4426	36	1802		
ndhB	-	95209	97423	775	682	758		
ndhA	-	127246	129313	479	1048	541		
ndhB	+	154623	156837	775	682	758		
ycf2	-	158189	164452	4426	36	1802		
rpl2	+	165080	166555	391	651	434		
trnK-UUU	-	1534	4092	37	2485	37		
trnG-UCC	+	9310	10083	32	682	60		
trnL-UAA	+	48521	49117	35	512	50		
trnV-UAC	-	52676	53316	39	567	35		
trnE-UUC	+	108435	109456	32	950	40		
trnA-UGC	+	109521	110405	37	812	36		
trnA-UGC	-	141641	142525	37	812	36		
trnE-UUC	-	142590	143611	32	950	40		

within a similar range to other *Ferula* species. In five *Ferula* species—*F. communis, F. transiliensis, F. gigantea, F. fedtschenkoana* and *F. renardii*—the *ndhF* gene extends into the SSC region by a maximum of 16 bp. On the other hand, in *F. olivacea, F. litwinowiana, F. sibirica* and *F. ovina,* the *ndhF* gene were positioned by 0-16 bp away from the Irb/SSC boarder.

Furthermore, the *trnH* gene on the LSC/Ira boundary extended 3-5 bp into the Ira region in all *Ferula* species except *F. olivacea* and *Daucus carota*. The *ycf1* gene had variable length among *Ferula* species, whereas *Daucus carota* had the smallest *ycf1* gene length in the Irb/SSC boundary region (Supplementary Figure F4). The inverted repeat (IR) region exhibited slight variation in size within the *Ferula* species.

3.10.2. Analysis of repeat elements

The pDNA of *F. communis* genome contained a total of 217 SSRs. The majority of these SSRs were located in the LSC/SSC regions, accounting for 82.01% of the total SSRs (Figure 8a). Mono-, di-, tri-, and tetranucleotide SSRs represented 65%, 26%, 1.84%, and 4.15% of the total SSRs, respectively. Hexanucleotide SSRs were found to be very rare in the pDNA genome of *F. communis* (Figure 8c). Among the mono-nucleotide SSRs, A/T repeats constituted the vast majority, accounting for 61.75%, while C/G mono-nucleotides were rarely present (3.22%). Among di-nucleotide SSRs, AT/TA repeats were the most common, accounting for 20.28% (Figure 8b).

Apart from SSRs, I analyzed the repeat sequences of *F. communis* pDNA genome using REPuter and Tandem Repeats Finder, and classified the sequence repeat motif into three categories: forward, palindromic, and tandem repeats. In the *F. communis* pDNA genome, 33 forward repeats, 28 palindromic repeats, and 24 tandem repeats were identified, including two with 11 and 16 indels (Supplementary Table 1: S11).

3.10.3. Sequence divergence and divergence hotspot regions

The result showed that the IR region was the most conserved region compared to the LSC and SSC among the nine pDNA genomes of studied Ferula species. A similar pattern was observed between *Daucus carota* and *F. communis*. However, the most hypervariable regions ($\pi > 0.006$) within *Ferula* species were identified in the intergenic spacer regions, such as *rps16-trnQ, psbI-trnS-trnG, atpH-atpI,*

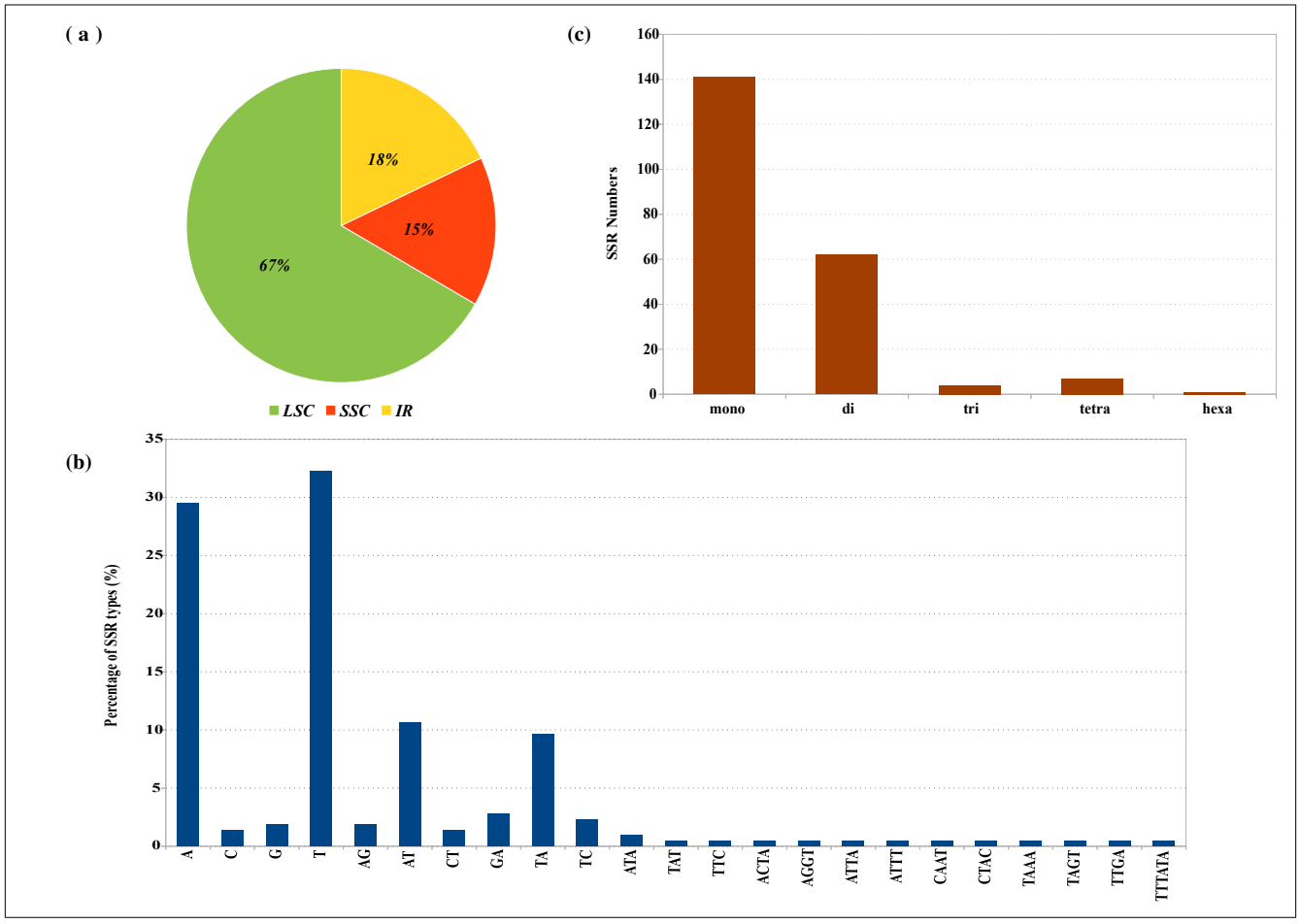

Figure 8. Distribution of simple sequence repeats (SSRs) in the *Ferula communis* pDNA genome. (a) The percentage of SSRs in LSC, SSC, and IR regions, (b) Percentage of particular SSR motifs in the genome, (c) Number of SSR types detected.

rpoB-trnC, trnC-petN-psbM, rps4-trnT-trnL, ndhC-trnV, psbH-petB, ycf1-ndhF, ndhF-rpl32 and *ndhH-rps15-ycf1*. Among these intergenic spacers, nine of them were found in the LSC region, and two of

them are located in the SSC region of the pDNA. Additionally, three genes, *clpP*, *rpl16* located in LSC as well as *ycf1* located in SSC exhibited high nucleotide diversity (Figure 9).

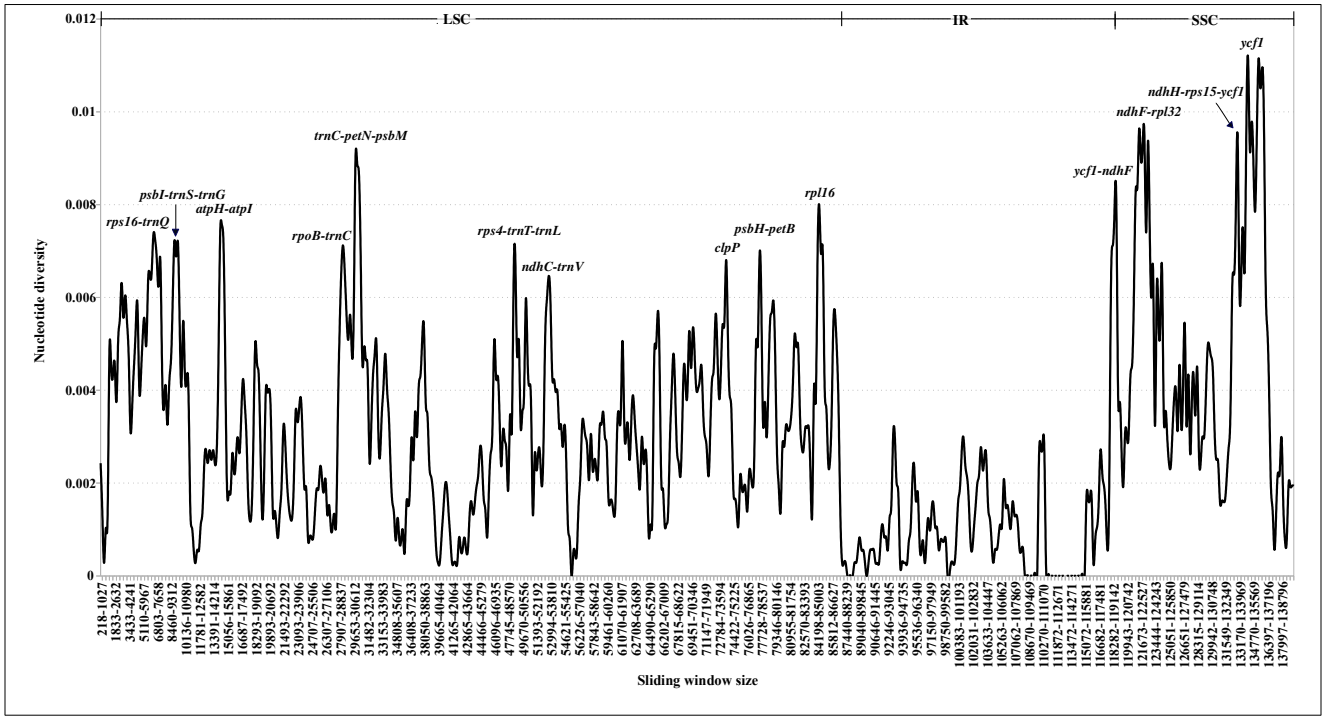

Figure 9. Sliding window analysis of the pDNA genome of nine *Ferula* species (window length: 800 bp, step size: 200 bp).

3.10.4. Positive selection analysis

To assess the selective pressure on the protein-coding genes, the non-synonymous (dN) to synonymous (dS) substitution rates were analyzed for 79 CDS within nine *Ferula* species. The traditional model that considers single ω ratio across the gene indicated that the *ccsA* and *ycf2* genes are the two positively selected genes in *Ferula* having dN/dS ratio greater than 1.00 (Figure 10). The *atpE*, *matK*, and *rps8* genes have dN/dS ratio between 0.60 and 1.00. All genes, with the exception of *ndhB*, *rpoC2*, *rps12,* and *rps15,* are under purifying selection with dN/dS ratio less than 0.60 (Figure 10). However, the site specific model using the Bayes Empirical (BEB) analysis identified seven genes with positively

selected sites within *Ferula* species (Supplementary Table 1: S12). These genes include *ndhF*, *rpoC2*, *ccsA*, *matK*, *rpl32*, and two genes of unknown function, *ycf1* and *ycf2*.

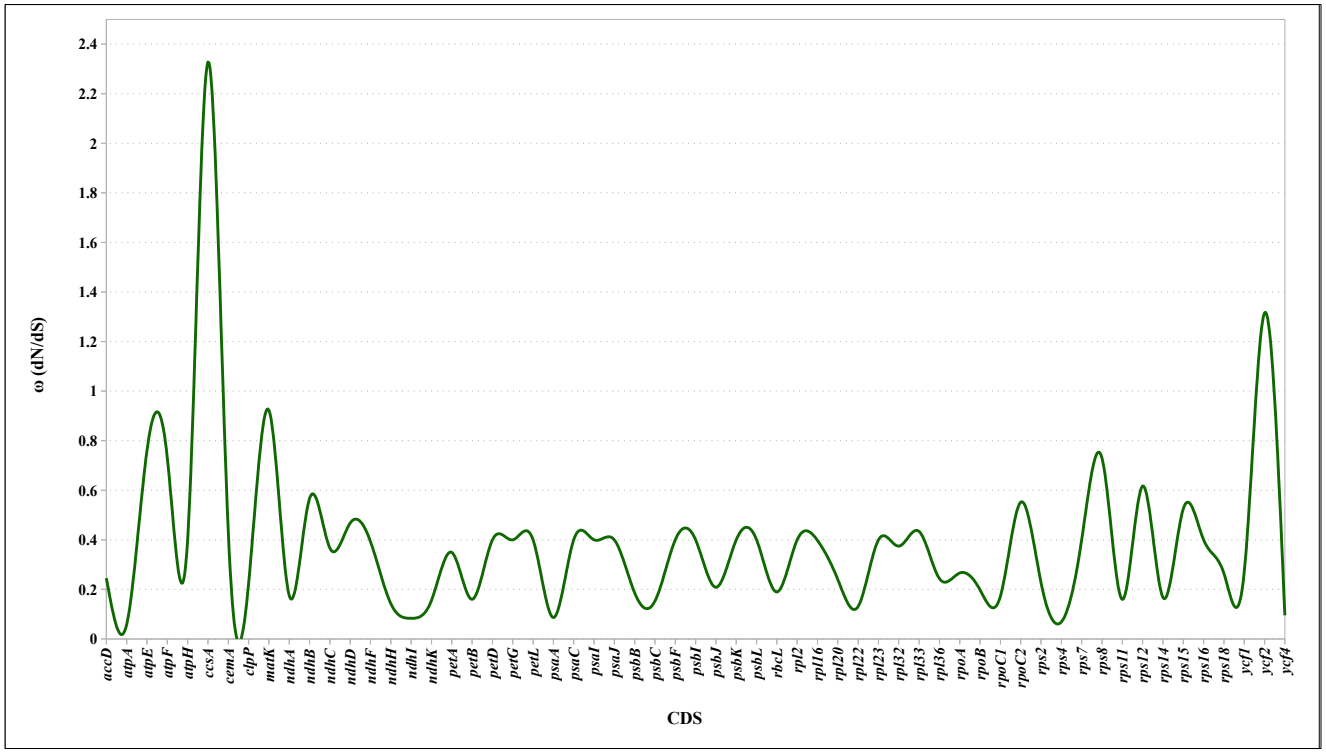

Figure 10. Non-synonymous to synonymous substitution within nine *Ferula* species.

Specifically, these models of positive selection identified *rpoC2* as having 27 positively selected sites, followed by *matK* with 20, and *ndhF* with 18. Additionally, *ycf1* had 16, while both *ccsA* and *ycf2* had 15 positively selected sites each, and *rpl32* has two positively selected sites (Supplementary Table 1: S12). However, the likelihood ratio between the tested models (M1a vs. M2a, and M7 vs. M8) provided strong evidence for the presence of positively selected codon sites only in the *ccsA* gene (Supplementary Table 1: S13).

3.11. The mitochondrial genome of *Ferula communis*

3.11.1. Mitochondrial genome assembly and annotations

The mitochondrial genome assembled using GetOrganelle with different assembly parameters resulted in two mtDNA genome sizes: 250,058 bp and 250,278 bp (Supplementary Table S14). Among the

assembly parameters used, the k-mer size, regardless of the number of round for assembly extension ('R') and word size ('w'), had a significant impact on the assembly result. Specifically, a k-mer size of 125 with different combination of extension rounds ('-R 30' and '45'), and word sizes ('-w 69, 83, 105, 120, 125') yielded the largest mitochondrial genome assembly of 250, 278 bp, with or without a reference genome used. The assembly with the longest k-mer was selected for downstream analysis.

However, instead of a single circular mitochondrial genome, all assembly parameter result produced a genome with 16 scaffolds, with a GC content of 46%. On the other hand, the NOVOplasty pipeline was unsuccessful in assembling the mtDNA genome of *F. communis*.

Compared to other mitochondrial genomes within the Apiaceae family, the mitochondrial genome of *F. communis* was found to be larger only than that of *Corindrum sativum* (82.9 kb), falling within a similar size range (212–463.79 kb) of other four compared mitochondrial genomes, such as *Daucus carota*. Using two mitochondrial genome annotation software (GeSeq and Mitofy), 37 protein-coding genes, which were classified into seven groups: NADH dehydrogenase (9 genes), ATP synthase (5), cytochrome C biogenesis (4), cytochrome C oxidase (3), ribosomal protein (SSU) (8), ribosomal protein (LSU) (3), transport membrane protein (1), mutarase (1), ubiquinol cytochrome c reductase (1), 3 rRNA, and 20 tRNA gene were identified (Supplementary Table 1: S14). The functional annotation and position of the annotated genes are presented in Figure 11.

Among the identified mitochondrial genes, eight protein-coding and three tRNAs had introns of different length (Figure 11). For instance, the *nad1* gene had an intron that is 58,268 bp long, while the *trnL-GAU* gene had an intron length of 15 bp. The *nad7* gene had two copies in the mtDNA genome, and each copy contains four introns, followed by *nad4* and *nad2* with three and two introns, respectively (Table 13).

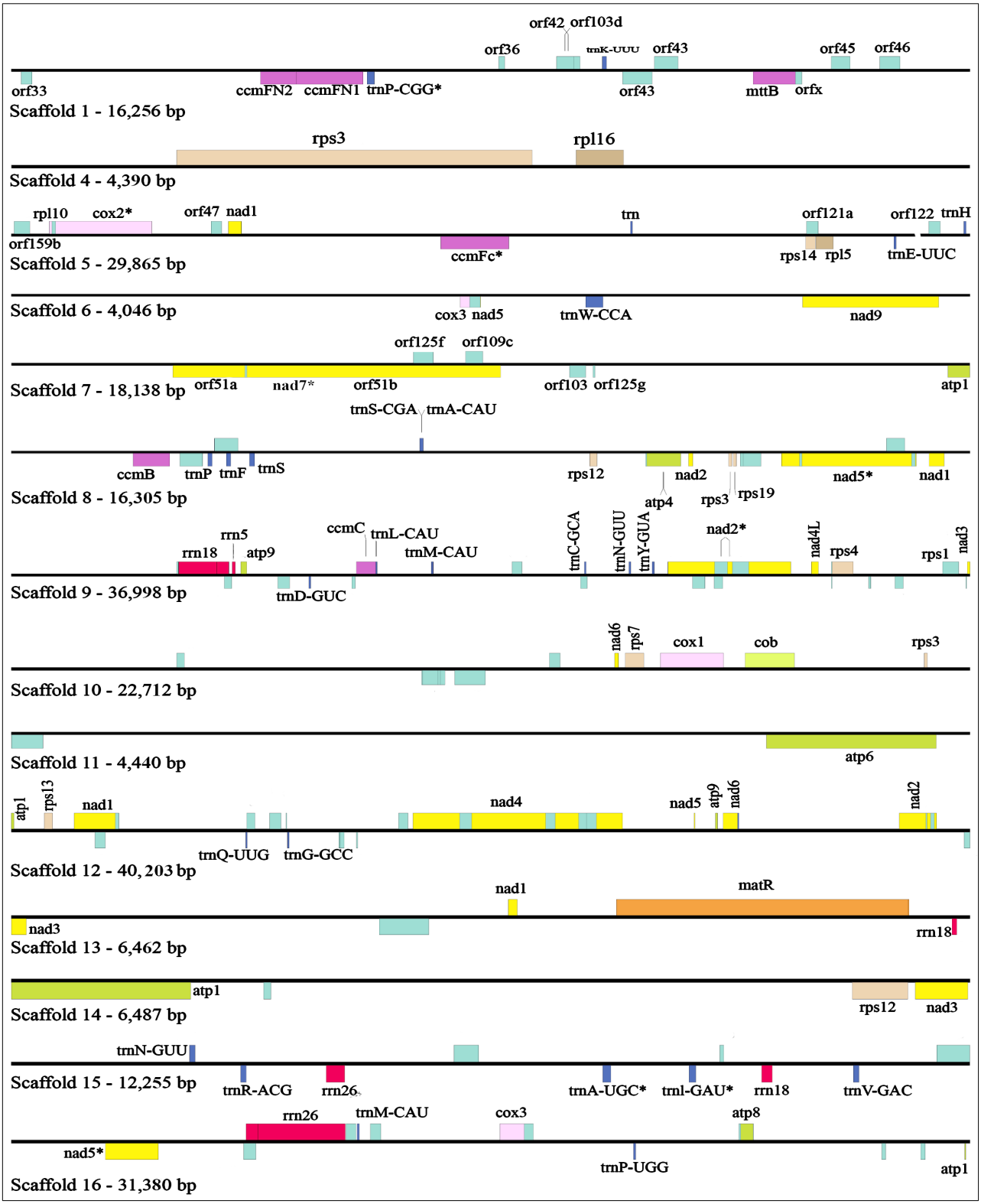

Figure 11. Map of mitochondrial genome for 14 scaffolds of *Ferula communis.* Genes, marked by asterisk, posses introns.

3.11.2. Repeat sequence analysis

Using the web-based Microsatellite MISA tool, a total of 183 SSRs were identified in the mtDNA genome of *F. communis*, and the proportion of different forms were shown in the Figure 12. The

Table 13. *Ferula communis* mitochondrial genes with introns

Genes	Intron		
	Start	End	Length
rpl5	32,486	68,760	36,274
cox2	45,273	46,613	1,340
cox2	46,924	48,004	1,080
atp1	79,927	84,097	4,170
nad1	50,875	109,143	58,268
nad2	150,504	152,816	2,312
nad2	153,391	154,896	1,505
nad4	183,752	185,154	1,402
nad4	185,671	188,820	3,149
nad4	189,241	191,989	2,748
nad5	109,633	109,700	67
nad7	93,120	94,020	900
nad7	94,090	95,423	1,333
nad7	95,890	96,865	975
nad7	97,112	98,910	1,798
trnA-UGC	214,227	214,303	76
trnL-GAU	215,338	215,353	15
trnP-CGG	6,039	6,125	86

highest percentage of SSRs in the *F. communis* mtDNA genome was observed in the monomer and dimer forms, accounting for 82% of the total SSRs, while the hexamer form of SSRs was the least frequently identified, representing only 1.09% (Figure 12). Among the monomer repeats, A/T repeats constituted 78% (58 out of 74 monomer SSRs), and among the dimeric SSRs, AG/CT repeats were the most frequent, accounting for 61%. Additionally, the *F. communis* mtDNA genome contained 5 trimeric, 20 tetrameric, 5 pentameric, and 2 hexameric SSRs (Supplementary Table 1: S16). Furthermore, the mtDNA genome of *F. communis* harbored a total of six tandem repeats (Table 14). The lengths of these repeats varying from 14 to 71 bp, and their similarity exceeded 90%.

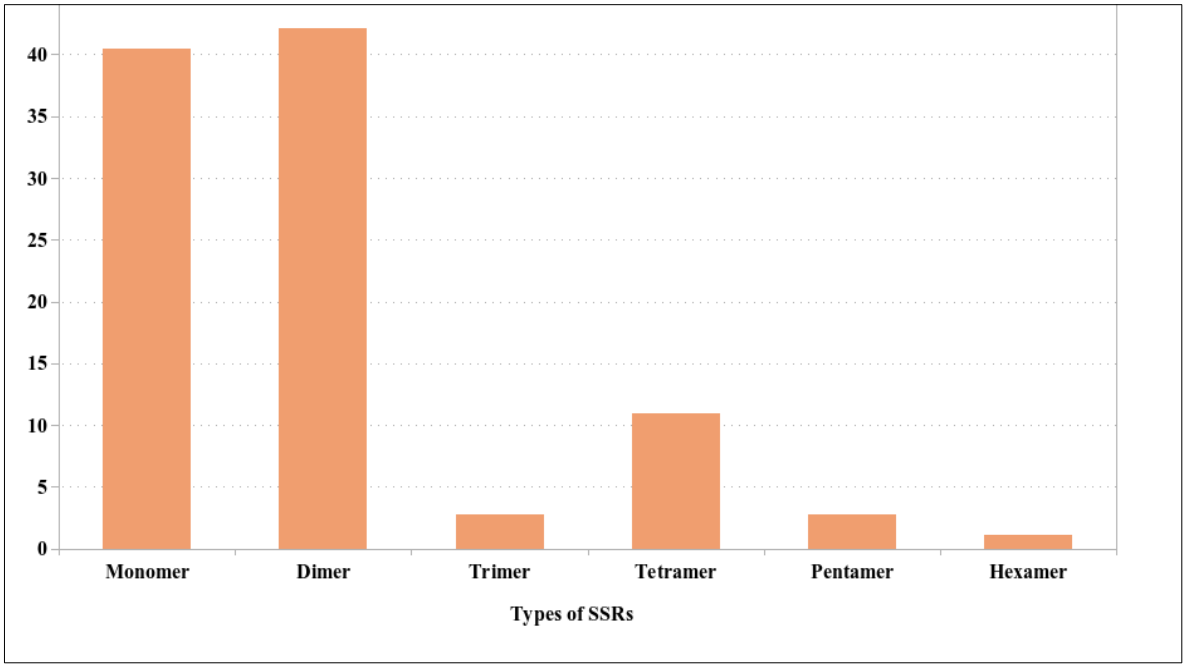

Figure 12. Types of SSRs and their distribution percentage in the *Ferula communis* mitochondrial genome.

The non-tandem repeat analysis with the REPuter software detected 50 repeat sequences with a length equal to or greater than 20 bp. Among these, 28 are forward, and 22 are palindromic repeats. The longest forward repeat is 250 bps, while the longest palindromic repeat is 157 bps. The majority of the non-tandem repeats are found in the 125-141 bp length range (Table 15, Supplementary Table 1: S17).

Table 14. Tandem repeat sequence in the *Ferula communis* mitochodrial genome

SN	Size	Repeat sequence	Copy	Similarity %	Start	End
1	26	TATGACACCAAGAATGTACAAGCGAG	2	91	39,579	39,628
2	18	GCAGCTAACCCCCCCATC	3	94	62,523	62,574
3	71	TGAAGAAAACAGACTTTAGCAAGGTGGTTAAGGTAG CTCAGCTGGTTAGAGCAAAGGACTGAAAATATCCT	2	91	121,168	121,303
4	14	ATAATAATATATAT	2	100	141,055	141,081
5	14	AATAATATATAAAC	2	90	141,057	141,090
6	32	TGGTTTTTTCATGTTGTCAAAGAGTTGAACAA	2	97	210,164	210,229

3.11.2. RNA editing site analysis

The analysis results from PREPACT identified 385 RNA editing sites within 25 protein-coding genes in the *F. communis* mtDNA genome. The highest number of RNA editing sites was recognized in *nad4L* (41), followed by *nad2* (31), *nad7* (27), and lowest number of RNA editing sites were observed in *atp9* (3) and *rps7* (2) genes, respectively (Supplementary Table 1: S18).

Table 15. Distribution of non-tandem repeats in the *Ferula communis* mitochondrial genome

SN	Repeat length (bp)	Types of repeats	
		Forward	Palindromic
1	59-75	3	5
2	76-91	6	2
3	92-108	2	0
4	109-124	3	2
5	125-141	13	12
6	> 142	1	1

Of these RNA editing sites, 70.69% (275) were located on the second codon position and 29.30% (114) occurred on the first position of the triplet (Supplementary Table S19 and S20). Among second position modifications, 23.58% (91) changed the amino acid serine (S) to leucine (L), 23.06% (89) proline (P) to leucine (L), 16.84% (65) serine (S) to phenylalanine (F) (Supplementary Table 1: S19). The results also suggested that the RNA editing might lead to the premature termination of protein-coding genes, and this phenomenon is likely to occur in *atp9* and *atp6-2,* where the triplets coding the amino acid glutamine (Q) changed to stop codon (Supplementary Table 1: S21).

4. Discussion

4.1. The *Ferula communis* genome assembly methods, metrics and assembly quality

The nuclear genome of *F. communis* was assembled using various genome assembly pipelines, which encompassed short-read, long-read, and hybrid assemblers. Although the same sequences were used for genome assembly, varying assembly results were observed. For example, in terms of genome contiguity, the long-read genome assembler Flye stood out. Meanwhile, the hybrid assembler DBG2OLC surpassed others in recovering highly conserved gene completeness, as evidenced by the BUSCO genes. On the other hand, the most fragmented genome with the lowest assembly metrics emerged from short-read assemblers such as SOAPdenovo2 and Meraculous-2D, as well as the hybrid genome assembler WENGAN (Table 6). Several factors contribute to these disparities, among them the unique algorithmic approaches employed by genome assemblers and their varying strategies for managing repetitive regions (Schatz et al., 2010; Jung et al., 2020).

Specifically, regarding algorithmic methodology, both short-read and hybrid assemblers initially utilized the de Bruijn graph algorithm. This algorithm has its limitations, a major one being that it decomposes reads into shorter k-mers and employs a k-mer coverage cutoff to eliminate low-frequency k-mers. Such a process can result in information loss and potential misassembly of the genome since rare but crucial sequences may be discarded (Sze et al., 2017). Moreover, unlike the overlap graph algorithm where each read is represented by a single node, the de Bruijn graph algorithm produces multiple nodes for each read. These nodes may not create a linear path when integrating edges from other reads (Schatz et al., 2010), causing the de Bruijn graph to misalign with the original reads. This misalignment can produce paths within the graph that are unsupported by the underlying reads (Myers, 2005). For instance, if a k-mer appears identically in the middle of two reads but these reads do not have any other overlapping sections, the resulting de Bruijn graph would feature a branching node

rather than two separate paths (Schatz et al., 2010). This methodology also consolidates multiple instances of identical repeats into one path within the assembly graph (Kolmogorov et al., 2019). Consequently, short-read assemblers may struggle with intricate genomic regions rich in extensive, large repetitive areas (Liao et al., 2019), especially when the length of the repetitive sequences surpasses the read length (Schatz et al., 2010). Such limitations can produce incomplete or fragmented assemblies, as evidenced by the results obtained from short-read and hybrid assemblers for the *F. communis* genome.

These limitations associated with short-read assembly were alleviated using long reads generated by the MinION platform (Sohn and Nam, 2016). Excluding NextDenovo, both Flye and Wtdbg2 surpassed short-read assemblers and the hybrid assembler WENGAN in performance. These long-read assemblers harness the power of long-read sequences to adeptly resolve intricate and repetitive genome regions (Murigneux et al., 2020). While NextDenovo follows a similar approach, it integrates an error correction step, filtering out numerous reads before addressing large repeats. Among the best assemblers, Flye demonstrates superior contiguity and completeness compared to Wtdbg2. This distinction can be ascribed to Flye assembler's unique methodology. Flye initiates the assembly by crafting disjointigs—concatenations of several independent genomic segments. Subsequently, it builds a precise repeat graph, resolving repeats to generate the final contigs (Kolmogorov et al., 2019). In contrast, Wtdbg2 utilizes the fuzzy-Bruijn graph (FBG) algorithm, which is adapted from short-read assembly algorithms to accommodate the base calling inaccuracies in noisy long reads (McCartney et al., 2021). However, while it excelled in contiguity, the Flye assembler lagged in terms of genome completeness and overall genome size, for example, in comparison with the DBG2OLC-assembled genome. Yet, this assembly was less contiguous, yielding a substantial number of genome fragments. This outcome might stem from the algorithm's lower proficiency in anchoring long-reads to the backbone contigs generated by the short-read assembler which DBG2OLC utilizes. As in the case of

Wtdbg2, the Flye assembly surpassed DBG2OLC in terms of genome quality assessment metrics (Table 7).

The genome assembly of *F. communis* is, as of now, the sole documented genome within the genus *Ferula*. It shows a reasonably well N50 and achieves a BUSCO score of 93.2%, underscoring its satisfactory contiguity and completeness. Although a BUSCO score surpassing 90% is generally regarded as an indicator of a high-quality genome assembly, it is important to understand that completeness is not the exclusive criterion that ascertains whether a genome is categorized as “draft”, “reference”, or “complete/high-quality” (Wang and Wang, 2023). Evaluative criteria encompassing contiguity, completeness, and correctness also play a decisive role (Wang and Wang, 2023). Taken together, these metrics advocate for the classification of the *F. communis* assembly as a draft genome (refer to Table 6). This draft genome of *F. communis* aligns well with other published draft genomes as exemplified by *Eucalyptus pauciflora* (Wang et al., 2020). Of particular note are the main metrics procured for this species, including the number of contigs, N50, and other parameters such as the LAI index (9.32), BUSCO score (94.58%), structural variants (4,017), short-read mapping rate (94.92%), short-read error rate (0.0060), and both the long-read mapping rate and its error rate (91.49% and 0.1655, respectively). All these closely align with the draft genome metrics of *F. communis*. Nonetheless, to elevate the assembly quality, particularly in regions enriched in repetitive elements, incorporating advanced sequencing technologies such as the PacBio High-Fidelity (HiFi) long-reads—characterized by read lengths averaging 10-25 kb and an astonishing 99.9% single molecule read accuracy (Hon et al., 2020)—is the logical progression towards achieving a high-quality assembly at the chromosomal level.

4.2. Genome evolution in *Ferula communis*

4.2.1. *Ferula communis* and *Arabidopsis thaliana* genomes as examples of genome plasticity in angiosperms

Comparative gene annotation metrics between *F. communis* and *Arabidopsis thaliana* reveal notable evolutionary differences between the two plants, emphasizing the plasticity of angiosperm genomes. The genome of *F. communis* is significantly larger, mainly due to the expansion of gene families and transposable elements (Table 10). In contrast, the *Arabidopsis thaliana* genome has predominantly evolved through reduction, driven by small deletions primarily in non-coding DNA and transposons. Furthermore, Hu et al. (2011) demonstrated that the number of protein-coding families in *Arabidopsis thaliana* is also fewer than in its sister species, *Arabidopsis lyrata*. This disparity becomes even more evident when compared to *Ferula*, where the number of gene models is approximately double that of *Arabidopsis thaliana*. However, *Arabidopsis thaliana* possesses a greater average number of exons per gene and a longer average exon length. This aligns with a non-adaptive perspective where intron expansion is more probable in species with smaller effective population sizes (Lynch, 2002). While *F. communis* is an outcrossing species, *Arabidopsis thaliana* is self-pollinating, suggesting a potential decrease in the efficiency of natural selection against intron expansion in the latter. Notably, *Arabidopsis thaliana*, with its higher exon count, also exhibits an elevated rate of alternative splicing compared to *F. communis*. This is supported by its greater average number of transcripts per gene and a higher overall transcript relative to gene count (Table 8). The concurrent rise of splice variants with increasing exon numbers is believed to be a general trend (Kopelman et al., 2005). A positive correlation between intron length, exon count, and alternative splicing rate has also been documented in other plants, such as the soybean (Shen et al., 2014). Despite differences in exon-intron configurations, the average gene length is comparable in both species (Table 8). This supports the notion that there is not a significant correlation between exon length and CDS length across diverse organisms

(Koralewski and Krutovsky, 2011). A potential reason for this consistency is that the 3-dimensional protein structure and binding sites predominantly determine protein functionality. Therefore, variations in the lengths of coding sequences and transcripts might be limited (Koralewski and Krutovsky, 2011).

4.2.2. Transposable Elements (TEs) in *Ferula communis* as a main factor determining the genome size

The diversity in genome sizes observed among angiosperms is driven not only by whole-genome duplications (WGD) but also by dynamic processes involving the amplification of transposable elements (TEs) and TE-mediated recombination (Anderson et al., 2019). These processes contribute to the expansion and restructuring of angiosperm genomes. The amplification of TEs, coupled with recombination events, plays a pivotal role in the genomic plasticity and evolutionary potential of angiosperms. This leads to the generation of new genetic variations, thereby facilitating adaptation to new ecological niches (Kazazian, 2004; Hawkins et al., 2006).

TEs are instrumental in shaping the complex structure of plant genomes (Bennetzen, 2000). The disparity in genome sizes among angiosperms is largely due to the abundance of repetitive, non-coding DNA, including TEs, scattered throughout the genome (Finnegan, 1989; Gregory et al., 2007; Oliver et al., 2013). For instance, TEs constitute 86.65% of the *F. communis* genome, a proportion analogous to other angiosperms with large genomes (> 2 Gb) such as *Zea mays* (85%) (Schnable et al., 2009), *Panax ginseng* (79.52%) (Kim et al., 2018), and *Lactuca sativa* (74.24%) (Reyes-Chin-Wo et al., 2017). Within the Apiaceae family, TEs make up 46% of the *Daucus carota* genome (Iorizzo et al., 2016), 70.59% of *Corindrum sativum* (Song et al., 2020), and an astounding 92.91% of the *Apium graveolens* genome (Song et al., 2021). The percentage of TEs in the genomes of members of the Apiaceae family, as well as in other angiosperms, generally correlates with their genome size.

Gypsy and *Copia* represent the two most prevalent class I retrotransposons (LTR) elements in *F. communis*, mirroring trends seen in other species of the Apiaceae family (Iorizzo et al., 2016; Song et

al., 2020, 2021). This dominance of LTR TEs also manifests in *Lactuca sativa* (Reyes-Chin-Wo et al., 2017), *Panax ginseng* (Kim et al., 2018), and *Medicago ruthenica (Yin et al., 2021)*. (Yin et al., 2021). Several studies have highlighted recent expansion events of these families in various plants under different conditions (Ungerer et al., 2009; Belyayev et al., 2010; Belyayev, 2014). For instance, *Ty3/Gypsy*-like LTR retrotransposons underwent notable independent proliferation in three ancient hybrid sunflower species following their emergence, leading to a significant increase in nuclear DNA content (Baack et al., 2005; Ungerer et al., 2009). Similarly, self-pollinated *ddm1* mutant lines of *Arabidopsis thaliana* exhibited an increased copy number of both *Gypsy* and *Copia* retrotransposons, as well as the *Mutator* family DNA transposon (Tsukahara et al., 2009). In contrast, Zhang et al. (2020) conducted a study on four Brassicaceae species, investigating their transposition rates and the prevalence of *Gypsy* and *Copia*. Their findings suggest that the DNA loss counteracting TE insertions observed in *Arabidopsis thaliana* and *Eutrema salsugineum* mirrors a broader evolutionary pattern within the Brassicaceae family. The lack of *Gypsy* and *Copia* accumulation in *Schrenkiella parvula* (Brassicaceae) over the past two million years likely points to a significant drop in TE activity, which in turn has likely contributed to its reduced genome size (Zhang et al., 2020).

Furthermore, unclassified LTR transposable elements, which do not fit neatly into established families like *Gypsy* or *Copia*, were detected in the *F. communis* genome. These elements, present in other species such as the *Daucus carota* (Kwolek et al., 2022), *Oryza sativa* (Sasaki, 2005), and *Arabidopsis thaliana* genome (Lisch, 2013), add to the total TE content of the *F. communis* genome. Hence, the proliferation and accumulation of *Copia*, *Gypsy*, and unclassified LTR TEs have played a pivotal role in shaping the extensive genome size of *F. communis*.

Various processes can activate bursts and insertions of TEs. These include domestication (Naito et al., 2006), polyploidy (Kraitshtein et al., 2010; Kenan-Eichler et al., 2011), interspecific and intergenic hybridization (Ungerer et al., 2009; Kenan-Eichler et al., 2011), and niche shifts (Belyayev

et al., 2010). Such bursts of TEs can happen abruptly, leading to swift genomic alterations (Belyayev, 2014). Given *F. communis*'s adaptation to diverse habitats, including extreme environments, and considering its propensity for hybridization, the sudden increase in TEs in its genome may significantly enhance genetic diversity and genome flexibility. Interestingly, genome size variations within *Ferula* species do not align with changes in chromosome number. Increased TE activity is often linked to hybrid speciation in both animals and plants (Choudhury and Parisod, 2017). When activated, TEs can amplify genetic variance when it is most beneficial.

In addition to their contribution to genome size, the amplification and insertion of TEs near or within genes can significantly impact the functions of host plant genomes (Vukich et al., 2009). For instance, the *Tos17*, a *Gypsy*-like element, can insert itself near genes, leading to disruptions and mutations in rice cultivars (Miyao et al., 2003). This can result in changed gene expressions and phenotypic variations (Miyao et al., 2003). Similarly, the *Copia* retrotransposon ONSEN typically inserts close to genes, inducing heat-activated responses in species belonging to the Brassicaceae family (Ito et al., 2011; Pietzenuk et al., 2016). Instances like the insertion of the *Tcs1 Copia* element upstream of the *Ruby* MYB transcription factor, which regulates anthocyanin biosynthesis in some blood orange varieties, affect gene expression. Another case of *Ty1-Copia* elements involves the retrotransposon *Hopscotch*'s insertion in maize's regulatory region, influencing apical growth by reducing branching (Studer et al., 2011). Other TE families in maize have been observed to enhance gene expression in response to environmental stresses (Makarevitch et al., 2015).

Several studies suggest that closely related TEs may induce genome rearrangements during species diversification (Gray et al., 1996; Geurts et al., 2006; Zhang et al., 2006; Huang and Dooner, 2008). Such rearrangements can involve chromosomal breakage, deletion, duplication, inversion, and translocation. For instance, the *Ac/Ds* TEs in maize and *Arabidopsis thaliana* have been linked to chromosomal rearrangements (Xuan et al., 2011). The *Ds* TEs, specifically, have been identified as a

cause for chromosomal breakage in maize (McClintock, 1956). Although the *Ac/Dc* TEs were not detected in the *F. communis* genome, related TEs may induce similar genomic changes, possibly contributing to the B chromosome's formation in the *F. communis* genome (Sànchez-Cuxart and Mercè, 1998).

4.2.3. Extensive gene family expansions in the *Ferula communis* genome

The absence of a chromosome-level assembly for the *F. communis* genome hinders our ability to infer WGD events in this species. However, such events have been documented in other species within the Apiaceae family, including *Daucus carota* (Dc-α and Dc-β) (Iorizzo et al., 2016), *Apium graveolens* (Apiaceae-α and Apiaceae-ω) (Song et al., 2021), and *Corindrum sativum* (A-beta and A-alpha) (Song et al., 2020). These observations underscore the profound influence of WGD events on the evolution of this plant group. These events have led to increased gene diversity and changes in chromosome numbers. By providing new genetic material and opportunities for gene fractionation, WGD has played an integral role in shaping the genomes of these plants throughout their evolutionary history (Jiao et al., 2011).

Duplicated genes resulting from WGD can undergo functional divergence or acquire new functions, enabling plants to explore new ecological niches and adapt to diverse habitats (Moriyama and Koshiba-Takeuchi, 2018). In plant genomes, approximately 65% of annotated genes are found to have duplicate counterparts. A significant portion of these duplicated genes originates from WGD events (Panchy et al., 2016). The presence of these duplications results in two copies of a gene, providing an opportunity for one or both copies to evolve with fewer constraints. Occasionally, they acquire novel gene functions that contribute to adaptation (Panchy et al., 2016).

Comparative genome analysis among species within the Apiaceae family reveals that *F. communis*, possessing 68,318 annotated genes, has a higher gene count than *Apium graveolens* (31,326 genes), *Daucus carota* (32,113 genes), and *Corindrum sativum* (40,747 genes). A similar trend has been

observed within the Brassicaceae family. For instance, *Arabidopsis thaliana*, a model plant species, is known to have roughly 27,000 annotated genes (Swarbreck et al., 2007; Cheng et al., 2017a), whereas *Brassica oleracea* (cabbage) surpasses this number with over 59,000 genes (Parkin et al., 2014; Golicz et al., 2016). This disparity in gene counts can be attributed to several factors, one of the primary ones being the prevalence of gene family duplications. Such duplications can lead to an expansion of gene families and an overall increase in genome size (Wang et al., 2011).

The expansion of gene families has been observed among the compared species, including *F. communis*. For instance, the gene family encoding highly conserved defense-related proteins (GO:0006952) has undergone significant expansion. Within this category, the plant pathogenesis-related protein (PR-10) Betv1, which ranges in size from 15-17 kDa and is prevalent among dicotyledonous plants (Wen et al., 1997), and the disease resistance genes (R genes) that confer resistance to diverse pathogens, are noteworthy members of this gene family (Staskawicz et al., 1995). These gene families play a pivotal role in helping the plant defend against diseases triggered by various pathogens through a spectrum of mechanisms. Moreover, the gene family associated with salt stress response (GO:0009551) has also seen a substantial expansion. The expansion of these genes among the compared species aligns with expectations. To withstand pathogens and challenging environmental conditions, plants must modulate their metabolism by adjusting the expression of genes associated with disease resistance and stress tolerance (Dixit and Dhankher, 2011). Coping with multiple stress conditions can be achieved either by overexpressing transcription factors that regulate genes from various pathways or by overexpressing genes involved in abiotic stress signal perception and transduction (Kanneganti and Gupta, 2008). For example, studies have shown that members of the rice stress-associated protein (SAP) gene family are induced by a myriad of abiotic stresses, such as salt, drought, cold, desiccation, submergence, wounding, and heavy metals (Kanneganti and Gupta, 2008).

Similarly, the *Arabidopsis thaliana* stress-associated protein-10 (AtSAP10) gene exhibits differential regulation in response to heavy metals, heat, cold, and salt (Dixit and Dhankher, 2011).

Beyond the aforementioned gene families, other sets have also undergone significant expansion within the Apiaceae species. Notably, the gene family responsible for the ubiquitin-dependent protein catabolic process (GO:0006511) has seen considerable growth. The ubiquitin-mediated proteolysis carried out by the proteasome is a vital regulatory mechanism influencing numerous biological processes, such as seed size (Li and Li, 2014; Linden and Callis, 2020), flowering induction, and pathogen response (Linden and Callis, 2020). Additionally, the SCF (Skp1, cullin/CDC53, and F-box protein) complexes, which constitute the largest and most studied family of E3 ubiquitin-protein ligases (Zheng et al., 2002), are known for their roles in cell regulation, signal transduction, transcription, and other biological functions (Zheng et al., 2002). Within plants, SCF complexes regulate aspects such as auxin responses, jasmonate signaling (Zhao et al., 2003), flower development (Zhao et al., 2001), circadian clock regulation (Mizoguchi and Coupland, 2000), and gibberellin signaling (Sasaki et al., 2003). These gene families encompass a wide array of functions that modulate various developmental and physiological processes.

When the Apiaceae species are compared, a substantial gene expansion is evident in *F. communis* (supplementary Figure F5 and supplementary Table 1: S23). Notably, among the gene families exhibiting significant expansion in *F. communis*, I find those associated with systemic acquired resistance and salicylic acid-mediated signaling pathways (GO:0009862 and GO:0031347), as well as the vernalization response (GO:0010048). For a comprehensive list of gene families that displayed expansion, please refer to Supplementary Table 1: S23. Plant species, including *F. communis*, have evolved defense mechanisms against pathogens, with systemic acquired resistance (SAR) being a crucial aspect initiated through localized pathogen infection (Yuan et al., 2007). This process is characterized by the expression of pathogenesis-related (PR) genes in various tissues, widely

acknowledged as an effective form of broad-spectrum disease resistance (Sticher et al., 1997), and salicylic acid is used as a essential signal molecule for SAR, as identified in *Arabidopsis thaliana* and tobacco plant species (Yuan et al., 2007; Hermann et al., 2013). This suggests that the expansion of these gene families in *F. communis* could be indicative of disease resistance behavior through systemic acquired resistance (SAR) mediated by salicylic acid pathways.

In addition, the gene family associated with the vernalization response is among the significantly expanded genes in the *F. communis* genome. Vernalization, which triggers flowering in response to cold temperatures, serves as a prime example of epigenetic regulation in plants (Jean Finnegan et al., 2011). This phenomenon is observed in various species, including wheat varieties (Singh et al., 2013), *Arabidopsis thaliana* (Kim and Sung, 2013), and winter-annual plants. Exposure to cold induces genetic and epigenetic alterations that activate specific flowering-related genes (Jean Finnegan et al., 2011). By doing so, the vernalization response enables plants to time their flowering strategically during favorable conditions, enhancing their reproductive success and adaptability to evolving habitats (Kim and Sung, 2013). Consequently, the expansion of these gene families in *F. communis* plays a pivotal role in optimizing reproductive outcomes and adapting to seasonal variations, essential for the species to flourish across diverse ecological niches with varying seasonal patterns.

4.2.4. Evolution of orthologous genes in *Ferula* and allies

Closely related species typically share a significant number of genes. Intriguingly, *F. communis* shares more genes with coriandrum than with other closely related species, such as the *Daucus carota* and *Apium graveolens*. Notably, genes responsible for secondary metabolite production, especially those involved in terpenoid biosynthesis pathways, are of particular interest. Despite the diverse roles of terpenoids—serving as components of electron transfer systems, agents for protein modification, hormone activators, and antioxidants—which likely evolved early in green plant history, lineage-specific terpenoid biosynthesis pathways are associated with the formation of plant volatile substances

(Pichersky and Raguso, 2018). These substances play numerous roles in plant interactions with their environments, such as attracting pollinators and defending against pests (Falara et al., 2011). Several factors contribute to terpene diversity. One such factor is the enzymes known as terpene synthases (TPSs), which typically exist as a gene family, boasting 30–100 genes per genome. This extensive gene family offers a platform for the evolution of new terpenes through mutation and selection (Pichersky and Raguso, 2018). The diversity of these genes and their products is believed to stem from the evolutionary arms race between plants and specialized herbivores. Moreover, plants can repurpose these defensive compounds for other ecological functions, such as antimicrobial activity or pollinator attraction (Pichersky and Raguso, 2018).

A possible explanation for why *F. communis* and coriandrum share more terpenoid genes than they do with *Apium graveolens*, despite their distant phylogenetic relationship, might be the aquatic ancestry of *Apium graveolens*. The *Apium* genus, to which *Apium graveolens* belongs, predominantly comprises aquatic species. Because aquatic plants struggle to store terpenoids in secretory ducts and release them through epidermal appendages, they might decrease the number of genes essential in terrestrial habitats and simultaneously acquire new ones essential in aquatic environments (Lange, 2015). While the *Daucus carota*, a close relative of *Ferula*, also shares fewer terpenoid genes, a plausible reason could be the genome reduction in *Daucus carota*, which exhibits the most significant gene contractions (Figure 6).

4.2.5. Characterization of unique genes in *Ferula communis*

In every characterized genome exists genes that are unique—meaning they have no apparent homologs within the same genome or among phylogenetically close relatives (Llorente et al., 2000; Rubin et al., 2000). Specifically, approximately 3,325 gene families, comprising 15,737 genes specific to *F. communis*, have been identified. Among these gene families, gene ontology analysis identified certain genes with a known gene family name and function (Table 10). For instance, 45 genes associated with

pollen recognition (GO:0048544) have been pinpointed. These genes play a vital role in fertilization, aiding the recognition and interaction between male and female gametes. Such genes are especially pronounced in plants that possess mechanisms to deter self-fertilization. In the Brassicaceae family, for example, *Brassica oleracea* utilizes a self/non-self recognition system that prevents self-fertilization (Hinata et al., 1995). This genetic regulation is governed by multiple alleles located at a single locus, denoted as (S). In *Brassica oleracea*, there are 50 distinct S alleles, highlighting considerable genetic diversity (Hinata et al., 1995). The existence of these genes supports the self-incompatibility mechanism that specifically prohibits self-fertilization, fostering outcrossing and genetic diversity within *F. communis*.

Moreover, unique gene families in *F. communis* include signal transduction genes (GO:0007165) and genes that defend against fungus (GO:0050832). Two protein domains commonly observed in plants are the Toll/interleukin-1 receptor homology (TIR) and Barwin domains. TIR is a conserved domain comprising approximately 200 amino acids and is found in plants such as *Arabidopsis* and *Nicotina* species (Wan et al., 2019). In plants, TIR immune receptors perform a specialized function: they decompose a molecule named nicotinamide adenine dinucleotide (NAD+) during the plant's defense response to fungi. This results in the death of infected cells and amplifies the plant's disease resistance (Wan et al., 2019). Additionally, the Barwin domain, a protein made up of 125 amino acids, was initially isolated from barley seed aqueous extracts and is similar to wound-induced genes, 122 amino acids in length, in potatoes and pathogenesis-related proteins in tobacco. These are believed to participate in a shared defense mechanism in plants (Svensson et al., 1992). This suggests that the expansion of various gene families with diverse roles in *F. communis* allows the plant to counteract fungal diseases.

4.3. The structure and evolution of the *Ferula communis* plastid genome

4.3.1. Characteristics of the pDNA genome in *Ferula communis*

In angiosperms, the pDNA typically exhibits a highly conserved structure, gene content, and gene order (Dong et al., 2014). The pDNA of *F. communis* possesses the circular quadripartite structure common in angiosperms, composed of two inverted repeat regions, one SSC region, and one LSC region. This pattern aligns with other assembled pDNA genomes of *Ferula* species (Yang et al., 2022; Qin et al., 2023). However, two distinct haplotypes are present in the pDNA of *F. communis*. These structures have comparable sizes but vary in the gene orientation within the SSC region. While heteroplasmy, denoting the coexistence of two distinct structural forms in one organism, is documented in certain angiosperms (Stein et al., 1986; Martin et al., 2013), no biological mechanism elucidating its evolution has been discerned. Yet, it is postulated that the expansive IRs might mediate regular intermolecular recombination, sustaining roughly equal quantities of the two differing haplotypes in the SSC region (Stein et al., 1986).

The plastid genome size, gene order, and composition of *F. communis* analyzed in this study closely resemble those of other sequenced *Ferula* plastid genomes, with a few minor exceptions (Yang et al., 2022; Qin et al., 2023). Notably, the gene *ycf15*, which has an uncertain function and is reported in other *Ferula* species (Yang et al., 2022; Qin et al., 2023), is absent in *F. communis*. The pDNA genome of *F. communis* displays minor variations in the Irb/SSC regions compared to that of *Daucus carota* and other *Ferula* species. In some species, this gene extends into the Irb/SSC border, while in others, it is offset. Additionally, the LSC and IR regions intersect with the *rps19* gene of ribosomal protein in the *F. communis* pDNA genome. Variations, particularly in the intergenic spacer (IGS) regions and inverted repeat (IR) borders, are documented in the pDNA of other angiosperms (Li et al., 2020). While the fundamental structure of pDNA genomes largely remains stable, the boundaries between LSC and IR domains can undergo dynamic modifications, leading to changes in gene

adjacency (Peery, 2015). Such sequence alterations at the IR borders, contributing to variations and rearrangements in pDNA genome size, represent a general evolutionary trend (Khakhlova and Bock, 2006). Minor shifts, typically less than 100 bp, in the positions of the LSC/IR border are frequently observed throughout angiosperm evolution (Goulding et al., 1996). In contrast, significant constrictions, without complete IR loss, are considered rare occurrences (Guisinger et al., 2011).

When comparing the pDNA genomes of *Anthriscus* and *Daucus carota*, both members of the tribe Scandiceae, several intriguing differences emerge. The *Daucus carota* pDNA genome contains an additional open reading frame (ORF80) located in the IR region (Peery, 2015). Furthermore, two fragments, each approximately 1,439 base pairs and originating from the mitochondria, are present in the *Daucus carota*'s IR region of its pDNA (Ruhlman et al., 2006). These fragment sequences do not match any known plastid sequences (Goremykin et al., 2009). This finding was supported by Iorizzo et al. (2012), who reported a fragment of 1,452 bp in the *Daucus carota*. Several studies have pointed out the presence of mtDNA in the plastomes of various Apiaceae species, indicating that mtDNA introgression into the plastid genome is a common event within this family (Downie et al., 2014; Downie and Jansen, 2015). While this phenomenon is not apparent in the pDNA genome of *F. communis*, thorough comparative studies should be conducted to verify potential mtDNA introgressions in the pDNA of *Ferula* species as more genomes become sequenced.

4.3.2. Sequence divergence of the pDNA genome of *Ferula communis*

The pDNA genome of *Ferula* species displays significant variation, with a markedly higher sequence diversity in non-coding regions, especially the IGS, than in coding regions. These IGS regions predominantly reside in the LSC region of the pDNA genome, which is abundant in SSRs. The exact molecular mechanisms underlying the evolution of SSRs are not fully understood. Replication slippage, unequal crossing-over, and nucleotide substitution have been proposed as possible drivers of SSR variation. Yet, these explanations do not fully account for the evolution of SSRs (Levinson and

Gutman, 1987; Schlötterer and Tautz, 1992). Further research into the evolution of SSRs suggests that random point mutations might drive their development and that rare DNA polymerase slippage events can increase repeat numbers (Messier et al., 1996; Rose and Falush, 1998; Schlötterer, 2000). Due to their extensive variability within and among species and populations, SSRs are frequently employed in phylogenetic and population genetics studies (Zeb et al., 2020).

Among the various SSRs, mono-nucleotide repeats were most prevalent in *F. communis*, comprising 65% of all SSRs. The distributions of mono-, di-, and tetranucleotide repeats in *F. communis* mirror those in plants like *Daucus carota*, *Zea mays*, and other *Ferula* species. In contrast, tri- and hexa-nucleotide repeats appear less frequently. Selection pressure likely limits SSRs presence in coding regions because shifts in the reading frame can impede the proper function of proteins (Oliveira et al., 2006). However, if present, coding region SSRs often manifest as tri- or hexa-nucleotide repeats. Additionally, some SSRs, four to six nucleotides in length, span introns and IGS regions of pDNA genomes in both single-copy and inverted repeat regions across various land plant lineages (Borsch and Quandt, 2009). This uniformity suggests a shared mutational mechanism across all plastid genomes, a pattern not observed in other genomic compartments like the nuclear genome (Borsch and Quandt, 2009).

In our analysis of pDNA sequence variation, we identified regions with high variability in three coding genes (*clpP*, *rpl16*, and *ycf1*) and eleven non-coding regions in various *Ferula* species (Figure 8). These IGS regions have previously been employed in phylogenetic studies aiming to determine relationships within different clades of the Apiaceae family (Downie and Jansen, 2015; Mustafina et al., 2019; Park et al., 2019). Notably, the *trnH-psbA*, *trnS-trnG*, and *atpB-rbcL* segments have been used in phylogenetic studies on *Ferula*. However, these IGS regions have proven inadequate in clarifying the phylogenetic relationships among *Ferula*, largely due to idiosyncratic evolution causing issues with site homology (Piwczyński et al., 2018).

4.3.3. Positive selection on pDNA genes in *Ferula communis*

The dN/dS ratio, which contrasts non-synonymous to synonymous nucleotide substitutions, offers insights into the type and intensity of selective pressure on protein-coding genes. Utilizing the M0 codon model—a simple model assuming a uniform ω ratio across all sites and lineages—we found only two genes, *ccsA* and *ycf2*, undergoing positive selection. This suggests these genes might be evolving adaptively, likely due to specific amino acid alterations conferring selective benefits in particular habitats (Álvarez-Carretero et al., 2023). In contrast, genes *atpE*, *matK*, and *rps8* seemed to experience near neutral evolution, suggesting relaxed constraints on the functions of these genes. For the majority of genes, the dN/dS ratio was under 0.5, suggesting a predominant purifying selection stabilizing these genes (Abdullah et al., 2019).

The site-specific models reaffirmed the analysis of non-synonymous to synonymous substitutions, identifying the *ccsA* gene as having sites under positive selection. My findings contrast with those presented by Qin et al. (2023) for 22 plastid genomes of *Ferula*. They identified 12 genes under positive selection using the optimized branch-site model, but notably, the *ccsA* gene was not among them. This disparity may be due to the differences in our sample sizes. A larger dataset might lend stronger support to genes identified as positively selected under site-specific models. Nevertheless, the ccsA gene, essential for attaching heme to cytochrome c—a pivotal component of the electron transport chain crucial for efficient photosynthesis (Xie and Merchant, 1996; Stoebe et al., 1998), has shown signs of positive selection in various plant taxa, including the Brassicaceae genus *Cardamine* (Hu et al., 2015), Orchid species (Dong et al., 2018), and Japanese apricot (*Prunus mume* Seib. Et Zucc.) (Huang et al., 2022b).

4.4. The structure and evolution of the *Ferula communis* mitochondrial genome

4.4.1. Multi-partite structure of *Ferula communis* mitochondrial genome

The mitochondrial genome of *F. communis*, assembled from high-quality Illumina short-reads, aligns in gene composition with other mitochondrial genomes in the Apiaceae family, encompassing protein-coding, tRNA, and rRNA genes. However, instead of the traditionally expected master circular mtDNA, the *F. communis* assembly revealed 16 non-circular scaffolds. Although the 'master circle' perspective is still prevalent among biologists, contemporary research emphasizes the vast diversity and complexity of plant mitochondrial DNA. This includes linear molecules, head-to-tail concatemers, and branched-linear formations (Backert et al., 1996; Bendich, 1996; Oldenburg and Bendich, 1998, 2001; Backert and Börner, 2000). In specific instances, like in *Chenopodium album*, unique configurations such as sigma-like molecules have been identified (Backert et al., 1996, 1997). Research indicates the remarkable adaptability of plant organelle genomes, showing they can undergo swift structural shifts, including organizational and compositional changes (Cheng et al., 2017b). The variability in plant mtDNA structures highlights their inherent flexibility (Cheng et al., 2017b), with genomes capable of swift structural modifications.

The mtDNA genomes of angiosperms typically contain repeated sequences that are recombinationally active, which often lead to rearrangements (Lonsdale et al., 1984; Palmer and Shields, 1984; Sloan, 2013; Cole et al., 2018). Research on the *Arabidopsis* mtDNA genome has offered insights into the nature and genetic drivers of these rearrangements (Shedge et al., 2007; Arrieta-Montiel et al., 2009). For example, Davila et al. (2011) observed sequence deletions and asymmetrical mitochondrial recombination in both wild-type *Arabidopsis* and *msh1* mutants. The *msh1* mutant, defined by a mutation in the MSH1 gene supervising mitochondrial genome structure (Abdelnoor et al., 2003), displays a process called substoichiometric shifting. This results in a significant increase in the copy number of the modified mitochondrial genome sections, often changing

the plant's phenotype (Mackenzie, 2023). Recombination in mtDNA largely arises from repeat sequence interactions. Many angiosperm mitochondrial genomes are rich in repeat sequences, with intermediate-length repeats (50-500 base pairs) especially prone to recombination (Maréchal and Brisson, 2010; Woloszynska, 2010). The MSH1 gene in *Arabidopsis* ensures that the mitochondrial DNA does not undergo excessive exchanges at intermediate repeats; these exchanges are infrequent (Davila et al., 2011). Conversely, the *msh1* mutant had forty-seven recombination repeat pairs, ranging from 50 to 556 bp, with sequence identities as low as 85% (Davila et al., 2011).

The mtDNA genome of *F. communis* harbors intermediate repeats within specific size ranges, potentially prompting recombination events. This could explain the non-circular structure identified. Large repeats tend to undergo regular, reversible recombination events, fostering interconversions between DNA molecules. They are also seen as probable locations for recombination-driven DNA replication (Zaegel et al., 2007). Conversely, the *F. communis* mtDNA genome is rich in monomer and dimer forms of short repeated sequences. These sequences experience occasional, irreversible recombinations, producing novel, stable DNA configurations that remain distinct from the primary genome (Woloszynska, 2010). Such repeat sequences in plant mitochondrial genomes are essential for both immediate and long-term structural adaptations (Kozik et al., 2019).

Additionally, several plant mitochondria across species have been identified to house small, linear, plasmid-like molecules autonomously. Occasionally, these molecules integrate into the mitochondrial genome, as seen in *Zea mays* (Allen et al., 2007) and *Daucus carota* (Iorizzo et al., 2012b). This integration can hinder the genomes from forming circular structures. To determine if the structure of *F. communis*'s mtDNA genome stems from repeat recombination or the incorporation of these plasmid-like entities, further research is necessary.

4.4.2. Significance of RNA editing in the mitochondrial genome of *Ferula communis*

RNA editing in plant mitochondrial genomes involves the post-transcriptional modification of RNA molecules. During this process, specific cytosine (C) residues are converted to uracil (U), and in some cases, uracil (U) is converted to cytosine (C) at particular sites (Takenaka et al., 2013). This editing serves critical functions, including error correction in the DNA template, expanding the genetic code, and ensuring accurate protein synthesis (Bentolila et al., 2008). Most RNA-editing events occur at the first or second positions of codons, leading to changes in the codon sequence and subsequently transcribed into precursor mRNA (pre-mRNA) (Takenaka et al., 2013).

In the mtDNA genome of *F. communis*, a comparable number of RNA editing sites have been identified, similar to those in the mtDNA genomes of *Arabidopsis thaliana* (Giegé and Brennicke, 1999), *Brassica napus* (Handa, 2003), and *Beta vulgaris* (Mower and Palmer, 2006). The majority of these editing sites are primarily located at the second (70.69%) and first (29.31%) positions of the triplet codon. A notable consequence of RNA editing is the alteration of specified amino acids, potentially changing the biochemical properties of mitochondrial proteins (Takenaka et al., 2013). In the *F. communis* mtDNA genome, RNA editing often increases the representation of hydrophobic amino acids. For example, the most frequent amino acid transitions—91 from serine (S) to leucine (L), and 65 serine (S) to phenylalanine (F)—result in hydrophobic amino acids. Overall, in the mtDNA genome of *F. communis*, 61.14% of the edits convert amino acids from hydrophilic to hydrophobic, while 31.09% involve changes between hydrophobic amino acids. Similar patterns have been observed in the mitochondrial genome of *Arabidopsis thaliana* (Giegé and Brennicke, 1999), suggesting that RNA editing enhances the hydrophobicity of mitochondrial proteins. Given that many genes in the mitochondrial genome code for membrane-bound proteins rich in hydrophobic amino acids (Jobson and Qiu, 2008), RNA editing appears to promote hydrophobicity. A pronounced selection for

maintaining hydrophobicity is evident in genomes primarily comprising genes that code for membrane-bound proteins (Jobson and Qiu, 2008).

Several studies have shown that disrupted RNA editing can detrimentally affect organelle biogenesis (Hao et al., 2021). For instance, the absence of the RNA editing protein PPR2263 in the cob transcript of maize resulted in the loss of a protein complex in the mitochondrial respiratory chain, causing inhibited growth in the mutant plant (Sosso et al., 2012). Also, the absence of the PPR-DWW subgroup protein, which facilitates C-to-U editing of several transcripts at *rpl16* in maize mitochondria, led to diminished embryo and endosperm development (Liu et al., 2013). This highlights the significance of RNA editing in regular plant growth and development. Although such effects have not been explicitly observed in the *F. communis* mtDNA, deeper investigations into this plant species' complex organelle genome are needed. While RNA editing might seem resource-intensive and expensive due to the allocation of nuclear genes and the use of genome space in plastids and mitochondria, the process might offer associated advantages (Takenaka et al., 2013). For instance, it enables genetic plasticity, allowing organisms to introduce changes in their gene expression without altering the DNA sequences. This feature is particularly important in environments where rapid adaptation is required in response to changing environmental cues (Takenaka et al., 2013). Additionally, RNA editing promotes protein diversification, resulting in the creation of multiple protein isoforms from a single gene, thereby expanding functional diversity, which is especially important when specific protein variants are necessary for different physiological conditions (Shikanai, 2015).

4.5. Conclusions

The results obtained from this research underscore the significance of *Ferula communis* as a model system in genomics and open up new avenues for exploring a plethora of genomic and evolutionary questions. Extensive gene duplications and the emergence of novel genetic elements within the genome of *F. communis* provide a solid foundation for investigating the mechanisms behind these evolutionary

processes. Especially interesting is the rapid evolution of genes responsible for the production of novel secondary metabolites, exemplified here by genes responsible for terpenoid biosynthesis. Furthermore, the role of transposons in shaping genome size evolution becomes an engaging area of exploration using *F. communis* as a model. This species offers valuable insights into how transposons have influenced genome expansion or contraction, and gene expression, shedding light on the factors that govern genomic plasticity.

The adaptation of *F. communis* to various habitats and the concurrent evolution of its plastid genes present another intriguing dimension of research. While the analysis of selection on plastid genes is frequently conducted by researchers, there often is not an adaptive or mechanistic explanation for the results obtained. The genus *Ferula*, with members inhabiting various, often extreme environments, can serve as a model system for seeking the evolutionary explanation for these patterns. Similarly, the organization and evolution of the mitochondrial genome, as well as RNA editing, offer another compelling area of research. The mitochondrial genome of *F. communis* has revealed distinctive features, such as a non-circular structure. Such structures are now being found in many other angiosperms, raising questions about the mechanisms keeping the mitochondrial genome coherent under strong recombination.

In conclusion, the insights gained from this research have broad implications for our understanding of the dynamics of genome evolution, adaptation to extreme environments, and the roles played by various genetic elements in shaping genome structure and function.

5. Bibliography

Abdelnoor, R. V., R. Yule, A. Elo, A. C. Christensen, G. Meyer-Gauen, and S. A. Mackenzie. 2003. Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. *Proceedings of the National Academy of Sciences* 100: 5968–5973.

Abdullah, I. Shahzadi, F. Mehmood, Z. Ali, M. S. Malik, S. Waseem, B. Mirza, et al. 2019. Comparative analyses of chloroplast genomes among three *Firmiana* species: Identification of mutational hotspots and phylogenetic relationship with other species of *Malvaceae*. *Plant Gene* 19: 100199.

Adams, K. L., Y.-L. Qiu, M. Stoutemyer, and J. D. Palmer. 2002. Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proceedings of the National Academy of Sciences* 99: 9905–9912.

Ajani, Y., A. Ajani, J. M. Cordes, M. F. Watson, and S. R. Downie. 2008. Phylogenetic Analysis of nrDNA ITS Sequences Reveals Relationships within Five Groups of Iranian Apiaceae Subfamily Apioideae. *Taxon* 57: 383–401.

Albert, B., B. Godelle, and P.-H. Gouyon. 1998. Evolution of the Plant Mitochondrial Genome: Dynamics of Duplication and Deletion of Sequences. *Journal of Molecular Evolution* 46: 155–158.

Allen, J. O., C. M. Fauron, P. Minx, L. Roark, S. Oddiraju, G. N. Lin, L. Meyer, et al. 2007. Comparisons Among Two Fertile and Three Male-Sterile Mitochondrial Genomes of Maize. *Genetics* 177: 1173–1192.

Alonge, M., L. Lebeigle, M. Kirsche, K. Jenike, S. Ou, S. Aganezov, X. Wang, et al. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* 23: 258.

Álvarez-Carretero, S., P. Kapli, and Z. Yang. 2023. Beginner's Guide on the Use of PAML to Detect Positive Selection K. Crandall [ed.],. *Molecular Biology and Evolution* 40.

Alverson, A. J., D. W. Rice, S. Dickinson, K. Barry, and J. D. Palmer. 2011. Origins and Recombination of the Bacterial-Sized Multichromosomal Mitochondrial Genome of Cucumber. *The Plant Cell* 23: 2499–2513.

Alverson, A. J., X. Wei, D. W. Rice, D. B. Stern, K. Barry, and J. D. Palmer. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution* 27: 1436–1448.

Amiryousefi, A., J. Hyvönen, and P. Poczai. 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes J. Hancock [ed.],. *Bioinformatics* 34: 3030–3031.

Anderson, S. N., M. C. Stitzer, A. B. Brohammer, P. Zhou, J. M. Noshay, C. H. O'Connor, C. D. Hirsch, et al. 2019. Transposable elements contribute to dynamic genome content in maize. *The Plant Journal* 100: 1052–1065.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Website http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arrieta-Montiel, M. P., V. Shedge, J. Davila, A. C. Christensen, and S. A. Mackenzie. 2009. Diversity of the *Arabidopsis* Mitochondrial Genome Occurs via Nuclear-Controlled Recombination Activity. *Genetics* 183: 1261–1268.

Asaf, S., A. L. Khan, A. R. Khan, M. Waqas, S.-M. Kang, M. A. Khan, S.-M. Lee, and I.-J. Lee. 2016. Complete Chloroplast Genome of *Nicotiana otophora* and its Comparison with Related Species. *Frontiers in Plant Science* 7.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25: 25–29.

Aury, J.-M., and B. Istace. 2021. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genomics and Bioinformatics* 3.

Baack, E. J., K. D. Whitney, and L. H. Rieseberg. 2005. Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. *New Phytologist* 167: 623–630.

Backert, S., and T. Börner. 2000. Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant *Chenopodium album* (L.). *Current Genetics* 37: 304–314.

Backert, S., P. Dörfel, R. Lurz, and T. Börner. 1996. Rolling-Circle Replication of Mitochondrial DNA in the Higher Plant *Chenopodium album* (L.). *Molecular and Cellular Biology* 16: 6285–6294.

Backert, S., R. Lurz, O. A. Oyarzabal, and T. Börner. 1997. High content, size and distribution of single-stranded DNA in the mitochondria of *Chenopodium album* (L.). *Plant Molecular Biology* 33: 1037–1050.

Baidouri, M. El, and O. Panaud. 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biology and Evolution* 5: 954–965.

Banasiak, Ł., M. Piwczyński, T. Uliński, S. R. Downie, M. F. Watson, B. Shakya, and K. Spalik. 2013. Dispersal patterns in space and time: A case study of apiaceae subfamily apioideae. *Journal of Biogeography* 40: 1324–1335.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

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

Baucom, R. S., J. C. Estill, C. Chaparro, N. Upshaw, A. Jogi, J.-M. Deragon, R. P. Westerman, et al. 2009a. Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome H. S. Malik [ed.],. *PLoS Genetics* 5: e1000732.

Baucom, R. S., J. C. Estill, J. Leebens-Mack, and J. L. Bennetzen. 2009b. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Research* 19: 243–254.

Beier, S., T. Thiel, T. Münch, U. Scholz, and M. Mascher. 2017. MISA-web: A web server for microsatellite prediction. *Bioinformatics* 33: 2583–2585.

Belyayev, A. 2014. Bursts of transposable elements as an evolutionary driving force. *Journal of Evolutionary Biology* 27: 2573–2584.

Belyayev, A., R. Kalendar, L. Brodsky, E. Nevo, A. H. Schulman, and O. Raskina. 2010. Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. *Mobile DNA* 1: 6.

Bendich, A. J. 1996. Structural Analysis of Mitochondrial DNA Molecules from Fungi and Plants Using Moving Pictures and Pulsed-field Gel Electrophoresis. *Journal of Molecular Biology* 255: 564–588.

Bennetzen, J. L. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology* 42: 251–269.

Bennetzen, J. L., and H. Wang. 2014. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology* 65: 505–530.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27: 573–580.

Bentolila, S., L. E. Elliott, and M. R. Hanson. 2008. Genetic Architecture of Mitochondrial Editing in *Arabidopsis thaliana*. *Genetics* 178: 1693–1708.

Benton, M. J., P. Wilf, and H. E. Sauquet. 2021. Tansley review The Angiosperm Terrestrial Revolution and the origins of modern biodiversity.

De Bie, T., N. Cristianini, J. P. Demuth, and M. W. Hahn. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22: 1269–1271.

Blum, M., H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* 49: D344–D354.

Bocchieri, E. 1988. Silene valsecchiae e *Ferula arrigonii*, due specie nuove della Sardegna. *Boll. Soc. Sarda Sci. Nat.* 26: 305 – 310.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Bone, T. S., S. R. Downie, J. M. Affolter, and K. Spalik. 2011. A phylogenetic and biogeographic study of the genus *Lilaeopsis* (Apiaceae Tribe Oenantheae). *Systematic Botany* 36: 789–805.

Borsch, T., and D. Quandt. 2009. Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Systematics and Evolution* 282: 169–199.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* 3.

Brůna, T., A. Lomsadze, and M. Borodovsky. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics* 2.

Bruneau, A., J. J. Doyle, and J. D. Palmer. 1990. A Chloroplast DNA Inversion as a Subtribal Character in the Phaseoleae (Leguminosae). *Systematic Botany* 15: 378.

Buchfink, B., C. Xie, and D. H. Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59–60.

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

Cai, Z., M. Guisinger, H.-G. Kim, E. Ruck, J. C. Blazier, V. McMurtry, J. V. Kuehl, et al. 2008. Extensive Reorganization of the Plastid Genome of *Trifolium subterraneum* (Fabaceae) Is Associated with Numerous Repeated Sequences and Novel DNA Insertions. *Journal of Molecular Evolution* 67: 696–704.

Calviño, C. I., F. E. Teruel, and S. R. Downie. 2016. The role of the Southern Hemisphere in the evolutionary history of Apiaceae, a mostly north temperate plant family. *Journal of Biogeography* 43: 398–409.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 1–9.

Campbell, M. S., C. Holt, B. Moore, and M. Yandell. 2014. Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics* 2014: 4.11.1-4.11.39.

Carbon, S., E. Douglass, B. M. Good, D. R. Unni, N. L. Harris, C. J. Mungall, S. Basu, et al. 2021. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* 49: D325–D334.

Carta, A., G. Bedini, and L. Peruzzi. 2020. A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytologist* 228: 1097–1106.

Chan, P. P., B. Y. Lin, A. J. Mak, and T. M. Lowe. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* 49: 9077–9096.

Chan, P. P., and T. M. Lowe. 2019. tRNAscan-SE: Searching for tRNA genes in genomic sequences. Methods in Molecular Biology, 1–14.

Chandler, G. T., and G. M. Plunkett. 2004. Evolution in Apiales: Nuclear and chloroplast markers together in (almost) perfect harmony. *Botanical Journal of the Linnean Society* 144: 123–147.

Chapdelaine, Y., and L. Bonen. 1991. The wheat mitochondrial gene for subunit I of the NADH dehydrogenase complex: A trans-splicing model for this gene-in-pieces. *Cell* 65: 465–472.

Chat, J., S. Decroocq, V. Decroocq, and R. J. Petit. 2002. A case of chloroplast heteroplasmy in kiwifruit (*Actinidia deliciosa*) that is not transmitted during sexual reproduction. *Journal of Heredity* 93: 293–300.

Chen, W., N. VanOpdorp, D. Fitzl, J. Tewari, P. Friedemann, T. Greene, S. Thompson, et al. 2012. Transposon insertion in a cinnamyl alcohol dehydrogenase gene is responsible for a brown midrib1 mutation in maize. *Plant Molecular Biology* 80: 289–297.

Chen, Y., F. Nie, S. Q. Xie, Y. F. Zheng, Q. Dai, T. Bray, Y. X. Wang, et al. 2021. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* 12.

Cheng, C., V. Krishnakumar, A. P. Chan, F. Thibaud-Nissen, S. Schobel, and C. D. Town. 2017a. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal* 89: 789–804.

Cheng, N., Y. Lo, M. I. Ansari, K. Ho, S. Jeng, N. Lin, and H. Dai. 2017b. Correlation between mtDNA complexity and mtDNA replication mode in developing cotyledon mitochondria during mung bean seed germination. *New Phytologist* 213: 751–763.

Chiu, W.-L., W. Stubbe, and B. B. Sears. 1988. Plastid inheritance in Oenothera: organelle genome modifies the extent of biparental plastid transmission. *Current Genetics* 13: 181–189.

Choudhury, R. R., and C. Parisod. 2017. Jumping genes: Genomic ballast or powerhouse of biological diversification. *Molecular Ecology* 26: 4587–4590.

Clark, J. W., and P. C. J. Donoghue. 2017. Constraining the timing of whole genome duplication in plant evolutionary history. *Proceedings of the Royal Society B: Biological Sciences* 284: 20170912.

Cole, L. W., W. Guo, J. P. Mower, and J. D. Palmer. 2018. High and Variable Rates of Repeat-Mediated Mitochondrial Genome Rearrangement in a Genus of Plants M. Purugganan [ed.],. *Molecular Biology and Evolution*.

De Coster, W., S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven. 2018. NanoPack: visualizing and processing long-read sequencing data B. Berger [ed.],. *Bioinformatics* 34: 2666–2669.

Dalechamps, J. 1586. Historia generalis plantarum... Historia generalis plantarum, 1107. Roville, Lyon, France.

Daniell, H., C. S. Lin, M. Yu, and W. J. Chang. 2016. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biology* 17.

Davila, J. I., M. P. Arrieta-Montiel, Y. Wamboldt, J. Cao, J. Hagmann, V. Shedge, Y.-Z. Xu, et al. 2011. Double-strand break repair processes drive evolution of the mitochondrial genome in *Arabidopsis*. *BMC Biology* 9: 64.

Degtjareva, G. V., M. D. Logacheva, T. H. Samigullin, E. I. Terentieva, and C. M. Valiejo-Roman. 2012. Organization of chloroplast *psbA-trnH* intergenic spacer in dicotyledonous angiosperms of the family umbelliferae. *Biochemistry (Moscow)* 77: 1056–1064.

Dettori, C. A., M. C. Loi, S. Brullo, P. Fraga i Arguimbau, E. Tamburini, and G. Bacchetta. 2016. The genetic diversity and structure of the *Ferula communis* L. complex (Apiaceae) in the Tyrrhenian area. *Flora: Morphology, Distribution, Functional Ecology of Plants* 223: 138–146.

Dettori, C. A., S. Sergi, E. Tamburini, and G. Bacchetta. 2014. The genetic diversity and spatial genetic structure of the Corso-Sardinian endemic *Ferula arrigonii* Bocchieri (Apiaceae). *Plant Biology* 16: 1005–1013.

Dierckxsens, N., P. Mardulyn, and G. Smits. 2016. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45: 18.

Divashuk, M. G., G. I. Karlov, and P. Y. Kroupin. 2020. Copy number variation of transposable elements in *Thinopyrum intermedium* and its diploid relative species. *Plants* 9.

Dixit, A. R., and O. P. Dhankher. 2011. A Novel Stress-Associated Protein 'AtSAP10' from *Arabidopsis thaliana* Confers Tolerance to Nickel, Manganese, Zinc, and High Temperature Stress A. Rahman [ed.],. *PLoS ONE* 6: e20921.

Dodsworth, S., M. W. Chase, and A. R. Leitch. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* 180: 1–5.

Dong, W.-L., R.-N. Wang, N.-Y. Zhang, W.-B. Fan, M.-F. Fang, and Z.-H. Li. 2018. Molecular Evolution of Chloroplast Genomes of Orchid Species: Insights into Phylogenetic Relationship and Adaptive Evolution. *International Journal of Molecular Sciences* 19: 716.

Dong, W., H. Liu, C. Xu, Y. Zuo, Z. Chen, and S. Zhou. 2014. A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: A case study on ginsengs. *BMC Genetics* 15: 1–8.

Downie, S. R., and R. K. Jansen. 2015. A Comparative Analysis of Whole Plastid Genomes from the Apiales: Expansion and Contraction of the Inverted Repeat, Mitochondrial to Plastid Transfer of DNA, and Identification of Highly Divergent Noncoding Regions. *Systematic Botany* 40: 336–351.

Downie, S. R., D. S. Katz-Downie, and M. F. Watson. 2000. A phylogeny of the flowering plant family apiaceae based on chloroplast DNA *rpl16* and *rpoC1* intron sequences: Towards a suprageneric classification of subfamily apioideae. *American Journal of Botany* 87: 273–292.

Downie, S. R., R. M. Peery, and R. K. Jansen. 2014. Another first for the Apiaceae: evidence for mitochondrial DNA transfer into the plastid genome. *İstanbul Eczacılık Fakültesi Dergisi* 44: 131–144.

Doyle, J. J., J. L. Doyle, and J. D. Palmer. 1995. Multiple Independent Losses of Two Genes and One Intron from Legume Chloroplast Genomes. *Systematic Botany* 20: 272.

Drude, C. G. O. 1898. Umbelliferae. Die natürlichen Pflanzenfamilien, 63–250. Wilhelm Engelmann, Leipzig.

Drula, E., M.-L. Garron, S. Dogan, V. Lombard, B. Henrissat, and N. Terrapon. 2022. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* 50: D571–D577.

Eickbush, T. H., and H. S. Malik. 2007. Origins and Evolution of Retrotransposons. Mobile DNA II, 1111–1144. Wiley.

Emms, D. M., and S. Kelly. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16: 157.

Emms, D. M., and S. Kelly. 2018. STAG: Species Tree Inference from All Genes. *BioRxiv*: 1–29.

Emms, D. M., and S. Kelly. 2017. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution* 34: 3267–3278.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, et al. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE* 7: 47768.

Evgen'ev, M. B., and I. R. Arkhipova. 2005. Penelope-like elements – a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenetic and Genome Research* 110: 510–521.

Evgen'ev, M. B., H. Zelentsova, N. Shostak, M. Kozitsina, V. Barskyi, D.-H. Lankenau, and V. G. Corces. 1997. Penelope , a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. *Proceedings of the National Academy of Sciences* 94: 196–201.

Falara, V., T. A. Akhtar, T. T. H. Nguyen, E. A. Spyropoulou, P. M. Bleeker, I. Schauvinhold, Y. Matsuba, et al. 2011. The Tomato Terpene Synthase Gene Family. *Plant Physiology* 157: 770–789.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, et al. 2014. Pfam: the protein families database. *Nucleic Acids Research* 42: D222–D230.

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

Fleischmann, A., T. P. Michael, F. Rivadavia, A. Sousa, W. Wang, E. M. Temsch, J. Greilhuber, et al. 2014. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany* 114: 1651–1663.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* 117: 9451–9457.

Folk, R. A., C. M. Siniscalchi, and D. E. Soltis. 2020. Angiosperms at the edge: Extremity, diversity, and phylogeny. *Plant Cell and Environment* 43: 2871–2893.

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

Gao, L., Y.-J. SU, and T. WANG. 2010. Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. *Journal of Systematics and Evolution* 48: 77–93.

Garcia, S., F. Galvez, A. Gras, A. Kova ik, and T. Garnatje. 2014. Plant rDNA database: update and new features. *Database* 2014: bau063–bau063.

Garrido-Ramos, M. A. 2015. Satellite DNA in Plants: More than Just Rubbish. *Cytogenetic and Genome Research* 146: 153–170.

Di Genova, A., E. Buena-Atienza, S. Ossowski, and M.-F. Sagot. 2021. Efficient hybrid *de novo* assembly of human genomes with WENGAN. *Nature Biotechnology* 39: 422–430.

Geurts, A. M., L. S. Collier, J. L. Geurts, L. L. Oseth, M. L. Bell, D. Mu, R. Lucito, et al. 2006. Gene mutations and genomic rearrangements in the mouse as a result of transposon mobilization from chromosomal concatemers. *PLoS Genetics* 2: 1413–1423.

Giegé, P., and A. Brennicke. 1999. RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. *Proceedings of the National Academy of Sciences* 96: 15324–15329.

Gil, J., Y. Um, S. Kim, O. Kim, S. Koo, C. Reddy, S.-C. Kim, et al. 2017. Development of Genome-Wide SSR Markers from *Angelica gigas* Nakai Using Next Generation Sequencing. *Genes* 8: 238.

Golicz, A. A., P. E. Bayer, G. C. Barker, P. P. Edger, H. Kim, P. A. Martinez, C. K. K. Chan, et al. 2016. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications* 7: 13390.

Goltsman, E., I. Ho, and D. Rokhsar. 2017. Meraculous-2D: Haplotype-sensitive Assembly of Highly Heterozygous genomes.

Goodwin, T. J. D., and R. T. M. Poulter. 2004. A New Group of Tyrosine Recombinase-Encoding Retrotransposons. *Molecular Biology and Evolution* 21: 746–759.

Goremykin, V. V., P. J. Lockhart, R. Viola, and R. Velasco. 2012. The mitochondrial genome of *Malus domestica* and the import-driven hypothesis of mitochondrial genome expansion in seed plants. *Plant Journal* 71: 615–626.

Goremykin, V. V., F. Salamini, R. Velasco, and R. Viola. 2009. Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. *Molecular Biology and Evolution* 26: 99–110.

Goulding, S. E., R. G. Olmstead, C. W. Morden, and K. H. Wolfe. 1996. Ebb and flow of the chloroplast inverted repeat. Springer-Verlag.

Gray, Y. H. M., M. M. Tanaka', J. A. Sved, and J. Sved. 1996. Present address: Department of Biological Sciences.

Gregory, T. R., J. A. Nicol, H. Tamm, B. Kullman, K. Kullman, I. J. Leitch, B. G. Murray, et al. 2007. Eukaryotic genome size databases. *Nucleic Acids Research* 35: D332–D338.

Greiner, S., P. Lehwark, and R. Bock. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research* 47: 59–64.

Guisinger, M. M., J. V. Kuehl, J. L. Boore, and R. K. Jansen. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: Rearrangements, repeats, and codon usage. *Molecular Biology and Evolution* 28: 583–600.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072–1075.

Gutman, B. L., and K. K. Niyogi. 2009. Evidence for base excision repair of oxidative DNA damage in chloroplasts of *Arabidopsis thaliana*. *Journal of Biological Chemistry* 284: 17006–17012.

Han, Y., S. Qin, and S. R. Wessler. 2013. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genomics* 14: 71.

Handa, H. 2008. Linear plasmids in plant mitochondria: Peaceful coexistences or malicious invasions? *Mitochondrion* 8: 15–25.

Handa, H. 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. *Nucleic Acids Research* 31: 5907–5916.

Hansen, A. K., L. K. Escobar, L. E. Gilbert, and R. K. Jansen. 2007. Paternal, maternal, and biparental inheritance of the chloroplast genome in *Passiflora* (Passifloraceae): implications for phylogenetic studies. *American Journal of Botany* 94: 42–46.

Hao, W., G. Liu, W. Wang, W. Shen, Y. Zhao, J. Sun, Q. Yang, et al. 2021. RNA Editing and Its Roles in Plant Organelles. *Frontiers in Genetics* 12.

Hawkins, J. S., H. Kim, J. D. Nason, R. A. Wing, and J. F. Wendel. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Research* 16: 1252–1261.

Hermann, M., F. Maier, A. Masroor, S. Hirth, A. J. P. Pfitzner, and U. M. Pfitzner. 2013. The *Arabidopsis* NIMIN proteins affect NPR1 differentially. *Frontiers in Plant Science* 4.

Hernandez, D., P. François, L. Farinelli, M. Østerås, and J. Schrenzel. 2008. *De novo* bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research* 18: 802–809.

Hinata, K., M. Watanabe, S. Yamakawa, Y. Satta, and A. Isogai. 1995. Evolutionary aspects of the S-related genes of the *Brassica* self-incompatibility system: synonymous and nonsynonymous base substitutions. *Genetics* 140: 1099–1104.

Hirsch, C. D., and N. M. Springer. 2017. Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1860: 157–165.

Hoff, K. J., A. Lomsadze, M. Borodovsky, and M. Stanke. 2019. Whole-genome annotation with BRAKER. Methods in Molecular Biology, 65–95.

Hon, T., K. Mars, G. Young, Y.-C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, et al. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data* 7: 399.

Hozza, M., T. Vinař, and B. Brejová. 2015. How big is that genome? Estimating genome size and coverage from k-mer abundance spectra. *In C. Iliopoulos*, *S. Puglisi*, and *E. Yilmaz* [eds.], Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 199–209. Springer International Publishing, Cham.

Hu, J., J. Fan, Z. Sun, and S. Liu. 2020. NextPolish: a fast and efficient genome polishing tool for long-read assembly B. Berger [ed.],. *Bioinformatics* 36: 2253–2255.

Hu, S., G. Sablok, B. Wang, D. Qu, E. Barbaro, R. Viola, M. Li, and C. Varotto. 2015. Plastome organization and evolution of chloroplast genes in *Cardamine* species adapted to contrasting habitats. *BMC Genomics* 16: 306.

Hu, T. T., P. Pattyn, E. G. Bakker, J. Cao, J.-F. Cheng, R. M. Clark, N. Fahlgren, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* 43: 476–481.

Huang, J. T., and H. K. Dooner. 2008. Macrotransposition and other complex chromosomal restructuring in maize by closely linked transposons in direct orientation. *Plant Cell* 20: 2019–2032.

Huang, N., F. Nie, P. Ni, X. Gao, F. Luo, and J. Wang. 2022a. BlockPolish: accurate polishing of long-read assembly via block divide-and-conquer. *Briefings in Bioinformatics* 23.

Huang, X., D. Coulibaly, W. Tan, Z. Ni, T. Shi, H. Li, F. Hayat, and Z. Gao. 2022b. The analysis of genetic structure and characteristics of the chloroplast genome in different Japanese apricot germplasm populations. *BMC Plant Biology* 22: 354.

Hubley S, G. P. 2013. RepeatMasker Open-4.0. Website http://www.repeatmasker.org.

Iorizzo, M., S. Ellison, D. Senalik, P. Zeng, P. Satapoomin, J. Huang, M. Bowman, et al. 2016. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics* 48: 657–666.

Iorizzo, M., D. Grzebelus, D. Senalik, M. Szklarczyk, D. Spooner, and P. Simon. 2012a. Against the traffic. *Mobile Genetic Elements* 2: 261–266.

Iorizzo, M., D. A. Senalik, D. Grzebelus, M. Bowman, P. F. Cavagnaro, M. Matvienko, H. Ashrafi, et al. 2011. *De novo* assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC Genomics* 12.

Iorizzo, M., D. Senalik, M. Szklarczyk, D. Grzebelus, D. Spooner, and P. Simon. 2012b. *De novo* assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. *BMC Plant Biology* 12.

Ito, H., H. Gaubert, E. Bucher, M. Mirouze, I. Vaillant, and J. Paszkowski. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* 472: 115–119.

Iwata, H., and O. Gotoh. 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research* 40: e161–e161.

Jansen, R. K., C. Kaittanis, C. Saski, S.-B. Lee, J. Tomkins, A. J. Alverson, and H. Daniell. 2006. Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evolutionary Biology* 6: 32.

Jansen, R. K., and T. A. Ruhlman. 2012. Plastid Genomes of Seed Plants. *In* R. Bock, and V. Knoop [eds.], Genomics of Chloroplasts and Mitochondria, 103–126. Springer Netherlands, Dordrecht.

Jansen, R. K., M. F. Wojciechowski, E. Sanniyasi, S.-B. Lee, and H. Daniell. 2008. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). *Molecular Phylogenetics and Evolution* 48: 1204–1217.

Jean Finnegan, E., D. M. Bond, D. M. Buzas, J. Goodrich, C. A. Helliwell, Y. Tamada, J.-Y. Yun, et al. 2011. Polycomb proteins regulate the quantitative induction of VERNALIZATION INSENSITIVE 3 in response to low temperatures. *The Plant Journal* 65: 382–391.

Jiao, Y., and A. H. Paterson. 2014. Polyploidy-associated genome modifications during land plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 20130355.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Jin, J. J., W. Bin Yu, J. B. Yang, Y. Song, C. W. Depamphilis, T. S. Yi, and D. Z. Li. 2020. GetOrganelle: A fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biology* 21: 1–31.

Jobson, R. W., and Y.-L. Qiu. 2008. Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift? *Biology Direct* 3: 43.

Johnson, L. B., and J. D. Palmer. 1989. Heteroplasmy of chloroplast DNA in Medicago. *Plant Molecular Biology* 12: 3–11.

Joly-Lopez, Z., E. Forczek, D. R. Hoen, N. Juretic, and T. E. Bureau. 2012. A Gene Family Derived from Transposable Elements during Early Angiosperm Evolution Has Reproductive Fitness Benefits in *Arabidopsis thaliana* J. L. Bennetzen [ed.],. *PLoS Genetics* 8: e1002931.

Jon Palmer, and J. S. 2019. nextgenusfs/funannotate: funannotate.

Joyce, P. B. M., and M. W. Gray. 1989. Chloroplast-like transfer RNA genes expressed in wheat mitochondria. *Nucleic Acids Research* 17: 5461–5476.

Judd, W. S., C. S. Campbell, E. A. Kellogg, P. F. Stevens, and P. D. Cantino. 1999. Plant systematics. A phylogenetic approach. Third edit. [ed.],. Sinauer Associates.

Jung, H., M.-S. Jeon, M. Hodgett, P. Waterhouse, and S. Eyun. 2020. Comparative Evaluation of Genome Assemblers from Long-Read Sequencing for Plants and Crops. *Journal of Agricultural and Food Chemistry* 68: 7670–7677.

Kalia, P., M. Mangal, S. Singh, C. Chugh, S. Mishra, and S. Chaudhary. 2019. Morphological and molecular changes on cytoplasmic male sterility (CMS) introgression in Asiatic carrot (*Daucus carota* L.). *Planta* 250: 507–518.

Kanehisa, M. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28: 27–30.

Kanneganti, V., and A. K. Gupta. 2008. Overexpression of OsiSAP8, a member of stress associated protein (SAP) gene family of rice confers tolerance to salt, drought and cold stress in transgenic tobacco and rice. *Plant Molecular Biology* 66: 445–462.

Kapitonov, V. V., and J. Jurka. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics* 23: 521–529.

Katoh, K., and D. M. Standley. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30: 772–780.

Kazazian, H. H. 2004. Mobile Elements: Drivers of Genome Evolution. *Science* 303: 1626–1632.

Kelly, S., and P. K. Maini. 2013. DendroBLAST: Approximate Phylogenetic Trees in the Absence of Multiple Sequence Alignments M. Robinson-Rechavi [ed.],. *PLoS ONE* 8: e58537.

Kenan-Eichler, M., D. Leshkowitz, L. Tal, E. Noor, C. Melamed-Bessudo, M. Feldman, and A. A. Levy. 2011. Wheat Hybridization and Polyploidization Results in Deregulation of Small RNAs. *Genetics* 188: 263–272.

Kenneth, R. 2013. The Plant List. *Version 1.1. Published on the Internet; Clusia suborbicularis*. Website http://www.theplantlist.org/1.1/browse/A/Menispermaceae/Tiliacora/%0Ahttp://www.theplantlist.org/ [accessed 1 April 2022].

Khakhlova, O., and R. Bock. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant Journal* 46: 85–94.

Khan, A. L., S. Asaf, I. J. Lee, A. Al-Harrasi, and A. Al-Rawahi. 2018. First chloroplast genomics study of *Phoenix dactylifera* (var. Naghal and Khanezi): A comparative analysis. *PLoS ONE* 13.

Khan, A. L., S. Asaf, Lubna, A. Al-Rawahi, and A. Al-Harrasi. 2021. Decoding first complete chloroplast genome of toothbrush tree (*Salvadora persica* L.): insight into genome evolution, sequence divergence and phylogenetic relationship within Brassicales. *BMC Genomics* 22: 1–16.

Kim, B. Y., J. R. Wang, D. E. Miller, O. Barmina, E. Delaney, A. Thompson, A. A. Comeault, et al. 2021. Highly contiguous assemblies of 101 drosophilid genomes. *eLife* 10.

Kim, D.-H., and S. Sung. 2013. Coordination of the Vernalization Response through a VIN3 and FLC Gene Family Regulatory Network in *Arabidopsis*. *The Plant Cell* 25: 454–469.

Kim, N.-H., M. Jayakodi, S.-C. Lee, B.-S. Choi, W. Jang, J. Lee, H. H. Kim, et al. 2018. Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. *Plant Biotechnology Journal* 16: 1904–1917.

Kljuykov, E. V., M. Liu, T. A. Ostroumova, M. G. Pimenov, P. M. Tilney, and B. E. Van Wyk. 2004. Towards a standardised terminology for taxonomically important morphological characters in the Umbelliferae. *South African Journal of Botany* 70: 488–496.

Knip, M., S. de Pater, and P. J. Hooykaas. 2012. The SLEEPER genes: a transposase-derived angiosperm-specific gene family. *BMC Plant Biology* 12: 192.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37: 540–546.

Kopelman, N. M., D. Lancet, and I. Yanai. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature Genetics* 37: 588–589.

Koralewski, T. E., and K. V. Krutovsky. 2011. Evolution of Exon-Intron Structure and Alternative Splicing J. Valcarcel [ed.],. *PLoS ONE* 6: e18055.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5: 1–9.

Korotkova, N., L. Nauheimer, H. Ter-Voskanyan, M. Allgaier, and T. Borsch. 2014. Variability among the Most Rapidly Evolving Plastid Genomic Regions is Lineage-Specific: Implications of Pairwise Genome Comparisons in Pyrus (Rosaceae) and Other Angiosperms for Marker Choice D. P. Little [ed.],. *PLoS ONE* 9: e112998.

Korotkova, N., J. V. Schneider, D. Quandt, A. Worberg, G. Zizka, and T. Borsch. 2009. Phylogeny of the eudicot order Malpighiales: analysis of a recalcitrant clade with sequences of the petD group II intron. *Plant Systematics and Evolution* 282: 201–228.

Kozik, A., B. A. Rowan, D. Lavelle, L. Berke, M. E. Schranz, R. W. Michelmore, and A. C. Christensen. 2019. The alternative reality of plant mitochondrial DNA: One ring does not rule them all N. M. Springer [ed.],. *PLOS Genetics* 15: e1008373.

Kraitshtein, Z., B. Yaakov, V. Khasdan, and K. Kashkush. 2010. Genetic and Epigenetic Dynamics of a Retrotransposon After Allopolyploidization of Wheat. *Genetics* 186: 801–812.

Kramerov, D., and N. Vassetzky. 2005. Short Retroposons in Eukaryotic Genomes. *International Review of Cytology* 247: 165–221.

Krupinska, K., N. E. Blanco, S. Oetke, and M. Zottini. 2020. Genome communication in plants mediated by organelle–nucleus-located proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.

Kurtz, S. 2010. The Vmatch large scale sequence analysis software features of Vmatch. *Briefings in Bioinformatics* 11: 473–483.

Kurtz, S., J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. 2001. REPuter: The manifold applications of repeat analysis on a genomic scale.

Kurzyna-Młynik, R., A. A. Oskolski, S. R. Downie, R. Kopacz, A. Wojewódzka, and K. Spalik. 2008. Phylogenetic position of the genus *Ferula* (Apiaceae) and its placement in tribe Scandiceae as inferred from nrDNA ITS sequence variation. *Plant Systematics and Evolution* 274: 47–66.

Kwolek, K., P. Kędzierska, M. Hankiewicz, M. Mirouze, O. Panaud, D. Grzebelus, and A. Macko-Podgórni. 2022. Diverse and mobile: eccDNA -based identification of carrot low-copy-number LTR retrotransposons active in callus cultures. *The Plant Journal* 110: 1811–1828.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. Initial sequencing and analysis of the human genome.

Landis, J. B., D. E. Soltis, Z. Li, H. E. Marx, M. S. Barker, D. C. Tank, and P. S. Soltis. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany* 105: 348–363.

Lange, B. M. 2015. The Evolution of Plant Secretory Structures and Emergence of Terpenoid Chemical Diversity. *Annual Review of Plant Biology* 66: 139–159.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.

Laslett, D., and B. Canback. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* 32: 11–16.

Lax, A. R., K. C. Vaughn, S. O. Duke, and J. E. Endrizzi. 1987. Structural and physiological studies of a plastome cotton mutant with slow sorting out. *Journal of Heredity* 78: 147–152.

Lee, C. S., and S. R. Downie. 2006. Phylogenetic relationships within Cicuta (Apiaceae tribe Oenantheae) inferred from nuclear rDNA ITS and cpDNA sequence data. *Canadian Journal of Botany* 84: 453–468.

Lee, H.-L., R. K. Jansen, T. W. Chumley, and K.-J. Kim. 2007. Gene Relocations within Chloroplast Genomes of Jasminum and Menodora (*Oleaceae*) Are Due to Multiple, Overlapping Inversions. *Molecular Biology and Evolution* 24: 1161–1180.

Leebens-Mack, J. H., M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, S. W. Graham, I. Grosse, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.

Lei, W., D. Ni, Y. Wang, J. Shao, X. Wang, D. Yang, J. Wang, et al. 2016. Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of Astragalus membranaceus. *Scientific Reports* 6: 21669.

Leitch, A. R., and I. J. Leitch. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist* 194: 629–646.

Lenz, H., A. Hein, and V. Knoop. 2018. Plant organelle RNA editing and its specificity factors: enhancements of analyses and new database features in PREPACT 3.0. *BMC Bioinformatics* 19: 255.

Levinson, G., and G. A. Gutman. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4: 203–221.

Li, H. 2012. Exploring single-sample SNP and INDEL calling with whole-genome *de novo* assembly. *Bioinformatics* 28: 1838–1844.

Li, J., D. F. Xie, X. L. Guo, Z. Y. Zheng, X. J. He, and S. D. Zhou. 2020. Comparative analysis of the complete plastid genome of five *Bupleurum* species and new insights into DNA barcoding and phylogenetic relationship. *Plants* 9.

Li, N., and Y. Li. 2014. Ubiquitin-mediated control of seed size in plants. *Frontiers in Plant Science* 5.

Liao, X., M. Li, Y. Zou, F.-X. Wu, Yi-Pan, and J. Wang. 2019. Current challenges and solutions of *de novo* assembly. *Quantitative Biology* 7: 90–109.

Librado, P., and J. Rozas. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–1452.

Linden, K. J., and J. Callis. 2020. The ubiquitin system affects agronomic plant traits. *Journal of Biological Chemistry* 295: 13940–13955.

Lisch, D. 2013. How important are transposons for plant evolution? *Nature Reviews Genetics* 14: 49–61.

Lisch, D., and R. K. Slotkin. 2011. Strategies for Silencing and Escape. The Ancient Struggle Between Transposable Elements and Their Hosts. 1st ed. Elsevier Inc.

Liu, J.-X., Q. Jiang, J.-P. Tao, K. Feng, T. Li, A.-Q. Duan, H. Wang, et al. 2021. Integrative genome, transcriptome, microRNA, and degradome analysis of water dropwort (*Oenanthe javanica*) in response to water stress. *Horticulture Research* 8: 262.

Liu, J., L. Shi, J. Han, G. Li, H. Lu, J. Hou, X. Zhou, et al. 2014. Identification of species in the angiosperm family Apiaceae using DNA barcodes. *Molecular Ecology Resources* 14: 1231–1238.

Liu, Y.-J., Z.-H. Xiu, R. Meeley, and B.-C. Tan. 2013. Empty Pericarp5 Encodes a Pentatricopeptide Repeat Protein That Is Required for Mitochondrial RNA Editing and Seed Development in Maize. *The Plant Cell* 25: 868–883.

Llorente, B., P. Durrens, A. Malpertuy, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, et al. 2000. Genomic Exploration of the Hemiascomycetous Yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Letters* 487: 122–133.

Lomsadze, A. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research* 33: 6494–6506.

Lonsdale, D. M., T. P. Hodge, and C. M.-R. Fauron. 1984. The physical map and organisation of the mitochondrial genome from the fertile cytoplasm of maize. *Nucleic Acids Research* 12: 9249–9261.

Luehrsen, K. R., and V. Walbot. 1990. Insertion of Mu1 elements in the first intron of the Adh1-S gene of maize results in novel RNA processing events. *Plant Cell* 2: 1225–1238.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, et al. 2015. Erratum to 'SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler' [GigaScience, (2012), 1, 18].

Lynch, M. 2002. Intron evolution as a population-genetic process. *Proceedings of the National Academy of Sciences* 99: 6118–6123.

Mackenzie, S. A. 2023. The mitochondrial genome of higher plants: a target for natural adaptation. *CABI Books*.

Macko-Podgorni, A., A. Nowicka, E. Grzebelus, P. W. Simon, and D. Grzebelus. 2013. DcSto: carrot Stowaway-like elements are abundant, diverse, and polymorphic. *Genetica* 141: 255–267.

Mahapatra, K., S. Banerjee, S. De, M. Mitra, P. Roy, and S. Roy. 2021. An Insight Into the Mechanism of Plant Organelle Genome Maintenance and Implications of Organelle Genome in Crop Improvement: An Update. *Frontiers in Cell and Developmental Biology* 9.

Majoros, W. H., M. Pertea, and S. L. Salzberg. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20: 2878–2879.

Makarevitch, I., A. J. Waters, P. T. West, M. Stitzer, C. N. Hirsch, J. Ross-Ibarra, and N. M. Springer. 2015. Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress M. Freeling [ed.],. *PLoS Genetics* 11: e1004915.

Manchanda, N., J. L. Portwood, M. R. Woodhouse, A. S. Seetharam, C. J. Lawrence-Dill, C. M. Andorf, and M. B. Hufford. 2020. GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics* 21: 193.

Marçais, G., and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27: 764–770.

Marchi, A., G. Appendino, I. Pirisi, M. Ballero, and M. C. Loi. 2003. Genetic differentiation of two distinct chemotypes of *Ferula communis* (Apiaceae) in Sardinia (Italy). *Biochemical Systematics and Ecology* 31: 1397–1408.

Mardanov, A. V., N. V. Ravin, B. B. Kuznetsov, T. H. Samigullin, A. S. Antonov, T. V. Kolganova, and K. G. Skyabin. 2008. Complete Sequence of the Duckweed (*Lemna minor*) Chloroplast Genome: Structural Organization and Phylogenetic Relationships to Other Angiosperms. *Journal of Molecular Evolution* 66: 555–564.

Maréchal, A., and N. Brisson. 2010. Recombination and the maintenance of plant organelle genome stability. *New Phytologist* 186: 299–317.

Martin, G., F.-C. Baurens, C. Cardi, J.-M. Aury, and A. D'Hont. 2013. The Complete Chloroplast Genome of Banana (*Musa acuminata*, Zingiberales): Insight into Plastid Monocotyledon Evolution J. G. Umen [ed.],. *PLoS ONE* 8: e67350.

Mayrose, I., S. H. Zhan, C. J. Rothfels, K. Magnuson-Ford, M. S. Barker, L. H. Rieseberg, and S. P. Otto. 2011. Recently Formed Polyploid Plants Diversify at Lower Rates. *Science* 333: 1257–1257.

McCartney, A. M., E. Hilario, S. Choi, J. Guhlin, J. M. Prebble, G. Houliston, T. R. Buckley, and D. Chagné. 2021. An exploration of assembly strategies and quality metrics on the accuracy of the rewarewa (*Knightia excelsa*) genome. *Molecular Ecology Resources* 21: 2125–2144.

McClintock, B. 1956. Controlling elements and the gene. *Cold Spring Harbor symposia on quantitative biology* 21: 197–216.

McCoy, S. R., J. V Kuehl, J. L. Boore, and L. A. Raubeson. 2008. The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. *BMC Evolutionary Biology* 8: 130.

Messier, W., S.-H. Li, and C.-B. Stewart. 1996. The birth of microsatellites. *Nature* 381: 483–483.

Mhiri, C., F. Borges, and M.-A. Grandbastien. 2022. Specificities and Dynamics of Transposable Elements in Land Plants. *Biology* 11: 488.

Miklenić, M. S., and I. K. Svetec. 2021. Palindromes in DNA—a risk for genome stability and implications in cancer. *International Journal of Molecular Sciences* 22: 1–19.

Millen, R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, et al. 2001. Many Parallel Losses of *infA* from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. *The Plant Cell* 13: 645–658.

Miyao, A., K. Tanaka, K. Murata, H. Sawaki, S. Takeda, K. Abe, Y. Shinozuka, et al. 2003. Target Site Specificity of the Tos17 Retrotransposon Shows a Preference for Insertion within Genes and against Insertion in Retrotransposon-Rich Regions of the Genome. *The Plant Cell* 15: 1771–1780.

Mizoguchi, T., and G. Coupland. 2000. ZEITLUPE and FKF1: novel connections between flowering time and circadian clock control. *Trends in Plant Science* 5: 409–411.

Moon, E., T.-H. Kao, and R. Wu. 1987. Rice chloroplast DNA molecules are heterogeneous as revealed by DNA sequences of a cluster of genes. *Nucleic Acids Research* 15: 611–630.

Morison, R. 1672. Plantarum umbelliferarum distributio nova. Plantarum umbelliferarum distributio nova., Theatro Sheldoniano, Oxford, UK.

Moriyama, Y., and K. Koshiba-Takeuchi. 2018. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Briefings in Functional Genomics* 17: 329–338.

Mower, J. P., and J. D. Palmer. 2006. Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*. *Molecular Genetics and Genomics* 276: 285–293.

Murigneux, V., S. K. Rai, A. Furtado, T. J. C. Bruxner, W. Tian, I. Harliwong, H. Wei, et al. 2020. Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience* 9.

Mustafina, F. U., D. K. Yi, K. Choi, C. H. Shin, K. S. Tojibaev, and S. R. Downie. 2019. A comparative analysis of complete plastid genomes from *Prangos fedtschenkoi* and *Prangos lipskyi* (Apiaceae). *Ecology and Evolution* 9: 364–377.

Myers, E. W. 2005. The fragment assembly string graph. *Bioinformatics* 21: ii79–ii85.

Naito, K., E. Cho, G. Yang, M. A. Campbell, K. Yano, Y. Okumoto, T. Tanisaka, and S. R. Wessler. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proceedings of the National Academy of Sciences* 103: 17620–17625.

Nowicka, A., E. Sliwinska, D. Grzebelus, R. Baranski, P. W. Simon, T. Nothnagel, and E. Grzebelus. 2016. Nuclear DNA content variation within the genus *Daucus* (Apiaceae) determined by flow cytometry. *Scientia Horticulturae* 209: 132–138.

Oda, K., K. Yamato, E. Ohta, Y. Nakamura, M. Takemura, N. Nozato, K. Akashi, et al. 1992. Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA. *Journal of Molecular Biology* 223: 1–7.

Okonechnikov, K., A. Conesa, and F. García-Alcalde. 2015. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*: btv566.

Okonechnikov, K., O. Golosova, and M. Fursov. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28: 1166–1167.

Oldenburg, D. J., and A. J. Bendich. 2015. DNA maintenance in plastids and mitochondria of plants. *Frontiers in Plant Science* 6.

Oldenburg, D. J., and A. J. Bendich. 2001. Mitochondrial DNA from the liverwort *Marchantia polymorpha*: circularly permuted linear molecules, head-to-tail concatemers, and a 5′ protein 1 1Edited by N.-M. Chua. *Journal of Molecular Biology* 310: 549–562.

Oldenburg, D. J., and A. J. Bendich. 1998. The structure of mitochondrial DNA from the liverwort, *Marchantia polymorpha*. *Journal of Molecular Biology* 276: 745–758.

Oliveira, E. J., J. G. Pádua, M. I. Zucchi, R. Vencovsky, and M. L. C. Vieira. 2006. Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology* 29: 294–307.

Oliver, K. R., J. A. McComb, and W. K. Greene. 2013. Transposable elements: Powerful contributors to angiosperm evolution and diversity. *Genome Biology and Evolution* 5: 1886–1901.

Olmedilla, A., D. Delcasso, M. Delseny, and A.-M. Cauwet-Marc. 1985. Variability in giant fennel (*Ferula communis*, Umbelliferae): Ribosomal RNA nuclear genes. *Plant Systematics and Evolution* 150: 263–274.

Ou, S., J. Chen, and N. Jiang. 2018. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*.

Ou, S., W. Su, Y. Liao, K. Chougule, J. R. A. Agda, A. J. Hellinga, C. S. B. Lugo, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* 20: 275.

Palmer, J. D. 1985. Comparative organization of chloroplast genomes. *Annual review of genetics* 19: 325–354.

Palmer, J. D., and L. A. Herbon. 1988. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution* 28: 87–97.

Palmer, J. D., and C. R. Shields. 1984. Tripartite structure of the *Brassica campestris* mitochondrial genome. *Nature* 307: 437–440.

Palmer, J. D., and W. F. Thompson. 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* 29: 537–550.

Palumbo, F., G. Galla, N. Vitulo, and G. Barcaccia. 2018. First draft genome sequencing of fennel (*Foeniculum vulgare* Mill.): identification of simple sequence repeats and their application in marker-assisted breeding. *Molecular Breeding* 38: 122.

Palumbo, F., A. Vannozzi, and G. Barcaccia. 2021. Impact of genomic and transcriptomic resources on apiaceae crop breeding strategies. *International Journal of Molecular Sciences* 22.

Panahi, M., Ł. Banasiak, M. Piwczyński, R. Puchałka, A. A. Oskolski, and K. Spalik. 2015. Phylogenetic relationships among Dorema, Ferula and Leutea (Apiaceae: Scandiceae: Ferulinae) inferred from nrDNA ITS and cpDNA noncoding sequences. *Taxon* 64: 770–783.

Panchy, N., M. Lehti-Shiu, and S.-H. Shiu. 2016. Evolution of Gene Duplication in Plants. *Plant Physiology* 171: 2294–2316.

Park, I., S. Yang, W. J. Kim, J. H. Song, H. S. Lee, H. O. Lee, J. H. Lee, et al. 2019. Sequencing and comparative analysis of the chloroplast genome of angelica polymorpha and the development of a novel indel marker for species identification. *Molecules* 24.

Parkin, I. A., C. Koh, H. Tang, S. J. Robinson, S. Kagale, W. E. Clarke, C. D. Town, et al. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biology* 15: R77.

Paysan-Lafosse, T., M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, et al. 2023. InterPro in 2022. *Nucleic Acids Research* 51: D418–D427.

Peery, R. 2015. Understanding angiosperm genome interactions and evolution: insights from sacred lotus (University of Illinois at Urbana-Champaign).

Pellicer, J., M. F. Fay, and I. J. Leitch. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* 164: 10–15.

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

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

Pevzner, P. A., and H. Tang. 2001. Fragment assembly with double-barreled data. *Bioinformatics* 17: S225–S233.

Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* 98: 9748–9753.

Pichersky, E., and R. A. Raguso. 2018. Why do plants produce so many terpenoid compounds? *New Phytologist* 220: 692–702.

Pietzenuk, B., C. Markus, H. Gaubert, N. Bagwan, A. Merotto, E. Bucher, and A. Pecinka. 2016. Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. *Genome Biology* 17: 209.

Pimenov, M. G., and M. V. Leonov. 1993. The Genera of Umbelliferae: A Nomenclator. Kew Bulletin, 592.

Piwczyński, M., P. Trzeciak, M.-O. Popa, M. Pabijan, J. M. Corral, K. Spalik, and A. Grzywacz. 2021. Using RAD seq for reconstructing phylogenies of highly diverged taxa: A test using the tribe Scandiceae (Apiaceae). *JSE Journal of Systematics and Evolution*.

Piwczyński, M., D. Wyborska, J. Gołębiewska, and R. Puchałka. 2018. Phylogenetic positions of seven poorly known species of *Ferula* (Apiaceae) with remarks on the phylogenetic utility of the plastid *trnH-psbA*, *trnS-trnG*, and *atpB-rbcL* intergenic spacers. *Systematics and Biodiversity* 16: 428–440.

Plohl, M., N. Meštrović, and B. Mravinac. 2014. Centromere identity from the DNA point of view. *Chromosoma* 123: 313–325.

Plohl, M., N. Meštrović, and B. Mravinac. 2012. Satellite DNA evolution. Genome Dynamics, 126–152.

Plunkett, G. M., and S. R. Downie. 2000. Expansion and contraction of the chloroplast inverted repeat in Apiaceae subfamily Apioideae. *Systematic Botany* 25: 648–667.

Plunkett, G. M., and S. R. Downie. 1999. Major lineages within Apiaceae subfamily Apioideae: A comparison of chloroplast restriction site and DNA sequence data.

Plunkett, G. M., M. G. Pimenov, J.-P. Reduron, E. V. Kljuykov, B.-E. van Wyk, T. A. Ostroumova, M. J. Henwood, et al. 2018. Apiaceae. *In* J. W. Kadereit, and V. Bittrich [eds.], Flowering Plants. Eudicots, 9–206. Springer International Publishing, Cham.

del Pozo, J. C., and E. Ramirez-Parra. 2015. Whole genome duplications in plants: an overview from *Arabidopsis*. *Journal of Experimental Botany* 66: 6991–7003.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* 26: 1641–1650.

Qin, H.-H., J. Cai, C.-K. Liu, R.-X. Zhou, M. Price, S.-D. Zhou, and X.-J. He. 2023. The plastid genome of twenty-two species from *Ferula*, *Talassia*, and *Soranthus*: comparative analysis, phylogenetic implications, and adaptive evolution. *BMC Plant Biology* 23: 9.

Que, F., X.-L. Hou, G.-L. Wang, Z.-S. Xu, G.-F. Tan, T. Li, Y.-H. Wang, et al. 2019. Advances in research on the carrot, an important root vegetable in the Apiaceae family. *Horticulture Research* 6: 69.

Raubeson, L. A., and R. K. Jansen. 2005. Chloroplast genomes of plants. Plant diversity and evolution: genotypic and phenotypic variation in higher plants, 45–68. CABI Publishing, UK.

Rawlings, N. D., A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn. 2018. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research* 46: D624–D632.

Reduron, J. P. 2021. Carrots and Related Apiaceae Crops: Taxonomy, Origin and Importance of The Apiaceae Family. *CAB International 2021*. 2: 1–8.

Renaut, S., H. C. Rowe, M. C. Ungerer, and L. H. Rieseberg. 2014. Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 20130345.

Reyes-Chin-Wo, S., Z. Wang, X. Yang, A. Kozik, S. Arikit, C. Song, L. Xia, et al. 2017. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications* 8: 14953.

Rice, A., L. Glick, S. Abadi, M. Einhorn, N. M. Kopelman, A. Salman-Minkov, J. Mayzel, et al. 2015. The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytologist* 206: 19–26.

Rice, D. W., A. J. Alverson, A. O. Richardson, G. J. Young, M. V. Sanchez-Puerta, J. Munzinger, K. Barry, et al. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. *Science* 342: 1468–1473.

Rizzi, R., S. Beretta, M. Patterson, Y. Pirola, M. Previtali, G. Della Vedova, and P. Bonizzoni. 2019. Overlap graphs and de Bruijn graphs: data structures for *de novo* genome assembly in the big data era. *Quantitative Biology* 7: 278–292.

Rose, O., and D. Falush. 1998. A threshold size for microsatellite expansion. *Molecular Biology and Evolution* 15: 613–615.

Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. Gabor, Miklos, C. R. Nelson, I. K. Hariharan, et al. 2000. Comparative Genomics of the Eukaryotes. *Science* 287: 2204–2215.

Ruhlman, T., S. B. Lee, R. K. Jansen, J. B. Hostetler, L. J. Tallon, C. D. Town, and H. Daniell. 2006. Complete plastid genome sequence of *Daucus carota*: Implications for biotechnology and phylogeny of angiosperms. *BMC Genomics* 7: 1–13.

Sabir, J. S. M., D. Arasappan, A. Bahieldin, S. Abo-Aba, S. Bafeel, T. A. Zari, S. Edris, et al. 2014. Whole Mitochondrial and Plastid Genome SNP Analysis of Nine Date Palm Cultivars Reveals Plastid Heteroplasmy and Close Phylogenetic Relationships among Cultivars F. C. C. Leung [ed.],. *PLoS ONE* 9: e94158.

Sahebi, M., M. M. Hanafi, A. J. van Wijnen, D. Rice, M. Y. Rafii, P. Azizi, M. Osman, et al. 2018. Contribution of transposable elements in the plant's genome. *Gene* 665: 155–166.

Sànchez-Cuxart, A., and B. C. Mercè. 1998. Estudi biosistemàtic de les poblacions de *Ferula communis* del NE de la península Ibèrica i de les illes Balears. *Acta Botanica Barcinonensia* 45.

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

Sasaki, A., H. Itoh, K. Gomi, M. Ueguchi-Tanaka, K. Ishiyama, M. Kobayashi, D.-H. Jeong, et al. 2003. Accumulation of Phosphorylated Repressor for Gibberellin Signaling in an F-box Mutant. *Science* 299: 1896–1898.

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

Sayed-Ahmad, B., T. Talou, Z. Saad, A. Hijazi, and O. Merah. 2017. The Apiaceae: Ethnomedicinal family as source for industrial uses. *Industrial Crops and Products* 109: 661–671.

Sayers, E. W., E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Research* 50: D20–D26.

Scarpino, S. V., D. A. Levin, and L. A. Meyers. 2014. Polyploid Formation Shapes Flowering Plant Diversity. *The American Naturalist* 184: 456–465.

Schatz, M. C., A. L. Delcher, and S. L. Salzberg. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research* 20: 1165–1173.

Schlötterer, C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma* 109: 365–371.

Schlötterer, C., and D. Tautz. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* 20: 211–215.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, et al. 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* 326: 1112–1115.

Schuster, W., and A. Brennicke. 1994. The plant mitochondrial genome: Physical structure, information content, RNA editing, and gene migration to the nucleus. *Annu.Rev.Plant Physiol.Plant Mol.Biol.* 45: 61–78.

Sedlazeck, F. J., H. Lee, C. A. Darby, and M. C. Schatz. 2018a. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* 19: 329–346.

Sedlazeck, F. J., P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, and M. C. Schatz. 2018b. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 15: 461–468.

Shedge, V., M. Arrieta-Montiel, A. C. Christensen, and S. A. Mackenzie. 2007. Plant Mitochondrial Recombination Surveillance Requires Unusual RecA and MutS Homologs. *The Plant Cell* 19: 1251–1264.

Shen, Y., Z. Zhou, Z. Wang, W. Li, C. Fang, M. Wu, Y. Ma, et al. 2014. Global Dissection of Alternative Splicing in Paleopolyploid Soybean. *The Plant Cell* 26: 996–1008.

Shi, L., H. Chen, M. Jiang, L. Wang, X. Wu, L. Huang, and C. Liu. 2019. CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Research* 47: W65–W73.

Shikanai, T. 2015. RNA editing in plants: Machinery and flexibility of site recognition. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1847: 779–785.

Shneyer, V. S., G. P. Borschtschenko, and M. G. Pimenov. 1995. Immunochemical appraisal of relationships within the tribe Peucedaneae (Apiaceae). *Plant Systematics and Evolution* 198: 1–16.

Siljak-Yakovlev, S., F. Pustahija, E. M. Šolić, F. Bogunić, E. Muratović, N. Bašić, O. Catrice, and S. C. Brown. 2010. Towards a Genome Size and Chromosome Number Database of Balkan Flora: C-Values in 343 Taxa with Novel Values for 242. *Advanced Science Letters* 3: 190–213.

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

Simon, P. W. 2021. Carrot (*Daucus carota* L.) Breeding. Advances in Plant Breeding Strategies: Vegetable Crops, 213–238. Springer International Publishing, Cham.

Simpson, J. T., and R. Durbin. 2010. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 26: i367–i373.

Simpson, J. T., and R. Durbin. 2012. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research* 22: 549–556.

Simpson, J. T., and M. Pop. 2015. The Theory and Practice of Genome Sequence Assembly. *Annual Review of Genomics and Human Genetics* 16: 153–172.

Singh, S., A. Singh, N. Jain, G. Singh, A. Ahlawat, and I. Ravi. 2013. Molecular characterization of vernalization and photoperiod genes in wheat varieties from different agro-climatic zones of India. *Cereal Research Communications* 41: 376–387.

Skippington, E., T. J. Barkman, D. W. Rice, and J. D. Palmer. 2015. Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all *nad* genes. *Proceedings of the National Academy of Sciences* 112.

Sloan, D. B. 2013. One ring to rule them all? Genome sequencing provides new insights into the 'master circle' model of plant mitochondrial DNA structure. *New Phytologist* 200: 978–985.

Small, I. D., M. Schallenberg-Rüdinger, M. Takenaka, H. Mireau, and O. Ostersetzer-Biran. 2020. Plant organellar RNA editing: what 30 years of research has revealed. *The Plant Journal* 101: 1040–1056.

Smith, D. R., and P. J. Keeling. 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences of the United States of America* 112: 10177–10184.

Sohn, J., and J.-W. Nam. 2016. The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics*: bbw096.

Soltis, P. S., and D. E. Soltis. 2021. Plant genomes: Markers of evolutionary history and drivers of evolutionary change. *PLANTS, PEOPLE, PLANET* 3: 74–82.

Song, X., P. Sun, J. Yuan, K. Gong, N. Li, F. Meng, Z. Zhang, et al. 2021. The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. *Plant Biotechnology Journal* 19: 731–744.

Song, X., J. Wang, N. Li, J. Yu, F. Meng, C. Wei, C. Liu, et al. 2020. Deciphering the high-quality genome sequence of coriander that causes controversial feelings. *Plant Biotechnology Journal* 18: 1444–1456.

Sosso, D., S. Mbelo, V. Vernoud, G. Gendrot, A. Dedieu, P. Chambrier, M. Dauzat, et al. 2012. PPR2263, a DYW-Subgroup Pentatricopeptide Repeat Protein, Is Required for Mitochondrial *nad5* and cob Transcript Editing, Mitochondrion Biogenesis, and Maize Growth. *The Plant Cell* 24: 676–691.

Spalik, K., S. R. Downie, and M. F. Watson. 2009. Generic delimitations within the Sium alliance (Apiaceae tribe Oenantheae) inferred from cpDNA *rps16*-5′*trnK* (UUU) and nrDNA ITS sequences. *Taxon* 58: 735–748.

Spooner, D. M., P. W. Simon, D. Senalik, and M. Iorizzo. 2019. Carrot Organelle Genomes: Organization, Diversity, and Inheritance. *In* P. Simon, M. Iorizzo, D. Grzebelus, and R. Baranski [eds.], 205–223. Springer International Publishing, Cham.

Spooner, D. M., M. P. Widrlechner, K. R. Reitsma, D. E. Palmquist, S. Rouz, Z. Ghrabi-Gammar, M. Neffati, et al. 2014. Reassessment of Practical Subspecies Identifications of the USDA *Daucus carota* L. Germplasm Collection: Morphological Data. *Crop Science* 54: 706–718.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stanke, M., O. Schöffmann, B. Morgenstern, and S. Waack. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7: 62.

Staskawicz, B. J., F. M. Ausubel, B. J. Baker, J. G. Ellis, and J. D. G. Jones. 1995. Molecular Genetics of Plant Disease Resistance. *Science* 268: 661–667.

Stein, D. B., J. D. Palmer, and W. F. Thompson. 1986. Structural evolution and flip-flop recombination of chloroplast DNA in the fern genus Osmunda. *Current Genetics* 10: 835–841.

Sticher, L., B. Mauch-Mani, and and J. Métraux. 1997. SYSTEMIC ACQUIRED RESISTANCE. *Annual Review of Phytopathology* 35: 235–270.

Stoebe, B., W. Martin, and K. V. Kowallik. 1998. Distribution and Nomenclature of Protein-coding Genes in 12 Sequenced Chloroplast Genomes. *Plant Molecular Biology Reporter* 16: 243–255.

Stoler, N., and A. Nekrutenko. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* 3.

Studer, A., Q. Zhao, J. Ross-Ibarra, and J. Doebley. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics* 43: 1160–1163.

Sun, H., J. DIng, M. Piednoël, and K. Schneeberger. 2018. FindGSE: Estimating genome size variation within human and *Arabidopsis* using k -mer frequencies. *Bioinformatics* 34: 550–557.

Sun, J., F. Lu, Y. Luo, L. Bie, L. Xu, and Y. Wang. 2023. OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Research*.

Svensson, B., I. Svendsen, P. Hoejrup, P. Roepstorff, S. Ludvigsen, and F. M. Poulsen. 1992. Primary structure of Barwin: a barley seed protein closely related to the C-terminal domain of proteins encoded by wound-induced plant genes. *Biochemistry* 31: 8767–8770.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, et al. 2007. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research* 36: D1009–D1014.

Sze, S.-H., J. J. Parrott, and A. M. Tarone. 2017. A divide-and-conquer algorithm for large-scale *de novo* transcriptome assembly through combining small assemblies from existing algorithms. *BMC Genomics* 18: 895.

Takenaka, M., A. Zehrmann, D. Verbitskiy, B. Härtel, and A. Brennicke. 2013. RNA Editing in Plants and Its Evolution. *Annual Review of Genetics* 47: 335–352.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, et al. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4: 1–14.

Tillich, M., P. Lehwark, T. Pellizzer, E. S. Ulbricht-Jones, A. Fischer, R. Bock, and S. Greiner. 2017. GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* 45: W6–W11.

Tilney-Bassett, R. A. E., and C. W. Birky. 1981. The mechanism of the mixed inheritance of chloroplast genes in Pelargonium. *Theoretical and Applied Genetics* 60: 43–53.

Tsuji, S., K. Ueda, T. Nishiyama, M. Hasebe, S. Yoshikawa, A. Konagaya, T. Nishiuchi, and K. Yamaguchi. 2007. The chloroplast genome from a lycophyte (microphyllophyte), *Selaginella*

uncinata, has a unique inversion, transpositions and many gene losses. *Journal of Plant Research* 120: 281–290.

Tsukahara, S., A. Kobayashi, A. Kawabe, O. Mathieu, A. Miura, and T. Kakutani. 2009. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* 461: 423–426.

Ungerer, M. C., S. C. Strakosh, and K. M. Stimpson. 2009. Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. *BMC Biology* 7: 40.

Unseld, M., J. R. Marienfeld, P. Brandt, and A. Brennicke. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genetics* 15: 57–61.

Vanneste, K., G. Baele, S. Maere, and Y. Van de Peer. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Research* 24: 1334–1347.

Vukich, M., T. Giordani, L. Natali, and A. Cavallini. 2009. Copia and Gypsy retrotransposons activity in sunflower (*Helianthus annuus* L.). *BMC Plant Biology* 9: 150.

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, and M. C. Schatz. 2017. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 33: 2202–2204.

Wakasugi, T., T. Tsudzuki, and M. Sugiura. 2001. The genomics of land plant chloroplasts: Gene content and alteration of genomic information by RNA editing. *Photosynthesis Research* 70: 107–118.

Wan, L., K. Essuman, R. G. Anderson, Y. Sasaki, F. Monteiro, E.-H. Chung, E. Osborne Nishimura, et al. 2019. TIR domains of plant immune receptors are NAD + cleaving enzymes that promote cell death. *Science* 365: 799–803.

Wang, D., Z. Zheng, Y. Li, H. Hu, Z. Wang, X. Du, S. Zhang, et al. 2021. Which factors contribute most to genome size variation within angiosperms? *Ecology and Evolution* 11: 2660–2668.

Wang, P., and F. Wang. 2023. A proposed metric set for evaluation of genome assembly quality. *Trends in Genetics* 39: 175–186.

Wang, W., A. Das, D. Kainer, M. Schalamun, A. Morales-Suarez, B. Schwessinger, and R. Lanfear. 2020. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing *de novo* assemblies. *GigaScience* 9.

Wang, W., and R. Lanfear. 2019. Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants B. Gaut [ed.],. *Genome Biology and Evolution*.

Wang, X.-J., Q. Luo, T. Li, P.-H. Meng, Y.-T. Pu, J.-X. Liu, J. Zhang, et al. 2022. Origin, evolution, breeding, and omics of Apiaceae: a family of vegetables and medicinal plants. *Horticulture Research* 9.

Wang, X., H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* 43: 1035–1039.

Wen, J., M. Vanek-Krebitz, K. Hoffmann-Sommergruber, O. Scheiner, and H. Breiteneder. 1997. The Potential ofBetv1Homologues, a Nuclear Multigene Family, as Phylogenetic Markers in Flowering Plants. *Molecular Phylogenetics and Evolution* 8: 317–333.

Wendel, J. F., S. A. Jackson, B. C. Meyers, and R. A. Wing. 2016. Evolution of plant genome architecture. *Genome Biology* 17: 37.

Wick, R. R., L. M. Judd, C. L. Gorrie, and K. E. Holt. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* 3.

Wick, R. R., M. B. Schultz, J. Zobel, and K. E. Holt. 2015. Bandage: Interactive visualization of *de novo* genome assemblies. *Bioinformatics* 31: 3350–3352.

Wicke, S., K. F. Müller, C. W. de Pamphilis, D. Quandt, N. J. Wickett, Y. Zhang, S. S. Renner, and G. M. Schneeweiss. 2013. Mechanisms of Functional and Physical Genome Reduction in Photosynthetic and Non-photosynthetic Parasitic Plants of the Broomrape Family. *The Plant Cell* 25: 3711–3725.

Wicker, T., H. Gundlach, M. Spannagl, C. Uauy, P. Borrill, R. H. Ramírez-González, R. De Oliveira, et al. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology* 19.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8: 973–982.

Wojewódzka, A., J. Baczyński, Ł. Banasiak, S. R. Downie, A. Czarnocka-Cieciura, M. Gierek, K. Frankiewicz, and K. Spalik. 2019. Evolutionary shifts in fruit dispersal syndromes in Apiaceae tribe Scandiceae. *Plant Systematics and Evolution* 305: 401–414.

Woloszynska, M. 2010. Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes-though this be madness, yet there's method in't. *Journal of Experimental Botany* 61: 657–671.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106: 13875–13879.

Wu, S., B. Han, and Y. Jiao. 2019. Genetic Contribution of Paleopolyploidy to Adaptive Evolution in Angiosperms. *Molecular Plant* 13: 59–71.

Wu, Z., X. Liao, X. Zhang, L. R. Tembrock, and A. Broz. 2022. Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. *Journal of Systematics and Evolution* 60: 160–168.

Wu, Z., D. B. Sloan, C. W. Brown, M. Rosenblueth, J. D. Palmer, and H. C. Ong. 2017. Mitochondrial retroprocessing promoted functional transfers of *rpl5* to the nucleus in grasses. *Molecular Biology and Evolution* 34: 2340–2354.

Xie, Z., and S. Merchant. 1996. The Plastid-encoded *ccsA* Gene Is Required for Heme Attachment to Chloroplast c-type Cytochromes. *Journal of Biological Chemistry* 271: 4632–4639.

Xuan, Y. H., H. L. Piao, B. Il Je, S. J. Park, S. H. Park, J. Huang, J. B. Zhang, et al. 2011. Transposon Ac/Ds-induced chromosomal rearrangements at the rice OsRLG5 locus. *Nucleic Acids Research* 39: 1–12.

Yang, L., O. Abduraimov, K. Tojibaev, K. Shomurodov, Y.-M. Zhang, and W.-J. Li. 2022. Analysis of complete chloroplast genome sequences and insight into the phylogenetic relationships of *Ferula* L. *BMC Genomics* 23: 643.

Yang, M., X. Zhang, G. Liu, Y. Yin, K. Chen, Q. Yun, D. Zhao, et al. 2010. The Complete Chloroplast Genome Sequence of Date Palm (*Phoenix dactylifera* L.) J. H. Badger [ed.],. *PLoS ONE* 5: e12762.

Yang, Y., P. Sun, L. Lv, D. Wang, D. Ru, Y. Li, T. Ma, et al. 2020. Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nature Plants* 6: 215–222.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Yang, Z., and R. Nielsen. 2002. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Mol. Biol. Evol* 19: 908–917.

Ye, C., C. M. Hill, S. Wu, J. Ruan, and Z. Ma. 2016. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports* 6.

Ye, C., and Z. (Sam) Ma. 2016. Sparc: A sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ* 2016.

Ye, C., Z. S. Ma, C. H. Cannon, M. Pop, and D. W. Yu. 2012. Exploiting sparseness in *de novo* genome assembly.

Yi, T. S., G. H. Jin, and J. Wen. 2015. Chloroplast capture and intra- and inter-continental biogeographic diversification in the Asian - New World disjunct plant genus Osmorhiza (Apiaceae). *Molecular Phylogenetics and Evolution* 85: 10–21.

Yin, M., S. Zhang, X. Du, R. G. Mateo, W. Guo, A. Li, Z. Wang, et al. 2021. Genomic analysis of *Medicago ruthenica* provides insights into its tolerance to abiotic stress and demographic history. *Molecular Ecology Resources* 21: 1641–1657.

Yuan, Y., S. Zhong, Q. Li, Z. Zhu, Y. Lou, L. Wang, J. Wang, et al. 2007. Functional analysis of rice NPR1 -like genes reveals that OsNPR1 / NH1 is the rice orthologue conferring disease resistance with enhanced herbivore susceptibility. *Plant Biotechnology Journal* 5: 313–324.

Zaegel, V., B. Guermann, M. Le Ret, C. Andrés, D. Meyer, M. Erhardt, J. Canaday, et al. 2007. The Plant-Specific ssDNA Binding Protein OSB1 Is Involved in the Stoichiometric Transmission of Mitochondrial DNA in Arabidopsis. *The Plant Cell* 18: 3548–3563.

Zeb, U., W. L. Dong, T. T. Zhang, R. N. Wang, K. Shahzad, X. F. Ma, and Z. H. Li. 2020. Comparative plastid genomics of Pinus species: Insights into sequence variations and phylogenetic relationships. *Journal of Systematics and Evolution* 58: 118–132.

Zerbino, D. R., and E. Birney. 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* 22: 2472–2479.

Zhang, J., F. Zhang, and T. Peterson. 2006. Transposition of reversed Ac element ends generates novel chimeric genes in maize. *PLoS Genetics* 2: 1535–1540.

Zhang, S.-J., L. Liu, R. Yang, and X. Wang. 2020. Genome Size Evolution Mediated by Gypsy Retrotransposons in Brassicaceae. *Genomics, Proteomics & Bioinformatics* 18: 321–332.

Zhang, T., Y. Fang, X. Wang, X. Deng, X. Zhang, S. Hu, and J. Yu. 2012. The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: Insights into the evolution of plant organellar genomes. *PLoS ONE* 7: 30531.

Zhao, D., A. A. Ferguson, and N. Jiang. 2016. What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1859: 366–380.

Zhao, D., W. Ni, B. Feng, T. Han, M. G. Petrasek, and H. Ma. 2003. Members of the *Arabidopsis*-SKP1-like Gene Family Exhibit a Variety of Expression Patterns and May Play Diverse Roles in Arabidopsis. *Plant Physiology* 133: 203–217.

Zhao, D., Q. Yu, M. Chen, and H. Ma. 2001. The ASK1 gene regulates B function gene expression in cooperation with UFO and LEAFY in Arabidopsis. *Development* 128: 2735–2746.

Zheng, N., B. A. Schulman, L. Song, J. J. Miller, P. D. Jeffrey, P. Wang, C. Chu, et al. 2002. Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase complex. *Nature* 416: 703–709.

Zhou, J., X. Gong, S. R. Downie, and H. Peng. 2009. Towards a more robust molecular phylogeny of Chinese Apiaceae subfamily Apioideae: Additional evidence from nrDNA ITS and cpDNA intron (rpl16 and rps16) sequences. *Molecular Phylogenetics and Evolution* 53: 56–68.

Zhu, C.-R., J. Xu, M.-L. DU, L.-H. Wang, C. Sui, and J.-H. Wei. 2019. Genome survey analysis and SSR loci mining of *Bupleurum falcatum*. *Zhongguo Zhong yao za zhi Zhongguo zhongyao zazhi China journal of Chinese materia medica* 44: 3960–3966.

Župunski, V., F. Gubenšek, and D. Kordis. 2001. Evolutionary Dynamics and Evolutionary History in the RTE Clade of Non-LTR Retrotransposons. *Molecular Biology and Evolution* 18: 1849–1863.

Appendix

Appendix 1. Glossary of terms

Contig	A continuous sequence of bases that is generated by an assembly algorithm (Simpson and Pop, 2015). It represents a segment of genetic material where the original sequence has been reconstructed by the assembler.
De Bruijn graph	A graph-based data structure used in bioinformatics for sequence assembly, named after the Dutch mathematician Nicolaas Govert de Bruijn. In this approach, each read is fragmented into a series of overlapping *k*-mers. These individual *k*-mers are then incorporated into the graph as vertices, and *k*-mers stemming from adjacent positions in a read are connected by edges (Simpson and Pop, 2015).
De novo	A Latin term that describes the process of creating or determining something, like as assembling a sequence without relying on prior knowledge or a reference.
k-mer	A short sequence of DNA bases of length k extracted from DNA reads (Rizzi et al., 2019), and play a crucial role in genome assembly by facilitating the identification of overlaps and enabling the reconstruction of longer DNA sequences.
K-mer walking	A technique used to fill gaps or resolve uncertainties in the sequence by systematically examining overlapping k-mers. The process involves extending the known sequence by one k-mer at a time, using the overlapping k-mers found in the raw data. This allows the assembler to move step by step, filling in gaps and refining the assembly in regions where the initial sequence was ambiguous

	or incomplete (Goltsman et al., 2017).
Overlap-Layout-Consensus (OLC) graph	The simplest graph-based model, where individual reads are represented as vertices, and overlapping reads are connected by edges. The process begins by identifying pairs of reads that overlap with each other. Next, a graph is built, and reads are organized and oriented in a suitable order (layout). Finally, a consensus sequence is generated by combining the ordered and oriented reads.
Palindromic sequence	Refers to a DNA sequence that contains two identical or highly similar inverted repeats. These repeats can either be adjacent to each other or separated by a spacer region, and exhibit a symmetry where the nucleotide sequence reads the same in both direction (Miklenić and Svetec, 2021).
Scaffold	Also referred as a supercontig, is formed by assembling a series of contigs using additional information, such as mate-pair data, long-reads, or a reference genome (Simpson and Pop, 2015). It serves as a larger structure that connects and organizes contigs, providing information about the relative ordering and orientation of the underlying sequences.
The FM index	A powerful tool in computational genomics that provides a space-efficient representation of DNA or protein sequences and enables efficient pattern matching operations on large-scale genomic data (Simpson and Durbin, 2010).

Appendix 2

Introduction to genome assembly algorithms

A critical challenge in genome assembly is the presence of repetitive DNA sequences within genomes. Many genomes contain segments that are repeated in similar or identical forms, complicating the assembly process. Two prominent paradigms, namely the greedy approach and graph-based approaches, have been employed to tackle the challenges of genome assembly.

Greedy approaches

One of the most straightforward approaches to genome sequence assembly involves iteratively connecting the reads based on the quality of their overlaps, with the highest-quality overlaps given priority (Rizzi et al., 2019). This process begins by joining the two reads that have the best overlap, considering factors such as overlap length or a more comprehensive quality measure that incorporates base quality estimates. This joining process continues until a predefined minimum quality threshold is reached. Consequently, the assembled sequences, known as contigs, grow either by incorporating new reads or by merging with previously constructed contigs (Simpson and Pop, 2015). Overlaps that conflict with existing contigs are disregarded. This strategy is referred to as greedy because it consistently makes the locally optimal choice at each step. Despite its simplicity, this approach, along with its variations, yields a reasonable approximation for achieving optimal assembly.

Although the greedy strategy initially achieved notable accomplishments, it possesses a significant drawback: due to its focus on local information, it struggles to effectively deal with repetitive regions within genomes. As the scale and intricacy of the sequenced genomes grew, more sophisticated graph-based algorithms emerged as a replacement for greedy approaches. These advanced algorithms provide improved modeling and resolution capabilities for highly repetitive genomic sequences.

Graph-Based approaches

In graph-based sequence assembly models, the sequence reads and their inferred connections are represented as vertices and edges in a graph. By traversing the graph, one can establish an order in which the reads can be assembled together. The goal of the sequence assembler is to identify a walk in the graph that optimally reconstructs the original genome while avoiding misassemblies caused by repeats. This approach aims to find the most accurate path that captures the true sequence while minimizing errors resulting from repetitive regions.

Overlap, Layout, and Consensus (OLC) graphs

In a basic graph-based model, each sequence read is treated as a vertex in the graph. If two vertices represent reads that overlap, they are connected by an edge. Alternatively, other formulations employ two vertices per read: one indicating the start of the read and the other representing its end. These vertices are linked by an edge that carries the sequence of the read (Myers, 2005). In this representation, overlaps are represented by edges connecting terminal vertices of different reads.

Regardless of the specific graph representation used, the assembly process typically consists of three main stages. First, the assembler identifies pairs of reads that overlap. Second, the graph is constructed, and an appropriate ordering and orientation (layout) of the reads are determined. Finally, a consensus sequence is computed based on the ordered and oriented reads (Simpson and Pop, 2015). The resulting set of consensus sequences is then output by the assembler as the sequence contigs. Assemblers that follow this approach are known as OLC assemblers, named after the three main stages of assembly: overlap, layout, and consensus. The overlap stage often demands substantial computational resources. In a naive approach, one could perform pairwise comparisons of all reads using dynamic programming to determine if each pair exhibits a substantial overlap (Rizzi et al., 2019).

De Bruijn graphs

In the de Bruijn graph-based assembly of whole-genome sequence data, each read is divided into a series of overlapping k-mers. These distinct k-mers are represented as vertices in the graph, and k-mers originating from neighboring positions in a read are connected by edges. The goal of the assembly problem is to find a path in the graph that visits each edge exactly once, known as an Eulerian path problem. However, in practice, sequencing errors and sampling biases introduce complexities to the graph, making it unlikely to obtain a complete Eulerian tour through the entire graph (Pevzner and Tang, 2001; Pevzner et al., 2001; Zerbino and Birney, 2008) Even if such a tour were found, it would not accurately represent the genome sequence due to the presence of repetitive regions. In fact, there can be an exponential number of possible Eulerian traversals of the graph, with only one being correct (Pevzner et al., 2001). Consequently, most assemblers strive to construct contigs that encompass the unambiguous and non-branching regions of the graph.

The de Bruijn graph approach offers a significant computational advantage compared to overlap-based assembly strategies. It eliminates the need to find overlaps between pairs of reads and thus avoids the computationally expensive dynamic programming procedures required for overlap identification. Instead, the graph structure implicitly represents the overlaps between reads. Constructing the graph can be done in two simple steps: first, the set of k-mers is extracted from the reads and added as vertices in the graph, and second, adjacent k-mers are extracted from the reads and added as edges (Rizzi et al., 2019).

To assemble large genomes using high-throughput sequencing data, further algorithmic development is necessary due to the substantial memory requirements of de Bruijn assembly. Since there is approximately one k-mer for each base in a genome, the de Bruijn graph for mammalian genomes can contain billions of vertices. The presence of sequencing errors exacerbates this issue, as each error can

introduce up to k erroneous k-mers, resulting in the expansion of the graph with additional vertices and edges. Consequently, efficiently representing the de Bruijn graph in minimal memory became a critical challenge in the field (Simpson and Pop, 2015).

String graphs

The de Bruijn graph exhibits an elegant feature where repetitive regions are condensed: all occurrences of a repeat are represented as a single segment in the graph with multiple entry and exit points. This allows for a concise representation of the genome's structure. In 2005, Myers (2005) discovered a similar property that could be achieved in overlap-based assembly methods by implementing two transformations on an overlap graph.

The first transformation involves removing contained reads, which are reads that are substrings of other reads. The second transformation involves eliminating transitive edges from the graph. The resulting graph, known as a string graph, shares many characteristics with the de Bruijn graph but without the need to break reads into k-mers. For instance, the Edena assembler applied the string graph approach to early short-read sequencing data (Hernandez et al., 2008). Further advancements in the field focused on developing efficient algorithms for constructing the string graph using the FM index (Simpson and Durbin, 2010), leading to memory-efficient assemblers for large genomes (Simpson and Durbin, 2010; Li, 2012).