The Study of Plant Genome Evolution by Means of Phylogenomics

Zhen Li

Promoter: Prof. Dr. Yves Van de Peer

Ghent University
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
VIB Center for Plant Systems Biology

Dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Science: Biochemistry and Biotechnology

Academic year: 2017 – 2018

Examination committee

Prof. Dr. Geert De Jaeger (chair)
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
Ghent University

Prof. Dr. Yves Van de Peer (promoter)
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
Ghent University

Prof. Dr. Ir. Steven Maere
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
Ghent University

Dr. Rolf Lohaus
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
Ghent University

Prof. Dr. Koen Geuten
Faculty of Science
Department of Biology
KU Leuven

Prof. Dr. M. Eric Schranz
Department of Plant Sciences
Biosystematics Group
Wageningen University

Acknowledgements

This thesis would not have been possible in its current form without the help and encouragement from many people, to whom my acknowledgements should apply. My deepest appreciation first applies to my promoter, Yves, who offered me the great opportunity to do my Ph.D. in Ghent five years ago. Thanks for the patient guidance and inspiring suggestions on the work, and also for the complete support for my career. My next appreciation goes to Lieven, Riet, and Rolf, who guided me throughout the work from scratch. Thanks for showing me how science works on the daily basis and how to go through the 'bad' times and embrace the "Eureka" moments. When I was writing the thesis, Rolf also gave me active encouragement and constant understanding, which relieved the stress of writing. Then I would like to express my most profound gratitude to my jury members for their time on reading the thesis and giving comments that helped improve the work.

My heartfelt thanks then go to my dear colleagues, Yao-Cheng, Stephane, Thomas, Shu-Min, Sri, and Phuong for their willingness of help in many ways. I also cherish all the good times and vivid memories shared with the biocomp group from the unforgettable brainstorms to the fun activities once in a while. They are indeed precious memories in the Ph.D. life of mine. I would also like to appreciate my dear friends, who live in Ghent, Xinyang, Meng, and Haolin, who live on the internet, Shi, Qin, Chao, and Yu-Fei, and who already moved from Ghent to the internet, Qian, Diya, Wei, Xiaohuan, Zhubing, and Hao. Thank you for listening to me, for accompanying me, and for sharing your colorful life with me. We are not all in the same time zones but thank you for always being there. In the end, I would like to dedicate my wordless thanks to my family – my parents and my elder sister – for their silent devotion and endless love that support me to go far.

Scope

Phylogenomics refers to an approach to integrating phylogenetic analysis and genome analysis. It is a recently developed area in genomics and has been broadly used to infer evolutionary relationships of genes and species, and to investigate evolutionary patterns and molecular adaptations. This thesis describes three studies in detail, which apply phylogenomic and other genomic approaches to understand the genome evolution of flowering plants and seed plants. It covers three contemporary topics in plant evolutionary biology. The first study in Chapter 2 explores a controversial question in the phylogeny of seed plants, that is the phylogenetic position of the gnetophyte clade, by developing and analyzing a taxonomically broad set of single-copy genes as phylogenetic markers. The second study in Chapter 3 investigates the process and patterns of the duplicated gene retention after gene and genome duplications in flowering plants by characterizing a large set of gene families that exist across angiosperms. The third study in Chapter 4 reports a newly sequenced genome of an early diverging orchid species, *Apostasia shenzhenica*. Using the sequenced orchid genomes and transcriptomes from all the five sub-families of Orchidaceae, the study infers the history of whole genome duplication(s) in extant orchids based on both comparative genomics and phylogenomics.

Summary

Phylogenomics is the application of phylogenetics in an era where advances in sequencing technology have enabled an unprecedented opportunity to obtain genomic data for most organisms on earth. It uses phylogenetic principles to make sense of the fast-accumulated genomic data by illustrating evolutionary relationships between genes, genomes, and species. After the release of the first plant genome of *Arabidopsis thaliana*, many plant genomes have been sequenced, let alone the ever-increasing number of plant transcriptomes. The accumulated sequence data allow us to understand the process of adaptation and diversification of plants by comparing genes and genomes from multiple species. Such comparisons require taken into account the evolutionary history of extant genomes, to which phylogenomics holds the promise to bring evolutionary insights, including phylogenetic relationships, evolutionary events, and their occurrence on the tree of life.

This thesis firstly describes how phylogenomics can help to resolve the phylogeny of seed plants, which remains contested, especially within the gymnosperms. In Chapter 2, using the available genomes and transcriptomes, we identified single-copy genes in a broad collection of seed plants and inferred the phylogenetic relationships between major seed plant taxa. This study aims to provide an extended phylogenetic toolkit for seed plants, assessing its ability for resolving seed plant phylogeny, and discussing potential factors affecting phylogenetic inference. In general, our phylogenomic analyses demonstrate that single-copy genes can uncover both recent and deep divergences of seed plant phylogeny.

The thesis then discusses how phylogenomics was used to identify gene duplications through the evolution of angiosperms and to establish a comprehensive overview of the ability of genes to be retained following gene duplications in Chapter 3. As an important mechanism, gene duplication underpins the increased novelty of plant genomes. After gene duplication, some duplicates tend to be retained, but some others will quickly get lost. This is a non-random process according to previous studies, which mostly have focused on individual species and have overlooked the influence of genomic context and the time of duplications. In the study, we carried out a phylogenomic analysis to identify duplication events and their timing during the evolution of angiosperms. The study focuses on gene families that are shared between 37 angiosperms in order to investigate the duplicate retention patterns after gene and genome duplications. We show a strikingly consistent pattern, with duplicates being either primarily lost or retained in particular gene families across all species. Such distinction between the two classes of gene families is also correlated with their functional roles. Also, the dosage-balance hypothesis may explain the extended periods in retaining duplicates after whole-genome duplication in the third class of gene families.

The ability to identify gene duplications allows phylogenomics to be readily adopted as an approach to discover whole-genome duplication. In Chapter 4, the thesis shows how to apply a phylogenomic approach in complement with other methods to identifying whole-genome duplication in the lineage of orchid (Orchidaceae). Orchids possess an extraordinary diversity among angiosperms. Previously sequenced genomes of orchids, *i.e.*, *Phalaenopsis equestris* and *Dendrobium catenatum*, suggest a whole-genome duplication occurred before their divergence. The genome of *Apostasia shenzhenica*, a species in the earliest branched lineage of extant orchids, queries the timing of the identified whole-genome duplication in Orchidaceae. Further on, we used age-based, collinear-based, and phylogenomic-based methods, to identify and circumscribe whole-genome duplication based on the available orchid genomes and transcriptomes covering all five subfamilies of Orchidaceae. We show that the previously identified whole-genome duplication is shared by all extant orchids and it occurred shortly before the divergence of Orchidaceae, suggesting a potential correlation between whole-genome duplication and orchid diversification.

Samenvatting

Phylogenomics is het toepassen van fylogenie op genomische data in een tijdperk waar, door de vooruitgang in de technologie om DNA te sequeneren, het mogelijk is geworden om bijna voor alle organismen ter wereld genomische data te verkrijgen. Het maakt daarvoor gebruik van fylogenetische principes om duidelijkheid te scheppen in de snel toenemende hoeveelheid genomische data, voornamelijk door het bepalen van evolutionaire relaties tussen genen, genomen en soorten. Nadat voor het eerst de genoom sequentie van een plant, *Arabidopsis thaliana*, was bepaald is dit daaropvolgend ook gedaan voor talrijke andere planten en dan spreken we nog niet over het aantal transcriptomen dat al bepaald is geweest. De zo verzamelde sequentie data stellen ons in staat om het proces van aanpassing en diversificatie van planten te begrijpen door middel van het met elkaar vergelijken van genen en genomen van meerdere soorten. Dergelijke vergelijkingen vereisen echter het in acht nemen van de evolutionaire voorgeschiedenis van bestaande genomen, iets waar fylogenomics belangrijke inzichten toe kan verschaffen, inclusief fylogenetische relaties en belangrijke evolutionaire gebeurtenissen alsook het positioneren op de boom des levens.

Deze thesis start dan ook met het beschrijven van hoe fylogenomics kan helpen om de fylogenie van de zaadplanten op te stellen, iets wat nog steeds bediscuteerd wordt, en dan voornamelijk voor de naaktzadigen. In hoofdstuk twee wordt dan besproken hoe we, gebruikmakend van de beschikbare genomen en transcriptomen, op zoek kunnen gaan naar single-copy genen in een brede collectie van zaadplanten en hoe we daaruit de fylogenetische relaties tussen de belangrijkste zaadplanten taxa kunnen afleiden. Deze analyse beoogt een uitgebreide fylogenetische toolkit aan te bieden voor zaadplanten, alsook te bestuderen hoe geschikt deze is om de fylogenie van de zaadplanten te bepalen en daarenboven de mogelijke factoren die fylogenetische interferentie beïnvloeden te bespreken. In het algemeen beschouwd kunnen we stellen dat onze fylogenetische analyse van single copy genen kan gebruikt worden om zowel recente als oudere divergentie in de zaadplanten bloot te leggen.

Vervolgens wordt besproken hoe phylogenomics kan gebruikt worden om gen-duplicaties te bepalen die gebeurd zijn doorheen de evolutie van de angiospermen alsook om inzicht te verkrijgen hoe gedupliceerde genen behouden kunnen blijven volgend op een duplicatie event. Dit is een belangrijk mechanisme om nieuwigheden te introduceren in plantengenomen. Na duplicatie zal voor de meerderheid van de gedupliceerde genen een kopij verloren gaan maar sommige zullen echter alle kopijen behouden. Vroegere studies, die echter enkel rekening hielden met individuele species, hebben uitgewezen dat dit geen willekeurig proces is maar zij hielden bovendien geen rekening met de genomische context noch met het tijdperk waarin de duplicatie zich voordeed. Hoofdstuk drie bespreekt een fylogenomische analyse uitgevoerd om dit soort duplicatie gebeurtenissen en hun timing te identificeren in de evolutie van de angiospermen. Hiervoor werden gen families geanalyseerd

die gedeeld worden tussen 37 verschillende angiospermen om zo de verschillende retentie van gedupliceerde genen na gen en genoom duplicatie te bekijken. Hieruit konden we een opvallend en consistent patroon afleiden waarin kopijen van genen voornamelijk behouden blijven of verloren gaan afhankelijk van de gen familie waartoe ze behoren wat daarenboven ook nog gecorreleerd is met de biologische functie van de gen familie. Er is echter nog een derde mogelijkheid waarbij duplicaten behouden blijven omwille van het dosis-balance effect aanwezig in bepaalde gen families.

Het gemak waarmee fylogenomics kan gebruikt worden om gen duplicaten op te sporen maakt die een toegankelijke techniek om ook volledige genoom duplicaties te analyseren. Vervolgens wordt in hoofdstuk vier besproken hoe fylogenetische analyses complementair zijn aan andere methoden om genoom duplicaties te identificeren in de orchideeën die een opmerkelijke diversiteit vertonen vergeleken met andere angiospermen. Eerder gesequeneerde genomen van orchideeën, namelijk *Phalaenopsis equestris* en *Dendrobium catenatum*, suggereerden dat er zich een genoom duplicatie heeft plaats gevonden voorafgaande aan hun divergentie. Analyse van het genoom van *Apostasia shenzhenica* stelt deze timing echter in vraag. Verder in datzelfde hoofdstuk gebruiken we op leeftijd-, collineariteit- en fylogenomics gebaseerde methoden om een genoom duplicatie te identificeren en te beschrijven op basis van de beschikbare orchidee-genomen en transcriptomen die alle vijf de subfamilies van Orchidaceae coveren. We toonden hiermee aan dat de eerder beschreven genoom duplicatie in de orchideeën gedeeld blijken te zijn door alle bestaande orchideeën en er aldus een mogelijke correlatie bestaat tussen de genoomduplicatie en de verbluffende diversiteit van orchideeën zoals we die nu kennen.

Abbreviations

AIC	Akaike Information Criterion
AICc	corrected Akaike Information Criterion
APG	Angiosperm Phylogeny Group
AU test	Approximate Unbiased (AU) tests
BD	birth-death
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
BP	bootstrap percentage
CI	confidence interval
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GC%	guanine-cytosine (GC) contents
GMM	Gaussian Mixture Modeling
GO	Gene Ontology
GOslim	Gene Ontology slim
Hi-C	chromosome conformation capture
HPD	Highest Posterior Density
IC	Internode Confidence
ICA	Internode Confidence All
K_S	number of synonymous substitutions per synonymous site
K_N	number of non-synonymous substitutions per non-synonymous site
KEGG	Kyoto Encyclopedia of Genes and Genomes
K-Pg	Cretaceous-Paleogene
LBA	long branch attraction
MCL	Markov cluster algorithm
ML	maximum likelihood
MP	maximum parsimony
MPI	Message Passing Interface
MSA	multiple sequence alignment
mya	million years ago
NGS	next generation sequencing
NJ	neighbour-joining
ORF	open reading frame
SD	standard deviation
SSD	small-scale duplication
SCP	single-copy percentage
WGD	whole-genome duplication

Table of contents

Chapter 1

Introduction

"If superior creatures from spaces ever visit earth, the first question they will ask, in order to assess the level of our civilization, is: 'Have they discovered evolution yet?'"

--Richard Dawkins

"The Selfish Gene"

1.1 Our view of life – a brief history of inferring the tree of life

1.1.1 Ideas before Charles Darwin

As long ago as the fourth century BC, the Greek philosopher Aristotle classified different groups of animals based on phenotypic characters in common sense, like tissues (blood or bloodless), habitats (land-living or water-living), reproduction forms (oviparous or viviparous), and so forth. After distinguishing different groups, he further conceived and assigned four attributes of the four basic elements – hot of fire, dry of air, moister of water, and cold of earth – to the characters, hence sorting them into an order of importance. For example, he thought hot is over cold and moister is over dry, so blood, considered to be both hot and moister, is of higher importance than other characters to distinguish animal groups. Animals with blood also rank higher than bloodless ones that seem to have less heat[1]. In such a way, a scale of importance was superimposed to the characters leading to a rank of potential "souls" of different animals. Indeed, in his book *On the Soul*, he clearly ranked human over animals and animals over plants based on their "souls", as plants have a vegetative "soul" to reproduce and grow, animals have sensitive "soul" to move and feel, and humans have rational "soul" to think and reflect[2].

Aristotle did not really propose a systematic classification of organisms[3], and his approach to classifying animals into groups actually only illustrates how he organized his physiological and ecological observations of animals[1]. However, his non-religious concept of ranking inspired scholastic philosophers in the Middle Ages. Through Arab's translation of Aristotle's works, the scholastic philosophers rediscovered his system and finally developed the great chain of being, which is often referred to as *scala naturae*. Inheriting the work from Aristotle but neglecting his practical and pragmatic spirit in *The History of Animals*[3], these theological philosophers deduced a linear hierarchical structure for major taxa of creatures, starting from God and progressing down to angelic beings, humanity, animals, plants, and non-living minerals. The logic behind the hierarchical structure in the *scala naturae* is the loss of "degree of perfection". For each link under the highest perfection, it loses an essential attribute of perfection step by step. For example, plants do not have motion and appetite compared with animals, while minerals have no life in contrast to plants[4].

In the 18th century, Carl Linnaeus, the father of systematics[3], first resolved a systematic classification for the subgroups in the links of the great chain of being below humanity. Before him, the classification of subgroups in each link was arbitrarily based on the will of people. He used binomial nomenclature to name species and introduced a hierarchical system to place the subgroups of animals, plants, and minerals. He classified animals into mammals (Quadrupedia), followed by avian (Aves), amphibian (Amphibia), fish (Pisces), insect (Insecta), and worms (Vermes). For plants, he used the structures of flowers, especially the number of stamens, and classified plants into 23 flowering groups and an extra flowerless group[5]. Linnaeus is probably the first naturalist who proposed that new species could be generated through hybridization of existing species, after thoroughly studying the floral structure of a *Linaria* species[6]. Even having such a radical hypothesis ahead his age, Linnaeus vehemently denied the idea that species could be modified time by time and the possibility that hybridization could apply to clades higher than genus. Linnaeus and his contemporaries also noticed that nature could barely fit into a linear arrangement as described in the *scala*

naturae[7]. He found that if there is any linear arrangement between species in plants and in animals, it barely connects "the most perfect Plants with Animals that are said to be the most imperfect", as indicated by the *scala naturae*, but joins "imperfect Animals and imperfect Plants"[7,8]. Augustin Augier, a French teacher and botanist, used an approach to associate and join two more perfect families of plants as two extremes through a less perfect family in the middle. The family in the middle could be further attached to another family that is immediately lower in a linear series[9]. He aimed at finding the single series imposed in the *scala naturae*, but ended with a tree-like structure of plants and "succeeded at least in making them all join by their bases" (Figure 1-1A)[9]. An American geologist, Edward Hitchcock, illustrated another tree-like view of changes of life in a textbook, *Elementary Geology*, written by himself, as a "paleontological chart"[10]. When arranging fossils from different geological strata, he uncovered a tree-like structure for both plants and animals in a geological time scale (Figure 1-1C). However, Hitchcock believed that the changes, which formed such a tree-like pattern, was from the deity rather than an inherent power of organisms[10].

It was Jean-Baptiste Lamarck, born nearly half a century before Hitchcock, who first admitted that species could change from one form to another. He conceived that the modifications of species are the results of their inherent power on pursuing complexity, as reflected in the ascending series through the *scala naturae*. In Lamarck's view, all the existing species first originated from inanimate beings independently and spontaneously, then they have been evolved from simple to complex because of what he called "*Le pouvoir de la vie*", *i.e.*, "complexifying force"[11]. He used a tree-like structure to illustrate his theory on the changes of different animal groups, in which dotted branches reflect the paths that how complex species evolved from simple ones (Figure 1-1B). The reason that we can observe all forms of species is, as Lamarck argued, that species originated from inanimate beings at different times, so the early born species could have a longer time to evolve complex forms than do species originated more recently[11].

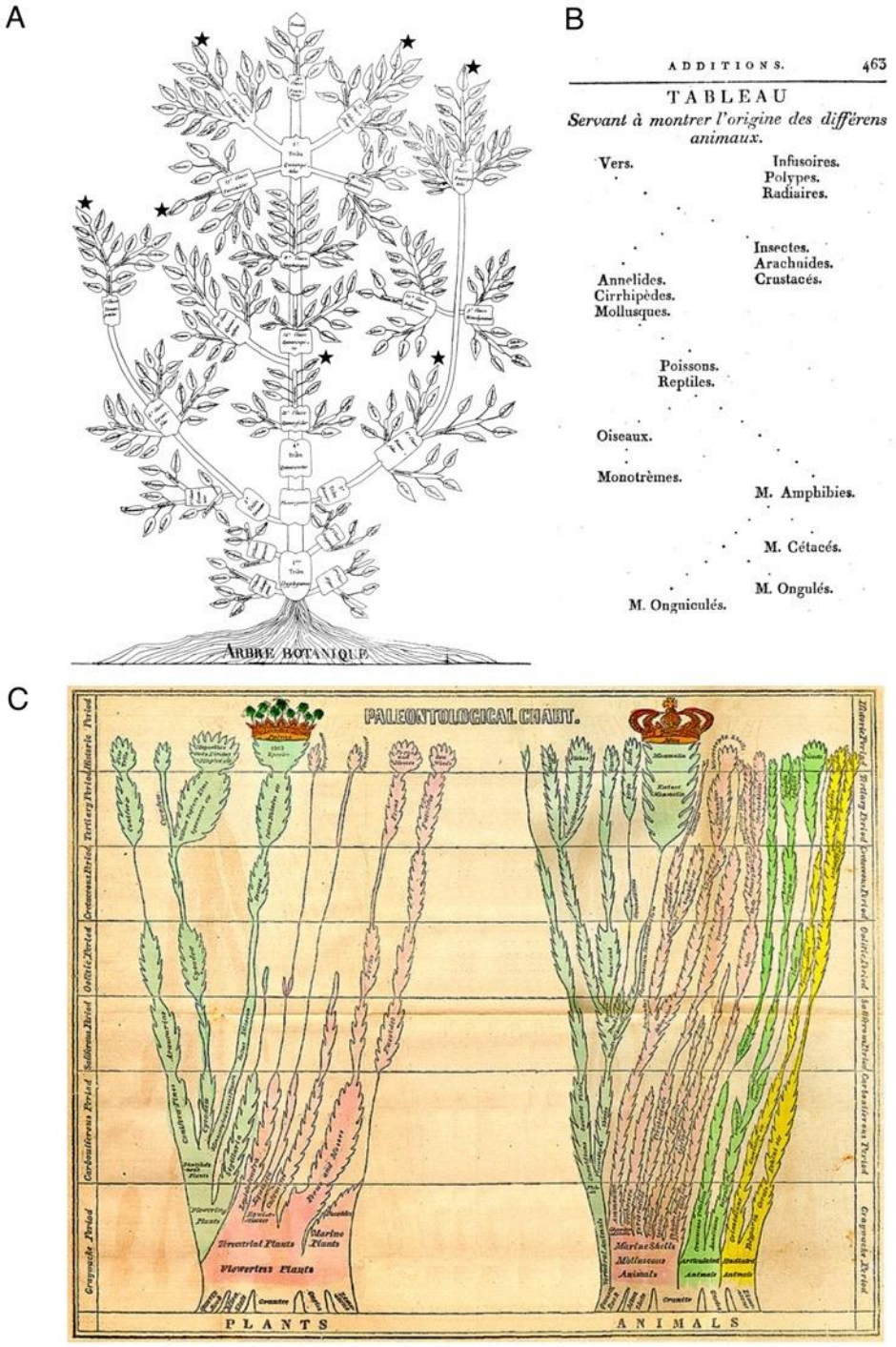


Figure 1-1 Tree-like structures of life before Charles Darwin.
(A) The tree-like diagram drawn by Augustin Augier based on his belief on the natural relationships of the plant kingdom (*Arbre Botanique* (1801); obtained via Stevens[9]); (B) The tree-like diagram showing the changes of animals from Jean-Baptiste Lamarck (*Philosophic Zoologique* (1809); obtained via Google books); and (C) The tree-like diagrams of fossils and living Plants (left) and Animals (right) with geological time (vertical axis) from Edward Hitchcock (*Elementary Geology* (1840); obtained via Wikipedia Commons).

1.1.2 On the origin of Darwin's tree of life

Apparently, Charles Darwin was not the first one who proposed to use a tree-like structure to symbolize relationships of species[7], but he has attributed evolutionary meanings to nodes, twigs, and branches on the tree-like structure with much more parsimony explanations of common ancestry and "descent with modification"[12]. In his book *On the Origin of Species* published in 1859, Darwin used the only diagram in the book to illustrate an imaginary tree with 11 species (A to L) in an assumed genus (Figure 1-2). In the diagram, horizontal lines denote a time scale of evolution regarding long generation time, which could be a thousand, ten thousand or even more generations. Taking species A as an example, a little bush formed by growing twigs represents different variants of species A. During a period of the 1st unit of the long generation time (I), all the variants of species A try to outcompete the surrounding twigs, eventually leaving two descendants, a^1 and m^1, at the time I. The two descendants also have their own variants (bushes), and each variant again tries to outperform their surrounding relatives (twigs). Such a process leads to the birth of a^2 from a^1 as well as the birth of m^2 and s^2 from m^1, with the extinction of other variants. Extinctions could happen to the variants (twigs) and also to branches composed of several continuous twigs, such as the branches leading to s^2, i^3, k^8, and l^8. After progressing transversely, at the 10th unit of the long generation time (X), three descendants, a^{10}, f^{10} and m^{10} become three modified descendants derived from species A. The similar progress can apply to species I which generates two modified descendants, w^{10} and z^{10}, at the 10th unit of the long generation time (X). Because each descendant over time is from a variant of its ancestor, they all have some modified traits in comparing with their ancestors. If the total generation time is long enough, these five modified descendants may accumulate enough characteristics that could be used to distinguish them from each other and from Species A or I, so new species could be generated in such a process. Noticing that it is possible for some species to have descendants without modified characteristics. For example, E^{10} and F^{10} keep the characters from their ancestor E and F, respectively. Seven of the 11 ancestral species (B, C, D, G, H, K, and L) have no descendants after such a long time because they are all extinct at some time points during the hypothetical process of evolution. If the same process continues, 14 modified descendants and one descendant with ancestral characters would be found at the time XIV[12].

Differences between the tree from Darwin and the "trees" before Darwin are shown in three aspects. First, internal nodes on a Darwin's tree represent the common ancestry of two branches but not an existing group with some ancestral characters, so it is not a "less perfect/complex" group as suggested in Augier's tree or Lamarck's tree (Figure 1-1A and B). This is a significant difference on the interpretation of the tree-like structure for life. Darwin indeed considered common ancestry as the primary evidence for evolution, because if all the species have a common ancestor, evolution must have occurred and hence derived all the different species as we see today[11].

Second, living groups with ancestral characters are viewed as the descendants of species that diverge "from a fork low down in a tree"[12] rather than as basal forms of species as links in the discontinuous creation of new species by metaphysical forces[7]. Darwin emphasized on the continuity of genealogy and attributed the existence of species to the survival of fatal competitions after ramifying out. This difference further suggests that extant species in Darwin's tree all have the same amount of time to evolve after their divergence from a common ancestor, so no species lineage is superior to another. However, in Lamarck's tree,

this is not the case, as all the species originated independently and followed the path in the *scala naturae*, in which each tip shows an advanced form of life, and each node is just an intermediate form[11].

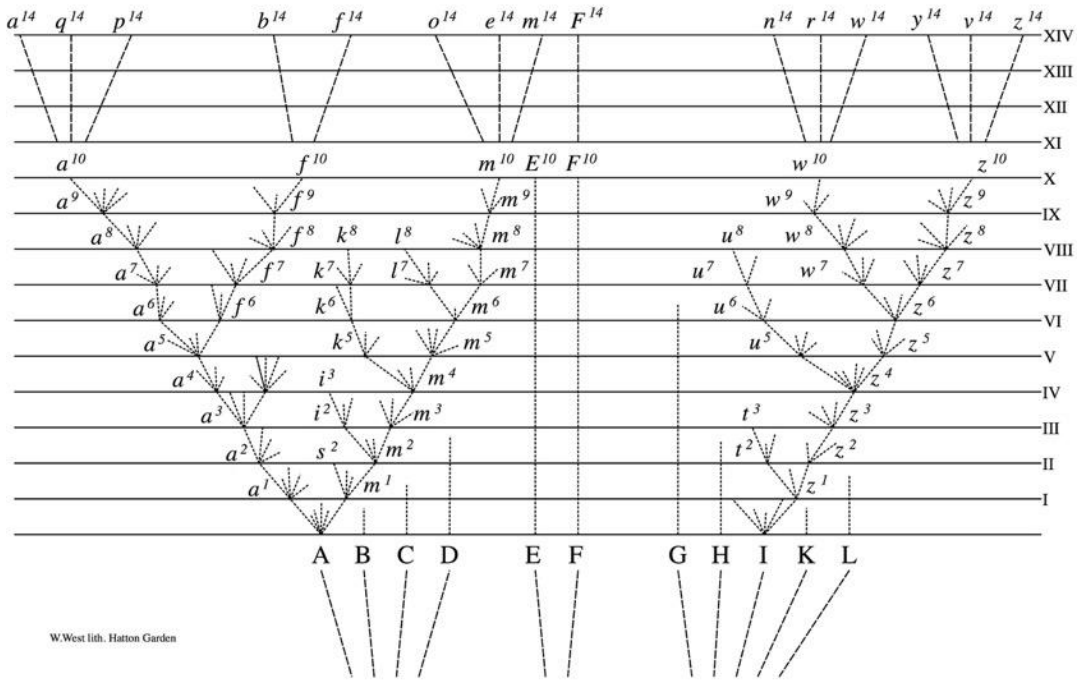


Figure 1-2 Charles Darwin's tree from *On the Origin of Species* (1859).
It is an imaginary tree and the only diagram in Darwin's famous book. See text for details.

Finally, species found in fossils could be placed on Darwin's tree in two different ways. In most cases, fossils are considered as the evolutionary dead end, so they form sister branches to the branches leading to extant species. They compose the terminated branches here and there in a Darwin's tree to represent that only a few branches now have modified descendants survived from the "great battle of life"[12]. In few cases, fossils could be placed on the internal nodes or branches leading to extant species to show the exact ancestor – descendant relationships. But it is really difficult to be assured that a fossil is an ancestor to a specific taxon because of the branching nature of trees and the incomplete records of fossils buried underground. In general, Darwin cherished that the metaphor of "the great Tree of Life" "largely speaks the truth", because it "fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications"[12].

1.1.3 Building the tree of life

Although heated debates on the force of evolution have continued since the publication of *On the Origin of Species*, using tree structures to represent life has been widely accepted along with the concept of common ancestry[11]. Ernst Haeckel was a pioneer, if he was not the first, who inspired by the conceptual diagram of the tree from Darwin (Figure 1-2), and further implemented the idea of the tree of life with actual species on the earth[13]. To use embryonic characters in ontogeny to build the tree of life, he applied his so-called biogenetic law, in which he claimed that the embryonic development of each species precisely recapitulates its evolutionary history. He inferred the ancestors of existing species and practically built several

trees of life[13,14]. It is not clear why, but Haeckel's trees of life reflect not only Darwin's concept on the common ancestry and ramifying branches but also the ladder-like structure usually found in the *scala naturae* (Figure 1-3)[11,13]. In his trees, each of the main groups forms a strong branch standing for the common ancestor, which further connects to a single common ancestor at the root. However, those trees also illustrate different groups in an order from primitive to complex, with species with "ancestral" characters closer to the root while species with "advanced" features closer to the top of the tree. Such an ordered tree of life seems to partially, if not intentionally, continue the *scala naturae*[11]. After one and a half century, Haeckel's biogenetics law and trees rarely appear in biological literature, but he left a heritage for us – the word *phylogenie*. Although he used the term to denote a series of morphological stages that passed through the evolutionary history of a given species instead of the trees he depicted, the meaning of the concept has been shifted since it was coined[7,15]. The current usage of the term phylogeny implies the evolutionary history of organism lineages through time represented by the genealogical relationships of organisms[7].

Inspired by Haeckel who illustrated numerous species in a format of branching trees, taxonomists have made a great of effort to reconcile the classification system from Linnaeus with the tree from Darwin, hence leading to a discipline called evolutionary systematics, or evolutionary taxonomy, or Darwinian classification. They have converted the static concept of species from Linnaeus into a dynamic concept of species by considering both the common ancestry and the ancestor – descendant relationship. Therefore, a phylogeny resulted from evolutionary systematics often includes both species formed through cladogenesis, *i.e.*, speciation splitting from a common ancestor, and anagenesis, *i.e.*, gradual changes of a descendant without branching the evolutionary line from its ancestor. The latter form is a significant hallmark of evolutionary systematics[16]. The system uses similarities to sort species into various taxonomic categories, which usually result in a species phylogeny because similar species can mostly represent the modified descendants of a common ancestor[16]. However, in high taxonomic groups with diverse species, similarity could be deceiving because of homoplasy due to convergent evolution. To resolve the ambiguity in such a methodology, fossils are critical evidence to uncover the real genealogy in a lineage by testing whether the similarity among diverse species is from homologs descended from a common ancestor. Fossil evidence is also essential for inferring the direct ancestor – descendant relationship, even though such fossils are scarce. In the early 20th century, the approach was integrated with modern evolutionary synthesis which reconciled Darwin's evolution and Mendel's genetics. The integration provides theoretical background and explanations on how a population of species could give rise to a population of new species through evolution[17]. A tree reconstructed by such an approach is often referred to as a Haeckelian dendrogram or Darwinian dendrogram in remembrance of the one who first coined the term phylogeny and the one who introduced the evolutionary theory, respectively[16].

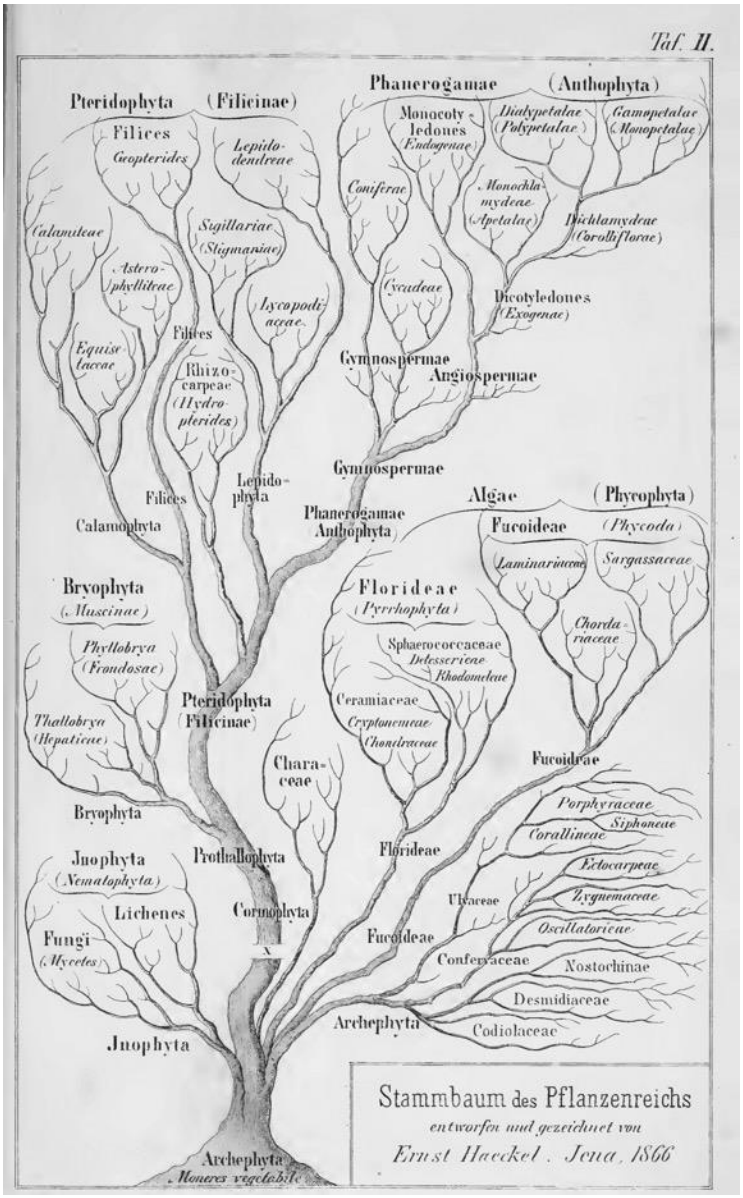


Figure 1-3 The tree of plants by Ernst Haeckel.
The figure is from *General Morphology of Organisms* (1866), obtained via Hossfeld *et al.*[14]. See text for details.

In contrast to baring both anagenesis and cladogenesis in mind, another group of evolutionary biologists keeps their eyes on the bifurcation pattern resulted from the branching nature of evolution as described by Darwin (Figure 1-2). Therefore, they described the relationships between either extant or extinct species in the way of nested sister groups. Any sister group shares a hypothetical most recent common ancestor that splits into the two clades, as a way to illustrate the common ancestry with modified descendants. To distinguish the approach from evolutionary systematics, this method is called phylogenetic systematics[16]. Willi Henning, in 1950, proposed a series of operational principles to classify species based on the relatedness, *i.e.*, the recency of common ancestry measured by acquired characters during evolution[16]. Henning has introduced four important terms: *plesiomorphy* as an ancestral trait, *apomorphy* as a derived trait, as well as *symplesiomorphy* and *synapomorphy* as shared ancestral and derived traits, respectively. A synapomorphy is a trait that only species within a specific group share and inherited from their common ancestor, while other

species do not have. A group of species originated from a common ancestor can be classified into one group, namely monophyletic group (monophyly), if and only if they have synapomorphic traits. In contrast, if a group of species originated from a common ancestor but some of them form monophyletic groups due to synapomorphic traits, the group of species excluding the species in the monophyletic groups is said to constitute a paraphyletic group (paraphyly). Sympleimorphic traits unite a paraphyletic group. A phylogenetic tree resulted from Henning's approach is often referred to as a cladogram. It is not shown the complete evolutionary process because it only partially coincides with Darwin's tree and relates species by the common ancestry[16]. However, the formalized methodology is able to create alternative evolutionary hypotheses that could be further verified by increasing evidence from morphologies and fossils. In fact, our picture of life is much clearer than the era of Darwin, because Henning's principles have been prevalently adopted by molecular phylogenetics[18] and made a pronounced shift from evolutionary systematics to phylogenetic systematics. Recently, various sequencing technologies have alleviated the pain in collecting molecular data and enables a tremendous increase in applications of molecular phylogenetics through integrating evolutionary analysis and genomic analysis, giving rise to phylogenomics. Although there are continuing controversies between evolutionary systematics and phylogenetic systematics, no approach seems to be superior to the other. To resolve the conflicts, considerable efforts have been put in integrating the two systems to build the tree of life[19-21]. In the rest of the thesis, phylogeny, phylogenetic, and phylogenetics only bear their meanings in the context of phylogenetic systematics.

Nowadays, evolutionary biologists could, optimistically, claim that it is possible to arrange thousands upon thousands of species on a single tree of life[22]. By integrating the accumulated data in phylogenetics, the TimeTree of Life (TTOL) probably generates the largest tree of life calibrated to divergence time[23]. Until 2017, the TTOL already has over 97,000 species from more than 3,000 phylogenetic studies and illustrates the tree of life with geological time, events, and climates, as an example illustrating orders of angiosperms shown in Figure 1-4[24].

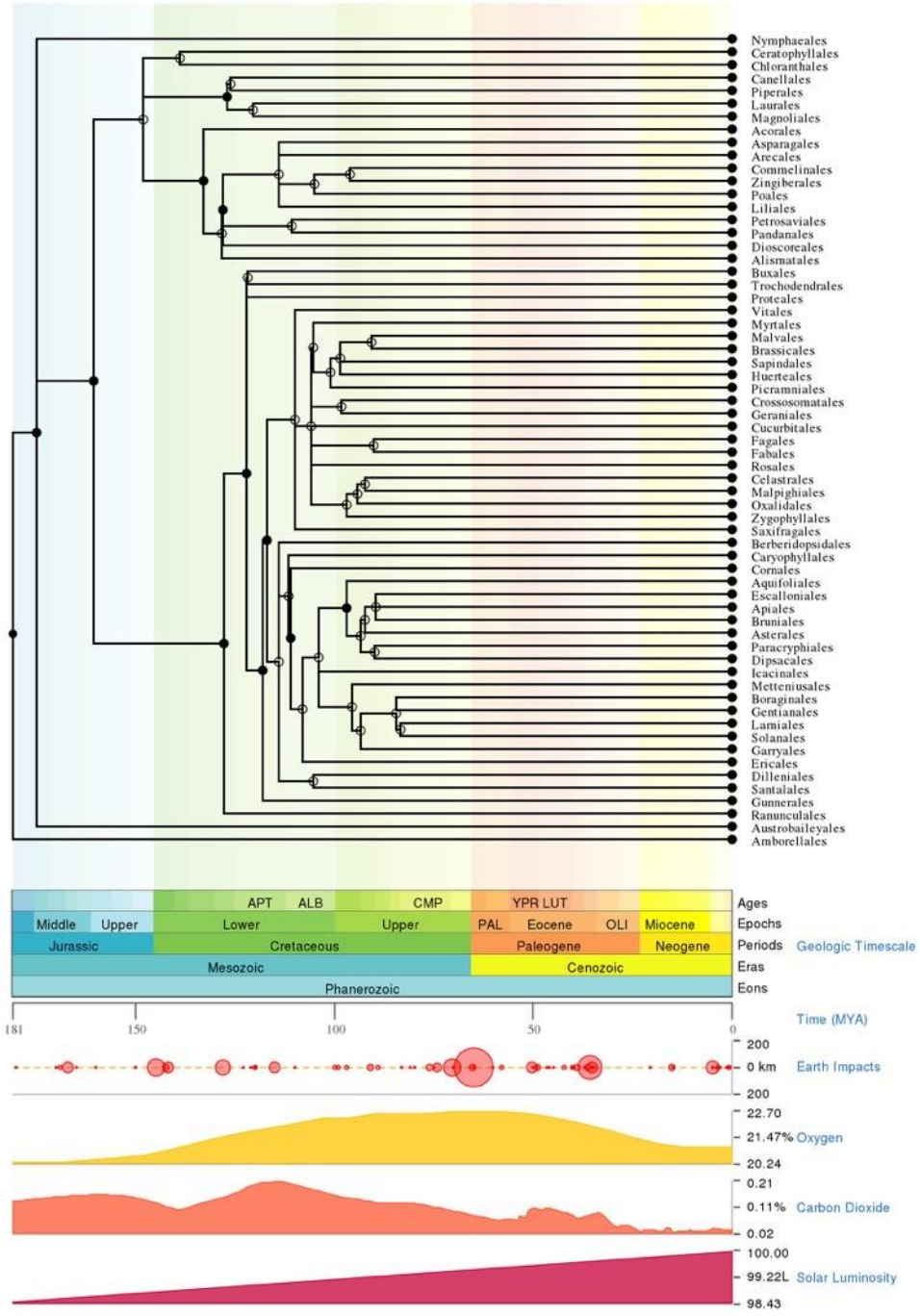


Figure 1-4 The phylogeny of flowering plants in the TimeTree of Life (TTOL).
Figure obtained by searching "angiosperms" on TTOL (www.timetree.org). Orders of extant angiosperms are illustrated in the phylogenetic tree, along with dynamics of earth, oxygen level, carbon dioxide level, and solar luminosity over a geological timescale.

1.1.4 Current views

Phylogenetic studies in the past half-century have tremendously improved our knowledge of the organization of life on the earth. First, it provides some views in depth on the major domains of the cellular life. Previously, the cellular life is divided into two domains, *i.e.*, eukaryote and prokaryote in a so-called universal tree, depending on whether a cell has a nucleus or not. Prokaryotes could be further classified as Bacteria and Archaea (Figure 1-5A).

However, two rounds of the growth of sequence data have dramatically changed our view of the universal tree. The first round occurred after the development of polymerase chain reaction and Sanger sequencing. Through sequencing ribosome RNAs and genes involved in information-processing machinery from the known cellular life at that time, an alternative picture of the universal tree shows that Archaea is, in fact, more closely related to Eukarya rather than to Bacteria, leaving "prokaryote" as a paraphyly instead of a monophyly (Figure 1-5B). Thus, it suggests that an assortment of three domains from three monophylies, *i.e.*, Bacteria, Archaea, and Eukarya, appears to be the universal tree[25]. Only after a decade, the second growth of sequence data has occurred when next generation sequencing has been broadly applied to biology. The three-domain universal tree has been further challenged by recent studies that investigate microbial diversity in unexamined environments. Not only do these studies illustrate the astonishing diversity of microbes[26], but they also suggest a two-domain universal tree with only two monophylies, *i.e.*, Bacteria and Archaea. In the two-domain universal tree, the original group of Archaea is no longer a monophyly which shares the most recent common ancestor with Eukaryota[27], because Eukaryota becomes a clade that nests within Archaea as a sister group to a recently discovered Archaea phylum Lokiarchaeota[26,28,29]. The two-domain universal tree further explains the similarity between eukaryotic cell and archaeal cell, in support of the hypothesis that archaeal host cells with mitochondrial endosymbiont gave birth to the ancestral eukaryotes[28,30]. The most recent views on the two-domain universal tree have triggered a series of debates in the areas of evolution and microbiology, so the thesis uses Bacteria, Archaea, and Eukarya as they are defined in the three-domain universal tree (Figure 1-5B). Although prokaryote is no longer a monophyletic group, it still denotes both Bacteria and Archaea in common usage.

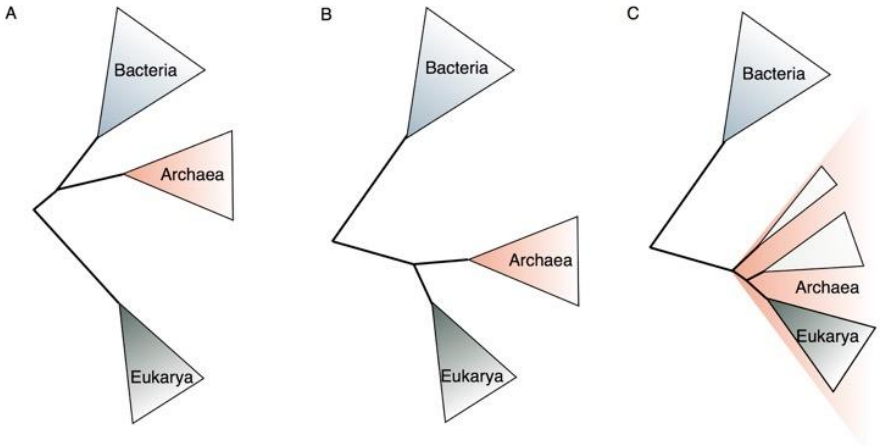


Figure 1-5 The shift of ideas on the universal tree.
Figures modified based on the figure from Pace[25]. See text for details.

Second, molecular phylogenetic studies together with morphological and paleontological studies have refined the classification of many groups of organisms. Among those, flowering plants, or angiosperms, are a group has been investigated by botanists for hundreds of years. The term 'angiosperms' was first coined by Paul Hermann in 1690 and maintained by Carl Linnaeus in his taxonomy system. Constituting approximately 300,000 species, flowering plants are the most diverse group of land plants[31]. Its diversity and sudden appearance in the fossil records led Darwin considered their evolution and diversification as an "abominable mystery"[32]. Classification of such a large group appeared to be difficult, so different systems

have been proposed based on morphological characters, which were at times subjective. In the 1990s, more and more studies based on molecular phylogenetics generated many congruent results for previous controversial groups, so an initiative of botanists has started collaborations to build a consistent classification system of angiosperms, which has resulted in the well-known Angiosperm Phylogeny Group (APG)[19]. The new system requires each classified group to be a monophyletic group but also tries to keep orders and families in the Linnaean system[19]. The major groups of angiosperms are then illustrated as a grade of isolated taxa of Amborellales, Nymphaeales, and Austrobaileyales (ANA-grade or ANITA-grade in previous) leading to the major radiation of angiosperms, mainly including a clade of all monocots, a clade of magnoliids and a large eudicot clade (Figure 1-6). The eudicot clade is composed of two large groups, Rosids mainly with Fabids and Malvids, and Asterids mainly with Campanulids and Lamiids. In 2016, by incorporating morphological and fossil evidence as well as molecular phylogenetic studies from chloroplast, mitochondrion, and ribosome DNA, the most recent APG IV has 64 orders with 416 families[20].

Figure 1-6 The phylogeny of angiosperms from APG IV.

The phylogenetic tree is modified from The Angiosperm Phylogeny Group[20]. The new classification system uses superrosids and superasterids to include additional orders in the two larger clades that are dominated by rosids and asterids, respectively.

Finally, molecular phylogenetic studies have demonstrated a potential replacement of the tree of life with an alternative metaphor, *i.e.*, the network of life. Even though it was

conceived many times independently even before the idea of the tree of life[7], the network of life has only come back in the late 1990s, explicitly after the discovery of the patchy distribution of genes among prokaryotes and incongruent phylogenetic trees for specific genes[33,34]. Prokaryotes, *i.e.*, Bacteria and Archaea, do exchange genes extensively through horizontal gene transfer via mobile genetic elements, like plasmids, bacteriophages, and transposons[35]. The observations have falsified the existence of a "solid" tree for all life forms on the earth, because genomes of the cellular organisms, especially for Bacteria and Archaea, are composed of genes from a somehow shared gene pool[34]. Therefore, a set of conserved genes could not represent the evolutionary history of all genes in different genomes, whereas a network of genetic exchanges is possibly an adequate representation. The network of life is also a potentially useful metaphor to represent the process of hybridization between species, which is not uncommon in plants and animals[36], suggesting the tree of life alone cannot describe the evolution of all the life forms on the earth. However, the network of life could not deny the tree-like structure of the genealogy of every single gene[11,34]. Besides, horizontal gene transfer can but rarely appear in multi-cellular eukaryotes[37,38], while species hybridization resulting in reproducible descendants is often limited within closely related lineages[11]. Therefore, the major clades of eukaryotes, like the ones in animals or plants, still have predominantly tree-like population histories in line with the tree of life. The thesis would focus on the evolutionary history of genes and species in flowering and seed plants, so it will discuss the effects of hybridization on the evolutionary history of genes but disregard horizontal gene transfer and the network of life in Bacteria and Archaea.

1.2 Phylogenomics – phylogenetics in the age of sequencing

1.2.1 Development of sequencing and its consequences for phylogenetics

Morphological characters were, not surprisingly, first used to infer phylogenetic relationships among different species, especially when taxonomists started combining the classification system of Linnaeus with the tree of life. Phylogenetic inference relies on characters with homologous relationship (homology), which is defined as the relationship of characters that have descended from a common ancestral character[39]. However, the number of homologs, *i.e.*, homologous characters, quickly became insufficient in morphologies to infer the phylogenies with closely related species, and even fewer of morphological characters could be used for microbes[40]. On top of that, some morphological characters are ambiguous in determining character states and in distinguishing homology and analogy. Unlike homology, which is the consequence of common ancestry, analogy is the existence of homoplasy resulted from convergent or parallel evolution. The consideration of analogy as homology would lead to incorrect phylogenetic inference as it puts characters that not share the most recent common ancestor together. Later development of molecular biology provided several molecular characters that could be used in phylogenetics, such as protein electrophoresis and DNA hybridization of homologous genes, *i.e.*, genes share the common ancestry. However, these methods were with difficulties to quantify character divergence in high resolution. The invention of DNA sequencing finally solved the above issues by supplying numerous characters with unambiguous states in the resolution of a single nucleotide.

DNA sequences have a few added advantages as homologous characters for phylogenetics comparing with morphological characters. First, each site of a DNA sequence is, in general, assumed to contribute independently to phylogenetic inference, even though these sites are physically connected in order[11]. Therefore, just a piece of DNA sequence generated by sequencing could considerably increase the number of characters used in phylogenetics. In addition, if a DNA sequence is from a protein-coding gene or a part thereof, it could be considered as a codon sequence or translated into an amino acid sequence. As the essence of evolution is different among DNA, codon, and amino acid sequences, such a simple conversion of DNA sequences leads to different types of characters for phylogenetics. The second advantage of DNA sequences is that the states of each site in a DNA sequence could be firmly determined as one of the four nucleobases, *i.e.*, adenine (A) and guanine (G) as purine bases, and thymine (T) and cytosine (C) as pyrimidine bases. Thus, there is no ambiguity in quantifying differences between DNA sequences, while unconfident measurements of morphological characters often confine their applications. Last but not least, the ubiquity of DNA, codon, or amino acid sequences in different species promise statistical models on sequence evolution can apply to the entire scope of life. Morphological characters often have a good chance not to exist in all the studied species causing issues in quantifying character differences appropriately.

With the above advantages, the invention and continuous development of DNA sequencing have tremendously accelerated molecular phylogenetics. After the release of the first human genome(s) in 2001[41-43], next generation sequencing (NGS) has revolutionized not only phylogenetics but nearly all the fields in biology[44], because of the steady decline of sequencing cost along with the increased throughput of sequencers (Figure 1-7). The low cost of sequencing in high throughput creates the possibility to generate sufficient data to cover each sample comprehensively in relatively short time and with low demand for labor[44]. In addition, NGS does not require *a priori* knowledge before sequencing. For example, it no more needs a specific sequencing primer, which is the crucial component but also limits broad applications of Sanger sequencing. Although some pilot investigations on features of sequences, such as their guanine-cytosine (GC) contents (GC%), would be appreciated before applying NGS, it is not difficult to switch to other samples or to make a decision for sequencing on the pilot studies. Therefore, we can sequence almost any samples based on scientific questions but are not limited by technical issues. The sequence samples nowadays are distributed more widely and deeply than ever, from all living species to lately extinct species, from thousands of individuals in one species to the whole ecological system[45,46].

The merits of NGS have led to an eruption of DNA sequences in recent years, offering particular promises in studying the population history of a species, the evolutionary relationships between different organisms, the genealogies of gene families, and so on so forth. Nowadays, several large sequencing projects aiming at exploring diverse species through genome and/or transcriptome sequencing have been initiated for thousands upon thousands of species (Table 1-1). A great number of DNA samples collected from unbiased surveys of environments in soil and marine are in the process of sequencing as well[26,29,46]. To analyze the fast-accumulating genomic data with an evolutionary perspective, phylogenomics was first coined in functional studies of genes based on sequence similarity[47] and further developed as molecular phylogenetics in a large-scale to reveal phylogenetic relationships among species and their evolutionary history[48].

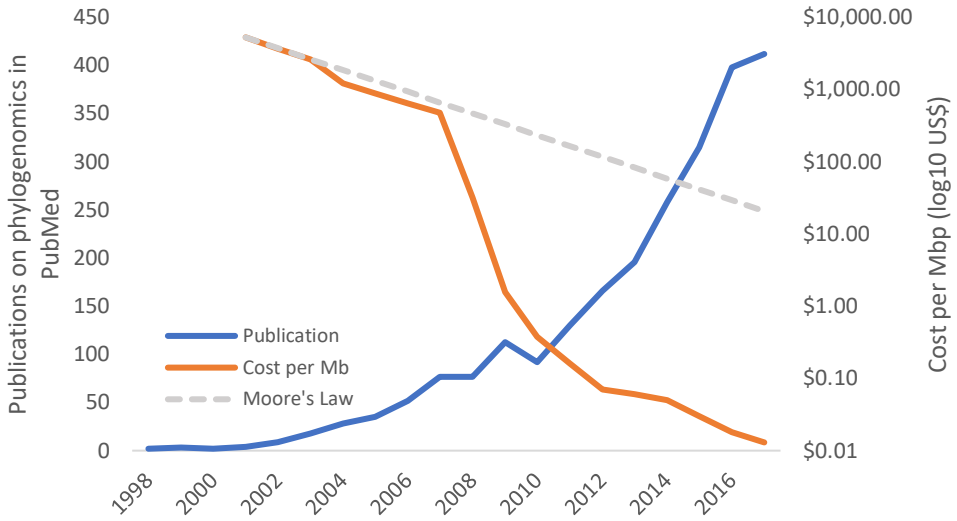


Figure 1-7 The reduction in sequencing costs and the increase of publications on phylogenomics in the past years.

The prices of sequencing are from National Human Genome Research Institute (www.genome.gov); the number of publications on phylogenomics was queried by searching 'phylogenom*' on PubMed (www.ncbi.nlm.nih.gov/pubmed).

Table 1-1 Sequencing projects exploring diverse species

Projects	Year	Goal
Genome 10K[49,50]	2009	To assemble genomes for a biospecimen collection of some 16,203 representative vertebrate species spanning evolutionary diversity across living vertebrates (ca. 60,000 species)
1000 Fungal Genomes[51]	2011	Department of Energy has embarked on a five-year project to sequence 1000 fungal genomes from across the Fungal Tree of Life.
1000 Insect Transcriptome Evolution (1KITE)	2012	1KITE aims at helping scientists uncover relationships among insects and tease apart the dates of origin of social behavior, parasitic behaviors, herbivory, flight, and so forth.
1000 plants (oneKP)[52]	2012	To generate large-scale gene sequences based on transcriptomes for over 1,000 species of plants
Fish T1K[53]	2013	To generate transcriptome sequences for 1,000 diverse species of living fishes
Bird 10K[54]	2015	To sequence and assemble draft genomes for about 10,500 extant bird species
10K Plant[55]	2017	To sequence at least 10,000 plant genomes representing every major clade of plants and eukaryotic microbes

1.2.2 From sequences to phylogenomics

In an era where advances in sequencing technology ensure that obtaining data from DNA is no longer a limiting procedure, the power of data analysis is still behind the ability of data generation[56]. The explosion of DNA sequences leads to two significant differences in molecular phylogenetics today in contrast to the days when sequences could only be obtained in a handful of species and gene loci. On the one hand, the data volume has increased dramatically with regard to the lengths of sequences and the number of taxa, adding considerable complexity in phylogenetic inference. On the other hand, a great number of phylogenetic trees from multiple loci are often reasonably requested to be inferred and interpreted in one study. Current phylogenomics often makes use of information in protein-coding genes from assembled genomes and transcriptomes. To obtain a complete genome and genes thereof were the primary purpose of DNA sequencing dating back to the development of Sanger sequencing in 1977[44,57]. Fortunately, we no longer need to assemble a genome by hand[57], thanks to the development of sophisticated algorithms on the *de novo* assembly of genomes[58] and transcriptomes[59,60]. Actually, the assembly of genomes and transcriptomes is still an active field in genomics and bioinformatics in order to obtain continuous and accurate genomic sequences and transcripts of genes. Comparing with analyzing the sequencing data directly, phylogenetic inference based on sequences is a down-stream analysis that mainly includes two steps: identification of homologous sequences and estimation of phylogenetic trees[40].

1.2.2.1 Handling sequence data for phylogenomics

Homologous identification using genomic data has been a long-lasting question[39], ever since the first few genomes were sequenced[61]. In general, homology in sequence data is determined by sequence similarity searching[40,62], followed by a technology, namely multiple sequence alignment (MSA), to identify homologous sites among different taxa. To identify homologous sequences, sequence similarity searching was developed to classify protein-coding genes into gene families that are likely to have shared common ancestral genes. From time to time, the criterion of similarity was somewhat arbitrary, before Tatusov *et al.*[61] firstly proposed a graph-based approach to solve the problem systematically with seven completed genomes at the time. The identified gene families include both orthologs and paralogs. Orthologs are genes in different species that shared a common ancestor through speciation, while paralogs are genes originated from gene duplications. As gene duplication can occur before and after speciation, paralogs can be further defined as out-paralogs and in-paralogs, respectively[39]. Because of the existence of in-paralogs, orthologous relationships of genes could exist between several genes in one species and several genes in another species, *i.e.*, a 'many-to-many' relationship. The identified homologous families in Tatusov *et al.*[61] are able to reflect the many-to-many orthologous relationship, so they are defined as orthologous groups. Later, Markov cluster algorithm (MCL) has been applied to the identification of orthologous groups to solve the low efficient approach from Tatusov *et al.*[61] and the limits of pairwise orthologous identification from Remm *et al.*[63]. TribeMCL[64] and OrthoMCL[65] are two widely used programs to identify orthologous groups (usage assessed by citations of 1,518 and 1,972 Web of Science citations at the time of writing this thesis) with precomputed all-against-all sequence searches by Basic Local Alignment Search Tool (BLAST)[66,67]. However, the vast scale of sequence data has severely affected the identification of orthologous groups[68]. To improve the analysis scalability to hundreds of plant genomes, OrthoFinder, a recent

developed program based MCL, has optimized the identification algorithms and abandoned the usage of MySQL database in OrthoMCL[69]. Sequence searching based on Hidden Markov Model has also been employed in recent studies to reduce time cost from the all-against-all BLAST search[68,70]. Apparently, sequence similarity is just a starting point to resolve the issues in the request for orthologous groups[71]. Approaches integrating phylogenetic inference, genomic synteny, and gene expression in closely related species have the potential to refine orthologous identification[72,73].

After identifying orthologous groups, the next step is to carry out MSA to determine homologous sites for phylogenetic inference. Starting from aligning multiple sequences by hand, to manually correcting the alignment generated by MSA aligners, the high throughput of current sequence data leaves little room to involve any labor efforts. MSA aligners have been developed based on different algorithms aiming at various data sets to solve the alignment issues[74]. For protein-coding genes used in phylogenomics, currently used MSA aligners could deal with most of the tasks through progressive alignment algorithms that add sequences one by one according to a guided tree inferred by the distance-based method. MSA aligners, such as ClustalW[75], T-Coffee[76], MAFFT[77], and MUSCLE[78], use the progressive alignment algorithms with various implementations. For the even larger size of sequences, derived progressive strategies, which use heuristic clustering methods rather than building a guided tree, are employed to direct the progress of adding sequences[74], like the algorithms implemented in MAFFT[77] and Clustal Omega[79]. SATé is an integrated MSA aligner that can generate MSA and phylogenetic tree simultaneously. It implements an iterating procedure that first builds an MSA and a maximum likelihood tree, and then optimizes the MSA based on the inferred tree[80,81]. Although no perfect tools can give absolutely accurate MSAs, most of them could do the work to identify homologous sites, especially when they are used followed by some alignment trimming procedures to remove low confident alignment regions and spurious homologous sites[82]. Among the available sequence trimming programs, like PAL2NAL[83] and Gblocks[82], trimAl is more suitable for preparing MSAs for many gene loci in phylogenomics, because it can automatically optimize the trimming parameters for MSA of each locus[84].

1.2.2.2 Phylogenetic inference based on sequences

With respect to phylogenetic inference, three main kinds of methods have been developed in order to make use of sequence data, including maximum parsimony approach, distance-based approach, and statistical approach. First and the most intuitive approach is the maximum parsimony (MP) approach. Inherited methods that use morphological characters, MP approach aims at finding phylogenetic trees that allow the minimum number of evolutionary changes of given characters, *i.e.*, sequences in our case. It reconstructs changes on each site of the sequences on different possible phylogenetic trees and selects the ones with the minimum number of changes. MP was widely adopted when formal probabilistic models of substitutions were yet proposed and when computers were not able to handle tentative computational tasks. However, MP is vulnerable to evolutionary rates in different lineages on a tree. Thus, it tends to select trees that group together fast evolving lineages often with long branches showing many changes, even when the two lineages are distantly related. Basically, the faster a lineage evolves, the higher opportunity it has had to acquire identical changes to other fast-evolving lineages purely by chance. The result is a phenomenon that distantly related long branches cluster as closely related branches, which

is often referred to as long branch attraction (LBA)[85]. MP is much more sensitive to LBA than other phylogenetic inference methods[86], but it is still a method that is used for morphological characters or serves as a starting point for statistical approaches, like its applications in RAxML[87].

Second, distance methods are based on the idea that if one can measure all the pairwise distances between different tips in a phylogenetic tree, the tree would have a fixed topology. The most straightforward measure of distance is the difference between two sequences (p-distance), but it does not make any correction for multiple substitutions at the same site. To deal with the issue, probabilistic models of substitutions of nucleotide, amino acid, and codon have been invented to estimate the expected number of substitutions per site as a function of substitution rate and time. Current probabilistic models are all time continuous Markov models, in which the occurrence of substitution at a site in a time window is only dependent on its state at that time but independent of how it has become such a state. Taking nucleotide substitution model as an example, Jukes and Cantor created a simple model (JC96) that allows one nucleotide converts into any of the other three nucleotides with the same probability[88]. The JC69 model was further complicated in the Kimura's model (K80)[89] by distinguishing the rates between transition, *i.e.*, substitutions from purine to purine or from pyrimidine to pyrimidine, and transversion, *i.e.*, substitutions from purine to pyrimidine or vice versa. Another improvement of the JC69 was implemented by considering unequal nucleotide frequencies in the Felsenstein's model (F81)[90]. A combination of the K80 and the F81 resulted in the model from Hasegawa, Kishino, and Yano (HKY85)[91]. The development of the nucleotide substitution models have been enhanced step by step, and it has eventually ended up with the general time-reversible (GTR) model with four transversion rates and two transition rates under unequal nucleotide frequencies[92]. For protein sequences and codon sequences, their substitution models are more complicated than the nucleotide models, because they have more possible states on each site[93], but the development of substitution models share the similar principles that formulated in the nucleotide substitution models.

Applying maximum likelihood (ML) estimation to these substitution models allows one to calculate a distance between any two sequences in the unit of the number of substitutions per site as a product of substitution rate and time[93]. After obtained distances, clustering algorithms, like neighbor-joining (NJ) and minimum-evolution, make use of a matrix of the pairwise distances of sequences to infer a phylogenetic tree. NJ would result in a tree that has the identical pairwise distances between taxa as the distances observed in the matrix. But minimum-evolution utilizes another strategy to optimize the tree topologies and the expected branch lengths to find a tree with a set of branch lengths that minimalizes the difference between the expected branch lengths and the observed branch lengths in condition of the shortest sum of expected branch lengths. Distance-based methods, especially NJ, are light in computation, so they are often employed as the methods to build starting trees for a heuristic search in statistical approaches. Minimum-evolution is less often used because it requires higher computational demands than NJ but has less statistical power than statistical methods as described below. Indeed, NJ is the most cited phylogenetic inference method[94] and had been widely applied to phylogenetics before statistical methods were freed from computational limitations. Nowadays, in cases when a probabilistic model of specific evolutionary changes is not established, such as the changes of rare genomic features, as long as a distance could be measured thereof, distance methods can be used infer phylogenetic trees. For example, overlapping genes, *i.e.*, adjacent genes that partially or

entirely overlap to each other in bacterial genomes, could be used to estimate distances between bacterial genomes and hence to infer their phylogenetic relationships[95]. Gene loss events are also used to estimate distances between co-linear regions within and between genomes to infer the order of whole-genome duplications and speciation events[96-98].

Third, the statistical approach on phylogenetics utilizes mathematical methods, such as ML inference and Bayesian inference, to infer not only the parameters in the substitution models but also tree topologies, branch lengths, or even other parameters. The formal ML approach for phylogenetics was established by Felsenstein in the 1970s[85], but its application with empirical data only became available in the phylogenetic community until 1990s[11]. In general, ML inference first uses a substitution model to calculate a probability (likelihood) for observing an aligned sequence matrix given a tree topology and a set of branch lengths. Then it tries to find the topology, the branch lengths, and the parameters in the substitution model that result in the highest likelihood. To calculate the likelihood for the complete sequence alignment matrix under a topology, the likelihood of each site (site likelihood) are first summed over all possible substitution scenarios at a specific site and then all the site likelihoods are multiplied[11]. To search a tree that most likely gives rise to the observed sequence alignment matrix, an algorithm in ML theoretically needs to explore all the possible trees with a fixed number of taxa. However, the demands of likelihood calculation and the number of possible trees increase dramatically with the increase of taxon number, so many heuristic algorithms have been developed. Major programs that perform ML phylogenetic inference have their own heuristic strategies to search for the ML tree, like those implemented in PhyML[99,100], RAxML/ExaML[87,101], FastTree[102], and IQ-Tree[103]. The performances of these programs concerning runtime and likelihoods of the best trees are comparable under benchmark tests based on simulated data[104] and empirical data[105]. They have significantly addressed the need for analyzing the ever-increasing data volume in the sequencing era. At present, ML has the ability to infer hundreds upon thousands of phylogenetic trees in reasonable time[99], and can even contribute to delivering gene trees in databases, like the ones in PhylomeDB[106] and PLAZA[107,108].

Bayesian inference in phylogenetics is a recently developed approach in the late 1990s[109,110]. It evaluates a phylogenetic tree by its posterior probability, *i.e.*, the probability if the tree is true, given a sequence alignment matrix, a substitution model, and prior probabilities. Among these, a prior probability is the probability of a random event that is assigned before any evidence is taken into account. For example, we could consider the prior probability distribution of tree topologies is from a uniform distribution, so they all have the same prior probability; or based on some *a priori* knowledge, we may consider a set of tree topologies is more likely than other tree topologies. A posterior probability is hence the conditional probability after the relevant evidence is taken into account. In other words, one can tell based on the posterior probability how confident it is for a phylogenetic tree resulting from Bayesian inference. However, the standard Bayesian equation can be hardly used to calculate posterior probability in phylogenetics, because the probability of observed data needs to sum over likelihoods of all possible priors of trees and parameters, which is almost impossible in a real phylogenetic analysis. The alternative approach is to use a computational algorithm, *i.e.*, Markov chain Monte Carlo (MCMC), to estimate the distributions of the posterior probabilities of different tree topologies, branch lengths, and parameters. For instance, the frequency of a specific tree topology after sufficient generations of MCMC equals to the posterior probability of the topology[11]. Although both ML and Bayesian inference need to

calculate the likelihood based on a specific substitution model, comparing with ML, Bayesian inference could explore larger spaces of parameters, so it could give different results to the ones inferred from ML sometimes. Bayesian inference is also able to incorporate more complicated models than ML, because of the relatively more efficient MCMC process than the ML search, even though the process still needs enormous computational resources. Bayesian inference programs in phylogenetics, like MrBayes[111], PhyloBayes[112], and BEAST[113,114], often attract phylogenetic community with specific substitution models and features that are able to be only implemented under the Bayesian scheme. For example, PhyloBayes is famous for its implementation of the CAT+GTR model, a Bayesian mixture model for across-site heterogeneities in amino acid substitutions[115]. BEAST is commonly used because of its extensibility for new functionalities from third-party developers and the feature has been further strengthened by incorporating a package management system into the latest release[114]. Nevertheless, complexities of usage also arise from the merit that Bayesian inference programs can accommodate complicated models. In addition, it seems to require more human interventions than other phylogenetic approaches, such as checking stationary and convergence of MCMC in an accepted Bayesian analysis, so projects involving inference of many gene trees are likely to use ML rather than Bayesian inference, although the burden of human intervention has been gradually laid down[111,113,114].

Furthermore, the most recently developed programs for phylogenetic inference offer solutions for the continuously increased data volume resulted from sequencing. Not only have some efficient searching algorithms for phylogenetic inference been implemented, such as the algorithms in ExaML[101] and IQ-Tree[103], but the fast development of computational ability also eases the pain of the high demands on computational time as well. For instance, popular ML or Bayesian phylogenetic inference programs utilize parallel computing to further reduce the amount of computational time with the implementation of the Message Passing Interface (MPI) in RAxML[87], PhyML[99], PhyloBayes[116], and BEAST[114], as well as the introducing of graphics processing units (GPUs) by the BEAGLE library[117].

Last but not least, together with the development of phylogenetic inference, the expansion of phylogenetic trees in both tree number and taxon size emphasize the overwhelming necessity for visualization, manipulation, and annotation of phylogenetic trees. A few web-based tools are available for users with limited experience, like PhyD3[118] in PLAZA4[108], while some programming libraries, like phytools[119], ETE 3[120] and ggtree[121], give experienced users extensible environments to illustrate and manipulate a large number of phylogenetic trees.

1.2.3 The origin of discordance in phylogenomics

Phylogenomics often involves sequences and the corresponding phylogenetic trees from multiple loci leading to a growth of heterogeneities in sequences and genealogies, especially when considering various evolutionary forces can act at different loci across lineages. Therefore, phylogenetic inferences of different loci very often result in incongruent trees with each other and/or with the species phylogeny[122,123]. The discordance might be partially originated from stochastic and systematic errors caused by non-phylogenetic signals[124]. Alternatively, specific evolutionary processes, such as incomplete lineage sorting, introgression, hybrid speciation, and gene and genome duplication, could also fundamentally underlie the discordance in phylogenomics. What makes the situation further complicated is the fact that it is usually challenging to disentangle the effects from the non-phylogenetic

signals and the genuine evolutionary processes, because they have no apparent characteristics. To some extent, modern evolutionary biologists need to be confronted with such a mosaic dataset that germinates from the sequencing soil. After assembly and annotation of the sequenced genomes or transcriptomes, proper procedures need to be applied to prepare the resulted sequences for reasonable phylogenetic inference, followed by interpretation of the derived phylogenetic trees in the light of the discordance in phylogenomics.

1.2.3.1 Discordance originated from model misspecifications

Technically, each step of phylogenetic inference could result in non-phylogenetic signals that lead to the discordance in phylogenomics. Some procedures of sequence analysis, such as identification of orthologous groups, alignment of homologous sequences, and not to speak taxon sampling, could all affect phylogenetic inference[124-126]. In addition to data processing, the overwhelming sequence data used in present phylogenomic analysis often violate the assumptions underneath current substitution models[127,128], due to deep phylogenetic scales and heterogeneous evolutionary rates among sites, genes, and lineages.

The deep phylogenetic scale of current data set is often considered as a major issue relevant to misspecification of substitution models with specific concerns on sequence compositions. As most substitution models assume that homologous sequences evolved from a common ancestral sequence through a globally stationary, time-reversible, and homogeneous way, the sequence compositions must reach the stationary equilibrium in such models. The descendant sequences hence should have similar sequence compositions as their common ancestor. However, to some extent, real biological sequences could hardly fit into this assumption, particularly when large molecular data sets from a broad taxonomy sampling is in an investigation. It has been well acknowledged that nucleotide compositions and amino acid usages vary dramatically across green plants and can mislead phylogenetic analyses specifically for deep phylogenetic relationships of green plants[127-129]. Furthermore, the scales of composition heterogeneity are different among sites. For example, Jeffroy *et al.*[122] have found that nucleotide compositions are more variable on the 3rd codon positions than on the 1st and 2nd codon positions, because of the mutation bias accumulated on the fast-evolving 3rd codon positions. The tree from Bayesian inference using the 3rd codon positions alone seems to be strongly correlated with the GC content (GC%) instead of the real phylogenetic relationship[122]. Similar patterns on 3rd codon positions are also found in other studies, indicating phylogenetic discordance is in part from violation of the same equilibrium frequencies in an alignment[127,130].

The current substitution models also assume that all sites in aligned sequences are homogenous instances of the same substitution process, which means all the sites should have constant substitution rates running along an underlying evolutionary tree. Apparently, such an assumption is not biologically realistic, because substitution rates could differ among different genes, different sites, and even different lineages due to the changes of selection pressure or life history during evolution. If the differences in evolutionary rates are only among genes, fitting substitution models gene by gene could be a solution. However, it is rational to imagine that sites in the same gene could have different evolutionary rates, for selection can act differently at each site. Inappropriate dealing with the rate variations across sites is indeed a source of inconsistent phylogenetic inference. A simple example is the

differences in substitution rates among the three codon positions. Because of codon degeneracy, third codon positions usually have higher substitution rates than the other two codon positions. To deal with such rate variations that we know *a priori*, we could treat them as different data sets to fit multiple substitution models. Nevertheless, treating rate variations based on *a priori* knowledge is not enough, for we usually have no knowledge on the rate differences across sites. To deal with such rate variations, Yang *et al.* proposed the "rate across site" model, in which the evolutionary rate at a site is considered as a random variable from a gamma distribution[93]. However, the "rate across site" model only loses the constant rate across sites but does not take into account the differences of equilibrium frequencies and relative substitution rates at each site[115]. Lartillot *et al.* has observed that some sites in an alignment of amino acid sequences tend to be replaced by fewer types of amino acids than the alignment of sequences simulated based on a common substitution model, suggesting that biochemical characteristics of amino acids often limit substitutions and some sites have higher probabilities of convergent substitutions than others[131]. If substitution models cannot adequately model the characteristic of amino acid substitutions, non-sister taxa with such substitution patterns would be inferred incorrectly as sister taxa, because they have a good chance to harbor multiple convergent substitutions[131]. Hence, a model called CAT considering site propensities, like its biochemical preferences, is somewhat realistic for substitutions of amino acid sequences.

Another source of evolutionary rate changes is dependent on both lineage and time because evolutionary constraints at some sites could be tuned in different lineages throughout time. Most of the current substitution models, including the ones considering rate variations across sites, apply a site-specific rate at a site to all the branches of an underlying tree. However, the evolutionary rate of a site could switch from a slow-evolving site to a fast-evolving site or vice versa in reality. In the context of a phylogenetic tree, such changes are reflected as shifts of evolutionary rates across branches, namely heterotachy[132]. It has been well recognized that heterotachy can mislead phylogenetic inference[86,133-135]. Some models have been proposed to accommodate heterotachy[18,136,137], but they are computationally expensive and may have the risk of data overfitting[93,136].

In the end, it should be noted that sequences in a phylogenomic analysis could be affected to different degrees by the various factors described above, so they may contain different amounts of non-phylogenetic signals depending on different mutation and/or fixation rates because of their functional roles. In addition, short sequences themselves are prone to random errors and lack of the ability to fit complex models simply due to limited sampling of characters[125], so phylogenetic trees generated from single locus would have higher chance to include stochastic errors.

1.2.3.2 Discordance originated from evolutionary processes

Apart from the issues in phylogenetic inference, the phylogenetic discordance between different gene loci does exist because they have undergone different evolutionary processes. The phylogenetic tree of a gene locus could be considered as a reflection of allele dynamics in populations of different species that are arranged by their evolutionary relationships[138]. In a population, a new allele is born through not only DNA mutations but also other processes like introgression and hybridization, which introduce alleles from other populations. In addition, gene duplication is also a force, which creates an extra gene locus and a new allele

simultaneously. Another potential force is horizontal gene transfer, but the thesis will not discuss it because of its rare occurrence in multi-cellular eukaryotes[37,38]. If we can keep a record of all the passages of alleles through reproduction in a population for a specific locus through time, when looking at the composition of alleles at a specific time point, we could coalesce any two alleles back to a common ancestor where the two alleles are created. By further arranging populations of species according to their evolutionary relationships, the existing alleles would hence further coalesce back in the assumed populations of common ancestors of the species[138]. As we often only sample one allele for each gene locus, the coalescent process of alleles would be considered as a gene tree. In below, the thesis will discuss how the incongruence between gene trees and between gene trees and species trees would appear in three possible evolutionary processes.

1.2.3.2.1 Incomplete lineage sorting

Lineage sorting is a process that gene lineages get lost over time in a population. Either reproduction failure or a failure of an allele to be passed to the next generation can cause lineage sorting. It is a result of Mendelian segregation in a random mating population, so lineage sorting occurs inevitably throughout evolution[11]. Lineage sorting in a population gives rise to the discordance of gene genealogies at different loci because of the random distribution of chromosomes and the recombination during meiosis. In other words, genes that are close to each other on the same chromosome tend to have similar gene genealogy, while genes that are on different chromosomes have a higher chance to have different gene genealogies[11]. As lineage sorting is inevitable, it may occur during a period of population split or speciation. If all alleles in a present population coalesce to a single ancestral allele before the population merges with another population lineage, *i.e.*, where the two populations split, the lineage sorting is claimed to be complete. Otherwise, coalescence of alleles to more than one ancestral alleles before merging with another population lineage leading to incomplete lineage sorting. Incomplete lineage sorting could produce gene trees with different topologies from the history of population split or speciation. In a case that two gene lineages each with a polymorphism allele created by a mutation exist in an ancestral population/species, while the population split or speciation occurs, the two descendant populations/species could inherit one or two of the alleles at the locus by chance. For example, we can assume population/species A has only one allele, while population/species A' has both of the alleles from the ancestral population. For the latter population/species A', if another population split or speciation occurs in a short period before the gene lineages of the two alleles become sorted completely into one gene lineage, the two alleles could be again transmitted into the further split populations/species, B and C. Afterwards, if each of B and C retains one of the two alleles respectively, we would have a gene tree with different history as the speciation process when sampling the genes (g_A, g_B, and g_C) in the three present populations/species (A, B, and C). Because the gene lineages have not sorted completely in the population/species A', the two genes, g_B and g_C, would not coalesce in the population/species A' but rather in the ancestor of A and A'. However, one of the genes, for example g_B, would first coalesce with g_A at the speciation or population split between A and A', and then further coalesce with g_C at the mutation that created the two polymorphism alleles. The gene tree is hence ((g_A, g_B), g_C), which is different from the population/species phylogeny as (A, (B, C)). Due to incomplete lineage sorting, gene trees may disagree with the species phylogeny, especially when a species tree has relatively short branches indicating species diverged in short time periods, because gene lineages have a high probability to coalesce on long branches[138].

1.2.3.2.2 Hybridization

The discordance in phylogenomics can also result from hybridization, which produces viable offspring from the process of mating between different species. The frequency of nature hybridization is not low, with roughly 10% of animal species and 25% of plant species that have been recorded to generate hybrids[36]. Hybridization mostly happens between closely related species from the same genus, but in some plants, for example in Orchidaceae, it could be observed even between species from different genera. There are two outcomes following hybridization, *i.e.*, introgression and hybrid speciation[6]. Introgression is the invasion of genetic materials from one parental species into the other. After hybridization, genetic materials from one parental species could integrate into the other parental species through repeated backcrossing between the interspecific hybrid and the latter parental species. In the other outcome, an interspecific hybrid may be reproductively isolated from its parental lineages and becomes a new species. This process is often referred to as hybrid speciation or lineage fusion[11]. Depending on whether hybridization increases ploidy or not, hybrid speciation is classified as allopolyploid and homoploid hybrid speciation, respectively[139]. Not only could the process of hybridization alter the hybrid genome with respect to gene expression, chromosomal structure, as well as genome size[6], but it also introduces alleles from other species. Therefore, in a phylogenomic analysis that includes descendants from either or both of the parental lineages, the introduced genes would have different genealogical history, because they are more closely related to the genes from the parental lineages that they originated, resulting in discordance among different gene loci. Furthermore, because a homoploid usually fails to form bivalents during meiosis due to the lack of homologous chromosomes[140], allopolyploid often accompanies the hybrid speciation in plants, which is either formed by genome duplication of a homoploid or by the fusion of unreduced gametes of two parental species[141]. Polyploid genome usually converts back into a diploid genome through a process referred to as diploidization. Many homeologous genes in the allopolyploid genome get lost during the diploidization process, similar to the trace following gene and genome duplication. Gene loss during the diploidization process hence would add another layer of discordance in phylogenomics on top of the discordance resulted from hybridization.

1.2.3.2.3 Gene and genome duplications

Gene duplication generates an extra copy of a segment of DNA in a genome through unequal crossing-over, DNA replication slippage, or retrotranspositions. Sometimes, the failure of separating homologous chromosomes properly during meiosis or mitosis may lead to the duplication of the chromosome(s). Rare catastrophe during meiosis could even cause unreduced gametes with an extra set of chromosomes, for example, a diploid gamete from a diploid species. The fusion of such unreduced gametes from the same species leads to genome duplication and forms polyploid with different numbers of chromosome sets, like tri-, tetra-, penta-, hexa-, and octapolyploid[141]. To distinguish polyploids formed by only one species instead of hybridization between two species as described above, these polyploids are named as autopolyploid. Auto- and allopolyploids could be born through other possible routes rather than the ways mentioned above[141], but in essence they both obtain extra sets of chromosomes. When disregarding the different origins of chromosomes in auto- and allopolyploid, they sometimes could be considered as the results of whole-genome duplication (WGD).

Gene and genome duplication could lead to the discordance among gene trees from different loci in phylogenomics, depending on the history of duplication and loss as well as speciation events. After gene duplication, most duplicated genes get lost and only a few of them would be retained[142]. Gene duplication and loss are a random process and have a comparable rate as nucleotide substitutions[143], so they can occur at any loci throughout evolution. The retention of duplicated genes varies among genes with regard to their roles in functions and the selection pressure acting on them. Therefore, the stochastic process of gene duplication followed by duplicated gene loss, entangling with speciation events during evolution, could result in different gene genealogies for different loci. For example, presuming two closely related species each has two homologous genes in their genomes, the gene trees of the four genes could be different depending on if there is one duplication occurred before the divergence of the two species or there are two independent duplications occurred after their divergence. This simple case only allows the occurrence of duplication within two species. Involving gene loss and more species would lead to more possible gene trees, for gene duplication and loss could behave differently in different lineages formed by speciation events.

For autopolyploid, all the genes get duplicated at once. After speciation, the variously possible retentions of duplicated genes in different lineages underlie a part of the discordance among gene trees. In these gene trees, all the retained paralogous genes originated from the autopolyploidization would coalesce precisely to the event where a genome gets duplicated. In contrast to autopolyploid in which all the chromosomes originate from one species, allopolyploid has chromosomes inheriting genetic backgrounds from its parental lineages, so the gene trees not only depends on the retained paralogous genes but also on whether the parental lineages are extinct or sampled in a phylogenomic study[144]. Therefore, the homeologous genes in an allopolyploid would coalesce to the divergence of the two parental lineages rather than the occurrence of the hybridization event. This makes it actually difficult to infer when the hybridization event occurs, especially when the descendants of the parental lineages may be extinct already[144]. In addition to the differences of coalesced point of duplicated/homeologous genes, auto- and allopolyploid also differ in their gene loss patterns. In autopolyploid, duplicated genes tend to lose evenly on the two homologous chromosomes, and the retained duplicated genes usually express around the same level. Whereas in allopolyploid, the subgenome from one parent tends to retain more homeologous genes than the other and to have higher expression level of its genes. The phenomenon is often referred to as genome dominance[145,146]. Although the mechanism underneath genome dominance is not fully resolved, it may result from the differences of epigenetic landscapes for the two subgenomes possibly mediated by transposon elements[145,147-149]. In a phylogenomic study with many species, the above factors related to gene and genome duplication could hence contribute to a large fraction of the incongruence between gene trees. This is particularly the case for phylogenomic studies in flowering plants, in which most lineages, if not all, have undergone at least one ancient WGD[150], giving rise to a considerable discordance in phylogenomic studies in angiosperms[151].

1.3 Research goals – using phylogenomic principles to study plant genome evolution

1.3.1 Reconstruction of phylogenetic relationships in seed plants

Dating back to the days when nucleotide sequences were first used to infer phylogenetic relationships, only one gene, usually the one that encodes small subunit ribosomal RNA, in different species was used[40]. In these days, gene trees were considered equivalent to species trees. Although the incongruence of gene trees has begun since more genes were taken into account, more sequences also provide enough data to obtain statistical support for phylogenetic relationships. To incorporate the added values of the increase of sequences, one of the primary goals for phylogenomics is to decipher the phylogenetic relationships on each part of the tree of life. Two main approaches have been proposed, including concatenating the MSAs of orthologous genes from different species into a very long super-gene ("super-matrix"), and extracting the phylogenetic information from different gene trees hence to infer the species history ("super-tree"). Both strategies have broad applications but also lead to some controversies on which is the optimal strategy[40].

The phylogeny of seed plants is very much a work in progress, and this study is trying to approach this controversial territory with newly sequenced and released transcriptomes in gymnosperms. Comparing with angiosperms, gymnosperms still lack molecular markers for broad comparisons. This study, thus, aims at extending the current phylogenetic markers of nuclear genes for seed plants and assessing its ability for resolving the phylogeny of seed plants. We used an approach based on Hidden Markov Model to identify single-copy genes among 31 gymnosperms and 34 angiosperms, followed by inferring the seed plant phylogeny by both super-matrix and super-tree methods. Single-copy genes are gene families that exist in most species and always return to single-copy status following gene and genome duplications. These nuclear genes are usually considered as good phylogenetic markers to solve unresolved phylogenetic questions raised by organelle genes because they inherit genetic information from both parents and have more sequence sites that could be used for phylogenetic analysis in comparing with organelle genes[152]. To deal with the heterogeneity of the identified phylogenetic markers in the super-matrix, different partitioning strategies were employed. Data partitioning takes advantage of the rich sampling of sequences but avoids the disadvantage of the increasing heterogeneity, because it mixes phylogenetic signals from different genes and, at the same time, estimates a set of parameters for each partition. PartitionFinder considers different genes and codon positions and uses ML to give a reasonable partitioning strategy that would most likely fit the data[153]. For the super-tree approach, we used the multispecies coalescent model, which accommodates the effects of incomplete lineage sorting on phylogenomic inference. It hence may help to resolve species phylogenies with short speciation radiation[138], such as its application in resolving the phylogeny of birds[154], and solving the deep divergence in angiosperms[155] and seed plants[156]. The other goal of the study is to investigate the potential factors that affect phylogenetic inference and potential methods for resolving it. To this end, we examined the various heterogeneities in the sequences underlying the different topologies inferred from different models and partitioning strategies.

1.3.2 Towards a better understanding on the retention of duplicated genes in angiosperm genomes

Gene duplication is an important mechanism for adding to genomic novelty. However, most of the duplicated genes get lost after duplication[142], hence, which genes are preserved following duplications is an important question. The gene retention after duplication events has been appreciated as a non-random process, in which certain duplicated genes seem to be more amenable to be retained than others [157-163]. The mode of duplication, either WGD or small-scale duplication (SSD), could influence the long-term survival of duplicated genes by reciprocal retention of different sets of genes[162,164]. In contrast, recent observations have found a specific set of genes that exist in all angiosperm genomes mainly in single-copy status[165-167], even though both WGDs and SSDs have frequently occurred in the flowering plant lineage[150,168]. The different fates of functional specific genes after duplication events hint that "gene duplicability", which is the ability of genes to be preserved following duplications, might be conserved across the angiosperm lineage.

To test the hypothesis by taking the phylogenetic context and the time of duplication events into account, we carried out a phylogenomic approach to assessed gene trees and gene retention patterns of 9,178 gene families shared between 37 flowering plant species. Twenty putative WGD events and numerous SSD events are hence covered in the study. Assessing the retention of duplicated genes across such a large number of genomes and duplication events further allows us to study the consistency of gene duplicability and provide an overview of gene duplicability across angiosperms.

1.3.3 Identification of whole-genome duplications in the lineage of orchids

Many lineages of present diploid species have undergone WGDs in their evolutionary history[150]. In flowering plants, the common ancestor of angiosperms might endure a WGD before the angiosperm radiation[169,170]. In addition, extra recent WGDs have been identified in many lineages of angiosperms, and many of them have been dated around the Cretaceous-Paleogene (K-Pg) boundary, suggesting ancient polyploids may help the ancestors of extant plants in many lineages today survive through extreme environments during the mass extinction at the K-Pg boundary around 66 million years ago (mya)[150,168,171,172]. WGD is often identified by comparing chromosomes in a genome to detect syntenic/co-linear regions resulted from WGD events[173-177]. To accelerate the identification of WGD, especially when genomes are unavailable, the age distribution of duplicated genes has also been employed in recent studies[178-180]. The time of WGD could also be estimated based on the age distribution of duplicated genes or a sophisticated phylogenomic dating based an approach under the Bayesian scheme[168].

Besides, recent advances in the sequencing technology have produced an enormous amount of sequence data for phylogenomics. Using gene trees and the corresponding species tree, one can test the hypothesis on the occurrence of WGD and its phylogenetic placement on the species phylogeny[98,169,180-183]. Indeed, Bowers *et al.* have identified a more ancient WGD in *Arabidopsis thaliana* when only a few plant genomes were released[181]. Later, Jiao *et al.*[169] have used gene trees to test different hypotheses and found two very ancient WGDs with one shared by all extant angiosperms and the other shared by all existing seed plants. Consistent with a reanalysis with a state-of-the-art molecular dating showing that the two WGDs could

not be entirely verified[184], the syntenic analysis in the genome of *Amborella trichopoda*, the early diverging angiosperm, only suggests one WGD event before the divergence of angiosperms[170]. To further substantiate WGDs, an approach integrated phylogenomics and other identification methods have been employed recently. The approach has added increase evidence for the ancestral WGD (γ) shared by the core eudicots[182] and the WGD (τ) shared by most of the monocots[98,185]. Except the ability to identify ancient WGDs, gene trees from phylogenomics also hold the promise to distinguish autopolyploid and allopolyploid by comparing topologies of gene trees under different hypotheses[144,186,187].

In this study, we integrated the approaches mentioned above to determine the existence of WGD(s) in the genome of *Apostasia shenzhenica*, an early diverged lineage in extant orchids (Orchidaceae). Previously published orchid genomes of *Phalaenopsis equestris* and *Dendrobium catenatum* illustrate a shared WGD occurred around 76 mya (with 72-81 mya as the lower and upper 90% confidence interval (CI)) before the divergence of the two species[188,189] (Figure 1-8). However, it is not clear if the WGD shared by *Phalaenopsis* and *Dendrobium* occurred before or after the divergence of all extent orchids, for the WGD falls into the initial divergence of extant orchid lineages as shown in Figure 1-8. Estimates for the crown age of extant orchids vary and range from 54 mya to 121 mya (Ramírez *et al.*[190]: 71–90 mya, youngest mean minus 1 standard deviation (SD) to oldest mean plus 1 SD; Gustafsson *et al.*[191]: 63–92 mya, 95% highest posterior density (HPD); Chen *et al.*[192]: 54–82 mya, 95% HPD; Chomicki *et al.*[193]: 75–121 mya, 95% HPD; Givnish *et al.*[194]: 80–100 mya, 95% CI; see also Figure 1-8). The *A. shenzhenica* genome hence allows us to test this by determining whether *A. shenzhenica* has a sign of WGD or not, and if the signal exists, whether it is the same WGD as identified in *Phalaenopsis* and *Dendrobium* or it is an independent WGD occurred in the lineage leading to *Apostasia* after the divergence of current orchids.

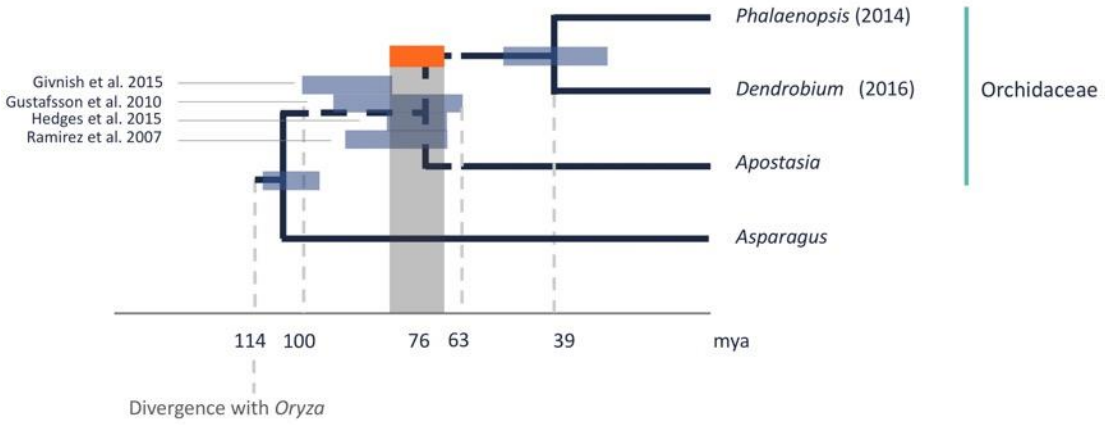


Figure 1-8 Estimates for ages of speciation events and WGD in Orchidaceae

The orange bar denotes the absolute age and 90% CI of the identified WGD in *Phalaenopsis*; the blue bars show the estimates of the speciation events on the tree (see the exact dates in text). Except for the divergence between *Apostasia* and other orchids, the divergence time between *Phalaenopsis* and *Dendrobium*, between Orchidaceae and *Asparagus*, and between Asparagales and *Oryza* are from TTOL (www.timetree.org).

Chapter 2

Single-copy genes as molecular markers for phylogenomic studies in seed plants

Zhen Li, Amanda R. De La Torre, Lieven Sterck, Francisco M. Cánovas, Concepción Avila, Irene Merino, José Antonio, Cabezas María, Teresa Cervera, Pär K. Ingvarsson, Yves Van de Peer

Genome Biology and Evolution **9**(5): 1130-1147 (2017)

Abstract

Phylogenetic relationships among seed plant taxa, especially within the gymnosperms, remain contested. In contrast to angiosperms, for which several genomic, transcriptomic and phylogenetic resources are available, there are few, if any, molecular markers that allow broad comparisons among gymnosperm species. With few gymnosperm genomes available, recently obtained transcriptomes in gymnosperms are a great addition to identifying single-copy gene families as molecular markers for phylogenomic analysis in seed plants. Taking advantage of an increasing number of available genomes and transcriptomes, we identified single-copy genes in a broad collection of seed plants and used these to infer phylogenetic relationships between major seed plant taxa. This study aims at extending the current phylogenetic toolkit for seed plants, assessing its ability for resolving seed plant phylogeny, and discussing potential factors affecting phylogenetic reconstruction. In total, we identified 3,072 single-copy genes in 31 gymnosperms and 2,156 single-copy genes in 34 angiosperms. All studied seed plants shared 1,469 single-copy genes, which are generally involved in functions like DNA metabolism, cell cycle, and photosynthesis. A selected set of 106 single-copy genes provided good resolution for the seed plant phylogeny except for gnetophytes. Although some of our analyses support a sister relationship between gnetophytes and other gymnosperms, phylogenetic trees from concatenated alignments without 3rd codon positions and amino acid alignments under the CAT+GTR model, support gnetophytes as a sister group to Pinaceae. Our phylogenomic analyses demonstrate that, in general, single-copy genes can uncover both recent and deep divergences of seed plant phylogeny.

2.1 Introduction

Seed plants originated about 370 million years ago (mya), and probably comprise 260,000 to 310,000 extant species[31,195]. Current seed plants consist of angiosperms (flowering plants) and gymnosperms, the latter of which are further subdivided into Cycadidae, Ginkgoidae, Gnetidae, and Pinidae[196]. Both morphological and molecular studies have clearly shown that angiosperms and gymnosperms are two monophyletic groups[197,198], but the relationship between the different clades in gymnosperms is less clear than in angiosperms[199], despite great efforts in resolving the phylogeny with diverse sets of molecular markers[156,200-202]. Particularly, the exact phylogenetic position of gnetophytes, a morphologically unique clade with accelerated molecular evolution rates, remains elusive[198]. Morphological studies, historically, agree that gnetophytes are a sister group of angiosperms (anthophyte hypothesis) (reviewed by Doyle[203]), because of obviously similar characteristics, such as, the existence of vessel elements and the simple, unisexual, flower-like reproductive organs. However, this hypothesis was later questioned on the basis of a flood of molecular data, with some providing support for gnetophytes as sister to the other seed plants (Gnetales – other seed plant hypothesis)[204] and others providing support for a sister group relationship with the other gymnosperms (Gnetales – other gymnosperms hypothesis)[200,205]. Still others provided support, usually based on mitochondrial or plastid genes, for gnetophytes as a sister group to conifers (Gnetifer hypothesis)[206], to one clade of conifers, *i.e.*, cupressophytes (Gnecup hypothesis)[156,202], or to the other conifer clade, *i.e.*, Pinaceae (Gnepine hypothesis)[134,135,201,207]. Also different approaches and data treatments yielded different phylogenetic placements of gnetophytes within the gymnosperms[70,135,201]. Besides the controversial systematic position of gnetophytes, *Ginkgo*, which is a monotypic genus of an ancient lineage that originated at least 270 mya, also has an ambiguous placement among the gymnosperms[198]. Some studies suggest *Ginkgo* as a sister group to a clade comprising conifers and gnetophytes[202,206,208]; whereas several recent phylogenomic analyses support a sister relationship between *Ginkgo* and cycads[70,156,205,209].

Increased species sampling could help resolving the evolutionary relationships within seed plants[210], but molecular markers for gymnosperms are still lacking to allow broad comparisons between taxa[202,205]. Single-copy gene families, or single-copy genes, have long been recognized as ideal molecular markers for inferring relationships of previously unresolved lineages[166,211,212]. Some characteristics, such as the uniqueness and high sequence conservation across species, allow single-copy genes to be straightforwardly amplified and sequenced. As nuclear genes, single-copy genes have bi-parental inheritance, unlike organelle genes that are mostly uniparentally inherited, so they may be better suited when dealing with hybridization, speciation, and incomplete lineage sorting of closely related species[152,166]. The use of multiple unlinked nuclear single-copy genes is more likely to reflect true species relationships and may solve incongruences between organelle genes[152,202,213].

Although widely applied to angiosperms[152,213,214], only a few single-copy genes have been used to resolve gymnosperm relationships[156,202,211]. In addition, current single-copy genes in gymnosperms were identified on the basis of those in angiosperms[70,211]. Whole genome sequencing can facilitate the identification of single-copy genes[167,215] but the huge genome sizes of gymnosperms (20-30 Gb) have greatly complicated their *de novo* sequencing[216]. As a consequence, only a few gymnosperm species have been sequenced so far[217-220]. However,

since single-copy genes are often more broadly expressed and at higher levels than non-single-copy genes[167,221], single-copy genes can be relatively easily detected by transcriptome sequencing, thereby simplifying the procedure to identify suitable molecular markers. In this study, using previously and newly developed genomic and transcriptomic data in 31 gymnosperms and 34 angiosperms, we identified single-copy gene families to increase the number of phylogenetic markers shared between gymnosperms (and between gymnosperms and angiosperms) that could be used for phylogenetic and comparative studies in seed plants.

2.2 Results

2.2.1 Transcriptome assembly and data integration

After assembly and removing redundant transcripts (Materials and Methods), we reconstructed 206,574 unigenes in *P. pinaster* and 121,938 unigenes in *P. sylvestris,* with an average length of 893 bp and 1,242 bp, respectively. Here a unigene is defined as transcripts from the same gene locus. For *P. glauca* and *P. sitchensis*, we integrated available public transcriptome data (Materials and Methods), which yielded 39,229 unigenes for *P. glauca* and 28,030 unigenes for *P. sitchensis*. TransDecoder predicted 20,434 to 76,426 open reading frames (ORFs) in the four species with around 57.3% to 68.5% of the ORFs having at least one Pfam domain (Table 2-1). For *P. abies* and *P. taeda*, we collected 54,381 proteins and 43,959 proteins from the two published conifer genomes, respectively[217,218]. Transcriptomes of another 25 gymnosperms were retrieved from public databases followed by removing redundant transcripts and predicting ORFs (Supplementary Table C-1).

Table 2-1 Transcriptome assembly and open reading frame (ORF) predictions

Species	# Transcripts	# ORFs	# ORFs with Pfam Domains
Pinus pinaster	206,574	76,426	43,771 (57.3%)
Pinus sylvestris	121,938	36,106	22,355 (61.9%)
Picea glauca	39,229	28,909	19,708 (68.2%)
Picea sitchensis	28,030	20,434	13,989 (68.5%)

2.2.2 Identification of single-copy genes in gymnosperms and angiosperms

Using OrthoMCL[65] and HMMER[222], we identified 3,072 single-copy genes in gymnosperms and 2,156 single-copy genes in angiosperms (Materials and Methods). Among these, 1,603 gene families were single-copy genes only found in gymnosperms, and 687 single-copy genes were specific to angiosperms. Additionally, 1,469 single-copy genes are shared between gymnosperms and angiosperms, so they are considered as the single-copy gene set representative for the seed plants.

Both missing data and whole-genome duplications complicate the identification of single-copy genes. First, as single-copy genes are usually conserved genes present in all seed plants by definition, species with incomplete annotations hamper the identification of conserved gene families and thus single-copy genes. Second, recent whole-genome duplications

resulted in a burst of recent duplicates, which decreases the number of identified single-copy genes. To explore the effects of missing data and genome duplication on the delineation of single-copy gene families, we performed *k*-means clustering on copy-number profiles of gymnosperms and angiosperms to cluster the species into two groups with similar profiles of copy numbers (Figure 2-1). Compared with angiosperms, we found that, in gymnosperms, the major factor affecting the identification of single-copy genes was missing data, as ten of the 31 gymnosperms showed serious incompleteness of gene space in the copy number profile (Figure 2-1A). These ten species had fewer proteins than the rest of the gymnosperms (*P* value = 3.78×10^{-5}, Wilcoxon rank sum test). In addition, for the 687 angiosperm specific single-copy genes, 586 of them were not conserved in gymnosperms according to our criterion (Materials and Methods), suggesting these conserved genes in angiosperms were either lost in some, if not all, gymnosperm lineages, or missed in their transcriptomes.

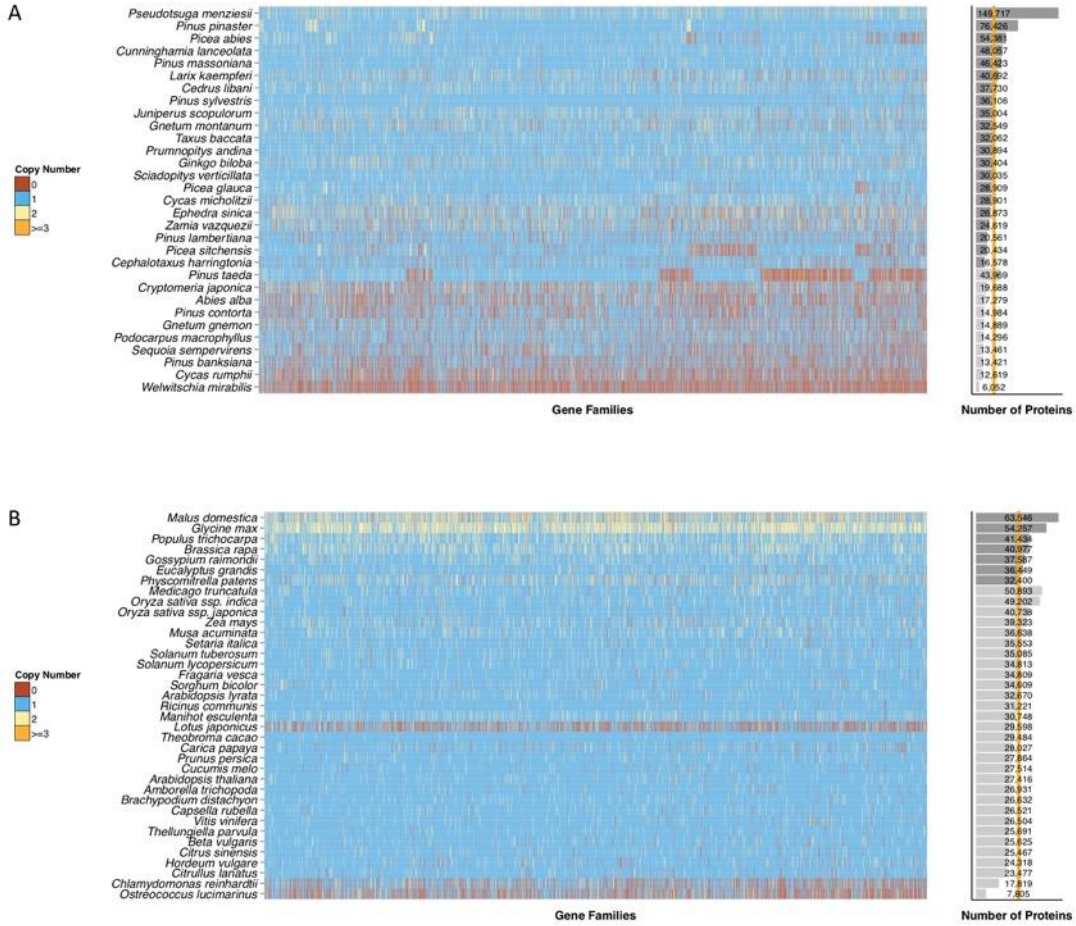


Figure 2-1 *k*-means clustering of copy number profiles for single-copy genes in gymnosperms (A) and angiosperms (B).

Rows represent species and columns represent gene families. In the copy number profiles, red denotes absence of genes in a gene family; blue denotes one copy; yellow denotes two copies; and orange denotes more than two copies in a gene family. The bar plot next to the copy number profile illustrates the number of proteins in each species with an orange line representing the average number of proteins. The dark and light gray bars distinguish the clusters identified by the *k*-means clustering.

For the copy number profile of angiosperms, the *k*-means clustering grouped species with recent whole-genome duplications together, indicating that species that have undergone

recent genome duplications still contain a large fraction of duplicated genes in the single-copy gene families (Figure 2-1B). For example, all seven species in the upper part of the copy-number profile, *i.e.*, *Malus domestica*, *Glycine max*, *Brassica rapa*, *Gossypium raimondii*, *Populus trichocarpa, Eucalyptus grandis* and *Physcomitrella patens,* have undergone lineage-specific whole-genome duplications[223-229]. On the contrary, the partial genome of *Lotus japonicus* and the small(er) proteome sizes of *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus* resulted in the absence of a large number of orthologous genes in these species.

2.2.3 Functional enrichment of single-copy genes

Single-copy genes are functionally biased toward certain conserved biological processes and organelle-related functions[166,167,215]. Since *A. thaliana* has been the most comprehensively annotated plant genome so far, we used *A. thaliana* genes to describe functions of single-copy genes for the angiosperms. GOSlim enrichment analysis revealed that the 2,156 single-copy gene families in angiosperms were often involved in photosynthesis, DNA metabolic processes, and cell cycle. Also, they were strikingly overrepresented in the plastid. On the other hand, single-copy genes of angiosperms were underrepresented in functional categories such as transcription factor activity, response to stimulus, and signal transduction (Figure 2-2). For the 3,072 single-copy gene families in gymnosperms, we used functionally annotated genes in *P. pinaster* to perform the GOSlim enrichment analysis, which, to some degree, suggested their similar functions as in angiosperms but with some exceptions, for example, lack of underrepresentation in response to stimulus, and extra overrepresentation in catabolic and lipid metabolic processes (Figure 2-2). We argue that the difference in the enrichment analyses between angiosperms and gymnosperms is largely due to the incompleteness of GOSlim annotations in *P. pinaster*, which only had 32,716 of the 76,426 (42.8%) genes that were annotated by at least one GOSlim term, whereas in *A. thaliana*, the percentage increased to 21,106 of 27,205 (77.6%) genes. A gene set with severely incomplete GO annotations could introduce systematic bias in the enrichment analysis. At last, the 1,469 single-copy gene families in seed plants were overrepresented or underrepresented in nearly identical functional categories as the ones in angiosperms, when using *A. thaliana* genes as representatives (Figure 2-2). The functions of single-copy genes in seed plants further confirm that these genes are involved in essential functions conserved across all seed plants and even throughout eukaryotes[167,215,230].

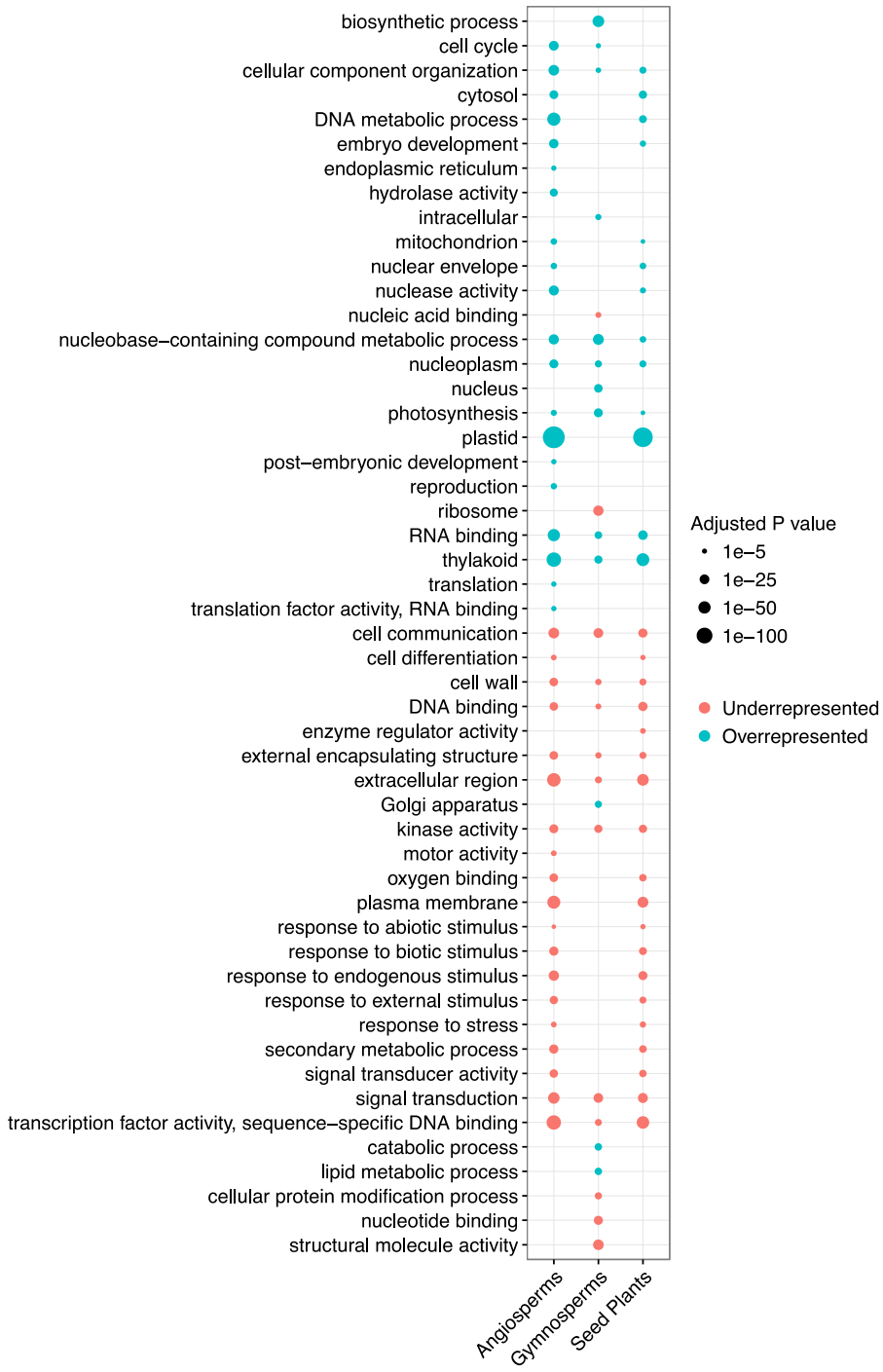


Figure 2-2 Gene Ontology Slim (GOSlim) enrichment analysis for single-copy genes in angiosperms, gymnosperms, and seed plants.
Dot size is representative for the statistical significance of overrepresented (green) and underrepresented (red) GOSlim terms.

2.2.4 Reconstructing seed plant phylogeny

We used both tree construction based on concatenated sequence alignments and multi-species coalescent approaches to reconstruct the phylogeny of seed plants based on 106 phylogenetic markers selected from the 1,469 single-copy genes in seed plants (Materials and

Methods). As 3rd codon positions have been known to affect the placement of gnetophytes[70], we built two different concatenated nucleotide sequence alignments from the 106 genes, one with and one without 3rd codon positions, named 'NT123' and 'NT12', respectively. Species trees were then inferred from the two alignments under the GTR+GAMMA model with different partitioning strategies (Materials and Methods). All of the inferred phylogenetic trees support a monophyletic origin for both extant gymnosperms and angiosperms (100% bootstrap percentage, BP)[70,156,200,231]. The angiosperm phylogeny is largely congruent with the APGIII tree[199] with *Amborella* as a sister group to the monocots and dicots (Figure 2.3 and Figure 2.4). The incongruence with respect to the position of the Malpighiales (*i.e.*, *P. trichocarpa*, *Ricinus communis*, and *Manihot esculenta*) between our phylogeny and the APGIII tree has long been recognized[129,200,232]. A hypothetical introgressive hybridization in the ancestral lineages of Fabidae and Malvidae has been proposed to explain a different ancestry of nuclear and chloroplast genes in extant Malpighiales[233].

For gymnosperms, the species trees inferred from NT123 and NT12 were largely similar except for some of the relationships within Pinaceae and cycads, and particularly the position of gnetophytes (Figure 2.3and Figure 2.4, and Supplementary Figures C-1 – C-5). For Pinaceae, the only difference concerned the genus *Pinus*. The NT123 alignment clearly distinguished between the two subgenera of *Pinus*, *i.e.*, subgenus *Strobus* (*Pinus lambertiana*) and subgenus *Pinus* (100% BP). The subgenus *Pinus* consists of the sections *Trifoliae* (*i.e.*, *P. taeda*, *Pinus contorta*, and *Pinus banksiana*) and *Pinus* (*i.e.*, *P. Pinaster*, *P. sylvestris*, and *Pinus massoniana*) (100% BP) as also observed in previous studies[234,235]. Trees inferred from the NT12 alignment had low bootstrap values for the genus *Pinus*, and incorrectly placed *Abies alba* (Figure 2.4), which was grouped with *Cedrus libani* as a sister to the other Pinaceae by the NT123 alignment (Figure 2.3), as expected based on morphological and molecular studies[236]. Both alignments show *Larix* and *Pseudotsuga* to form a clade with *Pinus* and *Picea* as a sister clade.

For cupressophytes, all topologies suggest that *Podocarpaceae* diverged first, followed by *Sciadopityaceae*, and then *Taxaceae* – *Cephalotaxaceae* as a sister to *Cupressaceae*. For *Ginkgo*, our phylogenetic analyses suggest that it belongs to a sister group of cycads (100% BP), in accordance with recent phylogenomic analyses[156,205,209], but in contrast to previous studies that support cycads as the sister lineage to the other gymnosperms[202,206,208].

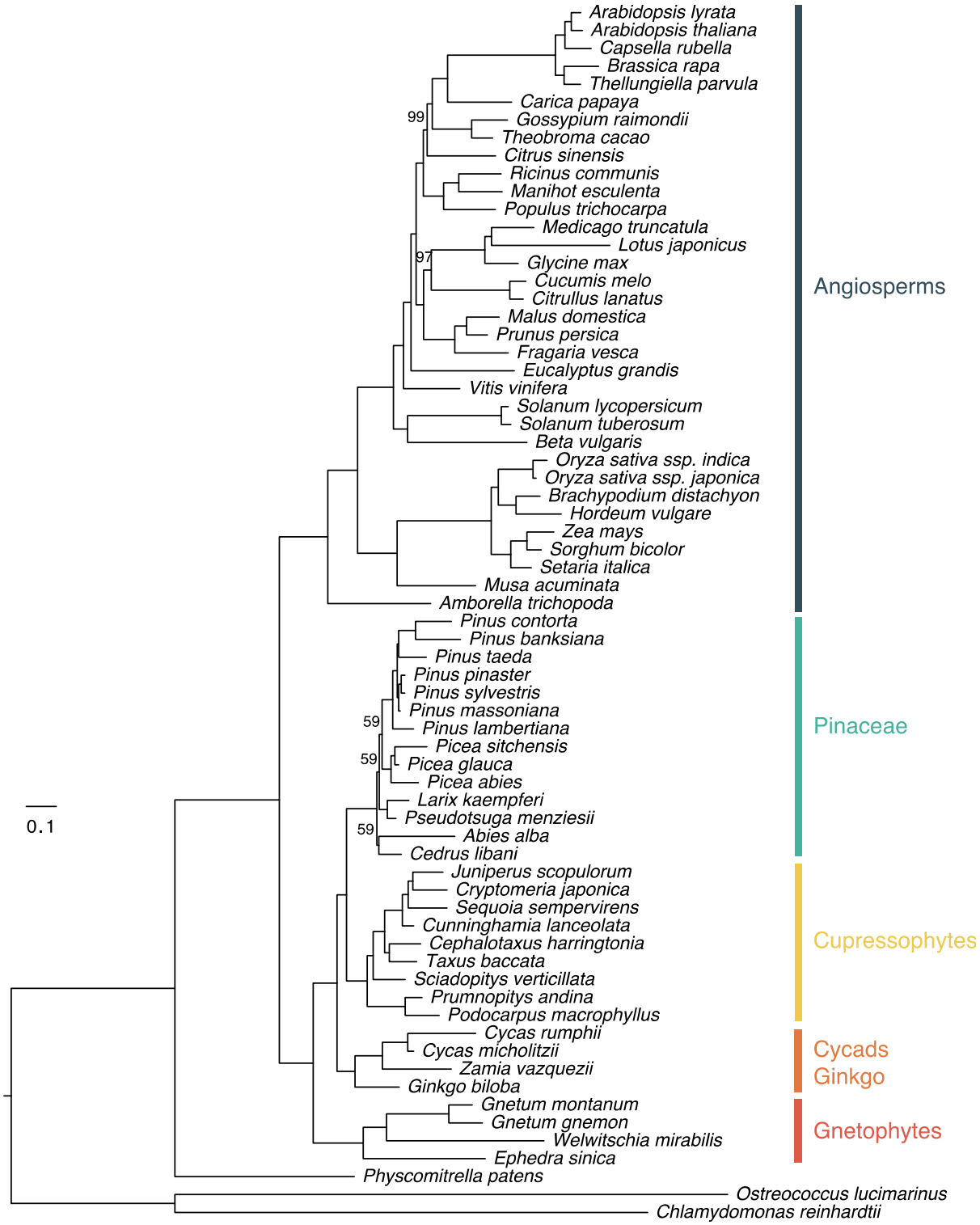


Figure 2-3 Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions, partitioned by PartitionFinder.

Bootstrap values less than 100% are shown on the specific branches. See Supplementary Figures C-1, C-2 and C-3 for maximum likelihood trees inferred from partitions based on codon positions.

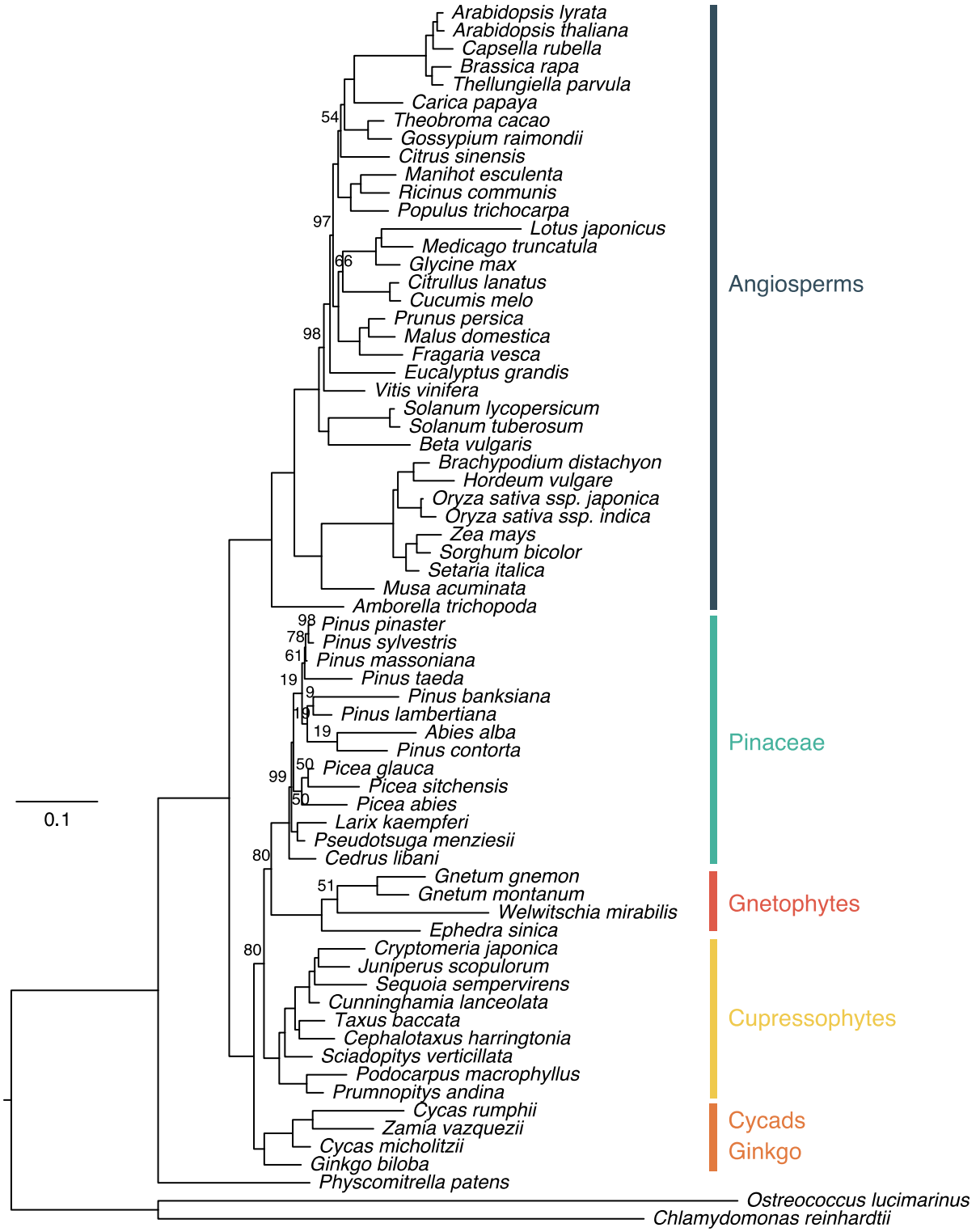


Figure 2-4 Maximum likelihood tree inferred from a concatenated alignment of 1st and 2nd codon positions for 106 single-copy genes in seed plants partitioned by PartitionFinder.
Bootstrap values less than 100% are shown on the specific branches. See Supplementary Figures C-4 and C-5 for maximum likelihood trees inferred from partitions based on codon positions.

2.2.5 The phylogenetic position of gnetophytes

Regarding the phylogenetic position of gnetophytes, NT123 and NT12 alignments gave contradictory results. In all species trees based on the NT123 alignment (Figure 2-3 and Supplementary Figures C-1 – C-3), gnetophytes were placed as a sister clade to the other gymnosperms (100% BP) in support of the 'Gnetales – other gymnosperms' hypothesis. Species trees based on the NT12 alignment, however, clustered gnetophytes with Pinaceae thus supporting the 'Gnepine' hypothesis (⩾73% BP, Figure 2-4 and Supplementary Figures C-4 and C-5). To obtain extra statistic support for the two alternative topologies instead of bootstrap values, we performed AU tests by CONSEL[237]. Based on per site log likelihoods for the two topologies, the NT123 alignment significantly rejected the 'Gnepine' topology (*P* value = 2×10^{-69} for three partitions by each codon position and *P* value = 6×10^{-36} for 52 partitions from PartitionFinder); notwithstanding, the NT12 alignment also rejected the 'Gnetales-other gymnosperms' topology (*P* value = 0.014 for two partitions by each codon position and *P* value = 0.028 for 37 partitions from PartitionFinder). We further inferred the species phylogenies based on the concatenated alignments of each codon position, named 'NT1', 'NT2', and 'NT3', to explore their contributions to the phylogenetic position of gnetophytes, independently. Interestingly, the NT3 alignment gave the same topology as the one based on the NT123 alignment and supported 'Gnetales – other gymnosperms' hypothesis with 100% BP (Supplementary Figure C-6). The NT1 and NT2 alignments both resulted in topologies similar to the one obtained from the NT12 alignment by supporting the 'Gnepine' hypothesis with 95% BP and 51%, respectively (Supplementary Figures C-7 and C-8). Our observations confirm that the inclusion of 3rd codon positions in the concatenated alignment indeed influences the phylogenetic position of gnetophytes in seed plant phylogeny as shown in previous phylogenomic studies[70].

For nucleotide sequences of protein-coding genes, most sites from 3rd codon positions are synonymous sites due to codon degeneracy. It has been acknowledged that 3rd codon positions not only can contribute to phylogenetic signal[129,238], but can also add noise to phylogenetic analysis because they quickly become saturated[239]. This might lead to problems when using stationary time reversible models, especially when dealing with deep phylogenetic relationships[128,134,135]. Therefore, we further investigated base compositional heterogeneity and lineage specific changes of evolutionary rates on different codon positions in the five concatenated alignments of nucleotide sequences. The GC content of the 106 phylogenetic markers at different codon positions were dissimilar in different species, and in particular the 3rd codon positions were more variable compared with the 1st and the 2nd codon positions (Supplementary Figure C-9). Pairwise comparisons of GC content among different species in the NT123, NT1, NT2, and NT3 alignments indicated that the NT123 and NT3 alignments exhibited significant compositional heterogeneity among different species (*P* value < 0.001, Wilcoxon test with Bonferroni correction). The differences were most outspoken in two sets of groups, *i.e.*, between the outgroup (two green algae and moss) and all seed plants, as well as between some angiosperms (especially Poaceae) and gymnosperms (Figure 2-5). However, significant differences in GC content in the NT1 and NT2 alignments almost only exist between the outgroup and seed plants. The pattern observed above still holds true after removing aligned codons that encode the same amino acids in the NT123 alignment (Supplementary Figures C-10 and C-11), suggesting that 3rd codon positions substantially contribute to the compositional heterogeneity in the NT123 alignment, while the base compositions of 1st and 2nd codon positions are in general very similar.

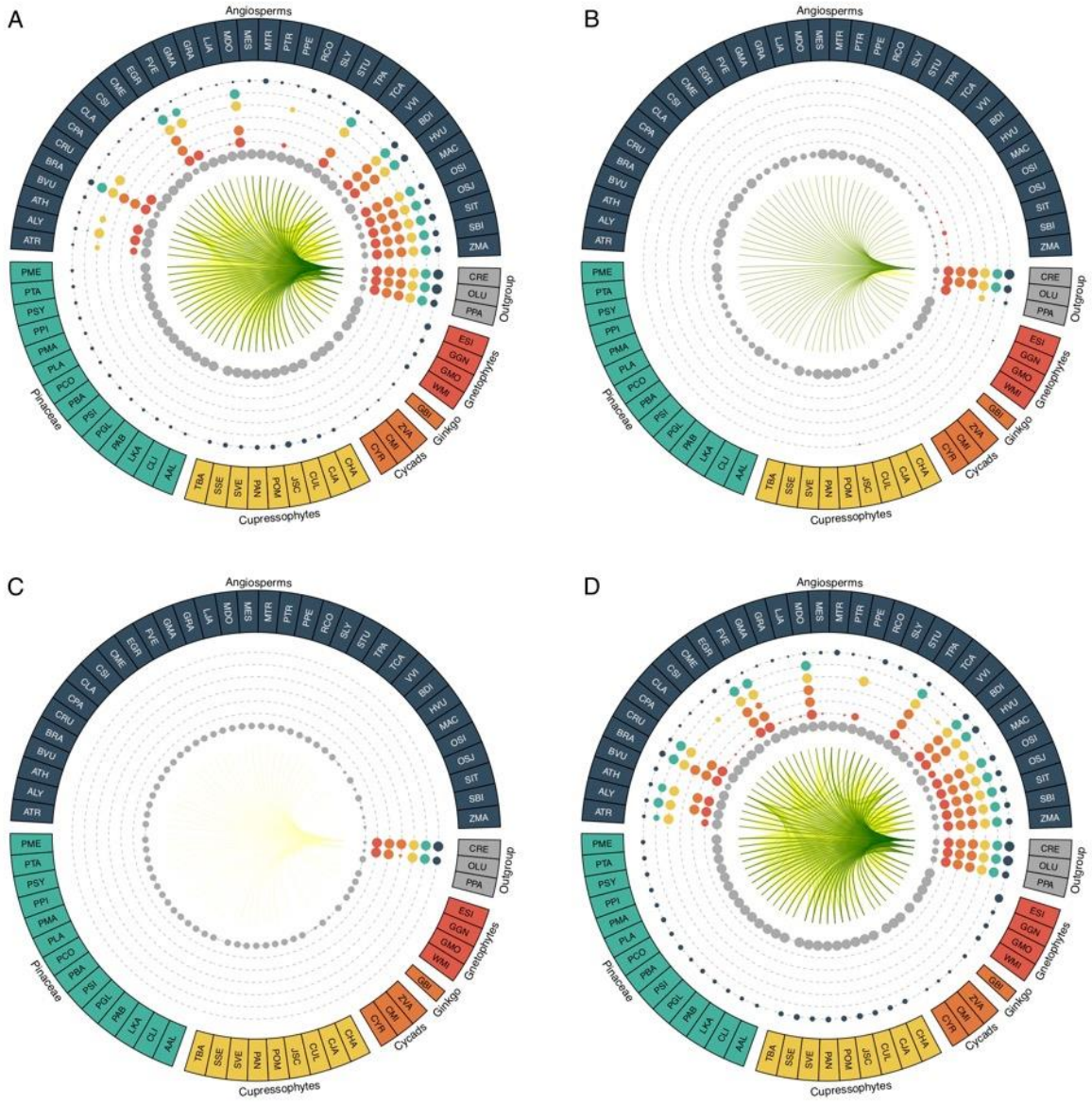


Figure 2-5 Comparison of GC content in the concatenated alignment (A) and at each codon position (B, C, and D) from 106 genes in 68 species.

Dot size correlates with the number of species in each lineage (group) that have a significantly different GC% (Wilcox test, *P* value < 1×10^{-3}) with the species compared with (colors of dots correspond to the compared lineages). Lines connecting any two species represent significant difference in GC content, with most significant in green and weakest in yellow (1×10^{-3}). The full names for the species can be found in Supplementary Table C-2.

Disparate evolutionary rates of different sites among lineages, known as heterotachy, violate the assumption of one set of branch lengths for all sites in the homogeneous models[134,135]. Using the ML phylogenies inferred from NT1, NT2, and NT3, we measured branch lengths from the most recent common ancestor for each of the five monophyletic groups (*i.e.*, angiosperms, gnetophytes, cycads and *Ginkgo*, cupressophytes, and Pinaceae) to every species in each group. As expected, the branch lengths were shorter for the trees inferred from 1st and 2nd codon positions than for the tree based on 3rd codon positions (Figure 2-6). An outspoken feature of the changes of branch lengths was their disproportional increase

from 1st and 2nd codon positions to 3rd codon positions in the five lineages, from angiosperms as the fastest clade, followed by gnetophytes, cycads and *Ginkgo*, cupressophytes, to Pinaceae as the slowest. The drastic increase of branch lengths of the tree based on 3rd codon positions for angiosperms and gnetophytes, compared with the relatively stable alteration in Pinaceae, indicate distinctive various evolutionary rates among codon positions in the five clades, which is a characteristic signal of heterotachy.

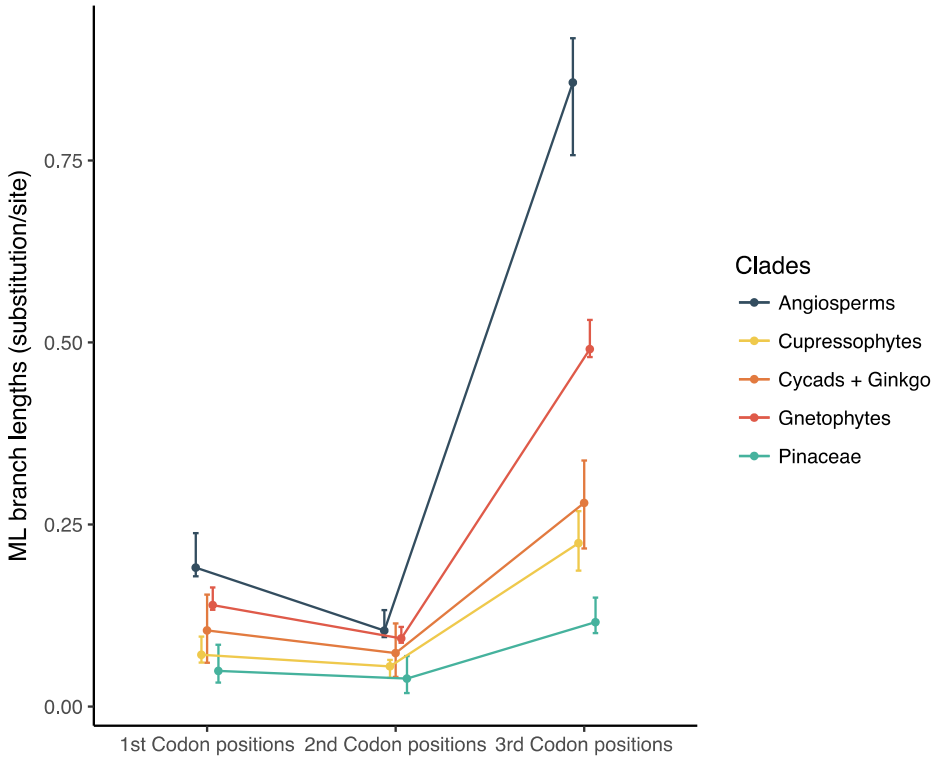


Figure 2-6 Lineage specific branch lengths

Lineage specific branch length estimates from each species to the most recent common ancestor of the five monophyletic groups (angiosperms, cupressophytes, cycads and Ginkgo, Gnetophytes, and Pinaceae), in trees inferred from sites at 1st, 2nd, and 3rd codon positions.

The elevated evolutionary rates of 3rd codon positions might suggest substitution saturation, so we used I_{SS} to characterize substitution saturation in the nucleotide alignments. If I_{SS} is close to 1 or greater than a critical I_{SS} ($I_{SS.C}$), the alignment is considered to exhibit substantial saturation[240]. Given its dependence on tree topologies, $I_{SS.C}$ is estimated under an extremely symmetrical ($I_{SS.C.Sym}$) as well as asymmetrical topology ($I_{SS.C.Asym}$). For the first two codon positions, either combined (NT12) or separate (NT1 and NT2), the I_{SS} values were significantly smaller than both $I_{SS.C,Sym}$ and $I_{SS.C,Asym}$ (P value < 1×10^{-4}, two-tailed t-test, Table 2.2), showing little evidence of substitution saturation on these sites. Nevertheless, for both alignments including 3rd codon positions (NT123 and NT3) I_{SS} were greater than $I_{SS.C,Asym}$ (P value < 1×10^{-4}, two-tailed t-test, Table 2-2), suggesting that sites from 3rd codon positions experienced substantially higher levels of substitution saturation than did sites from the 1st and 2nd codon

positions. As values of I_{SS} for NT123 and NT3 were smaller than $I_{SS.C,Sym}$, they may be only useful when the real topology is extremely symmetrical, but the real topology of the sampled species in this study is somewhere in between a symmetrical and an asymmetrical tree.

Table 2-2 The index of substitution saturation (I_{SS}) on concatenated nucleotide alignments and alignments of each codon position.

Dataset	# Sites	I_{SS}	$I_{SS.C.Sym}$	$I_{SS.C.Asym}$
Alignment with 3rd codon positions (NT123)	149,679	0.612	0.820*	0.605*
Alignment with 1st and 2nd codon positions (NT12)	99,786	0.521	0.819*	0.603*
Alignment of 1st codon positions (NT1)	49,893	0.551	0.818*	0.598*
Alignment of 2nd codon positions (NT2)	49,893	0.494	0.818*	0.598*
Alignment of 3rd codon positions (NT3)	49,893	0.796	0.818*	0.598*

*P value < 1×10^{-4}, two-tailed t-test

The above results clearly illustrate that sites from the 3rd codon positions have features typically found in fast evolving sites, which are distinguishable from sites of the first two codon positions. Since using 3rd codon positions solely can produce nearly identical phylogenies as those based on the NT123 alignment (Figure 2-3 and Supplementary Figure C-6), it is plausible to assume that inclusion of the 3rd codon positions in the concatenated alignment of nucleotide sequences leads to systematic bias in the phylogenetic analysis of seed plants, which constantly placed gnetophytes as a sister group to the other gymnosperms.

We further tested whether codon and amino acid substitution models are robust to the potential bias introduced by the 3rd codon positions. Unlike DNA substitution models, codon substitution models can explicitly describe synonymous and nonsynonymous substitutions and realistically estimate natural selection acting on protein-coding sequences. By separating the two types of substitutions with different rates, they are supposed to reflect both recent and early divergences[100,241]. Protein sequences, as the translated products of coding sequences, have been shown to be less affected by substitution saturation than nucleotide sequences[70], as they record nonsynonymous substitutions but ignore synonymous substitutions that may hamper phylogenetic inference due to substitution saturation[238]. As mostly synonymous sites, sites at 3rd codon positions may negligibly influence the phylogenetic placement of gnetophytes under the codon and amino acid substitution models. Therefore, trees built under the codon and amino acid models were expected to be congruent with those inferred from NT12 alignments and the GTR+GAMMA model. Surprisingly, the codon model and amino acid model both gave nearly identical ML trees as the topologies inferred from the NT123 alignment under the GTR+GAMMA model, highly supporting the 'Gnetales – other gymnosperms' hypothesis (Supplementary Figures C-12 and C-13). A similar topology has been suggested by Lee *et al.*[200] based on a concatenated amino acid matrix of nuclear genes, although all amino acid substitution matrices in Wickett *et al.*[70] strongly support a closer relationship between gnetophytes and conifers.

Since the propensities of amino acids play an important role in the evolutionary rates across sites, an effect not modeled by the discrete GAMMA distribution in our ML analysis, we used

the CAT and CAT+GTR model implemented in PhyloBayes-MPI to infer the phylogeny based on single-copy genes[112,115,116]. For computational reasons, the original alignment consisting of 49,893 sites was reduced to a shorter alignment with 7,562 sites (Material and Methods). The reduced alignment resulted in a similar ML topology as the original amino acid alignment under the JTT+I+GAMMA+F model (Supplementary Figure C-14). Interestingly, the CAT model supported the 'Gnetophytes – other gymnosperm' hypothesis (posterior probability = 0.98, Supplementary Figure C-15), while the CAT+GTR model supported the 'Gnepine' hypothesis (posterior probability = 0.86, Supplementary Figure C-16). Because the CAT model uses flat exchange rates that are not actually realistic, the CAT+GTR model is more appropriate for real biological data and is virtually always the model with the highest fit in PhyloBayes[112]. Amino acid compositions also exhibited compositional heterogeneity in a few species distributed across the phylogeny, as 'ppred' in PhyloBayes-MPI pointed out. *Physcomitrella patens*, *Medicago truncatula*, *Musa acuminata*, *Oryza sativa*, *Pinus taeda*, *Pinus banksiana*, and *Gnetum Montanum* rejected compositional homogeneity under the CAT+GTR model (posterior predictive $P < 0.05$). In summary, as the sites at 3rd codon positions were included in the 'codon' alignment and GC content is correlated with specific amino acid residues[129], the above results suggest that the codon model (GY) and the amino acid model (JTT+I+GAMMA+F and CAT) may fail to accommodate the systematic bias introduced by the 3rd codon positions, except for the CAT+GTR model.

2.2.6 Phylogeny based on multi-species coalescent model

Except for the analyses based on concatenated alignments, we also applied recently developed coalescent approaches implemented in STAR[242] and in ASTRAL-II[243], taking into account incomplete lineage sorting in gene trees. To further assess the effects of 3rd codon positions on the placement of gnetophytes, we built gene trees of the 106 different phylogenetic markers based on alignments with and without 3rd codon positions. The two sets of gene trees were named as 'GT123' and 'GT12', respectively. Coalescent analyses on GT123 from both STAR and ASTRAL-II were largely congruent with the ML phylogenies inferred from the NT123 alignment with both the DNA model, codon model, and amino acid model, hence in support of the 'Gnetales-other gymnosperms' hypothesis (100% BP, Supplementary Figure C-17 and C-18). Nevertheless, GT12 resulted in two different topologies with respect to gnetophytes. STAR fully supported the 'Gnetales-other gymnosperms' hypothesis (100% BP, Supplementary Figure C-19), but ASTRAL supported the 'Gnetifer' hypothesis (60% BP), which placed gnetophytes as a sister group to all conifers (Supplementary Figure C-20). However, the 'Gnetifer' topology was accepted by neither the NT123 alignment (*P* value = 2×10^{-11} for three partitions by each codon position and *P* value = 3×10^{-103} for 52 partitions by PartitionFinder) nor the NT12 alignment (*P* value = 1×10^{-47} for two partitions by each codon position, and *P* value = 0.001 for 37 partitions by PartitionFinder).

The phylogenetic signal in the two sets of gene trees was further measured by Internode Confidence (IC) and Internode Confidence All (ICA), which account for existed topological bipartitions in gene trees to estimate incongruence of phylogenetic signal[123,244,245]. We used IC and ICA to determine the incongruence in both GT123 and GT12 trees with respect to the three alternative topologies obtained from the phylogenomic analyses described above (Figure 2-7). Interestingly, both sets of gene trees have no prevalent bipartitions to support either cupressophytes (Figure 2-7A and Figure 2-7C) or gnetophytes (Figure 2-7B) as a sister group to Pinaceae, since the values of IC and ICA were extremely close to zero. However,

there was a slight phylogenetic signal to group gnetophytes within or with conifers from the GT12 gene trees inferred without 3rd codon positions (Figure 2-7B and Figure 2-7C, respectively). In contrast to the incompatible phylogenetic signals for the position of gnetophytes, both sets of gene trees exhibited a strong phylogenetic signal for *Ginkgo* as a sister group to cycads independent of the position of gnetophytes (Figure 2-7).

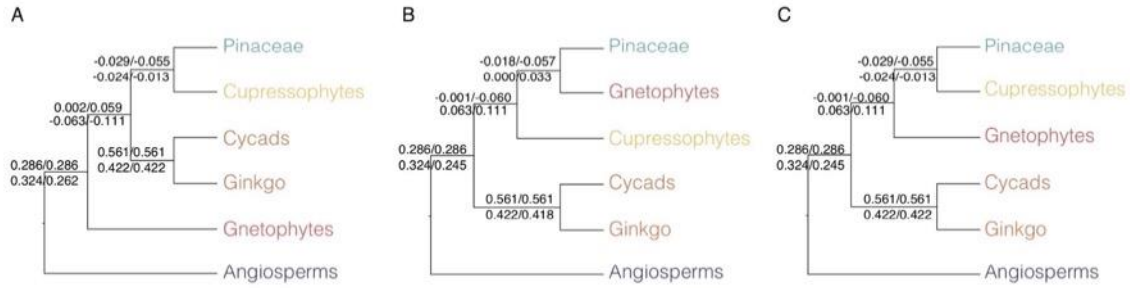


Figure 2-7 Internode Certainty (IC) and Internode Certainty All (ICA) estimated from gene trees of 106 phylogenetic markers for the deep divergence of seed plants.
(A) the 'Gnetales – other gymnosperms' hypothesis; (B) the 'Gnepine' hypothesis; (C) the 'Gnetifer' hypothesis. Numbers above branches represent IC and ICA estimated from the gene trees based on alignments with 3rd codon positions; numbers below branches represent IC and ICA estimated from the gene trees based on alignments without 3rd codon positions.

2.3 Discussion

2.3.1 Single-copy genes resolve the phylogeny of seed plants

Resolving the exact phylogeny of seed plants is fundamental to our understanding of the evolution, diversification, and colonization of major plant groups on Earth. Despite recent advances in sequencing technologies and great efforts to use diverse sets of molecular markers, the phylogenetic relationships among the five main seed plant lineages remain contested. Here, we have identified a set of 1,469 single-copy genes that are shared among 65 species comprising five seed plant lineages. This data set represents one of the most comprehensive comparative studies including gymnosperm species. With such a broad taxonomic sampling that includes all conifers (except Araucariaceae), cycads, *Ginkgo*, gnetophytes and angiosperms, our markers have the potential to unlock phylogenetic and evolutionary relationships in seed plants.

The phylogenetic markers developed here are effective markers for phylogenetic analyses in each lineage of seed plants. With different partitioning strategies and multi-species coalescent methods, the markers give clear phylogenetic relationships within angiosperms, Pinaceae, cupressophytes, cycads, and gnetophytes. The phylogenies, for instance, inferred from the NT123 alignment partitioned by PartitionFinder based on GTR+GAMMA model (Figure 2-3), based on codon substitution model (Supplementary Figure C-12), and based on multi-species coalescent models with GT123 (Supplementary Figure C-17 and C-18), all provide excellent examples of the applications of the 106 phylogenetic markers in all lineages

of seed plants. It is also interesting to note that 3rd codon positions of the phylogenetic markers have limited effects on such phylogenetic relationships within each clade. Although the position of *A. alba* in Pinaceae changes in a small fraction of the phylogenetic trees, this is probably due to the lack of species available in closely related genera to *Abies*, e.g. *Keteleeria*, *Pseudolarix*, *Nothotsuga*, and *Tsuga*.

Our phylogenetic markers have the further potential to resolve the deep divergence of seed plants. The only conflicting clade in this study remains the gnetophytes, which is notorious in almost all current phylogenomic analyses[70,135,156,198,201]. Some of our topologies, including the ones inferred from the NT123 alignment with substitution models of DNA, codons, and amino acids, as well as the coalescent based methods with exception of one ASTRAL-II analysis, all support the 'Gnetales – other gymnosperms' hypothesis with high bootstrap values. The 'Gnepine' topology is obtained by the amino acid alignment under the CAT+GTR model and the concatenated alignments of nucleotide sequences without 3rd codon positions (NT12, NT1, and NT2). The 'Gnetifer' hypothesis is only supported - with low bootstrap values - by one ASTRAL-II analysis based on GT12 and is rejected by AU tests accounting for the NT123 and NT12 alignments.

Removing 3rd codon positions in nuclear genes can change the position of gnetophytes as shown in this study and in Wickett *et al.*[70], and we found further evidence to argue that 3rd codon positions contribute to most of the compositional heterogeneity in the NT123 alignment and exhibit increase of evolutionary rates to different extents in different lineages of seed plants. Therefore, including 3rd codon positions in alignments of nuclear genes is most likely unfit for the GTR+GAMMA model and adds phylogenetic noise when dealing with the deep divergence of seed plants. Such noise may also pose problems for phylogenetic inference based on the amino acid and codon substitution models, which may explain the different observations reported by Lee *et al.*[200] and Wickett *et al.*[70]. It is worth noting that although it is computationally intensive, the CAT+GTR model is still among one of the most robust amino acid models when it comes to dealing with various phylogenetic noise. Last but not least, gene trees of the 106 phylogenetic markers indicate an inconsistent mixture of disparate phylogenetic signals on the related internode with respect to the positions of gnetophytes (Figure 2-7). The heterogeneous phylogenetic signals for the exact phylogenetic position of gnetophytes are consistent with the evolutionary history of gymnosperms, which endured several extinctions and recent radiations[198,246,247]. The lack of ancient diverged lineages in gymnosperms as well as the lack of exhaustive samples from fossil lineages may mislead current systematic studies.

With respect to the 'Gnetales – other gymnosperms' hypothesis, the 'Gnepine' hypothesis has been widely accepted when considering other molecular evidence except for molecular sequences. For example, both gnetophytes and Pinaceae lost some homologous genes in the chloroplast, such as the *rps16* gene and two introns of *clpP*[248]. Alternatively, the loss of non-homologous inverted repeats in Pinaceae and cupressophytes is not against the 'Gnepine' hypothesis[134]. Among those lost genes, the most striking example is the loss of all 11 plastid *ndh* genes in gnetophytes and Pinaceae, which is usually interpreted as a major synapomorphy for gnetophytes and *Pinaceae*[249]. However, like other plastid protein complexes, the NDH complex requires subunits encoded in both the plastid and the nucleus, so related genes would get lost coordinately. The pattern of loss of nuclear-encoded *ndh* genes is different in gnetophytes and Pinaceae, particularly for the retained *ndhS* gene in

Pinaceae[250]. Also, the loss of all plastid *ndh* genes is less likely an immediate but a continuous process, as many pseudogenes of *ndh* still exist in the chloroplast genome in extant Pinaceae[251]. Furthermore, convergent loss of *ndh* genes is not rare among seed plants. Several lineages in Orchidaceae and Geraniales also lost plastid and nuclear *ndh* genes, coordinately[250]. Therefore, the loss of *ndh* genes could be interpreted as compatible with both the 'Gnepine' or 'Gnetales – other gymnosperms' hypothesis.

Our result also confirms that *Ginkgo* and cycads form a monophyletic group, which is strongly supported by all phylogenomic topologies estimated in this study. Compared to previous studies, in which the sister relationship of *Ginkgo* and cycads depended on the presence or absence of gnetophytes and tree-building approaches used[209], our phylogenetic placement of *Ginkgo* is exceptionally solid. The gene trees of the 106 phylogenetic markers also show a definite preference for the monophyly, which is consistent with morphological traits such as haustorial pollen tube and motile sperm[198,200].

2.3.2 Limits and perspectives

We are well aware of the limitations of using draft genome assemblies and transcriptome data for the identification of single-copy genes. Single-copy gene families may suffer from the biased estimation of copy numbers due to gene predictions from draft assemblies[252] as well as artifacts of transcriptome assembly. Although transcriptome sequencing has considerably expanded our knowledge on the physiology and evolution of gymnosperms[253-256], they still often result in partial or redundant allelic transcripts, which may lead to erroneous copy number estimations because of the flawed construction of gene families. In fact, this is a more serious issue in gymnosperms than in angiosperms, because gymnosperms tend to have high heterozygosity[198], which could fail De Bruijn Graph-based assembly algorithms and leads to partial or redundant allelic transcripts[257].

Besides, the integration pipeline we used to remove redundancy can also bias copy number estimation through elimination of some recently duplicated genes. Because CD-HIT-EST collapses transcript sequences with similarities higher than 90%, not only different isoforms and allelic transcripts are removed, but possibly also some duplicated genes with high sequence similarity. However, a stringent cut-off of similarity may fail to deal with high allelic variation in gymnosperm sequences[198] and data from different samples. To a certain degree, the functional analysis of single-copy genes in seed plants resulted in similar functional categories as the single-copy genes in angiosperms[167,215] and other eukaryotes[230], suggesting the loose cut-off used here had only negligible effects.

The optimal solution to the problems described above are of course well-assembled gymnosperm genomes, but recently released conifer genomes are still extremely fragmented[217-220,258]. While the sequencing of some new gymnosperm genomes is in progress, the published ones are continuously being improved using more sophisticated assembly strategies or novel technologies, which yield longer reads and better genome assemblies[220]. All these efforts would further improve our knowledge on seed plant phylogeny, diversification, and their evolutionary history.

2.4 Materials and Methods

2.4.1 Plant material and cDNA libraries construction

Pinus pinaster seeds from the Oria provenance (Southern Spain) were germinated and grown at 20/24 °C with a 16/8 h photoperiod. Germinating seeds were watered twice a week with distilled water. One-month-old seedlings were used for cryosectioning and 0.5-cm tissue sections were processed for laser capture microdissection[259]. Tissues of *P. pinaster* were collected from cortex of hypocotyl, cortex of developing root, cortex of root, developing needle, mesophyll of cotyledon, mesophyll of new needle, pith hypocotyl, root apical meristem, shoot apical meristem, and vascular tissues of cotyledon, developing root, root, hypocotyl, and new needle. Pooled samples from needles, roots and stems from Galicia 1056xOria6 F1 progenies grown under different stress and hormone treatments were also included (Supplementary Table C-3). RNA isolation, cDNA synthesis, and construction of normalized cDNA libraries were performed following the protocol described by Cañas *et al.*[259].

Pinus sylvestris tissues represent different developmental stages during the development of zygotic embryogenesis. Zygotic embryos (E) and megagametophyte (M) samples were collected from immature cones and sorted separately into four different stages: early embryos (E1, M1), embryos at the stage of cleavage (E2, M2), dominant and subordinate embryos (E3DO, E3SU, M3) and dominant embryos before cotyledon differentiation (E4, M4) (Supplementary Table C-3). Total RNA was isolated by using the RNAqueous-Micro RNA isolation kit (Ambion) and its quality was verified by an Agilent 2100 BioAnalyzer System (Agilent Technologies) following manufacturer's instructions. Double-strand cDNA libraries were constructed by using the Mint-2 cDNA synthesis kit (Evrogen), followed by a reamplification step to incorporate the 454 pyrosequencing specific primers.

2.4.2 Transcriptome sequencing and *de novo* assembly

Transcriptome sequencing was performed using the GS-FLX+ platform with a GS-FLX Titanium kit, Roche Applied Sciences (Indianapolis, IN, USA) as described by Cañas *et al.*[259] (Supplementary Table C-3). We assembled transcriptomes of *P. pinaster* and *P. sylvestris* from the 454 sequencing reads using the Newbler software (v2.8.1). Before feeding reads to Newbler, we removed adapter sequences and reads shorter than 75 base pairs (bp) by SeqClean. Newbler then assembled all the remaining reads for *P. pinaster* and for *P. sylvestris*, until over-represented sequences were removed. CD-HIT-EST[260] then clustered the Newbler assemblies in each isogroup, which represents a unique transcriptional locus in the Newbler assemblies. In the end, we selected the longest transcript (at least 150 bp) as a unique representative for each isogroup.

In order to integrate public transcriptomes, we built an integration pipeline. SeqClean first screened the public data against the NCBI UniVec resource and retained transcripts longer than or equal to 150 bp. Next, public data was compared with the Newbler assemblies described above by CD-HIT-EST-2D[260] to add novel transcripts to our assemblies. Finally, CD-HIT-EST[260] selected a representative sequence from the clusters formed by the novel transcripts and the Newbler assemblies with 90% identity to remove redundant transcripts. For *P. pinaster*, we integrated 15,648 PlantGDB-assembled Unique Transcripts (PUTs, based on GenBank release 177)[261] and 210,513 unigenes from SustainPineDB[262]. For *P. sylvestris*, we

integrated 73,609 PUTs (based on GenBank release 187) and a set of 2,261 EST assemblies. With respect to *Picea glauca* and *Picea sitchensis*, only public transcriptomes are available, so we carried out CD-HIT-EST with 90% identity to construct non-redundant transcripts from 48,315 PUTs (based on GenBank release 175) and 27,660 FL-cDNAs[253] in *P. glauca* as well as 31,087 PUTs (based on GenBank release 183) and 13,197 EST assemblies in *P. sitchensis* [254].

We used TransDecoder (r20131117) to predict open reading frames (ORFs) in the transcripts of *P. pinaster*, *P. sylvestris*, *P. glauca* and *P. sitchensis* based on training sets built from protein-coding genes in *Picea abies*[217] and *Pinus taeda* [218]. We queried the transcripts from *P. pinaster*, *P. sylvestris*, *P. glauca* and *P. sitchensis* against the proteins from *P. abies* and *P. taeda* by BLASTX[66]. For each transcript, the complete ORF found within one High Scoring Pair was retained in the training sets. TransDecoder then used the training sets to build a Markov model and to predict ORFs with default parameters. Pfam (27.0) domains in the predicted ORFs were identified by HMMER embedded in TransDecoder.

2.4.3 Retrieval and integration of transcriptome data from public databases

We retrieved transcriptome data from another 25 gymnosperms that were stored in PlantGDB[261], oneKP[70], and TreeGenes (https://dendrome.ucdavis.edu/treegenes/). These data are fragmented and redundant, as they have been generated by different technologies and experiments (Supplementary Table C-1). To obtain a non-redundant set of transcripts for each species, we used SeqClean to remove NCBI UniVec vectors and poly-As from the downloaded transcripts. MIRA3 assembled ESTs into longer transcripts unless PUTs were available[263]. Next, we clustered transcripts in each species with 90% identity by feeding MIRA assemblies or PUTs, cDNAs, 454 assemblies, Transcriptome Shotgun Assemblies (TSAs), and oneKP assemblies to CD-HIT-EST[260], which produced a set of non-redundant representative sequences which were then further assembled by CAP3 into unigenes[264]. TransDecoder (r20131117) was applied to predict ORFs in a self-training mode, which used the 500 longest ORFs to train a Markov model for coding sequences. For angiosperms, we downloaded protein-coding genes for 34 angiosperms, one moss, and two green algae from PLAZA 3.0[107]. Green algae (*Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*) and moss (*Physcomitrella patens*) were used as outgroups in this study.

2.4.4 Identification of single-copy gene families

We started with building gene families in six conifers, *i.e.*, *P. pinaster*, *P. sylvestris*, *P. taeda*, *P. abies*, *P. glauca*, and *P. sitchensis*, because they, compared with other gymnosperms, have abundant genomic or transcriptomic data of outstanding quality. For instance, genes from *P. taeda* and *P. abies* were predicted based on genomes[217,218,258] and transcript sequences in *P. glauca* and *P. sitchens* were supplemented with Sanger reads based on BACs[253,254], while, because of their economic importance, high-coverage transcript data were generated for *P. pinaster* and *P. sylvestris* (European ProCoGen project; see www.procogen.eu for more information). Applying OrthoMCL[65] to these datasets, we obtained 32,017 multi-gene gene families comprised of 147,782 of the 259,547 input proteins (56.9%). To narrow down the search space for single-copy genes, we selected 11,152 gene families that were conserved throughout, and had low-copy number, in the six conifers. Furthermore, these gene families needed to be present in at least four of the six conifers and could have maximum two copies in two species.

To assign proteins from other species to the 11,152 gene families, we first used HMMER (v3.1b1)[222] to build an HMM profile for each of the gene families based on a multiple sequence alignment created by ClustalW (v2.1)[75] using parameters for amino acids as recommended by[265]. For every species, additional proteins were retrieved using a profile search against the HMM profiles with HMMSCAN. For each HMM profile, hits with E-values less than 1×10^{-10} were retained and their bit-scores were used to infer a cumulative probability distribution. The hits were assigned to a gene family accounting for 95% of the cumulative distribution (Supplementary Figure C-21A)[70]. Since the above described approach might fail to assign genes with similar sequences to the assigned hit at the 95% border, so we further assigned those genes to a gene family if their *E* values were similar enough (Δ*E* value < 1×10^{20}) to the hit with the smallest bit-score (Supplementary Figure C-21B).

After assigning additional genes to the initial gene families, we selected gene families according to species occurrence, *i.e.*, gene families had to be present in more than 20 (out of 31) gymnosperms and more than 30 (out of 37) species in PLAZA 3.0[107]. Afterwards, we removed gene families for which the single-copy percentage was less than 80%, which was defined as the fraction of species with exactly one copy in a gene family[215]. In the end, if more than five genes in a gene family were assigned to other gene families, we removed the gene family from further analysis. When fewer than five genes were assigned to other gene families, we reassigned these genes to the proper gene families according to the lowest E-value. Species occurrence and single-copy percentage were double checked for the modified gene families.

2.4.5 Gene Ontology enrichment analysis

Gene Ontology Slim (GOSlim) enrichment analyses were carried out by BiNGO (3.03) with a threshold of 0.01 for *P* values, which were corrected for multiple testing by Benjamini and Hochberg False Discovery Rate[266]. We used the *A. thaliana* annotation from TAIR (release 06/03/2016) and the *P. pinaster* annotation predicted by InterProScan (v5.15-54). GO terms for both species were mapped to GO slim plant by Map2Slim in OWLTools.

2.4.6 Selection of phylogenetic markers

To remove paralogs and to increase sequence sampling for phylogenetic analysis, we used the following procedure to find reciprocal best hits to select phylogenetic markers. Because HMMSCAN uses proteins to find matching HMM profiles and HMMSEARCH uses HMM profiles to find matching proteins, we carried out both of them sequentially. A pair of protein and HMM profiles was considered as each other's reciprocal best hit if they were the best match to each other. From the 1,469 single-copy genes in seed plants, we finally retained 106 such gene families that were present in 36 out of 37 species from PLAZA 3.0 and 30 out of 31 gymnosperms species for multiple sequence alignment. We used Muscle (v3.8.31) to align amino acid sequences[78] followed by trimal (v1.4) to remove low-quality alignment regions in a heuristic mode ('-automated1') and to back-translate the amino acid alignments into nucleotide sequence alignments[84].

2.4.7 Phylogenetic analyses

We employed different substitution models and partitioning strategies to reconstruct the phylogeny of seed plants. We built five sets of concatenated nucleotide sequence alignments: one with all codon positions (NT123); one with only the first two codon positions (NT12); and another three with each codon position separately (NT1, NT2, NT3). For the NT123 alignment, we partitioned it as: 1) one partition; 2) two partitions with 1st and 2nd codon positions as the first part, and 3rd codon positions as the second one; 3) three partitions with each codon positions; 4) 52 partitions by PartitionFinder (v1.1.1) given different genes and codon positions[153]. Similarly, the NT12 alignment was partitioned as: 1) one partition; 2) two partitions with 1st and 2nd codon positions; 3) 37 partitions by PartitionFinder given different genes and codon positions. RAxML (v8.2) was used to infer maximum likelihood (ML) trees based on the above-described concatenated alignments with different partitioning strategies under the GTR+GAMMA model[87]. The best ML tree was searched from optimizing every 5th bootstrap tree in 200 rapid bootstraps.

For the corresponding amino acid alignment of NT123, we first used ProtTest3 to select the best-fit model according to the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) score and the corrected AIC (AICc)[267]. The JTT+I+GAMMA+F model outperformed all the other models and was used in RAxML to search the ML tree with 200 rapid bootstrap analyses. For Bayesian reconstruction, we carried out PhyloBayes-MPI with the CAT and CAT-GTR model and a discrete gamma distribution with four rate categories. We ran two independent chains under each model and considered the chains to be converged when the 'maxdiff' parameter was less than 0.1 and the effective size greater than 300[112]. Due to limitations of computational resources, especially for the CAT+GTR model, the original amino acid alignment was trimmed by trimal with '-gt 0.9 –cons 10', followed by removing invariant sites and sequences from the two green algae.

In addition to the DNA and amino acid model, we selected the Goldman and Yang (GY) model[268] among several available codon models for the NT123 alignment, with codon frequency estimated by ML implemented in CodonPhyML (v1.0)[100]. The ratios of nonsynonymous to synonymous substitutions were drawn from a discrete gamma distribution with four rate categories. The ML tree was estimated from a BioNJ tree optimized by Nearest Neighbor Interchange and Subtree Pruning and Regrafting. Branch support values were represented by the SH-like approximate likelihood-ratio test[99] instead of traditional bootstrap values.

Two recently developed coalescent methods, *i.e.*, Species Tree estimation using Average Ranks of coalescence (STAR)[242] and Accurate Species Tree ALgorithm II (ASTRAL-II)[243], were used to infer the species phylogeny. For both coalescent analyses, we constructed a gene tree for each of the 106 phylogenetic markers by RAxML with the GTR+GAMMA model and 200 rapid bootstraps. To test the effects of 3rd codon positions, we built two sets of gene trees, one with (GT123) and the other without 3rd codon positions (GT12), for the coalescent analyses. Then the 106 gene trees were fed to STAR in an R package 'phybase' (v1.4) and ASTRAL-II (v4.10.0) to infer the species phylogeny under the multi-species coalescent model. To obtain branch support, we used bootstrap values that were obtained by bootstrapping both gene loci and the sequence alignment with 100 replicates and reconstructed 100 coalescent species trees for both analyses.

2.4.8 Estimate saturation of substitutions and Approximate Unbiased test

We determined an entropy-based index of substitution saturation (I_{ss}) for nucleotides using DAMBE5 for NT123, NT12, NT1, NT2, and NT3 alignments[240,269]. Two hundred replicates were performed with gaps treated as unknown states. Approximate Unbiased (AU) tests[237] were carried out by CONSEL (v0.20)[270] on both the NT123 and NT12 alignments with partitions by each codon position and partitions from PartitionFinder. RAxML was carried out to calculate per site log-likelihood values based on the GTR+GAMMA model[87].

2.4.9 Measurement of phylogenetic incongruence

Internode Confidence (IC) and Internode Confidence All (ICA) were estimated by RAxML with the two sets of gene trees based on the 106 phylogenetic markers[123,244]. The probabilistic and observed adjustment schemes were applied, because the gene trees contained both comprehensive and partial trees[245]. An IC/ICA value close to 1 means absence of conflicting bipartitions for a given internode, while a value close to zero suggests that incongruent bipartitions equally exist, and a value close to -1 indicates the lack of support for a given internode[244]. However, random gene trees always give (close-to) zero IC/ICA value due to the lack of phylogenetic information. To rule out possibility of the random effect, we simulated 1,000 random gene trees and compared the Robinson-Foulds distance between a species tree and the random gene trees, and the real gene trees, respectively. The gene trees of the 106 phylogenetic markers had significantly shorter Robinson-Foulds distances to the species tree than the random gene trees to the species tree (P value < 2.2×10^{-16}, Wilcoxon rank sum test), indicating that any conflicting bipartition that exists in the real gene trees is from actual phylogenetic signal.

2.5 Acknowledgements

This work was supported by the European Union Seventh Framework Programme (FP7/2007-2013) under the ProCoGen (Promoting Conifer Genomic Resources) project [FP7-KBBE-289841]; the Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks” Project of Ghent University [01MR0310W]; and the Spanish Ministry of Economy and Competitiveness [PinCoxSeq-AGL2012-35175 to M.T.C]. The computational resources (Stevin Supercomputer Infrastructure) and services used for PhyloBayes-MPI were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI.

2.6 Author contributions

Z.L., A.D.L.T., L.S., P.K.I., and Y.V.d.P. designed the study; Z.L. and L.S. performed data analysis on transcriptome assembly and phylogenomic analyses; F.M.C., C.A., J.A.C., and M.T.C. sequenced the transcrpitomes of *P. pinaster*; I.M. sequenced the transcriptomes of *P. sylvestris*; Z.L., L.S., and Y.V.d.P. wrote the paper with the assistance of the other co-authors.

Chapter 3

Gene duplicability of core genes is highly consistent across all angiosperms

Zhen Li*, Jonas Defoort*, Setareh Tasdighian, Steven Maere, Yves Van de Peer, Riet De Smet

*contributed equally

The Plant Cell 2016 **28**(2): 326-344 (2016).

Abstract

Gene duplication is an important mechanism for adding to genomic novelty. Hence, which genes undergo duplication and are preserved following duplication is an important question. It has been observed that gene duplicability, or the ability of genes to be retained following duplication, is a non-random process, with certain genes being more amenable to survive duplication events than others. Primarily, gene essentiality and the type of duplication (small-scale versus large-scale) have been shown in different species to influence the (long-term) survival of novel genes. However, an overarching view of 'gene duplicability' is lacking, mainly due to the fact that previous studies usually focused on individual species and did not account for the influence of genomic context and the time of duplication. Here, we present a large-scale study in which we investigated duplicate retention for 9,178 gene families shared between 37 flowering plant species, referred to as angiosperm core gene families. For most gene families, we observe a strikingly consistent pattern of gene duplicability across species, with gene families being either primarily single-copy or multi-copy in all species. An intermediate class contains gene families that are often retained in duplicate for periods extending to tens of millions of years after whole-genome duplication, but ultimately appear to be largely restored to singleton status, suggesting that these genes may be dosage balance-sensitive. The distinction between single-copy and multi-copy gene families is reflected in their functional annotation, with single-copy genes being mainly involved in the maintenance of genome stability and organelle function and multi-copy genes in signaling, transport and metabolism. The intermediate class was overrepresented in regulatory genes, further suggesting that these represent putative dosage-balance sensitive genes.

3.1 Introduction

Since the seminal work of Susumu Ohno[271], the importance of gene and genome duplication for evolution and adaptation has been well-appreciated. Indeed, ample examples of gene diversification following duplication have been described and 'gene duplicability', by which we mean the ability of genes to be preserved in a population following duplication, has been extensively studied[157,159,160,272-275]. Studies published on a large array of species seem to converge on the idea that some duplicated genes are more likely to be preserved in a population, and as such to potentially contribute to functional innovation, than other genes. One factor that seems to influence gene duplicability is the mode of duplication, as in several organisms that have undergone ancient whole-genome duplications (WGD) it has been shown that different sets of genes were retained following WGD and small-scale duplication (SSD) events[161-163,276-278].

Both SSDs and WGDs have occurred frequently in the flowering plant lineage, and in particular WGDs have happened at a much higher rate than in, for instance, fungi or animals[168,279]. Studying the *Arabidopsis thaliana* genome, it has been observed that certain sets of genes have almost exclusively duplicated through WGDs[162,163,277]. These genes have distinctive functional features, as they primarily encode transcription factors and components of multi-protein complexes, and are involved in development and in signaling pathways[161-163,277]. A potential explanation for this phenomenon is given by the 'gene dosage balance theory', which states that for many genes that participate in essential complex cellular networks or protein complexes, it is crucial that the stoichiometry between the gene products is maintained[165,276,278,280-282]. While WGD preserves the relative dosage between genes, the stoichiometry is disrupted when only one or few interaction partners are duplicated. In other plant species, vertebrate and unicellular organisms that have also undergone ancient WGDs, similar observations were made[278,283-286]. Hence, while gene loss following SSD is generally a relatively fast process, with average duplicate half-life estimates being in the range of a few million years[142], after WGD, a substantial set of genes is often retained in duplicate for a much longer time[162]. For instance, it is estimated that about 16% of the genes for *A. thaliana* are still present in duplicate following the most recent WGD that occurred about 49 mya (million years ago)[168], while 75% of the genes are still present in duplicate in soybean (*Glycine max*), which underwent a WGD approximately 13 mya[224]. Whether these genes will be retained indefinitely is still an unresolved question[287-289], although the lower numbers of retained genes reported for more ancient WGD events seems to suggest that, at least for a subset of genes, dosage constraints eventually get relaxed, leading to functional diversification or loss of these genes.

In stark contrast to observations of prolonged retention of a set of 'dosage-sensitive' genes are recent observations that a substantial fraction of 'core angiosperm genes', *i.e.*, genes that are present in all angiosperm genomes, occur as singletons throughout, suggesting that their duplication might be detrimental[165-167,290-292]. While these observations are not necessarily in contradiction with each other, as they likely concern different gene sets, an overarching picture that unifies the different observations regarding 'gene duplicability' is currently still missing. Specifically, the fact that most studies concerning 'gene duplicability' report species-specific patterns adds to the confusion, as genetic context, species biology, ecological

requirements at the time of duplication and the timing of the WGD event might greatly influence the observed duplicate retention patterns[293-296].

Here we undertake a large-scale comparative approach to determine whether patterns of gene duplicability can be generalized across diverse lineages. In particular, we investigate the duplicability of 9,178 core angiosperm genes identified across 37 different angiosperm genomes and covering 20 putative WGD events. For most gene families, our analyses reveal a striking non-random picture of gene duplicability, with the majority of the core genes occurring as single copies in almost all of the angiosperm genomes and a more restricted set of genes occurring in duplicate throughout. This pattern is supported by a strong functional dichotomy between both classes of gene families, with single-copy genes being involved in the maintenance of genome integrity and organelle function, and multi-copy genes being biased towards signaling, transport and metabolism. Next to these two extremes, we also identified an intermediate class of gene families that show a pattern of prolonged duplicate retention spanning several tens of millions of years following WGD, but appear to eventually also mostly return to singleton status. We hypothesize that dosage-balance constraints prolong duplicate retention in these particular gene families. Overall, we advocate that, at least for genes present in all angiosperms, the so-called core genes, selection plays an important role in the long-term preservation or non-preservation of duplicated genes, considering the highly non-random pattern that arises in this cross-species and cross-duplication event analysis.

3.2 Results

3.2.1 Core angiosperm gene families show a strong preference towards the single-copy state

We collected the protein coding sequences for 37 sequenced angiosperm genomes (Figure 3-1) and constructed gene families using OrthoMCL (see Materials and Methods). To ensure that each of these gene families traced back to a single angiosperm ancestral gene we further processed these gene families using phylogenetic tree construction followed by reconciliation of the gene trees and the species tree (see Materials and Methods). Of the 69,133 gene families that were obtained using OrthoMCL and verified by phylogenetic analysis, 9,178 belong to the angiosperm core genome, defined as that part of the genome containing genes present in all angiosperms, including the angiosperm ancestor. To accommodate for errors in genome annotation, the presence of partial genome sequences and errors in gene family construction and/or phylogenetic analysis, we allowed for gene families in this core set to be missing in up to five genomes (see Supplementary Figure D-1 for a justification of this threshold). This set of genes was used in this study for all subsequent analyses. For each gene family, we calculated the fraction of species for which the gene family contains exactly one copy, further referred to as 'Single-Copy Percentage' (SCP). For instance, a value of 0.7 means that for that particular gene family, 70% of the species examined have exactly one copy while 30% of the species have more than one copy. The distribution of the SCPs for all core gene families is depicted in Figure 3-2. As can be observed, the distribution is highly skewed towards high SCPs, with the mean of the distribution lying at 66.8% and the mode of the distribution at 87.5%. Furthermore, if we remove genomes that still have a high number of retained duplicates due to a recent (< 20 mya) WGD event (such as soybean, flax (*Linum*

usitatissimum), maize (*Zea mays*), and *Brassica rapa*, Figure 3-1), we observe an even stronger shift towards the single-copy state with the mode of the distribution being at 92.5% (Supplementary Figure D-2).

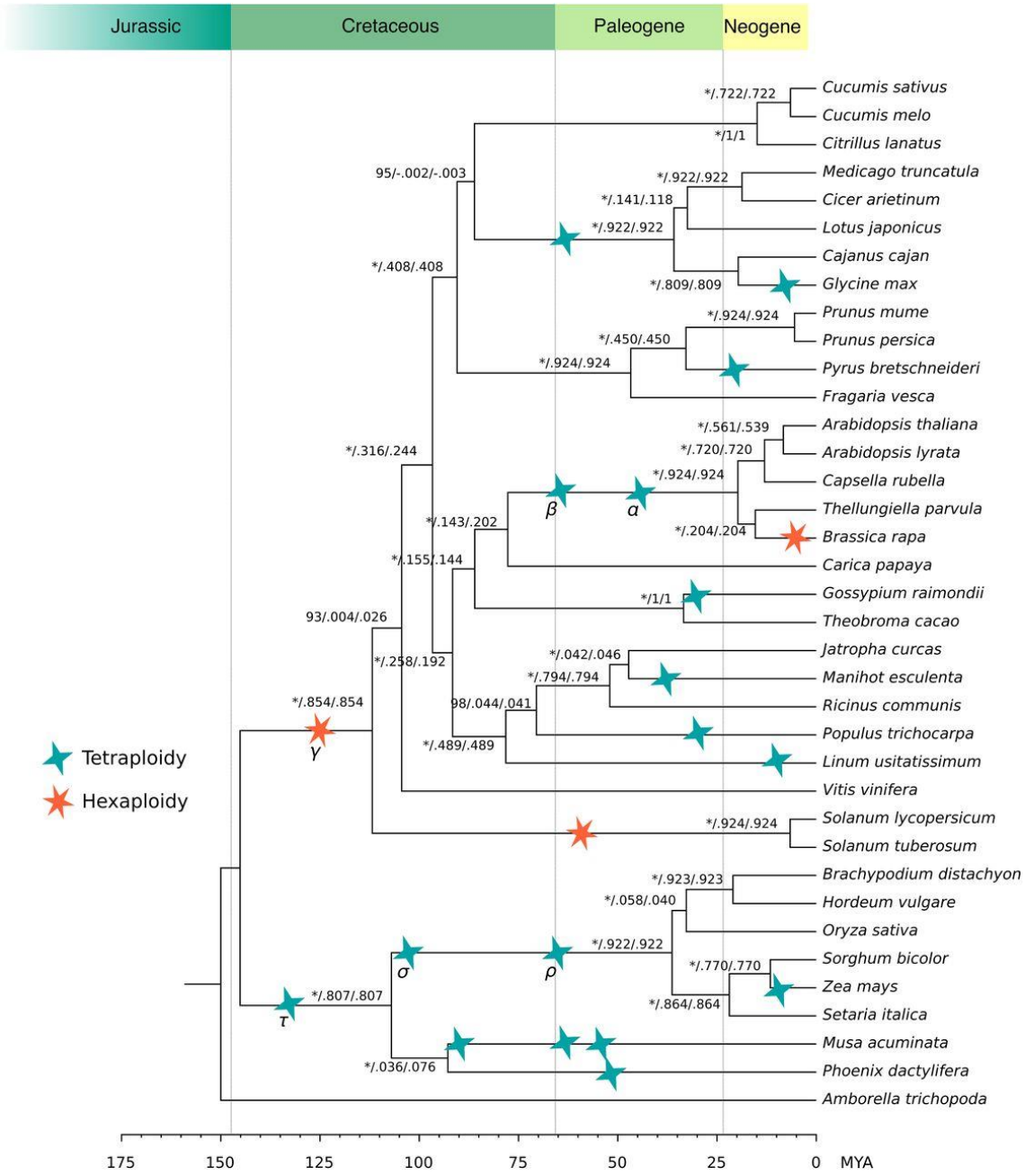


Figure 3-1 Angiosperm species tree.

Phylogenetic tree depicting the relationships amongst the 37 angiosperm genomes used in this paper. The tree topology was inferred from a concatenated alignment based on 107 almost single-copy gene families (see Materials and Methods). Numbers on the branches represent bootstrap supports, internode certainty (IC) and internode certainty all (ICA), respectively. Whole-genome duplication (WGD) events were inferred from literature[98,168] and are depicted by stars. Only WGD duplications were considered that are more recent than the angiosperm common ancestor.

Since the most likely outcome following gene duplication is duplicate loss, with average duplicate half-lives estimated at a few million years for SSDs[142], we have assessed whether our observations could be explained by simple stochastic gene duplication and loss dynamics.

Therefore, we simulated gene family copy-number evolution along the 37 species tree, using a probabilistic model in which SSD is modeled as a random birth-death (BD) process[297] and that takes into account known WGD events by assuming an instantaneous doubling (or triplication) of all genes, as in Rabier *et al.*[298] (see Materials and Methods). Using this model as a null hypothesis and using realistic rates of small-scale gene duplication and loss, λ, sampled from a normal distribution with mean μ = 0.53 and standard deviation σ = 0.156 duplications/losses per evolutionary time unit (see Materials and Methods), we generated gene counts at the leaves of the species tree for 9,178 × 1,000 = 9,178,000 simulated gene families. We observe that the SCP distribution under the null model has a mode of 22.5% on average, compared to 87.5% for the core angiosperm gene families and that both distributions are significantly different (*P* value < 2.2×10^{-16}, Wilcoxon rank-sum test) (Figure 3-2). Hence, under the neutral scenario of stochastic gene birth and death, there is no bias towards the single-copy state. We have repeated this analysis for different sampling distributions of λ-values and observed that the general trend of the distribution of SCPs for the simulated families remains similar, indicating that rejection of the null hypothesis is robust with respect to changes in the distribution of λ-values. Therefore, our observations suggest that gene families belonging to the so-called angiosperm core genome, *i.e.*, gene families present in all angiosperm genomes are skewed towards the single-copy state more strongly than expected under a random gene duplication-loss process and hence appear to be under (strong) selection to be single-copy.

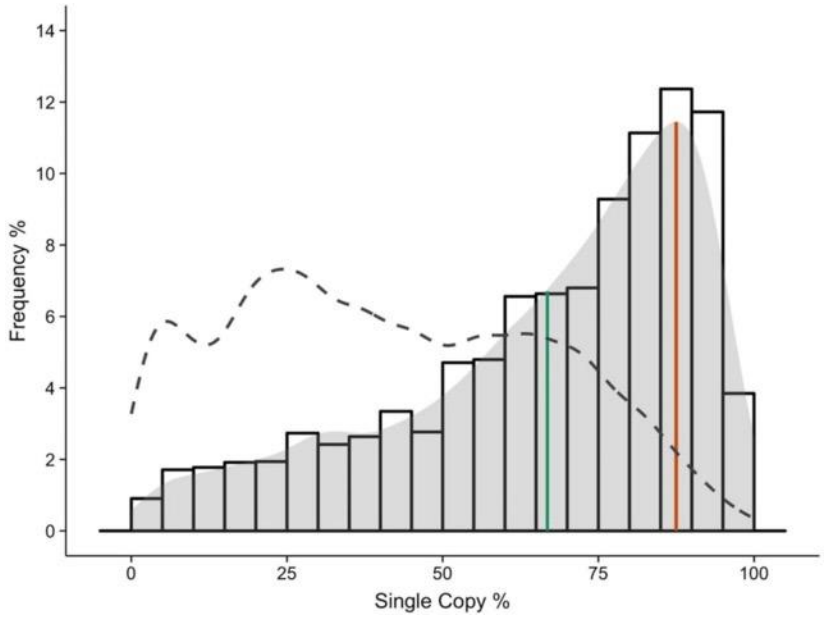


Figure 3-2 Overall distribution of single-copy percentage for all angiosperm core gene families.

The distribution depicts the degree to which the 9,178 core gene families are single-copy in the 37 angiosperm species investigated. The x-axis represents, for each gene family, the percentage of species with exactly one gene copy with respect to the total number of species in the family. The distribution illustrates a very strong tendency of angiosperm core gene families towards the single-copy state. The mode (87.5%) and the mean (66.8%) of the distribution are indicated by green and red lines, respectively. The observed distribution strongly deviates from the expected distribution under a stochastic duplicate birth-death model (depicted by dashed lines).

3.2.2 Homeologs are quickly lost following WGD

The observation that many core gene families are single-copy, in spite of the large number of both recent and ancient genome duplication events, seems to suggest that gene loss occurs relatively fast following WGD. The large number of WGD events in this study and their different ages (Figure 3-1) provide an excellent case to study duplicate retention following WGD[299].

To study the dynamics of duplicate gene retention in the core gene families, we first assessed the contribution of WGDs as compared to SSDs to duplicate retention in the core gene families. Specifically, we applied gene tree - species tree reconciliation to obtain predictions of duplication events and their associated timing for all gene families (see Materials and Methods). To this end, we classified each node in the species tree (Figure 3-1) as either being associated with WGD or SSD, based on whether WGD events have been predicted on the branch leading to the specific node (Supplementary Figure D-3). Then we compared the predicted numbers of duplication events at WGD nodes versus SSD nodes for both core and non-core gene families, the latter referring to gene families that arose more recently than the angiosperm common ancestor or that underwent massive gene loss in some species since speciation from the angiosperm common ancestor. For the core gene families, we estimated that in total 69.8% (65,531 out of 93,942 predicted duplication events) of the duplications could be attributed to WGDs, whereas for the non-core gene families this was only 34.6% (48,778 out of 140,786 predicted duplication events) (Supplementary Figure D-4). Hence, for core families, as compared to non-core gene families, the presence of duplicates seems to be biased towards WGD-associated gene duplication (P value < 2.2×10^{-16}, Fisher's exact test) (also see Supplementary Figure D-5). In further support of the hypothesis that core gene families were more heavily impacted by WGD than non-core gene families, we observed that K_S (number of synonymous substitutions per synonymous site)-based age distributions of duplicated gene pairs in the different species show clear peaks for the predicted WGD events if only duplicates from the core gene families are considered, while these peaks seemed to be absent for age distributions constructed for duplicates of non-core gene families (Supplementary Figure D-6). Hence, core gene families appear to be particularly suited to study duplicate preservation patterns following WGD.

We took advantage of the large number of WGD events and their different ages to study the dynamics of gene duplicate loss following WGDs. To this end, we assigned retained duplicates in the core gene families to the different WGD events or as being created by SSD based on a Gaussian Mixture Modelling (GMM) approach (see Materials and Methods). This way, for each species we obtained predictions of the timing (expressed in K_S-values) of the WGD events they experienced and the number of gene families with retained duplicates for each of the WGD events[178,179,300] (see Materials and Methods). We used these data to assess the relationship between the number of gene families with retained duplicates and the estimated timing of the WGD events. As can be seen in Figure 3-3, duplicate retention subsequent to WGD follows an L-shaped curve that can be approximated by a power-law function (see Materials and Methods), confirming common expectations that gene loss subsequent to WGD is initially fast and then slows down. A similar power-law pattern was recently also observed in a genome-wide analysis of duplicate retention following WGD for a more restricted set of genomes[299]. For ease of interpretation, we grouped the WGD events into three different sets according to the overall time frame during which the WGD event occurred. ‘Ancient’ refers

to the WGD events that have been predicted to have occurred at least 75 million years ago (Figure 3-1). This includes the ancient γ WGD event that is shared by all eudicots and the σ WGD event that is shared by the Poaceae. Using the mixture modelling approach, we could not find support for the predicted ancient τ event that is shared by all monocots [98]. 'K-Pg boundary' refers to WGD events situated at approximately the K-Pg (Cretaceous-Paleogene) boundary, which reflects a clustering of WGD events at approximately 50-70 mya [168]. Finally, the 'recent WGD' set includes the duplication events that are more recent than the K-Pg boundary (< 50 mya). In Figure 3-3, duplicate retention patterns associated with the 'recent WGD' events show a steep decline as a function of WGD age. Whereas on average 41.64% (SD 21.74%) of the core gene families retain duplicates for the recent WGD events, for the 'K-Pg boundary' WGDs the number of core gene families with retained duplicates has dropped to on average 16.04% (SD 7.48%), and for the 'Ancient set' this number further reduces to 8.37% on average (SD 2.24%).

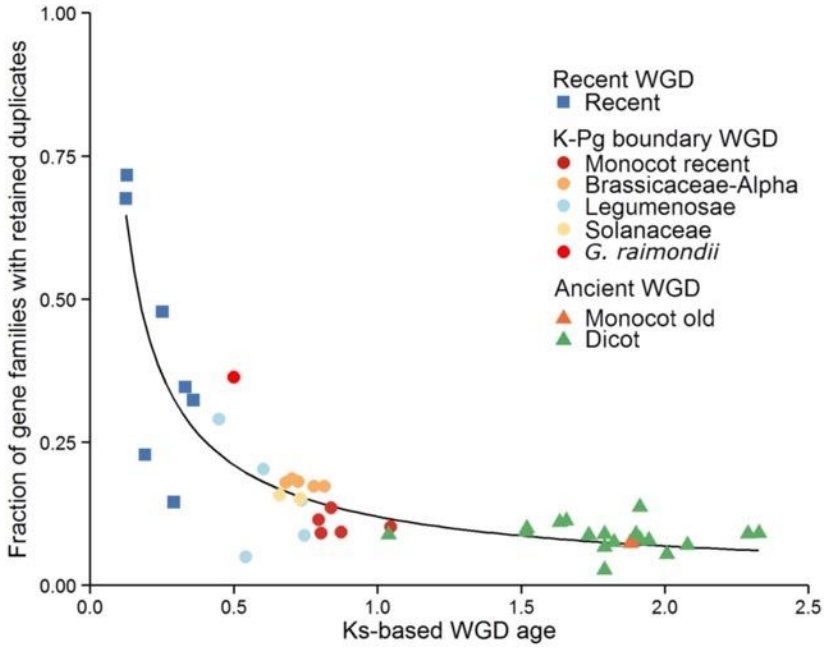


Figure 3-3 Duplicate gene retention in function of time since WGD.
Each dot represents the fraction of core gene families with retained duplicates following a specific WGD (y-axis), as a function of WGD age, expressed in K_S-units (x-axis). The timing of the WGD events and the particular gene families that retained duplicates following a specific WGD event were inferred by fitting Gaussian mixture models to K_S-age distributions for all 37 species separately (see Materials and Methods). As such, each point represents a species-specific estimate for a WGD and WGD events shared by multiple descendant species will be represented by multiple data points that cannot be regarded as being independent. SSD-related peaks and dubious WGD peak callings were omitted. Additional information on all the peaks can be found in Supplementary Table D-2 and Supplementary Figure D-7. A power-law function was fitted to the data (χ^2 goodness-of-fit = 0.77, p = 1).

The distinction between SSD and WGD duplicates in this paper are approximate and SSD numbers are likely underestimated by both strategies (GMM and reconciliation method), because some SSDs might be located on a WGD branch (gene tree – species tree reconciliation) or might be hidden under a WGD peak (GMM analysis). However, we do not expect this to have a large influence on the observations that core gene families in contrast

to non-core gene families are mainly duplicated by WGD nor on observed differences in gene duplicability patterns for different gene family groups (see further), as this underestimation likely affects all gene families equally.

3.2.3 Core gene families belong to different groups that reflect major differences in gene duplicability

Our global analyses on duplicate retention following WGD show that the majority of the angiosperm core gene families revert quickly to the single-copy state following WGD. Yet, the distribution in Figure 3-2 suggests that certain gene families revert faster to single-copy status than others. Therefore, we explored gene family specific differences in duplicate retention by constructing a copy-number profile matrix, which for each gene family lists the number of genes for a given species. We classified gene families into different groups based on an unbiased clustering of their copy-number profiles. By using a sub-sampling strategy in combination with clustering[301] (see Materials and Methods) we found that the data are best described by three stable clusters (Figure 3-4A, Supplementary Figures 8 and 9): Group 1 contains 5,097 gene families and covers 5,473 *A. thaliana* genes, Group 2 contains 2,832 gene families and covers 4,312 *A. thaliana* genes and Group 3 contains 1,249 gene families and covers 3,255 *A. thaliana* genes. The heatmap in Figure 3-4A clearly shows the overall tendency of gene families in Group 1 to occur as single copies. If duplicates are present these are mainly biased towards species with recent WGDs. Gene families within Group 2 show mainly duplicate retention for species that are associated with 'Recent' and 'K-Pg Boundary' WGDs, while being largely single-copy for species that only underwent 'Ancient' WGDs. The latter suggests that while duplicates for these gene families are in general preserved for prolonged times, they eventually largely return to single-copy status. Finally, gene families in Group 3 have retained duplicates for all species, also for the ones that only underwent 'Ancient' WGDs. We also observe that the outgroup species *Amborella trichopoda,* which has no evidence of WGDs postdating angiosperm diversification[170], seems to be singleton for most of the core gene families, further substantiating the above observations that core gene families mainly duplicate through WGDs. Investigating the SCPs for the gene families in the three groups confirms that gene families in the first group show a strong preference towards the single-copy state, whereas gene families in the third group represent gene families with a strong tendency to be multi-copy in the majority of the species. The SCP distributions for each of the three groups are significantly different (P value < 2.2×10^{-16} for all comparisons, Kruskal-Wallis test followed by Dunn's test with Benjamini-Hochberg multiple testing correction) and there is almost no overlap in SCPs for Group 1 and Group 3 (Figure 3-4B). We will further refer to the gene families in Group 1 as 'Single-copy', those in Group 2 as 'Intermediate' and those in Group 3 as 'Multi-copy'.

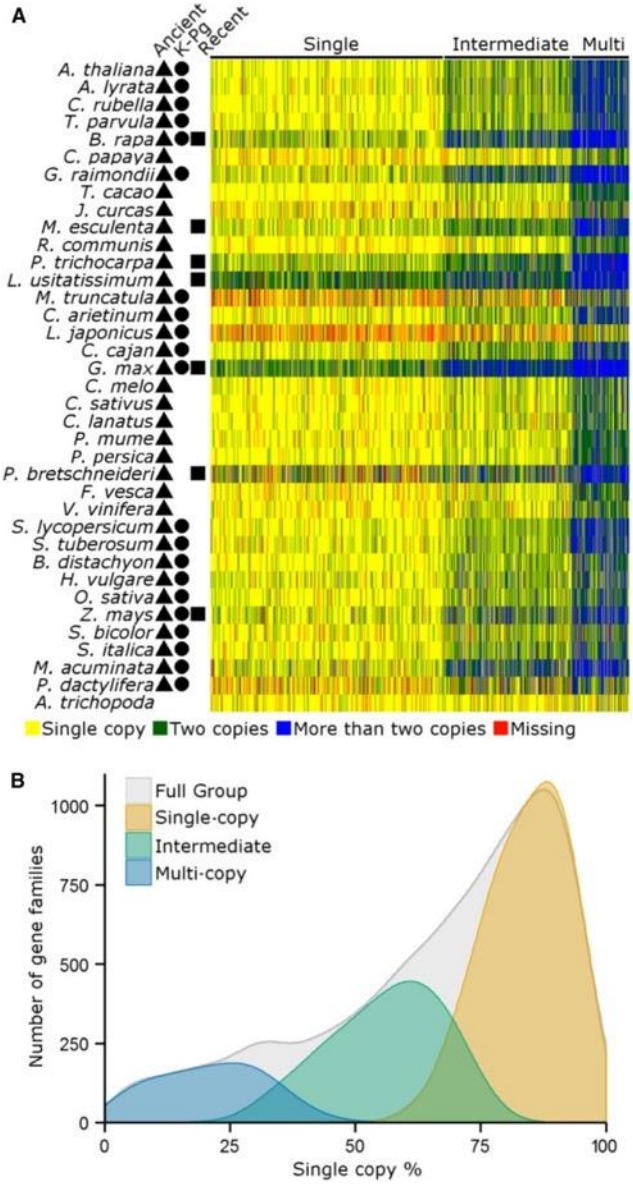


Figure 3-4 Core gene families partition into three groups based on clustering of the copy-number profile data.
(A) Heatmap of the clustered copy-number profile matrix. Rows represent species and columns represent the core gene families. Gene families (columns) are sorted according to the three different groups obtained by k-means clustering. Symbols indicate for each species whether WGD events that might have contributed to duplicates in the species fall into the 'Recent' (rectangle), 'K-Pg boundary' (circle) or 'Ancient' (triangle) category. (B) Single-Copy Percentage distributions for the gene families in each of the three different groups. The 'Cumulative' distribution shows the SCP distribution of all core gene families together (cfr. Figure 3-2).

Whereas the analyses described above clearly show differences in duplicate retention patterns for the different gene families, it does not provide direct information on the origin of the retained duplicates: e.g. are duplicates in the Multi-copy group also more ancient than those in the other two groups or is the increased number of species with duplicates in the Multi-copy group mainly due to recent lineage-specific expansions? Therefore, we investigated whether the copy-number patterns observed in Figure 3-4 are related to different ages of retained duplicates in the three groups by using duplication age predictions obtained by GMM of K_S-based age distributions and gene tree - species tree reconciliation (see Materials and Methods). The former approach (GMM modeling) provides us with species-specific estimates of duplication ages expressed on continuous time scales (K_S - values), whereas the latter approach (reconciliation) gives estimates of the absolute counts of duplication events on a gene family base. Hence, the GMM approach provides multiple estimates of duplicate retention per WGD for events with multiple descendant species, since the modeling is performed in a species-specific manner and as such predictions for the same event are obtained for the species separately. These predictions are not necessarily

independent since gene losses following duplication might have predated speciation. However, since K_S-values and also their distributions are not always comparable between species[302], the multiple estimates obtained for the same event in different species could not be collapsed. We used the GMM approach to study duplicate retention dynamics over time for gene families in the three different groups, similarly as we did above for the full set of core gene families (Figure 3-3). Overall, when comparing numbers of retained duplicates for the core gene families in function of the WGD ages we observe that gene families in the three different groups differ markedly in their duplicate retention dynamics over time (*P* value < 9.2×10^{-6} for all comparisons, Kruskall-Wallis test followed by Dunn's test with Benjamini-Hochberg multiple testing correction) (Figure 3-5A). In particular, we observe higher duplicate retention for all WGD event classes (*i.e.*, for 'Recent', 'K-Pg Boundary' and for 'Ancient' WGD events) for the core gene families in the Multi-copy group, whereas the proportion of core gene families in the Single-copy group with retained duplicates is consistently lower (Figure 3-5A). Next, we used the gene tree – species tree reconciliation approach to obtain absolute counts of predicted duplications and their corresponding ages for all core gene families and used this data to identify group-specific differences in duplicate retention for specific duplication age classes as compared to the full set of core gene families (Figure 3-5B). This shows that gene families in the Single-copy group seem to be specifically biased towards duplicates from the 'Recent' WGDs (*P* value = 3.55×10^{-137}, Fisher's exact test with Bonferroni multiple-testing correction), while duplicates from the 'K-Pg boundary' (*P* value = 5.79×10^{-83}, Fisher's exact test with Bonferroni multiple-testing correction) and 'Ancient' (*P* value = 6.36×10^{-98}, Fisher's exact test with Bonferroni multiple-testing correction) events are underrepresented. Duplicate retention for gene families in the Intermediate group is biased towards the 'K-Pg boundary' events (*P* value = 5.05×10^{-5}, Fisher's exact test with Bonferroni multiple-testing correction). Multi-copy gene families are enriched for duplicates from the 'Ancient' events (*P* value = 2.09×10^{-50}, Fisher's exact test with Bonferroni multiple-testing correction), while showing a deficit in duplications from the 'Recent' events (*P* value = 1.81×10^{-73}, Fisher's exact test with Bonferroni multiple-testing correction). SSDs are underrepresented in the Intermediate group (*P* value = 1.65×10^{-23}, Fisher's exact test with Bonferroni multiple-testing correction), while being overrepresented in the Multi-copy group (*P* value = 1.50×10^{-22}, Fisher's exact test with Bonferroni multiple-testing correction). A comparison of the relative number of duplications obtained for each duplication age class based on gene tree – species tree reconciliation and GMM of K_S-based age distributions provide consistent results (Supplementary Figure D-10). Despite these differences in duplicate retention for the three groups, all groups have retained more duplicates from the 'Recent' events, followed by the 'K-Pg boundary' and the 'Ancient' events (Figure 3-5A, B).

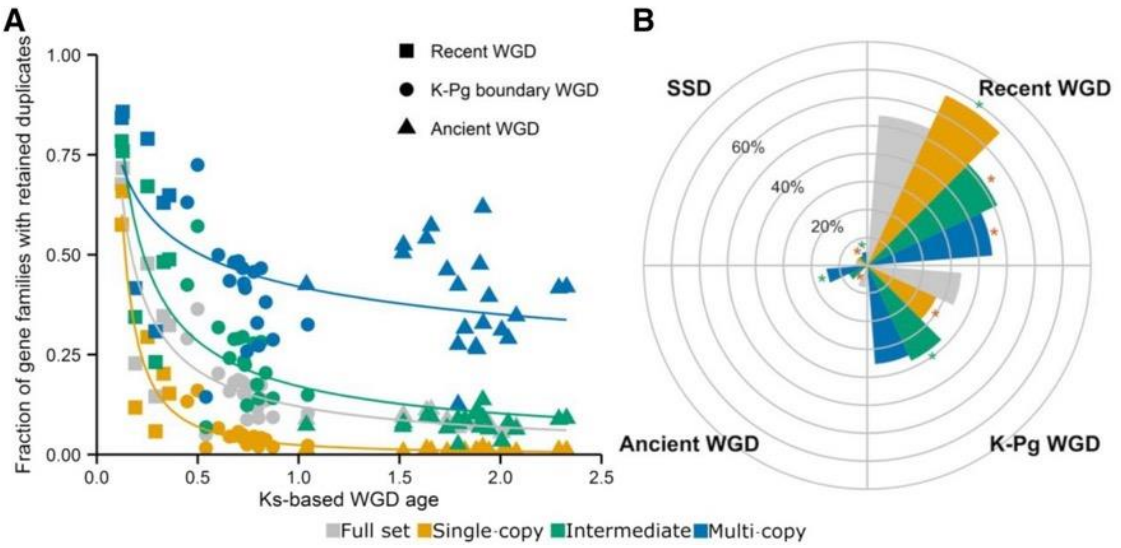


Figure 3-5 Analyses of duplication events of the three groups.
(A) For each of the clusters in Figure 3-4, power-law functions were fitted to the corresponding data points representing the fraction of core gene families with retained duplicates following a particular WGD (y-axis) as a function of WGD age (x-axis), as in Figure 3-3 (χ^2 goodness-of-fit Single-copy group = 0.52, p = 1; χ^2 goodness-of-fit Intermediate group = 1.38, p = 1; χ^2 goodness-of-fit Multi-copy group = 1.83, p = 1). The 'Full Set' curve corresponds to the curve represented in Figure 3-3. (B) Polar diagram depicting the fraction of duplication events in each gene family group belonging to either 'Recent', 'K-Pg boundary', 'Ancient' WGDs or 'SSD' events. Here, predicted duplication events were inferred based on gene tree-species tree reconciliation. Green and red asterisks denote statistically significant over- and underrepresentation, respectively, of duplicates of a certain class for a specific group, comparing each time the number of associated duplications for each group with that of the full set (grey bar) by Fisher's exact test. Similar results were obtained by using predicted duplication events inferred using Gaussian mixture modeling of K_S-distributions (Supplementary Figure D-10).

3.2.4 The partitioning in different groups is mirrored by gene function

We conducted a GOSlim enrichment analysis of the *A. thaliana* genes in the three different groups, revealing that the three different groups have a remarkably different functional composition (Figure 3-6A). The 'Single-copy' group is enriched for genes that function in organelles (e.g. 'mitochondrion', 'thylakoid' and 'photosynthesis') and that have to do with the maintenance of DNA repair and integrity (e.g. 'DNA metabolic process' and 'nucleobase-containing compound metabolic process'). An independent analysis of 2,090 nuclear-encoded chloroplast-targeted genes taken from The Chloroplast Function Database [303] supported the overrepresentation of genes with chloroplast-associated functions in this particular group (*P* value = 1.1×10^{-59}, Fisher's exact test with Bonferroni multiple-testing correction). No such overrepresentation was found for the 'Intermediate' and 'Multi-copy' groups (Supplementary Figure D-11). The 'Intermediate' group is enriched for genes that are involved in development ('multicellular organism development') and growth and regulation of transcription ('transcription factor activity' and 'chromatin binding'). This last observation was confirmed by an independent analysis of 1,795 putative transcription factors in *A. thaliana*[304], which showed that these genes were clearly overrepresented in the 'Intermediate' group (*P* value = 4.8×10^{-17}, Fisher's exact test with Bonferroni multiple testing correction) while not being enriched for the 'Multi-copy' group and being underrepresented in the 'Single-copy' group (Supplementary Figure D-12). The overrepresentation of regulatory genes in this group, together with the longer retention times for these gene families, suggests that this group mainly consists of dosage-balance sensitive genes[162,165,280,305]. We further

investigated this hypothesis by assessing the extent to which genes within this group are involved in protein interactions[276] and the contribution of WGD to duplicate retention for this specific group[162,163,276], which represent two characteristics, other than functional overrepresentation, associated with dosage-balance constraints. First, we observed that *A. thaliana* interacting protein pairs (see Materials and Methods) are indeed most overrepresented in the 'Intermediate' group, yet these results are only borderline significant following multiple testing correction (*P* value = 0.01, randomization test with Bonferroni multiple testing corrections) (Supplementary Table D-1). Second, while all core gene families duplicate preferentially by WGD, the 'Intermediate' group has a higher fraction of WGD-associated duplicates versus SSD-associated duplicates as compared to the 'Single-copy' group (*P* value = 2.96×10^{-17}, Fisher's exact test with Bonferroni multiple-testing correction) or 'Multi-copy' group (*P* value = 2.76×10^{-61}, Fisher's exact test with Bonferroni multiple-testing correction), as derived from the gene tree – species tree reconciliation predictions, strengthening our belief that the 'Intermediate' group contains dosage balance-sensitive gene families. Finally, 'Multi-copy' gene families are enriched for genes that appear to be involved in the interaction with the environment ('signal transduction', 'transport' and 'cell wall'), translation, and different metabolic processes ('carbohydrate and protein metabolic process', 'biosynthetic process' and 'catalytic activity').

We also analyzed a dataset that describes loss-of-function phenotypes for 2,400 *A. thaliana* genes[306] of which 1,521 are present in the core gene set. Genes within this dataset are placed in four different groups according to their knock-out phenotype. We find that the three core angiosperm groups show markedly different signatures with regards to their classification into LOF phenotype groups (Figure 3-6B). In particular, genes in the 'Single-copy' group are enriched for the 'Essential' category (*P* value = 1.44×10^{-39}, Fisher's exact test with Bonferroni multiple-testing correction), consisting of genes that are essential for early development and survival. On the other hand, essential genes are underrepresented in the 'Multi-copy' group. This is agreement with recent observations that lethal genes in *A. thaliana* usually lack duplicates in this particular genome[307]. Noteworthy, overrepresentation of essential genes in the 'Single-copy' group is not specifically due to the genes involved in DNA integrity within the single-copy set, but also organelle genes are associated with essentiality[306]. The 'Intermediate' set is enriched for genes of the 'Morphological' class (*P* value = 6.96×10^{-5}, Fisher's exact test with Bonferroni multiple-testing correction), which contains genes associated with clear morphological phenotypes, involved in reproduction and timing (e.g. flowering time, senescence), in agreement with the strong overrepresentation of developmental genes in this particular group. Finally, the 'Multi-copy' class is overrepresented for genes in the 'Cellular and Biochemical' group, *i.e.*, genes functioning in metabolism, or other biochemical pathways or showing phenotypic effects at the cellular level (*P* value = 1.14×10^{-6}, Fisher's exact test with Bonferroni multiple-testing correction) and 'Conditional' class, *i.e.*, genes that respond to biotic and abiotic stress (*P* value = 6.84×10^{-4}, Fisher's exact test with Bonferroni multiple-testing correction), consistent with GOSlim enrichment results. In summary, both the GOSlim enrichment analysis and the analysis of loss-of-function phenotype data indicate that the separation of core gene families into three different groups according to gene duplicability is mirrored by a separation of the gene families in the space of gene functions.

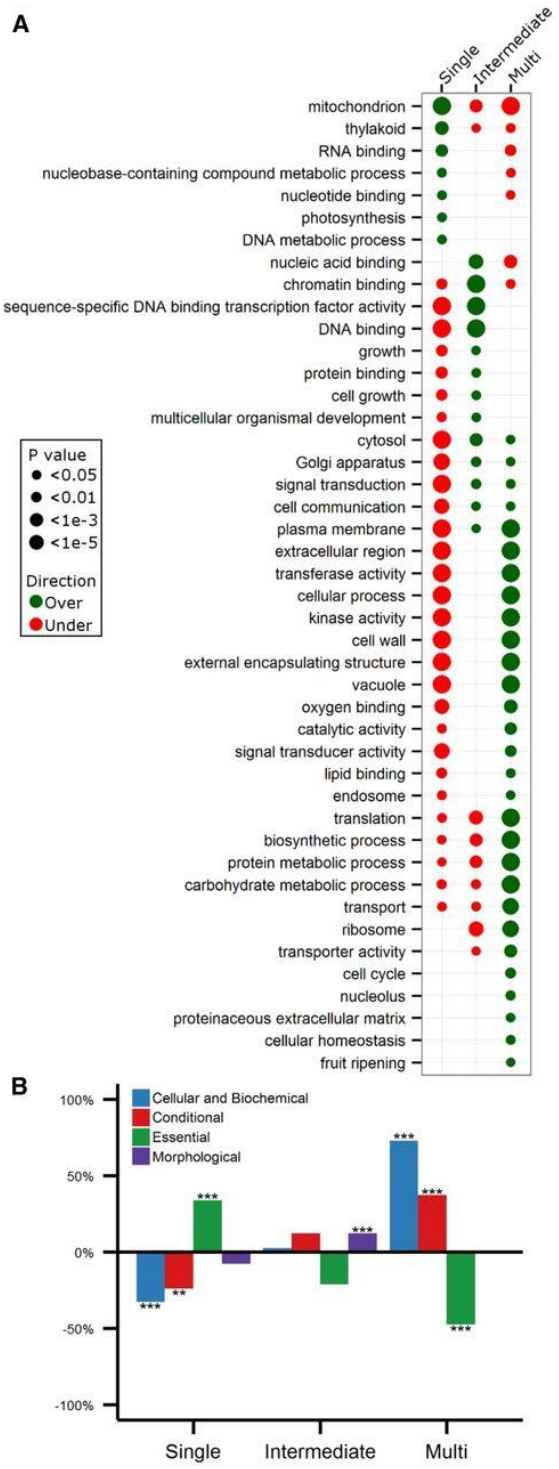


Figure 3-6 Functional analyses of the three different groups.

(A) GOSlim enrichments and underrepresentations calculated for the *A. thaliana* genes in each of the three gene family groups in Figure 3-4. Dot sizes are representative for the statistical significance of over- (green) or underrepresentation (red). (B) Enrichment analysis of the three gene family groups for knock-out mutant phenotype annotations[306]. Bars represent overrepresentation (positive values) or underrepresentation (negative values) of knock-out phenotypes belonging to any of four functional categories (bar colors). Asterisks denote significance levels as calculated by Fisher's exact test (***: *P value* < 0.001, **: *P value* < 0.05).

3.3 Discussion

We assessed duplicate retention patterns for 9,178 core angiosperm gene families (*i.e.*, gene families shared by all angiosperm species) in 37 angiosperm genomes, covering 20 putative WGD events. Assessing the retention of duplicated genes across such a large number of genomes and duplication events allows for replicated tests of gene duplicability, mitigating potential biases due to differences between individual species and WGDs[293-296]. In addition, because of the varied age range of the WGD events in our dataset and the observed large

contribution of WGD to the expansion of core gene families, we were able to compare duplicate retention patterns across WGD events of different ages.

We observe that gene duplicability is highly consistent across angiosperm genomes, with over 50% of the core angiosperm genes reverting quickly to single-copy status following duplication, whereas a much smaller set seems to occur in multiple copies throughout. An intermediate group is formed by putative dosage-balance sensitive genes that are maintained in duplicate for prolonged periods of time, but eventually mostly return to single-copy status. By showing that there is a clear distinction between genes that generally occur as a single-copy throughout and genes that show prolonged duplicate retention in the genome or that are retained 'indefinitely' following WGD, we reconcile previous observations on high numbers of single-copy genes shared across multiple angiosperm genomes, despite the many, often nested, WGD events they experienced[166,167,290,292], with observations that duplicates can be retained for long periods following WGD[162,163]. Previous, smaller-scale comparisons of duplicate retention following WGD in multiple plant species have observed strong differences between species[293,295]. These differences do most probably exist, yet, by focusing on a large number of species and a large number of WGD events we were able to retrieve dominant and striking patterns of gene duplicability that have remained concealed in smaller-scale comparisons. As our study only focused on core gene families, it is possible that important differences between species result from duplicate retention patterns in gene families that were not considered in this analysis. In addition, while here we showed that the overall duplicate retention tendency seems to be highly consistent across a large number of species and duplication events for the angiosperm core gene families, further detailed cross-species exploration of duplications in both core and non-core angiosperm gene families might reveal other parallelisms in duplicate retention that have remained concealed in this work. For instance, other works have shown that the mode of SSD (primarily tandem versus transposition-duplication) is also preserved cross-taxon for certain gene families[308-310].

We found that gene duplicability is highly associated with gene function, with single-copy genes being biased towards essential genes, functioning in genome integrity pathways and organelles, and multi-copy genes being biased towards functions involved in interactions with the environment. An evaluation of duplicate gene loss and retention patterns following the three successive WGDs in *A. thaliana* uncovered similar correlations between duplicate retention pattern and gene function as the ones observed here[162]. Here, we show that these function-retention patterns can be generalized across a large number of angiosperm genomes and WGD events. In addition, these patterns appear not to be limited to the plant kingdom: in a study focusing on the duplication history of genes across 17 ascomycete genomes, a similar functional separation was observed between genes that generally occur in duplicate and those that are single-copy in most ascomycetes[311]. Likewise, a large-scale analysis of prokaryotic genomes suggested that the number of genes functioning in DNA repair and replication remains relatively constant irrespective of genome size, whereas the number of transcription factors, genes involved in signaling and transporter genes, seems to increase with increasing genome size[312,313]. Consequently, patterns of duplicate retention and loss for core genes in angiosperms and other organisms appear to abide by general function-based rules.

The question remains what causes these specific duplication patterns to occur. Given the overall short half-lives of duplicate genes[142], one could speculate that the observed high

fraction of single-copy gene families and a more limited number of multi-copy gene families are caused by a stochastic gene duplication and loss process. We tested this hypothesis and found that stochastic birth-death processes cannot reproduce the observed duplicability distribution, which is heavily skewed towards single-copy gene families. In addition, the observed overall consistency of patterns across genomes and across large-scale duplication events and the functional enrichments observed for the various duplicability classes of gene families argue against such a random scenario. Considering the strong association with gene function, a possibility is that gene function directly or indirectly constrains gene duplicability. The observed patterns of gene duplicability are indeed consistent with the idea of the existence of a conserved core, that needs to remain untouched ('Single-copy' group), and the existence of processes that are more amenable to modifications and that might be responsible for adaptations to new environments and the evolution of distinct morphological features ('Multi-copy' group)[314]. Gene duplication in itself can indeed modulate gene function in a negative way and as such impact core gene function, by for instance increasing absolute gene dosage of genes with strict gene expression constraints[315], through the accumulation of mutations in duplicate copies with potential pleiotropic negative effects on wild-type fitness[167,316-318] or potential cytotoxic effects (e.g. protein misfolding)[319]. As a result, duplicates of genes sensitive to these processes might be eradicated quickly, also after WGD. On the other hand, repeated biased retention of certain duplicates for long periods of time ('Intermediate' group) or indefinitely ('Multi-copy' group) suggests a mechanism of duplicate retention other than sub-/neofunctionalization, which are in general assumed to be slow processes[320] and would not be expected to lead to repeated biased retention. Considering the primary role of WGD in duplicate retention of the core genes and the specific association of gene functions enriched in the 'Intermediate' and 'Multi-copy' group with previously defined putative dosage-balance sensitive genes[162,163], we hypothesize that dosage-balance constraints may have contributed to the prolonged retention of duplicate genes in these sets. Prolonged retention of duplicate genes, accompanied by gradual circumvention of dosage balance constraints, may increase the possibility that duplicate genes diversify and get permanently preserved[282,296]. Alternatively, duplicate genes could also be permanently retained through absolute dosage constraints replacing over time the relative dosage-balance constraints responsible for initial duplicate retention[296,321]. In our results, the 'Intermediate' group of gene families exhibits the hallmarks of dosage-balance constraints that wear off over time, leading to prolonged preservation and ultimately loss of duplicates. A subset of genes in the 'Multi-copy' group may also have been retained initially because of dosage-balance constraints and, in this instance, preserved indefinitely through other mechanisms; in particular transporters, signaling transducers and cell communication genes have been reported earlier as potentially dosage balance-sensitive[162,163]. On the other hand, the 'Multi-copy' set of gene families is also enriched in 'environmentally responsive' genes. Consequently, their repeated and biased retention following WGD might be a consequence of an increased adaptive advantage of polyploidy under environmental stress. Indeed, increasing evidence suggests that polyploids show wider environmental tolerance and higher levels of phenotypic plasticity than diploids[175,322-327]. In particular transporters and metabolic genes, enriched in the 'Multi-copy' class, have been identified before as putative driver genes explaining the increased tolerance of polyploids for environmental stress[326,328-330]. Despite the strong correlation between gene duplicability and gene function observed here, it remains to be further investigated which evolutionary mechanisms are responsible for the observed strong bias in duplicate retention patterns, and it remains to be established whether

gene function directly influences gene duplicability or whether biased gene retention could be a by-product of other evolutionary phenomena instead, such as for instance the preservation of intermolecular interactions (dosage balance) or sequence constraints related to high levels of gene expression[272,331]. In particular, since network structure is often believed to constrain protein evolution and to underlie complex phenotypic traits, future work into this direction might benefit from investigating gene duplicability in a network context (e.g. Bekaert *et al.*[321], D'Antonio *et al.*[332], Alvarez-Ponce *et al.*[333], Chae *et al.*, and Conant *et al.*[296])

3.4 Materials and Methods

3.4.1 Genome data

We employed protein-coding genes from 37 fully sequenced angiosperm genomes, 35 of which were used in Vanneste *et al.*[168]. Protein-coding sequences for *Amborella trichopoda*[170] and *Capsella rubella*[334] were retrieved from the Amborella Genome Database (http://www.amborella.org/) and Phytozome V10, respectively.

3.4.2 Gene family prediction

3.4.2.1 OrthoMCL

We identified gene families based on protein sequence similarities by OrthoMCL[65]. After all-against-all BLASTP searches, OrthoMCL was used to group proteins with high sequence similarity into gene families. An important parameter of OrthoMCL is the inflation parameter, which controls cluster tightness. We calculated gene families for different inflation parameter values (*i.e.*, 1.5, 2.0, 2.5, and 3.0) to assess its influence, and observed large variations in the number of gene families detected and their overall size. We decided to use the inflation parameter that gives on average the largest gene families (*i.e.*, 1.5), since the gene families are further processed by phylogenetic tree construction (and split up if necessary, see below). As such we obtained 69,133 multi-gene families.

3.4.2.2 Species tree construction

A species tree was constructed from a concatenated multiple sequence alignment inferred from 107 gene families that are present in all of the 37 angiosperm species and contain no more than 40 genes in total. The genes within these 107 gene families are on average longer than 150 amino acid residues. If a species had paralogs in a gene family, we only kept the paralog with the most orthologous hits in the gene family in the intermediate OrthoMCL results file. We used MUSCLE (3.8.31)[78] with default parameters to perform multiple sequence alignments for each gene family based on the amino acid sequences. We then used trimal (1.4) to remove low quality regions of the alignments based on an automatically selected threshold (-strictplus), which depends on a distribution of residue similarity inferred from multiple sequence alignment for each gene family[84]. Multiple sequence alignments of amino acid sequences were back-translated into alignments of codon sequences and were concatenated one by one into an integrated alignment. In the end, we obtained an alignment of 36,631 codons with 109,893 nucleotide sites.

To construct the species tree, we used CodonPhyML (1.0)[100] under three different codon models that differ in their instantaneous substitution rates between codons, being the Muse and Gaut (MG) model[335], the Goldman and Yang (GY) model[268] and the YAP model[336]. The stationary frequency of codons and the transition-transversion ratio were estimated by maximum likelihood. The different ratios of nonsynonymous to synonymous substitution rate (ω) over the sequence alignment were drawn from a discrete gamma distribution with three, four, or five classes. The parameters α and β of the gamma distribution were optimized by maximum likelihood. An initial tree was built using the BioNJ algorithm, based on the empirical model ECMK07. CodonPhyML then employs Nearest Neighbor Interchange (NNI) and Subtree Pruning and Regrafting (SPR) to optimize the tree topology. Branch lengths and model parameters are also fully optimized during this process.

Based on the different codon models and parameters described above, we obtained nine phylogenetic trees with identical topology but with slightly different branch lengths. The branch lengths of the different trees have no effects on the phylogenetic placement of WGDs (see 'Evolution of gene families under a stochastic birth-death null model' and Supplementary Figure D-18). We used the Akaike Information Criterion (AIC) to compare likelihoods for the different trees and selected the tree with the lowest AIC tree as the species tree in this study. This tree corresponds to the tree inferred under the MG model with five classes for ω.

We calculated bootstrap support values for all branches of the species tree by obtaining 100 bootstrap samples for the concatenated multiple sequence alignment and running CodonPhyML on each bootstrapped alignment using the same model and parameter settings as chosen for the species tree. The bootstrap values were added on each branch of the species tree by RAxML[87]. As an alternative support measure to the bootstrap we assessed the degree of congruence between the species tree topology and the topology of the 107 gene trees, also obtained using codonPhyML with the same parameter settings, for the gene families used for species tree construction. Specifically, using RAxML, we calculated two measures: (1) internode certainty (IC) and (2) IC All (ICA) that evaluate the support for an internode in the species tree by considering its frequency in the set of 107 gene trees[123,244]. An Internode Certainty value of one means that none of the gene tree topologies conflict with the species tree topology, whereas a value close to zero for internodes suggests that there is another possible bipartition that occurs with almost equal frequency to the inferred one. In the end, the species tree was rooted on the branch of the basal angiosperm species *A. trichopoda* and was visualized by FigTree (http://tree.bio.ed.ac.uk/software/figtree/). This obtained species tree is largely consistent with the APGIII tree[337] (see Supplementary Figure D-13 for a comparison).

3.4.2.3 Gene tree construction and reconciliation

Next, we implemented a pipeline to automatically construct phylogenetic trees for all 69,133 gene families and to test whether these trees could be traced back to a single angiosperm ancestral gene. We first removed 253 gene families with more than 200 genes because of the enormous computational resources required by large gene families. Then we built maximum likelihood phylogenetic trees for each of the remaining gene families with more than two genes. Multiple sequence alignments based on protein sequences were produced using MUSCLE with default settings[78] and were further trimmed by trimal in a heuristic automated approach (-automated1)[84]. The processed multiple sequence alignments were fed into PhyML

3.0[99] using the LG model with the equilibrium frequencies defined in the substitution model. The best trees produced from either Nearest Neighbor Interchange or Subtree Pruning and Regrafting were retained as maximum likelihood gene trees. To obtain branch support values for the gene trees, we used the SH-like approximate Likelihood-Ratio Test[338] instead of traditional bootstrap values because of its speed.

For 28,946 gene families with at least four genes from at least two different species we used gene tree-species tree reconciliation[339] to root the gene trees and to obtain estimates of duplication and speciation events along the gene tree. For the remaining 39,934 gene trees, prediction of duplication and speciation events is trivial (see below). Since the reconciliation process is error prone[340-342] and depends on the quality of the gene tree, species tree and the parameter settings of the reconciliation method we implemented a pipeline to mitigate these problems as much as possible: (1) Since PhyML does not explore the entire search space of possible tree topologies, we investigated whether alternative tree topologies with improved reconciliation duplication/loss costs, obtained by branch rearrangements of the original gene trees in the reconciliation strep (see below), had an increased likelihood under the multiple sequence alignment than the gene tree produced by PhyML. As such we obtained a reconciled gene tree that is maximally supported by both the reconciliation criterion (in this instance duplication/loss cost) and the multiple sequence alignment as described in Wu *et al.*[342]; (2) To deal with the problem of reconciliation solutions being dependent on the parameter settings we performed the reconciliation with a range of different parameter settings and we also considered multiple possible optimal reconciliations under the same parameter settings, if available. Since duplication/speciation events that were predicted for multiple parameter settings are assumed to be more reliable[341], we built a majority-rule consensus reconciliation in which we only retained duplication/speciation events supported by at least 50% of the reconciliations (see details on Gene tree-species-tree reconciliation pipeline in below).

If a duplication event was predicted at the Angiosperm-associated node, we split the phylogenetic tree into two subtrees (and hence also two associated gene families), ensuring that each subtree traced back to a single ancestral Angiosperm gene. With this procedure we obtained 11,131 gene families with gene trees tracing back to an angiosperm ancestral gene. From this set we removed the gene families that did not have gene copies for at least 32 out of 37 species (Supplementary Figure D-1), ending up with a final set of 9,178 core gene families.

For the remaining 39,934 gene families (*i.e.*, gene families with at least two species but no more than three genes or gene families that are only present in one species), we inferred duplication events by simply applying the following rules (see Supplementary Figure D-15). For gene families with only one species, after mid-point rerooting of the gene tree, each node in the tree represents a duplication node. For gene families with two genes, after mid-point rerooting of the gene tree, nodes were annotated as duplication nodes if the two genes were from the same species. For gene families with three genes we used the topology of the gene tree to infer the duplication events.

3.4.2.4 Gene tree-species-tree reconciliation pipeline

We used NOTUNG version 8[339] for reconciliation. NOTUNG is based on the maximal parsimony criterion and outputs the reconciled tree that minimizes the overall

duplication/loss cost. We first ran NOTUNG in the "rooting" mode, saving different trees with different optimal rootings under the given duplication/loss cost scheme. We then ran NOTUNG in the "reconcile" mode, again retaining different optimal reconciliation solutions. We also ran NOTUNG in the "rearrange" mode, which allows for weakly supported branches (provided by aLRT scores) to be rearranged. We used two different thresholds, a more stringent one in which only branches with an aLRT ≤ 0.5 could be rearranged and a more relaxed one in which rearrangements were not restricted by aLRT scores. Since running NOTUNG in the rearrange mode essentially modifies the unrooted tree topology, we used the CONSEL program[237] to select the tree topology that has the highest likelihood for the multiple sequence alignment. The motivation behind this whole procedure is to obtain the tree topology that both minimizes duplication-loss cost and has the highest likelihood for the multiple sequence alignment, as was proposed by Nguyen *et al.*[341]. We also performed tree reconciliation for different values for the duplication and loss cost parameters: (1,1),(1,2),(2,1),(2,2). Finally, we combined all "optimal" reconciliations according to the parsimony criterion and corresponding to the most optimal unrooted tree topology according to the multiple sequence alignment into one consensus reconciliation. NOTUNG predicts for each node in the gene tree whether it arose through duplication or speciation. We calculated two confidence scores for the predicted duplication events since these are further used for downstream analyses: (1) the duplication consistency score, which assesses the imbalance of the predicted duplication event by comparing the overlap in species on the daughter branches with their union; and (2) the annotation support score, which assesses the reliability of the duplication event based on the annotation or age given by NOTUNG to the duplication event. We noticed that there are duplication events with a high duplication consistency that seem to date back to the angiosperm common ancestor but that only encompass one monocot and one dicot species. Hence, we calculated the annotation support as the ratio of the total number of species associated with a duplication node in the gene tree to the expected number of species associated with that node in the species tree and deemed duplication events with low annotation support scores as being unreliable. In this article, we only considered duplication events exceeding a duplication consistency score of 0.2 and with an annotation support of at least 0.5. We found that the number of predicted duplication events stays relatively stable for duplication consistency scores up until 0.4 (Supplementary Figure D-16).

3.4.3 K_S-based age distributions

3.4.3.1 K_S-based estimation of timing of duplication

Estimates of K_S-values were obtained for all paralogous pairs associated with the predicted duplication events inferred by the gene tree – species tree reconciliation process. For cases where there are multiple possible pairs for a predicted duplication event, we calculated K_S-values for all possible gene pairs and selected the gene pair with the smallest K_S-value to represent the timing of the duplication event. For each paralogous gene pair we aligned the protein coding sequences using ClustalW[343] using parameter recommendations from Hall[265]. PAL2NAL[83] was used to back-translate the aligned amino acids into corresponding codons without gaps. Then codeml[268] from PAML[344,345] was used to obtain K_S-values for each gene pair using the GY model with stationary codon frequencies empirically estimated by the F3×4 model.

3.4.3.2 Gaussian Mixture Modeling of K_S-based age distributions

For each species in our dataset we fitted Gaussian mixtures to age distributions inferred from K_S-values[168,178,300], using the R-package 'mixtools'. We ignored K_S-values that exceeded 5.0. First, we determined for each age distribution the number of components (k) using the 'boot.comp' function. Specifically, we performed parametric bootstraps with 1000 bootstrap realizations of the likelihood ratio statistic for testing the null hypothesis of a k-component fit versus the alternative hypothesis of a (k+1)-component fit. For this test a significance level of 0.01 was used. For each age distribution we tested the presence of one to 6 components. The number of components determined in this first step was used to fit a mixture of Gaussian models to the K_S distribution, using the 'normalmixEM' function with the following parameters: k=k, maxit = 1×10^{30}, maxrestarts = 1×10^{3}, epsilon = 1×10^{-50}. We manually curated the obtained peaks, only further focusing on solid WGD peaks (Supplementary Figure D-16). Dispersed background peaks with mean μ >3 and model peaks with obvious misfits to the data were ignored for the purpose of duplication assignment. We assume that each remaining peak corresponds to a WGD event, except for the first peak, which likely consists of recent small-scale duplications[162]. A duplication was assigned to the peak that showed the highest probability density at the K_S value obtained for its representative paralog pair[162]. For each WGD, we obtain an associated estimate of the number of gene families with retained duplicates as the ratio of the number of core gene families with duplicates for that event to the total number of core gene families. Each peak was characterized by an age (expressed in K_S-values) that corresponded to the mean (μ) of the Gaussian mixture component (see Supplementary Table D-2 for detailed peak information). To assess duplicate retention in function of time since duplication we plotted duplicate retention associated with a certain WGD (y) in function of the predicted age of that event (x). We then fitted exponential and power-law functions to these data. Both functions have previously been used to describe the relationship between duplicate retention and time since duplication[142,162]. In all instances, the power-law fit was preferred over the exponential fit based on the χ^2 goodness-of-fit measure (Supplementary Figure D-17, Supplementary Table D-3).

3.4.4 Evolution of gene families under a stochastic birth-death null model

3.4.4.1 The null model

The null hypothesis describes the evolution of gene families along the phylogeny as a random birth-death (BD) process with equal rates of SSD gene duplication and loss per evolutionary time unit (unit branch length), λ, as proposed by Bailey[297]. Since WGDs violate the assumption of independency of duplication events in Bailey's BD model[297], we have placed these events as separate nodes on the branches of the species tree, similar to the strategy employed by Rabier *et al.*[298]. At WGD nodes, all gene family members are instantaneously duplicated (or triplicated, depending on the nature of the polyploidy event). As in the model of Rabier *et al.*[298], we assume that a given fraction of duplicates is lost very quickly after WGD, represented by an immediate loss rate parameter q in our model. The remaining WGD duplicates are lost over time at a loss rate λ, the same as for SSD duplicates. A full description of the model will be published elsewhere.

Our purpose is to use this BD model to generate gene counts at the leaves of the species tree for a number of simulated gene families and compare the Single Copy Percentage (SCP)

distribution of these simulated families to the SCP distribution observed for the core gene families. In each run, we simulated gene counts under the random BD model for 9,178 gene families, corresponding to the number of families in the core set. We performed 1,000 such runs and estimated the SCP null distribution as a kernel density function over the 9,178 × 1000 simulations.

For each simulated gene family, we sample a value for λ and q from predefined distributions (see below), and we assume that the root size - the gene count at the root of the species tree - is equal to 1. We start at the root and generate a gene count for each of the child nodes of the root through an MCMC process that samples a child node size from the node size probability distribution function described in the BD model[297]; 5,000 MCMC steps were used as burn-in to guarantee MCMC convergence to the stationary BD probability distribution. The same procedure is used for any further progeny node up to the leaf nodes, each time starting from the previously generated gene count at its parent node. At WGD nodes, the node size is multiplied after node size sampling with 1+ d.(1-q) to mimic the WGD effect, with d=1 for duplications and d=2 for triplications. In our simulations, we imposed the limitation of generating at least 32 non-zero gene counts at the leaves of the species tree, to be consistent with the fact that the core gene families studied were required to be present in at least 32 out of 37 species.

The q value to be used for a given duplicate birth-death simulation is uniformly sampled from the range [0-1], with 0 being complete retention and 1 complete loss of duplicates immediately after WGD (q is assumed to be the same for all WGDs across the tree, *i.e.*, it is assumed to be a property of the gene family). The λ-value to be used for a given simulation is sampled from a normal distribution with mean λ_{av} = 0.53 and standard deviation σ = 0.156. The rationale for sampling birth rates from this specific distribution is the following. We assume that the average duplication rate per gene, λ_{av}, is approximately equal to the average synonymous substitution rate per synonymous site[300], *i.e.*:

Equation 3.1

$$\begin{aligned}\lambda_{av} &= \frac{\text{average \# duplications/gene}}{t \text{ time unit}} \\ &\approx \frac{\text{average \# synonymous substitutions/syn. site}}{t \text{ time unit}} \\ &= \frac{\text{average } K_s}{t \text{ time unit}}\end{aligned}$$

where 't time unit' stands for the evolutionary time unit used in the species tree (where branch lengths are expressed in terms of the number of substitutions per codon t), *i.e.*, the evolutionary time needed to obtain one substitution per codon on average (unit branch length t=1). To assess approximately how many synonymous substitutions per synonymous site (K_S) are expected to occur per t time unit in an average plant DNA sequence, we inferred an average relationship between t and K_S from the following formula for the number of substitutions per codon t in a given sequence[346]:

Equation 3.2

$$t = \frac{(K_S \times S) + (K_N \times N)}{\frac{S+N}{3}}$$

with *S* and *N* the number of synonymous and non-synonymous sites in the sequence and K_S and K_N the number of synonymous and non-synonymous substitutions per (non)-synonymous site, respectively. Equation 3.2 can be rewritten as:

Equation 3.3

$$t = 3\ K_S \times\ (1 + \frac{\omega - 1}{\frac{S}{N} + 1})$$

with $\omega = K_N/K_S$ the ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, and S/N the ratio of synonymous sites to non-synonymous sites in a sequence. For both ω and S/N, we substitute genome-wide average estimates to obtain an approximate relationship between *t* and K_S for an average sequence evolving under average selective pressure. Taking S/N = 0.345 for the average codon[347], and taking an ω value of 0.5 on average (as observed for *Arabidopsis* duplicates in the K_S range [0,1][179]), the following estimate of *t* as a function of K_S is obtained for the average plant DNA sequence:

Equation 3.4

$$t \approx 1.884\ K_S$$

In other words, in one *t* time unit, $1/1.884 \approx 0.53$ synonymous substitutions are estimated to have accumulated per synonymous site on average. We use this estimate in equation (1) to obtain an estimate of the average duplication rate per gene $\lambda_{av} = 0.53/\text{gene}/(t\ \text{time unit})$. To assess how this λ_{av} estimate compares to literature estimates of duplication rates expressed per gene per million years, we used the average duplicate K_S and absolute age estimates for fairly recent WGDs (0 < K_S < 1, in the range where K_S estimates are reliable) reported by Vanneste *et al.*[168] to convert the resulting estimate $\lambda_{av} = 0.53/\text{gene}/(t\ \text{time unit}) = 1/\text{gene}/(K_S\ \text{time unit})$ to an estimate of the duplication rate expressed per million years (here, one K_S time unit is the evolutionary time it takes to obtain $K_S = 1$ on average, which corresponds to $1/0.53 \approx 1.884$ *t* time units according to equation (4)). By dividing the average WGD duplicate pair K_S estimates by twice the absolute WGD age estimates reported in Vanneste *et al.*[168] (note that the evolutionary time elapsed between WGD duplicates in My is twice the age of the WGD), and averaging over all WGDs, we get a K_S/My conversion factor of 0.00585, giving $\lambda_{av} = 0.00585/\text{gene/My}$, which is reasonably comparable to earlier estimates of duplications/gene/My across species[300,348]. With the average duplication rate λ_{av} in our tree estimated at 0.53/gene/(*t* time unit), we defined a λ-distribution around this value with standard deviation 0.156, so that more than 99% of the probability mass lies within the λ interval [0-1]. Qualitatively similar results were obtained with other λ_{av} values and λ-distribution shapes (results not shown).

3.4.4.2 Dating whole-genome duplications

To run the simulations described above, WGD events need to be added to the phylogenetic tree as new nodes with known branch lengths in terms of *t*, the number of substitutions per codon. To this end, for each of the WGDs, we averaged the *t* estimates for all (predicted) homeologs for which the K_S estimates fall within the WGD K_S range described in Vanneste *et al.*[168]. *t* and K_S estimates for all homeolog pairs were obtained using codeml(Goldman and

Yang 1994) as described in Vanneste *et al.*[168]. As we repeated this procedure for each species separately (except for *C. rubella* and *A. trichopoda*, which were not analyzed in Vanneste *et al.*[168]), multiple *t* estimates were obtained for shared WGDs. In this case, we used the average species-specific *t*-estimates to position a given shared WGD on the tree.

All of the resulting WGD estimates were positioned on the species phylogeny in a manner consistent with their taxonomic positioning reported earlier[168,169], except for the most recent WGDs in *Gossypium raimondii* and *Zea mays*, which were inferred by our *t*-estimation protocol to be positioned on older branches than the accepted ones, likely because of *t* and K_S estimation and averaging inaccuracies. In these cases, we positioned the WGD in the beginning of the branch reported in literature. See Supplementary Figure D-18 for the tree that was obtained using this approach.

3.4.5 Clustering of the copy-number profile matrix

To determine gene family-specific differences in duplicate retention, the gene family data was transformed into a count matrix, in which elements represent the number of gene copies for a certain gene family (columns) in a certain species (rows). To reduce the influence of outliers (families with lots of genes), we only used gene families with maximum three gene copies per species. We clustered this matrix in the direction of the gene families using ConsensusClusterPlus, which incorporates a subsampling approach to infer cluster number and cluster confidence[301,349]. This R implemented package was run using the following options: maxK = 8, reps=100, pItem=0.8, pFeature=1, *k*-means, inner linkage=average, final linkage=average, distance=pearson. A solution with three clusters was found to be optimal according to the built-in cluster stability criterion (Supplementary Figure D-8)[301]

3.4.6 Functional data

3.4.6.1 PPI data in A. thaliana

A compendium of protein-protein interactions in *A. thaliana* was constructed combining the following sources, BioGRID 3.2.110[350], CORNET(only experimentally validated interactions)[351], STRINGv9.1 (only category Binding)[352], EVEX (only category binding)[353] and a TAP dataset assembled from literature[304,354-367]. After removing redundancy and self-interactions this lead to a set with a total of 46,113 interactions between 9,813 proteins.

3.4.6.2 Enrichment of PPI, LOF, chloroplast genes and transcription factors

The Fisher's exact test was used to calculate if a class is overrepresented in a given set of genes. In order to test whether there are more protein interactions within a group than between a group, 1000 randomized interaction networks with the same degree distribution were constructed. For each group of genes, a z-score was obtained by comparing the number of protein interactions within the group based on the extant PPI network with the distribution of within-group interaction counts observed in the randomized networks. Z-scores were then converted into one-tailed *P* values.

3.4.6.3 Functional enrichment analysis

The BINGO 2.44 Cytoscape plugin[266] was used to calculate functional enrichment values for the set of *A. thaliana* genes. We used a *P* value threshold of 0.05 and *P* values were corrected for multiple testing using the Benjamini and Hochberg method[368].

3.5 Acknowledgements

We thank three anonymous reviewers for their useful comments. R.D.S. is a postdoctoral fellow of The Research Foundation - Flanders (FWO). Y.V.d.P acknowledges the Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks" Project (no 01MR0310W) of Ghent University and the European Union Seventh Framework Programme (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739–DOUBLE-UP. This project is supported by The Research Foundation – Flanders (FWO) (G008812N).

3.6 Author contributions

R.D.S. and Y.V.d.P. designed the study; R.D.S., Z.L., J.D.F. and S.T. performed research; Z.L., J.D.F. and R.D.S. designed and performed analyses on gene family data, gene family evolution and gene function; S.T. and S.M. designed and performed the modeling approach; R.D.S. wrote the paper with the assistance of the other co-authors.

Chapter 4

The *Apostasia* genome and the evolution of orchids

Guo-Qiang Zhang[*], Ke-Wei Liu[*], Zhen Li[*], Rolf Lohaus[*], Yu-Yun Hsiao[*], Shan-Ce Niu, Jie-Yu Wang, Yao-Cheng Lin, Qing Xu, Li-Jun Chen, Kouki Yoshida, Sumire Fujiwara, Zhi-Wen Wang, Yong-Qiang Zhang, Nobutaka Mitsuda, Meina Wang, Guo-Hui Liu, Lorenzo Pecoraro, Hui-Xia Huang, Xin-Ju Xiao, Min Lin, Xin-Yi Wu, Wan-Lin Wu, You-Yi Chen, Song-Bin Chang, Shingo Sakamoto, Masaru Ohme-Takagi, Masafumi Yagi, Si-Jin Zeng, Ching-Yu Shen, Chuan-Ming Yeh, Yi-Bo Luo, Wen-Chieh Tsai , Yves Van de Peer, and Zhong-Jian Liu

[*]contributed equally

Nature **549**(7672): 379-383 (2017).

Abstract

Constituting approximately 10% of flowering plant species, orchids (Orchidaceae) display unique flower morphologies, possess an extraordinary diversity in lifestyle, and have successfully colonized almost every habitat on Earth[194,369,370]. Here we report the draft genome sequence of *Apostasia shenzhenica*[371], a representative of one of two genera that form a sister lineage to the rest of the Orchidaceae, providing a reference for inferring the genome content and structure of the most recent common ancestor of all extant orchids and improving our understanding of their origins and evolution. In addition, we present transcriptome data for representatives of Vanilloideae, Cypripedioideae, and Orchidoideae and novel third-generation genome data for two species of Epidendroideae, covering all five orchid subfamilies. *A. shenzhenica* shows clear evidence of a whole-genome duplication, which is shared by all orchids and occurred shortly before their divergence. Comparisons between *A. shenzhenica* and other orchids and angiosperms also permitted the reconstruction of an ancestral orchid gene toolkit. We identify new gene families, gene family expansions and contractions, and changes within MADS-box gene classes, which control a diverse suite of developmental processes, during orchid evolution. This study sheds new light on the genetic mechanisms underpinning key orchid innovations, including the development of the labellum and gynostemium, pollinia, and seeds without endosperm, as well as the evolution of epiphytism; reveals relationships between the Orchidaceae subfamilies; and helps clarify the evolutionary history of orchids within the angiosperms.

4.1 Introduction

The Apostasioideae are a small subfamily of orchids that includes only two genera (*Apostasia* and *Neuwiedia*[194,372]), consisting of terrestrial species confined to the humid areas of Southeast Asia, Japan, and northern Australia[373]. Although Apostasioideae share some synapomorphies with other orchids (for example, small seeds with a reduced embryo and a myco-heterotrophic protocorm stage), they possess several unique traits, the most conspicuous of which is their floral morphology[374]. *Apostasia* has a non-resupinate, solanum-type flower with anthers closely encircling the stigma (including postgenital fusion), a long ovary, and an actinomorphic perianth with an undifferentiated labellum. Three stamens (two of which are fertile) are basally fused to the style, forming a relatively simple gynostemium, and the anthers contain powdery pollen (grains not unified into pollinia). These characteristics (Figure 4-1A) differ from those of other Orchidaceae subfamilies, which have three sepals, three petals (of which one has specialized to form the labellum), and stamens and pistil fused into a more complex gynostemium (Figure 4-1B) but are similar to those of some species of Hypoxidaceae (a sister family to Orchidaceae, in the order Asparagales).

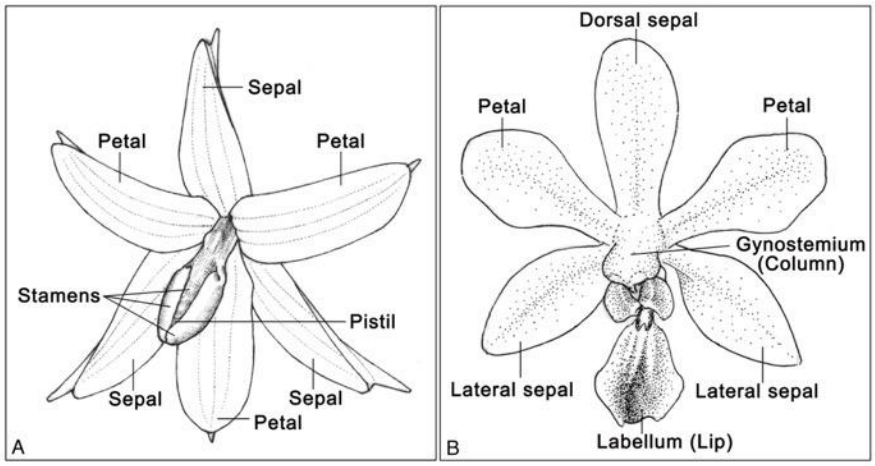


Figure 4-1 The morphology of orchid flowers
(A) Illustration of an *Apostasia* flower; (B) Illustration of a *Phalaenopsis* flower.

4.2 Results and Discussion

We sequenced the *A. shenzhenica* genome using a combination of different approaches; the total length of the final assembly was 349 Mb (see Materials and Methods and Supplementary Tables E-1–4). We confidently annotated 21,841 protein-coding genes, of which 20,202 (92.50%) were supported by transcriptome data (Supplementary Figure E-1 and Supplementary Table E-5). Using single-copy orthologues, we performed a BUSCO[375] assessment that indicated that the completeness of the genome was 93.62%, suggesting that the *A. shenzhenica* genome assembly is of high quality (Supplementary Table E.6). For comparative analyses, we also improved the quality of the previously published genome

assemblies of the orchids *Phalaenopsis equestris*[188] and *Dendrobium catenatum*[189] (see Materials and Methods and Supplementary Tables E-6 and E-7).

4.2.1 Evolution of gene families

We constructed a high-confidence phylogenetic tree and estimated the divergence times of 15 plant species using genes extracted from a total of 439 single-copy families (Figure 4-2and Supplementary Figure E-2). We undertook a computational analysis of gene family sizes (CAFÉ 2.2[376]) to study gene family expansion and contraction during the evolution of orchids and related species (Figure 4-2 and Supplementary Note E.1-1).

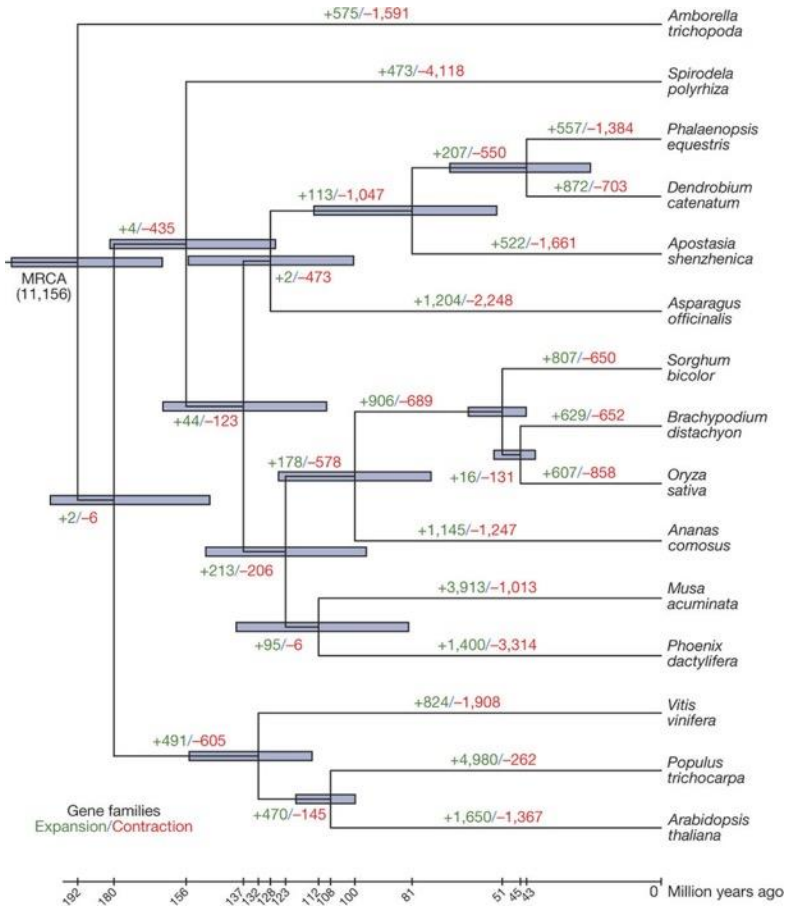


Figure 4-2 Phylogenetic tree showing divergence times and the evolution of gene family sizes.

The phylogenetic tree shows the topology and divergence times for 15 plant species. As expected, as a member of the Apostasioideae, *A. shenzhenica* is sister to all other orchids. In general, the estimated orchid divergence times are in good agreement with recent broad scale orchid phylogenies[194,370]. Divergence times are indicated by light blue bars at the internodes; the range of these bars indicates the 95% confidence interval of the divergence time. Numbers at branches indicate the expansion and contraction of gene families (see Materials and Methods and Supplementary Figure E-2). MRCA: most recent common ancestor. The number in parentheses is the number of gene families (11,156) in the MRCA as estimated by CAFE[376].

By comparing 12 plant species, we found 474 gene families (Figure 4-3) that appeared unique to orchids (Supplementary Note E.1-2). Gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis found these gene families to be specifically enriched in the terms '*O*-methyltransferase activity', 'cysteine-type peptidase activity', 'flavone and flavonol biosynthesis' and 'stilbenoid, diarylheptanoid and gingerol biosynthesis' (Supplementary Note E.1-2).

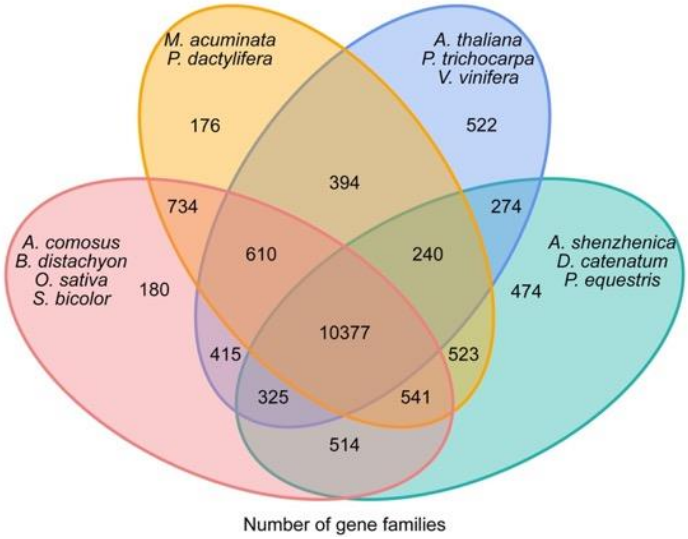


Figure 4-3 Venn diagram showing unique and shared gene families among members of Orchidaceae, dicots and Poaceae, and *M. acuminata* and *P. dactylifera*.
Numbers represent the number of gene families. Comparison of the four groups revealed 474 gene families unique to Orchidaceae which exist in all three Orchidaceae species. If we consider lineage specific gene families for each group (*i.e.*, gene families present in one or a few but not all species in a group), then there are 4,958 unique gene families for Orchidaceae, 7,503 for Poales, 4,494 for the dicots, and 1,560 for the group of *M. acuminata* and *P. dactylifera*.

4.2.2 Whole-genome duplication

4.2.2.1 K_S distributions and absolute phylogenomic dating

Analyses of the number of synonymous substitutions per synonymous site (K_S) in *A. shenzhenica* for both the whole paranome (the set of all duplicated genes in the genome, Figure 4-4A) and 'anchor' duplicates retained in co-linear regions only (*i.e.*, excluding duplicates from small-scale duplications, Figure 4-4B) consistently identified a clear peak of duplicates with a K_S value close to 1. The K_S distributions generated from the genomes of the epidendroid orchids *P. equestris*[188] and *D. catenatum*[189] and from transcriptomes of nine additional orchids (together covering all five orchid subfamilies) all show similar K_S peaks with K_S values of 0.7 to 1.1 (Supplementary Figure E-3). In the apostasioid *Neuwiedia malipoensis*, a prominent second peak at a much lower K_S value of about 0.25 may signify a more recent and *Neuwiedia*-specific second WGD. In contrast, the peaks at K_S values < 0.2 as apparent in the genomes of *D. catenatum*, *P. equestris* and *A. shenzhenica* stem from background (tandem) duplications and most likely do not signify additional recent WGDs.

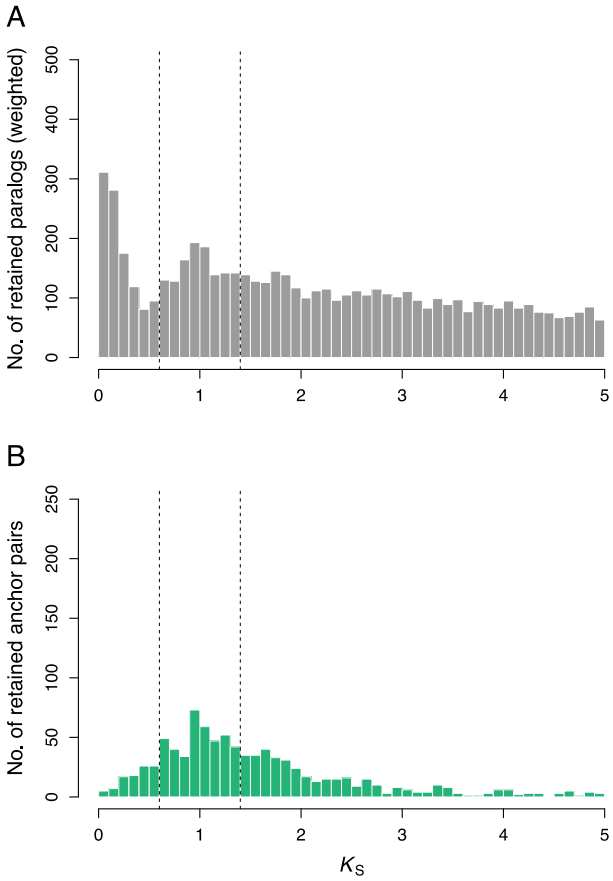


Figure 4-4 *A. shenzhenica* K_S-based age distributions.
(A) Distribution of synonymous substitutions per synonymous site (K_S) for the whole *A. shenzhenica* paranome. (B) Distribution of synonymous substitutions per synonymous site (K_S) for duplicated anchors found in collinear regions as identified by i-ADHoRe. A WGD event is identified in both distributions with its peak centred around a K_S of 1. The dashed lines indicate the K_S boundaries used to extract duplicate pairs for absolute phylogenomic dating of the WGD event (see Materials and Methods and Figure 4-5).

We constructed K_S-based age distributions of the one-to-one orthologues of *A. shenzhenica* and *Asparagus officinalis* (asparagus, Asparagaceae, a sister family to Orchidaceae in the order of Asparagales), *A. shenzhenica* and *P. equestris*, *A. shenzhenica* and *D. catenatum*, and *P. equestris* and *D. catenatum* (each representing the divergence event between the two respective species) and compared these to the distributions of the duplicated anchors from each of the three orchids for which we have a genome (Figure 4-6A). The peaks of the three anchor-pair distributions all have lower K_S values than the peak of the *A. shenzhenica–A. officinalis* orthologue distribution, indicating that the WGD signatures are specific to Orchidaceae and not shared with other non-orchid Asparagales (see also Supplementary Figures E-3 and E-4**)**. They also all have higher K_S values than the peak of the *P. equestris–D. catenatum* orthologue distribution, confirming a WGD event that is shared at least between these two species, as reported previously[189]. The anchor-pair distributions of *A. shenzhenica* and *P. equestris* are also slightly shifted towards higher K_S values compared with the *A. shenzhenica–P. equestris* and *A. shenzhenica–D. catenatum* orthologue distributions (each of which represents the divergence between the *A. shenzhenica* lineage and the rest of the Orchidaceae), whereas the anchor pair distribution of *D. catenatum* largely overlaps with these two orthologue distributions. *D. catenatum* likely has a slightly lower substitution rate, as hinted by the slightly "younger" peak of the *A. shenzhenica–*

D. catenatum orthologue distribution compared to the *A. shenzhenica–P. equestris* orthologue distribution (also compare the orchid–*A. officinalis* orthologue distributions in Supplementary Figure E-3**,** long-dashed versus dashed vertical yellow lines for *D. catenatum* and *P. equestris*). These patterns suggest that the WGD(s) and the initial orchid speciation events occurred relatively close in time and that the WGD events evident in the three orchid genomes (and, by extension, those in the orchid transcriptomes) could represent a single event, slightly older than their divergence, and thus shared among all orchids. However, the distance between the peaks is small, there is substantial overlap among the distributions and the number of anchor pairs we could extract is relatively low for all three species. In addition, some heterogeneity in substitution rate between these species is expected given the age of the events and this is apparent in the one-to-one orthologue K_S distributions with *A. officinalis* (yellow distributions and long-dashed vertical yellow lines versus dashed vertical yellow lines in each panel of Supplementary Figure E-3).

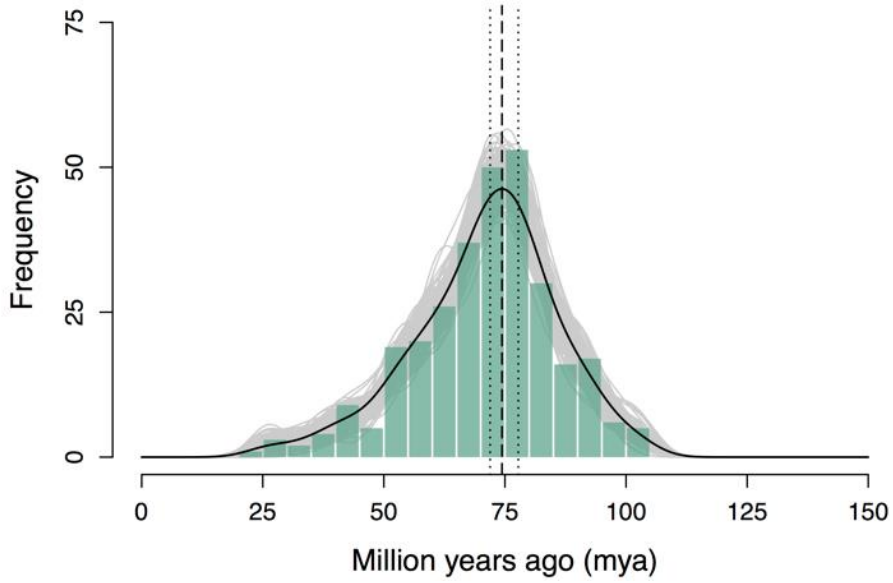


Figure 4-5 Absolute age of the *A. shenzhenica* WGD event.

Absolute age distribution obtained by phylogenomic dating of *A. shenzhenica* paralogues. The solid black line represents the kernel density estimates (KDE) of the dated paralogues, and the vertical dashed black line represents its peak at 74 mya, which was used as the consensus WGD age estimate. The grey lines represent density estimates from 2,500 bootstrap replicates and the vertical black dotted lines represent the corresponding 90% confidence interval for the WGD age estimate, 72–78 mya (see Materials and Methods). The histogram shows the raw distribution of dated paralogues.

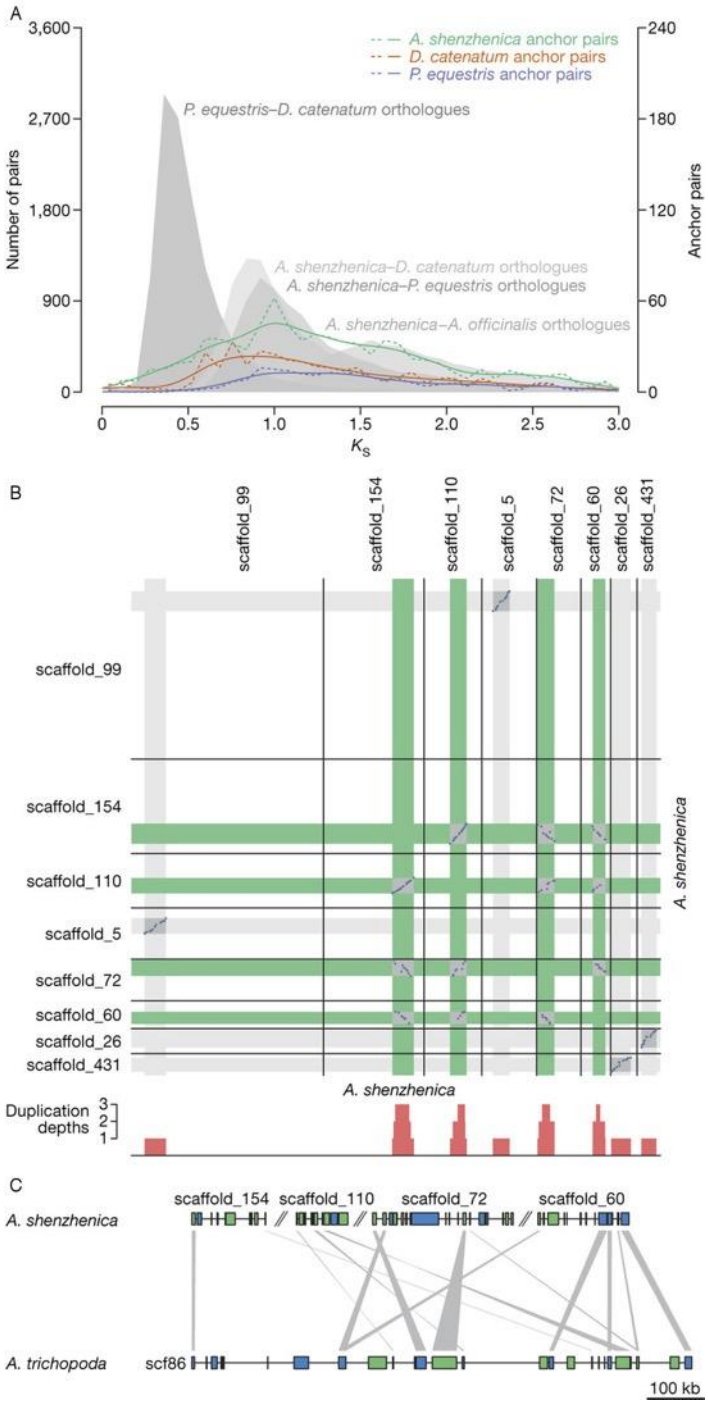


Figure 4-6 K_S and co-linearity analysis of the *A. shenzhenica* WGD. (A) Distribution of K_S for the one-to-one *P. equestris–D. catenatum*, *A. shenzhenica–D. catenatum*, *A. shenzhenica–P. equestris*, and *A. shenzhenica–A. officinalis* orthologues (filled grey curves and left-hand y-axis). Distribution of K_S for duplicated anchors found in co-linear regions of *A. shenzhenica* (green lines), *D. catenatum* (red lines), and *P. equestris* (blue lines). The filled grey curves and dashed coloured lines are actual data points from the distributions; the solid coloured lines are kernel density estimates (KDE) of the anchor-pair (duplicated genes found in co-linear regions) data scaled to match the corresponding dashed lines. All anchor-pair data are scaled up ×15 (right-hand y-axis) compared to the orthologue data. (B) Syntenic dot plot of the self-comparison of *A. shenzhenica*. Only co-linear segments with at least 15 anchor pairs are shown. The sections on each scaffold with co-linear segments are shown in grey. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each position; see Materials and Methods). The co-linear regions in green indicate the four co-linear segments that have a common orthologous co-linear segment in *A. trichopoda* as shown in (C). (C) Co-linear alignment of *A. shenzhenica* and *A. trichopoda*. The colours of genes in the alignment indicate gene orientation, with blue for forward strands and green for reverse strands. The grey links connect orthologues between *A. shenzhenica* and *A. trichopoda*. Scf86: scaffold00086 of the *A. trichopoda genome* (v1.0).

We performed absolute phylogenomic dating[168] of the WGD event identified from the *A. shenzhenica* genome to determine its age in relation to orchid phylogeny. Paralogous gene pairs present under the WGD peak in the *A. shenzhenica* K_S distributions (Figure 4-4A and B) were dated (see Materials and Methods) and the resulting absolute age distribution showed a peak at 74 million years ago (mya) with a 90% confidence interval of 72–78 mya (Figure 4-5). This estimate of the date of the WGD event in the *A. shenzhenica* lineage coincides with the date estimated for the WGD event in the *P. equestris* lineage (76 mya, 90% CI: 72–81 mya[188]). Estimates for the crown age of extant orchids, *i.e.*, the time of divergence of the *A. shenzhenica* lineage from the lineage that gave rise to the rest of the Orchidaceae, vary

widely and range from 54 mya to 121 mya (Ramírez *et al.*[190]: 71–90 mya, youngest mean minus 1 SD to oldest mean plus 1 SD; Gustafsson *et al.*[191]: 63–92 mya, 95% HPD; Chen *et al.*[192]: 54–82 mya, 95% HPD; Chomicki *et al.*[193]: 75–121 mya, 95% HPD; Givnish *et al.*[194,370]: 80–100 mya, 95% CI; see also Figure 4-2 and Supplementary Figure E-2), but again suggest that the initial divergence of extant orchid lineages and any single or multiple ancient orchid WGD event(s) likely occurred closely together in time. Some of these ranges would support the possibility of a single shared WGD in the most recent common ancestor of extant orchids indicated in the K_S-based age distributions, as discussed above (Figure 4-6A), though our date estimate for such an event would fall towards the lower end of most of those range estimates.

4.2.2.2 Co-linearity and synteny analyses

Co-linearity analysis (see Materials and Methods) of *A. shenzhenica* indicated that 43.85% of the genome retains paralogous genes from WGD events, with 7,271 (33.31%), 1,423 (6.52%), 669 (3.06%), and 210 (0.96%) genes on co-linear regions with two, three, four, and five paralogous segments, respectively, suggesting two WGD events that can be identified in the current *A. shenzhenica* genome (Figure 4-6B and Supplementary Figure E-5). One of the WGD events seems much more recent than the other because co-linear regions consisting of three and four segments are much rarer than those with two, which almost cover one third of the genome. The few co-linear regions with five segments (making up less than 1% of the genes) might be remnants of a more ancient third WGD event.

To circumscribe the two WGD events, we compared the genome of *A. shenzhenica* with the genomes of several other flowering plants: the earliest-diverging extant angiosperm *Amborella trichopoda*, the dicot *Vitis vinifera*, and the two monocots *Ananas comosus* (pineapple, order Poales) and *A. officinalis*. Due to the fragmented nature of the current genome assembly of *A. shenzhenica*, counting the number of co-linear segments in a comparison of *A. shenzhenica* with other species did not directly unveil the number of individual WGDs. We thus illustrated duplication depth, *i.e.*, the number of overlapping co-linear segments at a broader genomic region, by mapping such co-linear segments onto their corresponding orthologous regions in the other species (see Materials and Methods). The pairwise comparisons with *A. trichopoda* and *V. vinifera* both support at least two WGDs in *A. shenzhenica* (Supplementary Figures E-6 and E-7) consistent with the *A. shenzhenica* self-comparison (Figure 4-6B); for example, four paralogous segments in *A. shenzhenica* corresponded to one orthologous region in *A. trichopoda* (Figure 4-6C).

In comparison with *A. comosus*, for each co-linear segment in *A. shenzhenica* we found mostly up to four orthologous co-linear segments in *A. comosus* and *vice versa* for each *A. comosus* segment mostly up to four orthologous segments in *A. shenzhenica* (Figure 4-7). Such a 4:4 pattern is consistent with the two monocot WGDs that have been proposed in the evolutionary history of *A. comosus*, the τ WGD[98,185] shared by most monocots and the σ WGD shared by all Poales[185]. In some of these co-linear regions, four co-linear segments in *A. shenzhenica* corresponded to a specific set of four co-linear segments in *A. comosus* (on chromosomes LG4, LG13, LG18 and LG23) which have been shown to originate from one of the seven ancestral pre-τ-WGD chromosomes in monocots (Anc6) (Figure 4-7 and Figures 2 and 3c in Ming *et al.*[185]). This and the 4:4 co-linearity pattern indicate that *A. shenzhenica* followed a similar evolutionary trajectory with regard to WGDs as *A. comosus*, with one (Orchidaceae-)lineage-specific WGD in addition to the shared τ WGD. Consistently, by

tightening and relaxing the criterion influencing duplication depth (*i.e.*, the required number of anchors on the co-linear segments) we could distinguish these two WGDs (and a putative older third event) (Supplementary Figures E-8 and E-9). Similarly, the comparison between *A. shenzhenica* and *A. officinalis* showed much the same pattern as the comparison between *A. shenzhenica* and *A. comosus* (suggesting *A. officinalis* also has one independent Asparagaceae-specific WGD; also see Supplementary Figure E-4), but with less paralogous segments retained from the τ WGD (Supplementary Figure E-10). Together, these patterns of co-linearity suggest that the older of the two WGDs evident in *A. shenzhenica* is likely to be shared with *A. comosus* and *A. officinalis* (representing the τ WGD[98,185] shared by most monocots), and corroborate the idea that the younger WGD represents an independent event, specific to the Orchidaceae lineage.

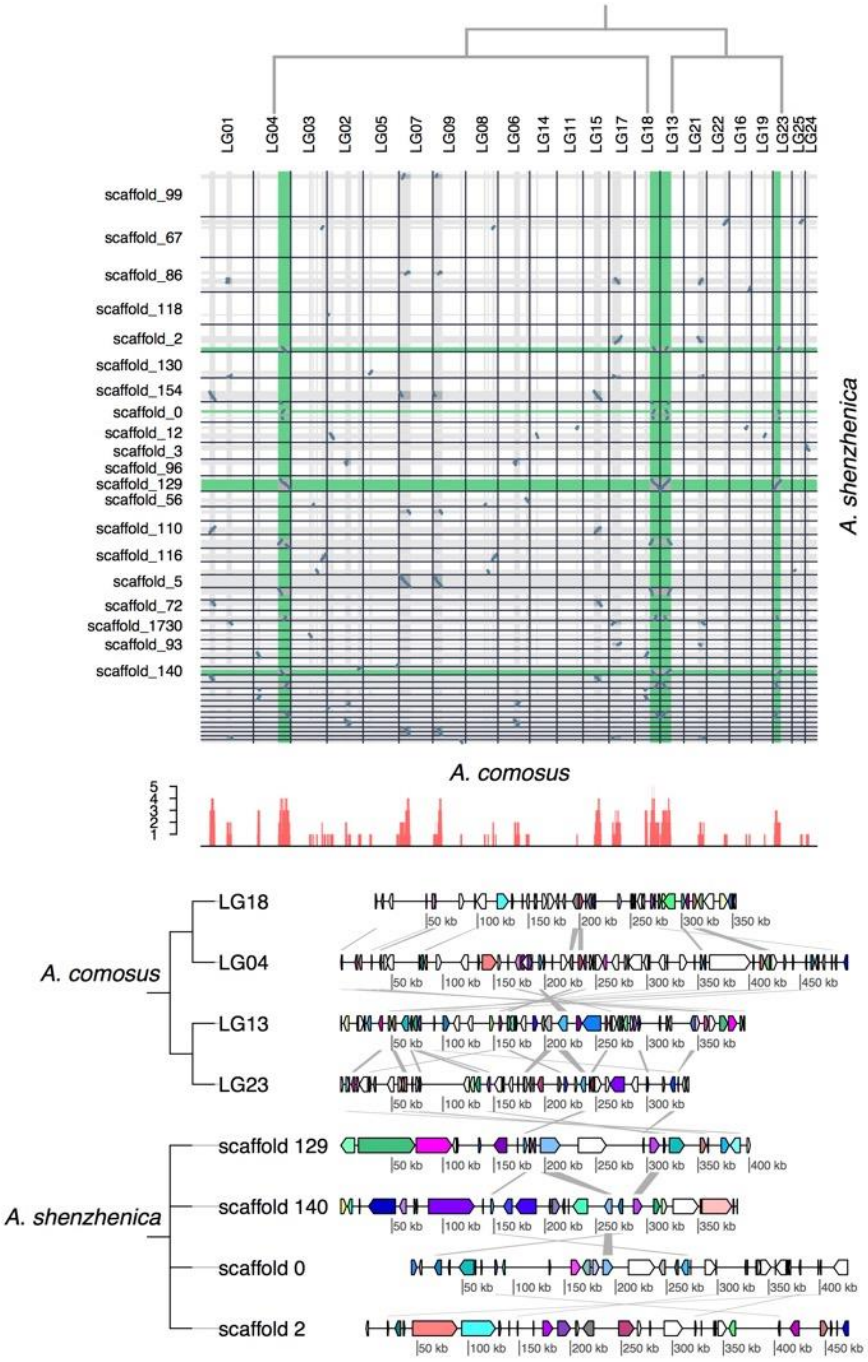


Figure 4-7 Synteny/co-linearity between *A. shenzhenica* and *A. comosus*.
Only collinear segments with at least 20 anchor pairs are shown. The sections on each scaffold with co-linear segments between *A. shenzhenica* and *A. comosus* are shown in grey. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Materials and Methods). Only connected collinear segments with at least 10 anchor pairs were used to calculate the duplication depths. The co-linear regions in green highlight the four collinear segments in *A. shenzhenica* that correspond to a specific set of four collinear segments in *A. comosus* which originated from one of the seven ancestral pre-τ-WGD chromosomes in monocots (Anc6)[185]. The phylogenetic tree on top of the dot plot indicates how Anc6 evolved into (segments of) the current four chromosomes in *A. comosus* (the pair of paired LG18 and LG04, and LG13 and LG23; Figure 2 in Ming *et al.*[185]) through two rounds of WGDs. Names of very small *A. shenzhenica* scaffolds are omitted for clarity. A part of the alignment of the co-linear segments between *A. shenzhenica* and *A. comosus* is shown below. The colors of genes in the alignment indicate anchor pairs with genes of the same color being homologous. The grey links connect anchor pairs between the two closest segments.

4.2.2.3 Gene tree analyses

To further corroborate the timing of the WGD events relative to orchid and monocot divergences, we constructed and analysed gene trees that included genes from 12 orchid species across the five subfamilies of Orchidaceae, four non-orchid Asparagales, three Poales, and *A. trichopoda* (see Materials and Methods). We selected the subset of trees that contained at least one pair of duplicated anchor genes from co-linear regions from one of the three orchids with complete genome information (*A. shenzhenica*, *D. catenatum* and *P. equestris*; *i.e.*, we only used gene families that also had 'spatial/structural' evidence of a WGD event) and mapped the coalescence points of these anchor pairs onto the species phylogeny (see Materials and Methods; see also Supplementary Figure E-11). Overall, the anchor-pair coalescence points in nearly half of such gene family trees, containing the majority of anchor pairs from *D. catenatum* and *P. equestris* and the largest fraction of anchor pairs from *A. shenzhenica*, mapped onto the orchid stem branch, providing support for a WGD event shared by all orchids (Figure 4-8). The anchor pairs in most of the remaining gene family trees mapped either onto the branch leading to all the included monocots, again providing support for the ancient τ WGD event shared by these monocots[98,185], or onto the two first diverging/descending branches from the orchid ancestor (Figure 4-8). A low number of anchor pairs from *D. catenatum* and *P. equestris* mapped along the internal orchid branches leading to Epidendroideae. However, even though the largest fraction of *A. shenzhenica* anchor pairs mapped onto the orchid stem branch (76), a substantial fraction of anchor pairs (70) actually mapped onto the Apostasioideae stem branch. These could be considered support for an additional WGD event in the Apostasioideae lineage. However, the co-linearity analyses discussed above suggest only one WGD event unique to Orchidaceae in the evolutionary history of *A. shenzhenica*. Therefore, we believe the large number of anchor points mapped on the Apostasioideae stem branch to be due to phylogenetic discordance as a result of the probably very short time interval between the shared WGD event and the divergence of Orchidaceae (Supplementary Note E.1-3).

We therefore find strong support for one WGD event that has been shared by all extant orchids, which is likely to be only slightly older than their earliest divergence. Dating of the WGD suggests that, as found for many other plant lineages, this WGD might be associated with the Cretaceous-Palaeogene boundary[150]. Furthermore, the WGD event might also be correlated with orchid diversification. Recently, Schranz *et al.*[377] proposed the radiation lag-time model based on the observations of tree imbalance in several angiosperm lineages that have been associated with paleopolyploidy. This model postulates that increases in diversification rates tend to follow WGD events, but only after lag times that may span millions of years[377]. Although this lag-time hypothesis is still controversial[150], we do find some support for it in the current study. While some subfamilies of orchids are relatively species-poor (Apostasioideae, Vanilloideae, and Cypripedioideae), the later-diverging subfamilies Orchidoideae and particularly Epidendroideae are known for an explosive radiation of novel species[194,369], which fits with the previously made observations that increases in diversification are only rarely perfectly associated with WGD events, but commonly follow them after a lag period[377,378].

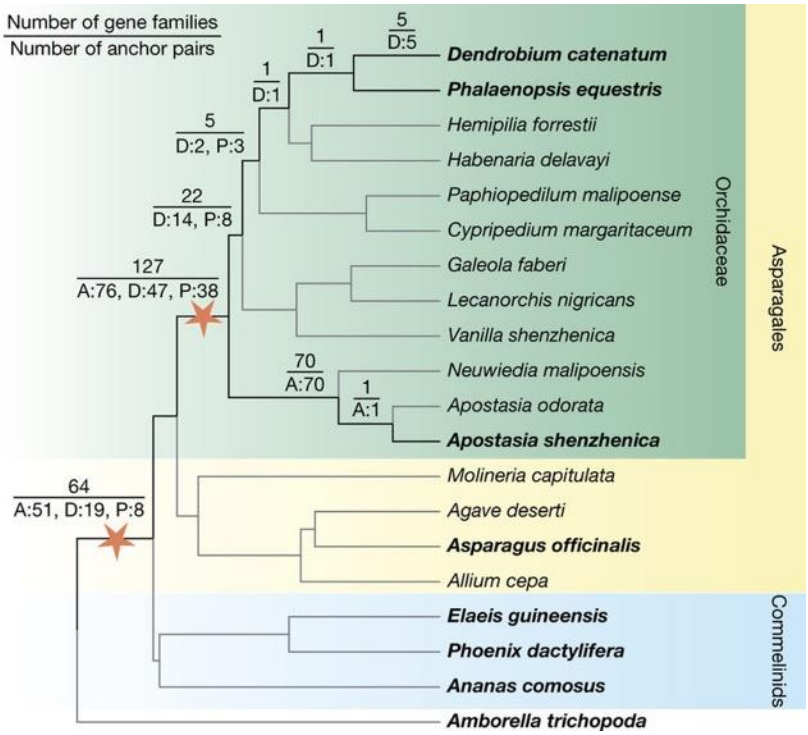


Figure 4-8 Phylogenomic analysis of orchid WGD events.
The numbers on the branches of the species tree indicate the number of gene families with one or more anchor pairs from at least one of the three orchids with genomes that coalesced on the respective branch (top), as well as the individual contributions of anchor pairs from the three orchids (bottom; A: *A. shenzhenica*; D: *D. catenatum*; and P: *P. equestris*). The two WGD events identified are depicted by stars. Species with published genomes are in bold. All the duplication events have bootstrap values over 80% (see Materials and Methods; for results for bootstrap values over 50% see Supplementary Figure E-12).

4.2.3 MADS-box genes and orchid morphological evolution

4.2.3.1 Labellum and gynostemium

Apostasia presents a number of characters that are plesiomorphic in orchids, such as an actinomorphic perianth with an undifferentiated labellum, a gynostemium with partially fused androecium and gynoecium, pollen that is not aggregated into pollinia, and underground roots for terrestrial growth[369,372-374]. The *A. shenzhenica* genome contains 36 putative functional MADS-box genes (Table 4-1, Supplementary Table E-8 and Supplementary Figure E-13), 27 of which are type II MADS-box genes (Table 4-1). Two type II MADS-box classes appear to be reduced: *A. shenzhenica* seems to have fewer genes in the B-AP3 (two members) and E classes (three members) than *P. equestris* (four B-AP3 and six E-class members) and *D. catenatum* (four B-AP3 and five E-class members) (Figure 4-9A). Previous studies have shown that expanded B-AP3 and E classes with members that have different expression patterns in floral organs are associated with the innovation of the unique labellum and gynostemium in orchids[188,379,380], and that duplicated *B-AP3* genes are responsible for the modularization of the perianth of orchid flowers[381]. We identified *B-AP3* genes from the transcriptomes of species of each of the orchid subfamilies and the B-class MADS-box genes from the floral transcriptome data of *Molineria capitulata*, a member of Hypoxidaceae that possesses a flower with petaloid tepals and powdery pollen (similar to that found in *Apostasia*). We found one member in each of the two B-AP3 subclades for both *A.*

shenzhenica and *M. capitulata*, but one or two members in each B-AP3 subclade for the other orchids (Supplementary Figure E-14). All these *B-AP3* genes are highly expressed in flower buds (Supplementary Figure E-14). These similarities suggest that the lower gene numbers in MADS-box B-AP3 and E classes in *Apostasia* represent an ancestral state, responsible for producing the plesiomorphic flower with an undifferentiated labellum and partially fused gynostemium. The B-AP3 and E classes may have expanded independently only in the non-apostasioid orchids or, alternatively, in the common ancestor of all extant orchids, possibly as a result of the shared orchid WGD, with subsequent loss of paralogous genes in *Apostasia* causing reversion to the ancestral state. The B-AP3 gene tree topology and some evidence from co-linearity analysis of orchid *B-AP3* genes (Supplementary Figure E-15) suggest the latter. We hypothesize that differential paralogue retention and subsequent sub- and neo-functionalization of B-AP3 and E-class members resulted in the derived labellum found in other orchids (Figure 4-9B).

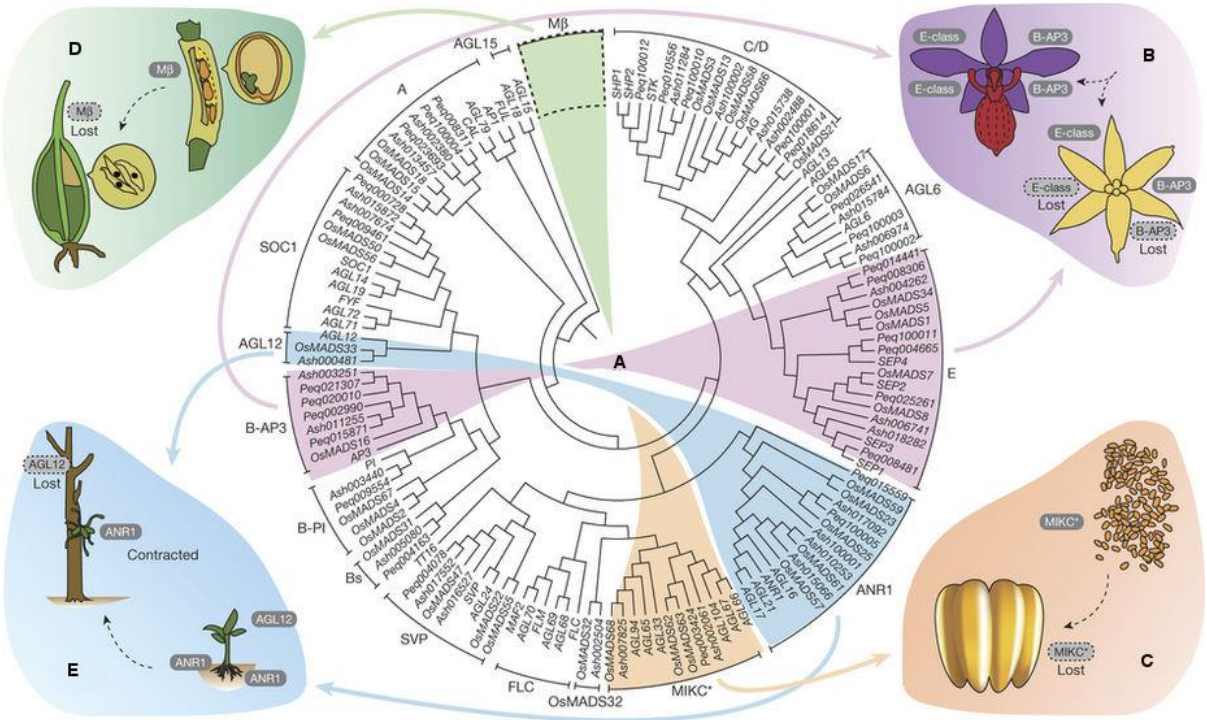


Figure 4-9 MADS-box genes involved in orchid morphological evolution.
(A) Phylogenetic analysis of MADS-box genes among *A. shenzhenica*, *P. equestris*, *O. sativa* and *Arabidopsis*. The B-AP3 and E-class, MIKC*, Mβ, AGL12 and ANR1 subclades are marked by purple, orange, green, and blue colour, respectively. (B) *A. shenzhenica*, with fewer B-AP3 class and E class MADS-box genes, keeps an undifferentiated labellum and partially fused gynostemium, while *P. equestris*, with more B-AP3 class and E class MADS-box genes, develops the specialized labellum and column (in red). (C) Loss of the P-subclade genes of MIKC* in *P. equestris* is likely to be related to the evolution of pollinia. (D) The failed development of endosperm in orchids might be related to the missing type I Mβ MADS-box genes (Supplementary Figure E-20). (E) *A. shenzhenica*, containing the AGL12 gene and expanded ANR1 genes, is a terrestrial orchid, while epiphytic orchids, such as *P. equestris*, have lost the AGL12 gene and some ANR1 genes.

4.2.3.2 The pollinium

The packaging of pollen grains into a compact unit known as the pollinium, specialized for transfer as a unit by pollinating vectors, was a key innovation in the evolutionary history of Orchidaceae and may have played a role in promoting the tremendous radiation of the

group[382]. In seed plants, the P- and S-subclades of MIKC*-type genes are major regulators of male gametophytic development[383,384]. The P-subclade, however, is absent in all orchids except *A. shenzhenica* (Supplementary Figure E-16). Gene expression analysis showed that, in orchids and *M. capitulata*, MIKC*-type genes are expressed in the pollinia or pollen, suggesting they play roles in its development (Supplementary Figure E-17). Although most orchids have a pollinium, *Apostasia* has scattered pollen, similar to *M. capitulata*, *Oryza sativa* (rice). Therefore, we propose that the loss of the P-subclade members of MIKC*-type genes is related to the evolution of the pollinium (Figure 4-9A, C and Supplementary Note E.1-4).

4.2.3.3 Seeds without endosperm

The reduction of seed volume and content to an absolute minimum is a pivotal aspect of Orchidaceae evolution: in all orchid species, endosperm is absent from the seed. Type I MADS-box genes are important for the initiation of endosperm development[385], and transcripts of type I Mα and Mγ MADS-box genes were found in developing seeds of *A. shenzhenica*, *P. equestris*, and *M. capitulata* (Supplementary Figures E-18 and E-19). Notably, the three orchid genomes do not contain any type I Mβ MADS-box genes (Figure 4-9A and Supplementary Figure E-20), which are found in *Arabidopsis*, *Populus trichocarpa* (poplar), *O. sativa* (Table 4-1), and in *M. capitulata* (Supplementary Figure E-21). The lack of endosperm in orchids might therefore be related to the missing type I Mβ MADS-box genes (Figure 4-9D).

4.2.3.4 Evolution of epiphytism

Orchids are one of very few flowering plant lineages that have been able to successfully colonize epiphytic or lithophytic niches, clinging to trees or rocks and growing in dry conditions using crassulacean acid metabolism[188,189,194]. The roots of epiphytic orchids, such as *Phalaenopsis* and *Dendrobium*, are extremely specialized and differ from the roots of terrestrial orchids such as *Apostasia*. These aerial roots develop the velamen radicum, a spongy epidermis that traps the nutrient-rich flush during rainfall, representing an important adaptation of epiphytic orchids[193,386,387]. The *Arabidopsis AGL12* gene is involved in root cell differentiation[388]. *A. shenzhenica* contains one AGL12 clade gene, as do *Arabidopsis* and rice. In addition, we found transcripts similar to *AGL12* in *M. capitulata*. In both *A. shenzhenica* and *M. capitulata*, these genes are highly expressed in root tissue (Supplementary Figure E-22). Notably, we did not find similar genes in epiphytic orchids, suggesting that the loss of these gene(s) may be involved in losing the ability to develop true roots for terrestrial growth (Figure 4-9E). *Utricularia gibba,* an asterid in the order Lamiales (only distantly related to the orchids) that lacks true roots, also lacks these AGL12 clade or similar genes[389]. The *Arabidopsis ANR1* gene is a key gene involved in regulating lateral root development in response to external nitrate supply[390]. We found that the MADS-box gene subfamily ANR1 is likely reduced in *P. equestris* (two members) and *D. catenatum* (three members), compared with four members in *A. shenzhenica* (Figure 4-9A): this is consistent with no development of lateral (aerial) roots in epiphytic orchids.

In conclusion, the genome sequence of *A. shenzhenica*, an orchid belonging to a small clade that is sister to the rest of Orchidaceae, provides a unique reference for studying orchid evolution. The genome reveals clear evidence of an ancient whole-genome duplication shared by all orchids, facilitates reconstruction of the ancestral orchid gene toolkit, and provides insights into many orchid-specific features such as the development of the labellum

and gynostemium, pollinia, and seeds without endosperm, as well as the evolution of epiphytism.

Table 4-1 MADS-box genes in the *A. shenzhenica*, *P. equestris*, *D. catenatum*, *P. trichocarpa*, *A. thaliana*, and *O. sativa* genomes.

Category	*A. shenzhenica*		*P. equestris*		*D. catenatum*		*P. trichocarpa**		*A. thaliana**		*O. sativa**	
	Functional	**Pseudo**	**Functional**	**Pseudo**	**Functional**	**Pseudo**	**Functional**	**Pseudo**	**Functional**	**Pseudo**	**Functional**	**Pseudo**
Type II (Total)	27	4	29	1	35	11	64	3	47	5	48	1
MIKC^{c}	25	3	28	1	32	9	55	2	43	4	47	1
MIKC*	2	1	1	0	3	2	2	0	2	0	1	0
Mδ	0	0	0	0	0	0	7	1	4	1	0	0
Type I (Total)	9	0	22	8	28	1	41	9	62	36	32	6
Mα	5	0	10	6	15	1	23	4	20	23	15	2
Mβ	0	0	0	0	0	0	12	5	17	5	9†	1
Mγ	4	0	12	2	13	0	6	0	21	8	8	3
Total	36	4	51	9	63	12	105	12	107	41	80	7

*Genes with stop codon in MADS-box domain were categorized as pseudogenes[391].

†Nine MADS-box genes belonging to the Mβ subgroup were identified[392].

4.3 Materials and Methods

4.3.1 Sample preparation and sequencing

For genome sequencing, we collected leaves, stems, and flowers from wild *A. shenzhenica*, a self-pollinating species found in southeast China[371] that has a karyotype of 2*N* = 2X = 68 with uniform small chromosomes (Supplementary Figure E-23). We extracted genomic DNA using a modified cetyltrimethylammonium bromide (CTAB) protocol. Sequencing libraries with insert sizes ranging from 180 bp to 20 kb (Supplementary Table E-1) were constructed using a library construction kit (Illumina). These libraries were then sequenced using an Illumina HiSeq 2000 platform. The 80.02-Gb raw reads generated were filtered according to sequencing quality, the presence of adaptor contamination, and duplication. Only high-quality reads were used for genome assembly.

Total RNA was extracted from this study's samples using the RNAprep Pure Plant Kit and genomic DNA contamination was removed using RNase-Free DNase I (both from Tiangen). The integrity of RNA was evaluated on a 1.0% agarose gel stained with ethidium bromide (EB), and its quality and quantity were assessed using a NanoPhotometer spectrophotometer (IMPLEN) and an Agilent 2100 Bioanalyzer (Agilent Technologies). As the RNA integrity number (RIN) was greater than 7.0 for all samples, they were used in cDNA library construction and Illumina sequencing, which was completed by Beijing Novogene Bioinformatics Technology Co., Ltd. The cDNA library was constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB) and 3 μg RNA per sample, following the manufacturer's recommendations. The PCR products obtained were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system. Library preparations were sequenced on an Illumina Hiseq 2000 platform, generating 100-bp paired-end reads.

4.3.2 Genome size estimation and preliminary assembly

The genome size of species in Apostasioideae is between 0.38 pg and 5.96 pg[393], which is relatively small compared to that of other subfamilies (ranging from 0.38 pg to 55.4 pg)[394]. To estimate the genome size of *A. shenzhenica,* we used reads from paired-end libraries to determine the distribution of *K*-mer values. According to the Lander–Waterman theory[395], genome size can be determined by the total number of *K*-mers divided by the peak value of the *K*-mer distribution. Given only one peak in the *K*-mer distribution, we found that *A. shenzhenica* has no heterozygosity (Supplementary Figure E-24). With the peak at the expected *K*-mer depth and the formula genome size = total *K*-mer/expected *K*-mer depth, the size of the haploid genome was estimated to be 471.0 Mb (haploid). We used ALLPATHS-LG software[396] and obtained a preliminary assembly of *A. shenzhenica* with a scaffold N50 size of 1.196 Mb and corresponding contig N50 size of 30.1 Kb.

4.3.3 PacBio library construction and sequencing and filling gaps

The preliminary assembly of *A. shenzhenica* and the previous published genome assemblies of *P. equestris*[188] and *D. catenatum*[189] were improved using PacBio and 10X Genomics Linked-Reads.

Genomic DNA was isolated from the leaves of *A. shenzhenica*, *P. equestris* and *D. catenatum*. For a 20-kb insert size library, at least 20 μg of sheared DNA was required. SMRTbell template preparation involved DNA concentration, damage repair, end repair, ligation of hairpin adapters, and template purification, and used AMPure PB Magnetic Beads. Finally, the sequencing primer was annealed and sequencing polymerase was bound to SMRTbell template. The instructions specified as calculated by the RS Remote software were followed. We carried out 20-kb single-molecule real-time DNA sequencing by PacBio and sequenced the DNA library on the PacBio RS II platform, yielding about 5.44 Gb (*A. shenzhenica*), 10.23 Gb (*P. equestris*) and 10.54 Gb (*D. catenatum*) PacBio data (read quality ≥ 0.80, mean read length of *A. shenzhenica* ≥ 7 Kb, of *P. equestris* and *D. catenatum* ≥ 10 Kb) (Supplementary Table E-2).

We used PBjelly software[397] to fill gaps with PacBio data. The options were "<blasr>-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore -500 -nproc 10 -noSplitSubreads</blasr>" for the protocol.xml file. Then, we used Pilon[398] with default settings to correct assembled errors. For the input BAM file, we used BWA to align all the Illumina short reads to the assembly and SAMTOOLS to sort and index the BAM file.

4.3.4 10X Genomics library construction, sequencing, and extending scaffolds

DNA sample preparation, indexing, and barcoding were done using the GemCode Instrument from 10X Genomics. About 0.7 ng input DNA with 50 kb length was used for GEM reaction procedure during PCR, and 16-bp barcodes were introduced into droplets. Then, the droplets were fractured following the purifying of the intermediate DNA library. Next, we sheared DNA into 500 bp for constructing libraries, which were finally sequenced on the Illumina HiseqXTen[399] (Supplementary Table E-3).

We used BWA mem to align the 10X Genomics data to the filled gaps assembly using default settings. Then, we used fragScaff[400] for scaffolding. The options were as follows: *A. shenzhenica* (stages1 "-m 3000 -q 30"; stages2 "-C 2"; stages3 "-j 1.25 -u 2"), *D. catenatum* (stages1 "-m 3000 -q 30"; stages2 "-C 1"; stages3 "-j 2 -u 2") and *P. equestris* (stages1 "-m 3000 -q 30"; stages2 "-C 1"; stages3 "-j 2 -u 2")[401].

The total length of the final assembly for *A. shenzhenica* was 349 Mb with a scaffold N50 size of 3.029 Mb and corresponding contig N50 size of 80.1 Kb. (Supplementary Table E-4). For the two previously published orchid genomes of *P. equestris* and *D. catenatum*, the scaffold N50 size as well as the completeness (see below) improved considerably: for *P. equestris*, the scaffold N50 size increased from 359.12 Kb[188] to 1.217 Mb and the corresponding contig N50 size from 20.56 Kb[188] to 45.79 Kb, while for *D. catenatum* the scaffold N50 size increased from 391.46 Kb[189] to 1.055 Mb, and the corresponding contig N50 size from 33.1 Kb[189] up to 51.7 Kb (Supplementary Table E-7).

4.3.5 Repeat prediction

A total of 146.65 Mb of repetitive elements occupying more than 42.05% of the *A. shenzhenica* genome were annotated using a combination of structural information and homology prediction[189]. Retrotransposable elements, known to be the dominant form of repeats in angiosperm genomes, constituted a large part of the *A. shenzhenica* genome and

included the most abundant subtypes, such as LTR/Copia (4.97%), LTR/Gypsy (11.84%), LINE/L1 (2.78%) and LINE/RTE-BovB (9.32%), among others. In addition, the percentage of *de novo* predicted repeats was notably larger than that obtained for homologous repeats based on Repbase[402], indicating that *A. shenzhenica* has multiple unique repeats compared with other sequenced plant species (Supplementary Table E-9).

4.3.6 Gene and non-coding RNA prediction

MAKER[403] was used to generate a consensus gene set based on *de novo* predictions from AUGUSTUS[404] and GlimmerHMM[405], homology annotation with the universal single-copy genes from CEGMA[406] and the genes from *Arabidopsis* (TAIR10) and another four sequenced monocots (*O. sativa*, *P. equestris*, *S. bicolor*, and *Zea mays*) using exonerate[407], and RNA-seq prediction by Cufflinks[408] and Tophat[409]. These results were integrated into a final set of protein-coding genes for annotation (Supplementary Table E-5). Using the same annotation pipeline as for *A. shenzhenica*, 29,545 and 29,257 protein-coding genes were predicted for *P. equestris* and *D. catenatum*, respectively (Supplementary Table E-7). *A. shenzhenica* was found to have a greater average gene length (here we considered the start and stop codons as the two boundaries for a gene) than most other sequenced plants, but this length was similar to that of *P. equestris* and *D. catenatum* (Supplementary Figure E-25 and Supplementary Table E-10), in both of which this is due to a long average intron length[188,189].

We then generated functional assignments of the *A. shenzhenica* genes with BLAST (version 2.2.28+) by aligning their protein-coding regions to sequences in public protein databases, including KEGG (59.3)[410], SwissProt (release 2013_06)[411], TrEMBL (release 2013_06)[412] and NCBI non-redundant protein database (20150617), and InterPro (v5.11-51.0)[413] is also used to provide function analysis (Supplementary Table E-11). We were able to generate functional assignments for 84.2% of the *A. shenzhenica* genes from at least one of the public protein databases (Supplementary Table E-11).

The tRNA genes were searched by tRNAscan-SE[414]. For rRNA identification, we downloaded the *Arabidopsis* rRNA sequences from NCBI and aligned them with the *A. shenzhenica* genome to identify possible rRNAs. Additionally, other types of non-coding RNAs, including miRNA and snRNA, were identified by using INFERNAL[415] to search from the Rfam database. In the end, we identified 43 microRNAs, 203 transfer RNAs, 452 ribosomal RNAs and 93 small nuclear RNAs in the *A. shenzhenica* genome (Supplementary Table E-12).

4.3.7 Transcriptome assembly

Before assembly, we got high-quality reads by removing adaptor sequences and filtered low-quality reads by using TRIMMOMATIC[416] from raw reads with parameters: ILLUMINACLIP:path/adaptor:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. The resulting high-quality reads were *de novo* assembled and annotated with the TRINITY program[417]. The commands and parameters used for running TRINITY were as follows: Trinity --seqType fq --JM 200G --left sample_1.fq --right sample_2.fq --normalize_by_read_set --CPU 32 --output sample --min_kmer_cov 2. Protein sequences and coding sequences of transcripts were predicted using TransDecoder (http://transdecoder.github.io), a software tool that identifies likely coding sequences from transcript sequences and compares the translated coding sequences with the PFAM domain

database[417]. For genes with more than one transcript, the longest one was used to calculate transcript abundance and coverage. Transcript abundance level was normalized using the Fragments Per Kilobase per Million mapped reads (FPKM) method.

Transcriptomes of *Agave deserti*[418] and *Allium cepa*[419] were downloaded from Dryad (h5t68) and NCBI (PRJNA175446), respectively. We removed the redundant unigenes in *A. cepa* by CD-HIT-EST with 99% identity and used TransDecoder to predict proteins with default parameters.

We carried out BLASTP (E-value $<1 \times 10^{-3}$) to search the best hits for the proteins predicted in the transcriptomes against a customized database, built with proteins from the genomes of *A. shenzhenica*, *P. equestris*[188], *D. catenatum*[189], and *A. officinalis* (GenBank accession number GCF_001876935.1) as well as public databases, such as NCBI Plant RefSeq (release 80), Ensembl (release 77), Ensembl Metazoa (release 24), Ensembl Fungi (release 24), and Ensembl Protists (release 24). Only plant-homologous proteins were retained in the transcriptomes to eliminate the effects of genes derived from commensal organisms, laboratory contaminants, and artefacts resulting from incorrect assembly (Supplementary Table E-13).

4.3.8 Gene family identification

We downloaded genome and annotation data of *A. trichopoda* (http://amborella.huck.psu.edu, version 1.0), *A. comosus* (GenBank accession numberGCF_001540865.1), *A. thaliana* (TAIR 10), *A. officinalis* (GenBank accession number GCF_001876935.1), *B. distachyon* (purple false brome; Phytozome v9.0), *M. acuminata* (http://ensemblgenomes.org, release-21), *O. sativa* (Nipponbare, IRGSP-1.0), *P. dactylifera* (http://qatar-weill.cornell.edu/research/datepalmGenome), *P. trichocarpa* (http://ensemblgenomes.org, release-21), *S. bicolor* (sorghum; Phytozome v9.0), *S. polyrhiza (common duckweed; http://www.spirodelagenome.org), and V. vinifera (Phytozome v9.0).* We chose the longest transcript to represent each gene and removed gene models with open reading frames shorter than 150 bp. Gene family clustering was performed using OrthoMCL[65] based on the set of 21,841 predicted genes of *A. shenzhenica* and the protein sets of the above ten other monocots, three dicots and the outgroup *A. trichopoda.* This analysis yielded 11,995 gene families in *A. shenzhenica* containing 18,268 predicted genes (83.6% of the total genes identified; orthologous genes in the 15 sequenced plant species are shown in Supplementary Figure E-26 and Supplementary Table E-14).

4.3.9 Phylogenetic tree construction and phylogenomic dating

We constructed a phylogenetic tree based on a concatenated sequence alignment of 439 single-copy gene families from *A. shenzhenica* and the 14 other plant species using MrBayes[420] software with GTR+Γ model (Figure 4-2). For the phylogenetic analysis incorporating ten additional transcriptome species (Supplementary Figure E-2), we first picked up the genes of *A. shenzhenica*, *D. catenatum*, and *P. equestris* in the single-copy gene families as seed genes, and then made a BLASTP alignment between the transcriptome unigenes and the seed sequences. For one single-copy family, if the three seed genes all had the identical best-hit to a unigene, this gene was identified as the orthologous gene to the gene family. With this method we found 132 single-copy gene families of the total 25 species,

then constructed the phylogenetic tree based on a concatenated sequence alignment of them using PhyML[99] with GTR+Γ model. Divergence times were estimated by PAML MCMCTREE[345]. The Markov chain Monte Carlo (MCMC) process was run for 1,500,000 iterations with a sample frequency of 150 after a burn-in of 500,000 iterations. Other parameters used the default settings of MCMCTREE. Two independent runs were performed to check convergence. The following constraints were used for time calibrations: (i) The *O. sativa* and *B. distachyon* divergence time (40–54 million years ago (mya))[421]; (ii) The *P. trichocarpa* and *A. thaliana* divergence time (100–120 mya)[228]; (iii) The monocot and eudicot divergence time with a lower boundary of 130 mya[422]; and (iv) 200 mya as the upper boundary for the earliest-diverging angiosperms[423].

4.3.10 Identification of WGD events in *A. shenzhenica* and phylogenomic analyses

K_S-based age distributions were constructed as previously described[179]. In brief, the paranome was constructed by performing an all-against-all protein sequence similarity search using BLASTP with an *E* value cutoff of 1×10^{-10}, after which gene families were built with the mclblastline pipeline (v10-201) (micans.org/mcl)[64]. Each gene family was aligned using MUSCLE (v3.8.31)[78], and K_S estimates for all pairwise comparisons within a gene family were obtained through maximum likelihood estimation using the CODEML program[268] of the PAML package (v4.4c)[345]. Gene families were then subdivided into subfamilies for which K_S estimates between members did not exceed a value of 5. To correct for the redundancy of K_S values (a gene family of n members produces $n(n-1)/2$ pairwise K_S estimates for $n-1$ retained duplication events), a phylogenetic tree was constructed for each subfamily using PhyML[99] under default settings. For each duplication node in the resulting phylogenetic tree, all m K_S estimates between the two child clades were added to the K_S distribution with a weight of $1/m$ (where m is the number of K_S estimates for a duplication event), so that the weights of all K_S estimates for a single duplication event summed to one. The resulting age distribution of the *A. shenzhenica* paranome is shown in Figure 4-4A.

Absolute dating of the identified WGD event in *A. shenzhenica* was performed as previously described[168,188]. In brief, paralogous gene pairs located in duplicated segments (anchors) and duplicated pairs lying under the WGD peak (peak-based duplicates) were collected for phylogenetic dating. Anchors, assumed to correspond to the most recent WGD event, were detected using i-ADHoRe (v3.0)[176,424]. Their K_S distribution is shown in Figure 4-4B. The identified anchors confirmed the presence of a WGD peak near a K_S value of 1 (the long tail and additional peaks in the anchor pair distribution are most likely due to small saturation effects[179] and the remnants of older WGD events in the monocot lineage, such as the τ WGD[98,185]). We selected anchor pairs and peak-based duplicates present under the WGD peak and with K_S values between 0.6 and 1.4 (dashed lines in Figure 4-4A, B) for absolute dating. For each WGD paralogous pair, an orthogroup was created that included the two paralogues plus several orthologues from other plant species as identified by InParanoid (v4.1)[425] using a broad taxonomic sampling: one representative orthologue from the order Cucurbitales, one from the Rosales, two from the Fabales, one from the Malpighiales, two from the Brassicales, one from the Malvales, one from the Solanales, two from the Poaceae (Poales), one from *A. comosus*[185] (Bromeliaceae, Poales), one from either *M. acuminata*[426] (Zingiberales) or *P. dactylifera*[427] (Arecales), and one orthologue from the Alismatales, either from *S. polyrhiza*[428] or *Zostera marina*[429]. In total, 85 orthogroups based on anchors and 230 orthogroups based on peak-based duplicates were collected. The node joining the two *A. shenzhenica* WGD

paralogues was then dated using the BEAST v1.7 package[113] under an uncorrelated relaxed clock model and an LG+G (four rate categories) evolutionary model. A starting tree with branch lengths satisfying all fossil prior constraints was created according to the consensus APGIV phylogeny[20]. Fossil calibrations were implemented using log-normal calibration priors on the following nodes: the node uniting the Malvidae based on the fossil *Dressiantha bicarpellata*[430] with prior offset = 82.8, mean = 3.8528, and SD = 0.5[431]; the node uniting the Fabidae based on the fossil *Paleoclusia chevalieri*[432] with prior offset = 82.8, mean = 3.9314, and SD = 0.5[433]; the node uniting the *A. shenzhenica* WGD paralogues with the other non-Alismatalean monocots based on fossil *Liliacidites*[190] with prior offset = 93.0, mean = 3.5458, and SD = 0.5[434]; and the root with prior offset = 124, mean = 4.0786, and SD = 0.5[435]. The offsets of these calibrations represent hard minimum boundaries, and their means represent locations for their respective peak mass probabilities in accordance with some recent and most taxonomically complete dating studies available for these specific clades[436]. A run without data was performed to ensure proper placement of the marginal calibration prior distributions[437]. The MCMC for each orthogroup was run for 10 million generations with sampling every 1,000 generations, resulting in a sample size of 10,000. The resulting trace files of all orthogroups were evaluated manually using Tracer v1.5[113] with a burn-in of 1,000 samples to ensure proper convergence (minimum ESS for all statistics was at least 200). In total, 303 orthogroups were accepted, and all age estimates for the node uniting the WGD paralogous pairs were then grouped into one absolute age distribution (Figure 4-5; too few anchor pairs were available to evaluate them separately from the peak-based duplicates), for which KDE and a bootstrapping procedure were used to find the peak consensus WGD age estimate and its 90% confidence interval boundaries, respectively. More detailed methods are available in Vanneste *et al.*[168].

To compare the relative timing of speciations and WGD event(s) in orchids based on K_S distributions, we first identified 839 anchors from *D. catenatum* and 355 anchors from *P. equestris* using i-ADHoRe 3.0 and calculated their K_S as described above. Identification of orthologues between *A. shenzhenica* and *A. officinalis*, *A. shenzhenica* and *P. equestris*, *A. shenzhenica* and *D. catenatum*, and *P. equestris* and *D. catenatum* was performed first by reciprocal BLASTP with *E* value $<1 \times 10^{-5}$ for proteins from the three orchids and asparagus, followed by sorting BLAST hits by bit-scores and *E* values. Reciprocal best hits in the four comparisons were selected as orthologues. In this way, we identified 9,142 orthologues between *A. shenzhenica* and *A. officinalis*, 10,699 orthologues between *A. shenzhenica* and *P. equestris*, 11,386 orthologues between *A. shenzhenica* and *D. catenatum*, and 13,139 orthologues between *P. equestris* and *D. catenatum*. For each pair of orthologues, ClustalW[343] alignment was carried out to perform sequence alignment using the parameter for amino acids recommended by Hall[265]. PAL2NAL[83] was then used to back-translate aligned protein sequences into codon sequences and to remove any gaps in the alignment. Estimates of K_S values were obtained from CODEML in PAML using the Goldman-Yang model with codon frequencies estimated by the F3 × 4 model.

We performed pairwise co-linearity analysis within *A. shenzhenica* and between *A. shenzhenica* and *A. officinalis*, *A. comosus*, *V. vinifera*, and *A. trichopoda*. Homologous pairs of *A. shenzhenica* and the above species were identified by all-against-all BLASTP (*E* value $<1 \times 10^{-5}$), followed by the removal of weak matches by applying a *c*-score of 0.5 (indicating their BLASTP bit-scores were below 50% of the bit-scores of the best matches)[96]. Then, i-

ADHoRe 3.0 was used to identify co-linear segments with parameters as described above except using 'level_2_only = FALSE', enabling the functionality to detect highly degenerated co-linear segments resulting from more ancient large-scale duplications (this is achieved by recursively building genomic profiles based on relatively recent co-linear segments). All co-linear dot plots were drawn by selecting co-linear segments according to a specified required number of anchor pairs (given in the figure legend of each of the dot plots). For the comparisons between *A. shenzhenica* and the chromosome-level assembled genomes (*A. officinalis*, *A. comosus*, and *V. vinifera*) we retained co-linear segments with at least ten anchor pairs (Figure 4-7 and Supplementary Figures E-7, E-9, and E-10). For the comparisons with fragmented genomes, like *A. trichopoda*, and the self-comparison of *A. shenzhenica*, we kept co-linear segments with five anchor pairs (Figure 4-6B and Supplementary Figures E-5 and E- 6). The start and end boundaries of selected co-linear segments were used to define broader regions containing such segments on the chromosomes or scaffolds by further connecting co-linear segments if they overlapped with each other. Then, duplication depths, *i.e.*, the number of connected co-linear segments overlapping at each position of a broader region, were illustrated in the margins of the plots by mapping the connected co-linear segments over each other. The number of anchors required in the co-linear segments could affect the duplication depth in such a way that increasing the number of anchors required tends to remove co-linear segments originating from more ancient WGD(s) due to increased gene loss.

To identify the duplication events that resulted in the 1,488 anchor pairs in *A. shenzhenica*, the 839 anchor pairs in *D. catenatum*, and the 355 anchor pairs in *P. equestris*, we performed phylogenomic analyses employing protein-coding genes from 20 species, including 12 orchids across all five subfamilies of Orchidaceae (the three orchids with genomes (*A. shenzhenica*, *D. catenatum* and *P. equestris*) plus nine orchid transcriptomes (Supplementary Table 13)), four non-orchid Asparagales (*A. officinalis* (genome), *M. capitulata* (Supplementary Table 13), *A. deserti*[418] and *A. cepa*[419]), three commelinid monocots (*Elaeis guineensis*, *P. dactylifera*, and *A. comosus*), and *A. trichopoda*. OrthoMCL (v2.0.9)[65] was used with default parameters to identify gene families based on sequence similarities resulting from an all-against-all BLASTP with *E* value $<1 \times 10^{-5}$. Then, 1,101 of the 2,582 anchor pairs with K_S values greater than five were removed. If the remaining anchors fell into different gene families, indicating incorrect assignment of gene families by OrthoMCL, we merged the corresponding gene families. In this way, we obtained 32,217 multi-gene gene families. Next, phylogenetic trees were constructed for the subset of 777 gene families with no more than 300 genes that had at least one pair of anchors and one gene from *A. trichopoda*. Multiple sequence alignments were produced by MUSCLE (v3.8.31) using default parameters. These were trimmed by trimAl (v1.4)[84] to remove low-quality regions based on a heuristic approach (-automated1) that depends on a distribution of residue similarities inferred from the alignments for each gene family. RAxML (v8.2.0)[87] was then used with the GTR+Γ model to estimate a maximum likelihood tree starting with 200 rapid bootstraps followed by maximum likelihood optimizations on every fifth bootstrap tree. Gene trees were rooted based on genes from *A. trichopoda* if these formed a monophyletic group in the tree; otherwise, mid-point rooting was applied. The timing of the duplication event for each anchor pair relative to the lineage divergence events was then inferred using the following approach (Supplementary Figures E-11): we first mapped internodes from a gene tree to the species phylogeny according to the common ancestor of the genes in the gene tree. Each internode of the gene tree was then

defined as either a duplication node, a speciation node, or a 'dubious' node. A duplication node is a node that shares at least one pair of paralogues, a speciation node is a node that has no paralogues and is consistent with divergence in the species phylogeny, and a 'dubious' node is a node that has no paralogues and is inconsistent with divergence in the species phylogeny. Then, if a pair of anchors coalesced to a duplication node, we traced back its parental node(s) until we reached a speciation node in the gene tree. In this way, we circumscribed the duplication event as between these two nodes with the duplication node as the lower bound and the speciation node as the upper bound on the species tree. If the two nodes were directly connected by a single branch on the species tree, the duplication was thus considered to have occurred on the branch. To reduce biased estimations, we used the bootstrap value on the branch leading to the common ancestral node of an anchor pair as support for a duplication event. In total, 628 anchor pairs in 493 gene families coalesced as duplication events on the species phylogeny, and duplication events from 318 anchor pairs in 262 gene families (or from 448 anchor pairs in 367 gene families) had bootstrap values greater than or equal to 80% (or 50%).

4.3.11 Evolution and expression analysis of orchid MADS box genes

We identified candidates of MADS-box genes by searching the InterProScan[413] result of all the predicted *A. shenzhenica* proteins. The candidates of MADS-box genes were further determined by SMART[438], which identified MADS-box domains comprised by 60 amino acids. The protein-sequence set of the MADS-box gene candidates was BLAST against the assembled *A. shenzhenica* transcriptomes with the TBLASTN program. The matched transcript sequences were then assembled with the candidates of MADS-box genes using Sequencher v5.1 (Gene Codes Corp.) and the exon structure of the final MADS-box genes was manually edited. In the end, we aligned all the identified MADS-box genes using the ClustalW program[343]. An unrooted neighbour-joining phylogenetic tree was constructed in MEGA5[439] with default parameters.

4.3.12 Transcriptomic analysis of other orchids

In addition, 53 more transcriptomes derived from 9 more taxa and 8 tissues (flower bud, anther, pollinium, shoot, stem, leaf, aerial root and root) (Supplementary Table E-13) were sampled to investigate the roles of the genes that may be important for the evolution of orchid traits. The gene expression levels were indicated by FPKM on the longest assembled transcript.

4.4 Acknowledgements

We acknowledge support from The Funds for Environmental Project of Shenzhen, China (no. 2013-02); The 948 programme of the State Forestry Administration P. R. China (no. 2011–4–53); The Funds for Forestry Science and Technology Innovation Project of Guangdong, China (no. 2016KJCX025; no. 2013KJCX014-05); Fundamental Research Project of Shenzhen, China (no. JCYJ20170307170746099; no. JCYJ20150403150235943); and the teamwork projects funded by Guangdong Natural Science Foundation (no. 2017A030312004) Z.-J.L. Y.V.d.P. acknowledges the Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides

to networks' Project (no. 01MR0310W) of Ghent University and the European Union Seventh Framework Programme (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739–DOUBLEUP.

4.5 Author contributions

Z.-J.L. managed the project; Z.-J.L., G.-Q.Z., Y.V.d.P., K.-W.L., W.-C.T., Y.-B.L. and C.-M.Y. planned and coordinated the project; Z.-J.L., K.-W.L., W.-C.T., Y.V.d.P., R.L. and Z.L. wrote the manuscript; Z.-J.L., L.-J.C., S.-C.N., M.W., G.-H.L., X.-J.X., H.-X.H., J.-Y.W., S.-J.Z. and L.P. collected and grew the plant material; Q.X., Z.-J.L., W.-C.T., K.-W.L., L.-J.C., X.-Y.W. and M.L. prepared samples; G.-Q.Z., Z.-J.L. and Y.-Q.Z. sequenced and processed the raw data; Z.-W.W., Z.-J.L., G.-Q.Z., K.Y., S.F., N.M., S.S., M.O.-T., M.Y. and C.-M.Y. annotated the genome; Z.-J.L., Z.-W.W., W.-C.T. and G.-Q.Z. analysed gene families; Z.L., R.L., Y.V.d.P. and Y.-C.L. conducted whole-genome duplication analysis; W.-C.T., Y.-Y.H., Z.-J.L., C.-M.Y., S.-B.C., W.-L.W., Y.-Y.C.,C.-Y.S. and K.-W.L. conducted the MADS-box gene analysis; Z.-J.L., G.-Q.Z., Y.-Q.Z., K.-W.L., S.S., M.O.-T. and C.-M.Y. conducted transcriptome sequencing and analysis.

Chapter 5

Concluding remarks and future perspectives

“It is said that despite its many glaring (and occasionally fatal) inaccuracies, The Hitchhiker's Guide to the Galaxy *itself has outsold the* Encyclopedia Galactica *because it is slightly cheaper, and because it has the words 'DON'T PANIC' in large, friendly letters on the cover.”*

--Douglas Adams

“The Hitchhiker's Guide to the Galaxy”

5.1 Concluding remarks

Phylogenomics, in a broad sense, has extended molecular phylogenetics and enlightened genomics with evolutionary perspectives[440], just as the memorable quote saying, "nothing in biology makes sense except in the light of evolution"[441]. The development of sequencing technologies enables the fast accumulation of genomic data in many current fields of biology, into which phylogenomics has been introduced. Indeed, it is not surprising to see the 'shadows' of phylogenomics behind functional genomic studies nowadays. For instance, phylogenomics allows transferring the knowledge of genes from well-studied species to their orthologs in under-examined species, as orthologous genes hold the idea that they perform similar, if not the same, functions in different species[68]. Although orthologous genes could be identified through sequence similarity searching, a phylogenetic perspective of orthologous genes provides robustness to distinguish orthologs and paralogs by explicitly addressing gene duplication and loss[72]. This is not the only example, because the integration of phylogenomics into other fields of biology has never stopped. Among these, if inferring phylogenies for each part of the tree of life is a natural extension of molecular phylogenetics in the era of genomics, inferring the evolutionary history of gene families and providing evidence for the divergence time of species and genes are the combination between phylogenomics and comparative genomics. Because gene trees are the results of gene genealogies, their conflicting topologies and branching patterns could also be used to estimate population histories of species, such as effective population size and population dynamics, leading phylogenomics into the field of population genomics[138,442,443].

The thesis has shown that multiple single-copy genes identified from transcriptomes and genomes can resolve the phylogeny of seed plants when taking care of the sites that violate the homogeneous assumption of current substitution models (Chapter 2). The phylogenetic context with a broad taxonomy sampling also enables us to further investigate the differences in evolutionary rates among major lineages of seed plants (Appendix B.1). Then, it has shown that the inference of gene duplications based on phylogenomics and comparative genomics could shed light on the general consistency of the biased gene retention following gene and genome duplications. Core gene families in angiosperms either primarily remain in single-copy or multi-copy status in all species. The existence of an intermediate group of gene families has also been observed, which mostly consist of regulatory genes. These genes seem to survive for a long time following WGD, suggesting that they may be dosage balance sensitive (Chapter 3). The conjunction is mostly confirmed by a study using the same data set to study gene families that retained reciprocally after WGDs and SSDs based on a stochastic model taking into account both discrete WGDs and continuous SSDs. Both the intermediate group and the multi-copy group are enriched with gene families retained reciprocally following WGDs. The study also suggests that the multi-copy group includes some gene families that tend to be retained after SSDs rather than WGDs, in agreement with our observation that both duplicates from SSDs and ancient WGDs are enriched in the multi-copy group (Appendix B.2). This might be the case for the reason that the expansions of multi-copy gene families could be a result of both SSDs and WGDs. For example, the MADS-box gene family in flowering plants may have expanded by a tandem duplication predating the divergence of seed plants with subsequent polyploidy events in angiosperms[149]. The phylogenomic approach to determine the timing of gene duplications has also been employed to sort paralogous genes showing coordinately diverged expression patterns into the WGDs

occurred over the evolution of *A. thaliana* (Appendix B.3). The last study in the thesis has shown that phylogenomic analysis could supplement, or to some extent improve upon, the identification of WGD in extent orchids (Chapter 4). A similar approach has also been applied to examine competing hypotheses for the placement of a recent WGD in the genome of wild olive along with a framework for testing phylogenetic hypotheses (Appendix B.4).

For phylogenomics, the taxonomy and sequence sampling in the thesis reflects a general trend in the field, that phylogenomic studies may easily dive into dozens to hundreds of species with thousands of gene sequences at present or in the near future. To deal with those large data sets, we have used various state-of-the-art methods not only from phylogenomics but also from comparative genomics. Most potential issues on the methods and data have been discussed after each study, so the following paragraph will generally discuss sequence and taxonomy sampling in the three studies. For sequence sampling, data missing in an alignment matrix is a potential issue in phylogenomic analysis. Data missing is due to erroneous sample preparation and/or sequencing, fragmented assembly of genome and transcriptome, or some other artificial factors. It should appear as unknown states in an alignment matrix, but it is actually difficult to distinguish data missing from genuine alignment gaps resulted from sequence evolution, like insertion and deletion or gene merging and split. Therefore, most phylogenetic tools treat data missing and alignment gaps in the same way[87,345]. The treatment may have less affected with sequences obtained from Sanger sequencing because it usually can generate complete sequences in each species if sequencing primers work well. However, NGS is more vulnerable to the factors causing data missing because of its short reads. In the study of the seed plant phylogeny in Chapter 2, we have 19% of the data accounted as gaps or missing data, which is comparable to or less than other recent phylogenomic studies, for example 37.1% gaps for the nucleotide genes and 14.1% gaps for the plastid genes in Xi *et al.*[156] or 40% gaps for the stringent set of nucleotide genes in Wickett *et al.*[70]. Although data missing has sometimes been used to argue for the possible reasons causing incongruent phylogenies[198], reducing gaps from 19% to 7.6% did not affect the position of gnetophytes in our ML analyses based on the amino acid alignment. The relationships between data missing and its effects on phylogeny are still unclear so far, with one study suggesting its ignorable effects in large data sets[444], and another study indicating its effects depending on the patterns of data missing[445]. For the studies in the thesis, we tried to trim sequence alignments to remove regions that were affected by spurious gaps to reduce the effects of data missing. However, it would be worth exploring in detail how exactly gene trees would be affected by data missing taking into account the limited number of sites at a locus.

The taxonomy sampling in the thesis, especially for Chapter 3, is limited by the available genome sequences at the time. The species sampling in Chapter 3 shows a bias towards eudicot genomes. Although further including newly sequenced monocot species may not affect the general conclusions in Chapter 3, it may affect the classification of some gene families that are at the boundaries of the three groups, likely for gene families that may contribute to the differences between monocots and eudicots. Also, the relatively sparse taxonomy sampling of flowering plant genomes in Chapter 3 could not be used to distinguish ancient autopolyploid and allopolyploid, which may lead to underestimating their different effects on gene retention during diploidizations. Although this topic is probably already out of the scope of the study, with the increasing ability of species sampling and growing power to distinguish autopolyploid and allopolyploid by phylogenomics[144,186,187], we may gain

particular perspectives on the evolutionary process following polyploidizations and test existing hypotheses about ancient polyploidy events and speciation. When considering species sampling, we also need to take the quality of genomes and/or transcriptomes into account. The qualities of the assembly and annotation of genomes and transcriptomes are different from studies to studies. Although most genomes are of reasonable quality in the study in Chapter 3, some of them, like *Medicago truncatula* and *Lotus japonicas*, performed like outliers in the analysis. The fact that they are partial genomes[446,447] is a factor that we did consider when selecting the core gene families. In Chapter 4, the genomes of two published orchids, *i.e.*, *P. equestris* and *D. catenatum*, have been improved with a similar strategy as the one used in assembling the *A. shenzhenica* genome. However, because of their larger genome sizes than *A. shenzhenica*, the improved genomes of *P. equestris* and *D. catenatum* still cannot provide sufficient co-linear evidence for the identified WGD in extant orchids. This suggests that even we can obtain sequences from almost any species as we required, to bring them into comparable qualities still demands different efforts.

In general, we are now in the era where sequence data are produced at a rate much faster than the speed that we can digest[56]. Although novel DNA sequencing technologies will further enlarge the gap between data generation and data analysis, they also promise to solve some issues that we are confronted with today, like genome continuity, structural variation calling, or spatiotemporal changes of transcriptomes, by supplying extra information like long sequencing reads with low error rates[44]. However, the data volume may be increased both by total amounts of sequences in one sample and by extensive sampling from different taxonomies. In any case, sequence data could be accumulated in a way that is out of our imagination, just as it has done to biologists before the start of the Human Genome Project. A large amount of sequence data is often considered as an opportunity to phylogenomics because large data sets are conceivably able to fit complex models and to increase statistical confidence[448]. However, to some extent, the amount of new data coming our way could be overwhelming. It has been evident that our current methods sometimes run short of ways in dealing with the quickly accumulated data as discussed below.

5.2 Tremendous amounts of sequences

The increased ability of sequencing has dramatically improved sequencing depth for a sample and hence dramatically expanded the number of sequence sites available for phylogenomic studies. Although studying phylogenomics needs to be aware of potential issues with large datasets, the ever-growing volume of data has brought about many unexpected outcomes. More sequence sites ideally could improve our confidence in the process of statistical inference, like estimating parameters during phylogenetic inference. However, this requires appropriate specification of models for sequence evolution. Otherwise, the increasing amounts of sequence sites tend to shelter the effects of model misspecifications. Because a majority of current phylogenomic analyses are based on overly simplistic models that cannot accommodate the realistic process of sequence evolution[128], the growing volume of sequence sites would incur extra risks on increasing precision of an inaccurate estimate. Here, the precision shows our confidence on an estimate, while the accuracy reflects how our estimate is close to the reality[124]. A simple example from Kumar *et al.*[126] shows such risks by using the JC model to estimate the distance between sequences simulated under the GTR model with

an imposed distance. The distributions of the estimated JC distances based on various lengths of sequences all have the same mean value, which is different from the imposed distance used for the simulation. However, the standard deviation of the estimated JC distances reduces when the sequences have more sites, suggesting our confidence increases on the incorrect estimate when the sequence gets longer. In practice, a similar situation could be spotted in the commonly used bootstrap analysis with long sequence alignments. Especially for phylogenetic trees that are inferred based on the concatenated alignments of multiple gene loci, certain systematic biases are sometimes reinforced with spuriously high bootstrap values[449] or posterior probabilities[450]. What makes the situation more imperative is the observation that removing a few genes or a few sites from an alignment could reduce the bootstrap support of the tree inferred from the alignment or even turn the topology into another topology with high bootstrap support[451].

Analyzing sequences of each gene locus, on the other hand, may sometimes suffer from the fact that short sequence alignments have insufficient sites for complex models. However, they also tend to be less affected by the overconfidence on the phylogenetic inference. This suggests that gene trees with high bootstrap values[123] or posterior possibilities often have strong phylogenetic signals to solve previous incongruent branches. As current sequencing technologies could obtain sequences from many gene loci, reporting gene trees can illustrate the strength of phylogenetic signals[123] and uncover the discordant gene histories formed during evolution[452]. In future, the performances of gene tree inference under various models may need to be explored. It is still unclear how model misspecification would affect gene tree inference, although most studies make use of an identical model to infer all of the gene trees. Simulating gene trees and sequences may allow us to explore such questions with specified models and controlled factors. The comparisons between simulated and empirical data can pinpoint potential issues in empirical phylogenomic studies.

5.3 Broad and deep taxonomic sampling

The advance of sequencing technology allows us to sample species from organisms in a broad range to individuals existing in a population. The lack of broad sampling of species has long been considered as a deficiency, especially when dealing with LBA[453]. Because LBA is from mistakenly recognizing homoplastic traits as shared derived traits (synapomorphies), sampling species in the lineage leading to a long branch could supplement phylogenetic inference with evident information to distinguish genuine synapomorphies from false signals. However, including more species may result in a higher complexity in phylogenomic analysis. An increasing number of species would not only impede orthologous identification and multiple sequence alignment[68] but also increase the chance to violate current models by introducing more data heterogeneities than before. Most of the substitution models applied to present data have been developed for handling a small number of sequences under specific homogenous assumptions. Although nearly no sequences could meet the assumption of homogeneity, heterogeneities become so severe in broadly sampled taxonomies that could violate the current models by differences in character compositions and variations in substitution rates across lineages. In fact, adding slowing evolving species to subdivide long branches could not always alleviate LBA but sometimes reduce the accuracy of phylogenetic inference[454].

The increasing depth of species sampling, notably dense sampling from closely related species, could be used to study the phylogenetic discordance caused by different evolutionary processes, such as incomplete lineage sorting and introgression[455]. For example, introgression on specific loci across genomes can be distinguished from incomplete lineage sorting by the ABBA-BABA statistics[456,457] using single nucleotide polymorphisms across genomes from closely related populations or species. It also enables phylogenomics to revisit the acknowledged evolutionary events identified previously with sparse samplings of species, as shown in the study that has distinguished paleoallopolyploid and paleoautopolyploid for the WGD identified in *Saccharomyces cerevisiae*[186]. The increasing dense sampling of yeasts has helped to find that paralogs from the paleopolyploidization mostly coalesce to a branch prior to the branch following which all the species shared the WGD. This indicates that the WGD is likely an allopolyploidization event from two closely related species, so the coalesced branch of the paralogous genes is, in fact, the time when the two parental species split. In general, incomplete sampling or species extinction can complicate the process of distinguishing ancient autopolyploid and allopolyploid, which add new complications to investigations on the effects of paleopolyploidy[144]. The dense sampling could also cause unexpected issues in phylogenomics, such as resulting in discordance among gene trees by adding extra species. Gatesy *et al.* have shown that adding orthologous genes from a relatively distant species to 106 congruent gene trees from five species has led to incongruent topologies because of introducing a new long branch[458]. Alternatively, adding lineages with species breaking branches of a tree could decrease internal branch lengths in both gene trees and the species tree, and again increase gene tree discordance due to the short divergence time[138]. It may cause problems, unexpectedly when most of the evolutionary processes behind the discordance are not modeled formally. Indeed, short and deep branches usually have incongruent phylogenetic signals that may require appropriate practices and models to address[123,154].

Therefore, species sampling might be worth being explored on its effects for gene trees and species trees, as increased species sampling may work for or against a phylogenomic analysis. Unfortunately, many factors could affect species sampling, for example, extinction, geographical distributions, and abundance of different species as well as various conservation policies. In most of the time, we cannot know *a priori* the effects of species sampling before we actually do a phylogenomic analysis. The ease of sequencing in future may enable us to replenish species sampling immediately after a pilot study. Before that, available integrated databases with thousands of phylogenomic studies, such as TTOL, may provide some hints on species sampling for us to avoid missing critical lineages for specific biological questions.

5.4 Potential issues in sequencing

The recently produced genomes and transcriptomes are often accorded far more confidence than they warrant. Extra efforts have always been required to improve the output by dealing with either the short sequencing reads generated by NGS or the high error rate in single molecular sequencing. In particular, for NGS, the short sequencing reads cause a series of issues on assembly and annotation of genomes and transcriptomes. Genome sequencing usually promises to provide a complete gene catalog for downstream analysis[459], but incomplete genome assemblies produced by short sequencing reads could result in

fragmented gene models and an incorrect number of genes[252]. Some released genomes still need to be improved by complementing additional contiguity information from genetic and physical mapping, long read sequencing technologies, and/or recently developed genome-wide chromosome conformation capture (Hi-C)[44]. Compared with genome sequencing, the above issues are more severe in transcriptome sequencing. To obtain a gene catalog of a species, *de novo* transcriptome sequencing takes less time and cost much less than *de novo* genome sequencing, so it has become a commonly used approach. However, transcriptome sequencing can only produce an inventory of expressed genes in a given tissue, at a given time, and under a specific condition, hence usually leading to an incomplete gene set. Although combining transcriptomes from different tissues under several conditions can complement with each other, the strategy may still fail to deal with low-expressed genes and chimeric assemblies from different isoforms. The incomplete data sets from either genomes or transcriptomes could disturb orthologous identification in studies of phylogenetic relationships[445,460] or mislead gene family history in phylogenomic analyses[252]. Various assembly chimeras could erroneously introduce gaps during MSA hence adding complexity or shortening sequence alignment for phylogenetic inference. Together with the incompleteness of data, unusually high gene numbers are occasionally observed in transcriptomes. They are often overlooked because assembled genes from transcriptomes are often considered as fragmented. Indeed, fragmented genes could burst the number of predicted genes in transcriptomes or even in genomes[217], but the surprising number of genes could also come from commensal organisms and laboratory contaminants. To deal with this, known proteins of the sampled lineage deposited in public databases are employed to filter out artifacts from incorrect assemblies. However, the procedure is just a compromise between the current transcriptome assembly algorithms and the requirement on species and sequence sampling in phylogenomic analyses.

5.5 DON'T PANIC

Despite many issues in phylogenomics caused by nowadays sequencing technologies, the flood of sequence data available today is still good news if we consider that it grants us the possibility to choose the data that we need to answer biological questions, but not to use all the generated data like previously[40,124,125]. Although selecting data would sometimes have a risk of tuning data for specific conclusions unintentionally, what we could do is to lower such a risk by keeping a reliable record and establishing effective communication on data processing. It has become an open solution for every area that needs to analyze large datasets. In phylogenomics, the careful collection of data could minimize systematic errors from model misspecification and maximize the phylogenetic signals[124]. Many studies have shown that data filtering with objective criteria is an efficient way to keep a sufficient number of sites hence entailing sound results in phylogenetic inference. For example, Zhong *et al.*[135,201] have shown that removing non-time reversible sites could change the phylogenetic placement of Gnetophytes. The similar pattern is also observed in Chapter 2 and by Wickett *et al.*[70] after removing the fast-evolved 3rd codon positions. Salichos and Rokas[123] have argued to remove genes with weak phylogenetic signals, which are reflected as gene trees with low bootstrap values, to reduce incongruences on short and deep branches. Hahn[340] has suggested collapsing branches with low bootstrap values to avoid biases in gene tree – species tree

reconciliations. Accordingly, bootstrap values are commonly considered as a cut-off in the phylogenomic methods for WGD identification, as implemented by Jiao *et al.*[169,182], McKain *et al.*[183], and in Chapter 4. Mirarab *et al.*[154] have even removed gene trees with likely incorrect topologies according to a statistical binning of gene tree topologies. Apparently, these criteria could be barely formalized into general rules, but it is not a surprise that data selection has become a principle for phylogenomics, although it apparently depends on questions, data quality, as well as characteristics of the studied lineages.

The discordance in phylogenomics that has been uncovered gradually by the vast amounts of data also drives us to build sophisticated models in further to handle the potentially complex signals harbored in sequence data[124,461]. Many currently heated debates in plant phylogenies have been lasted for a long time, such as the phylogenetic placement of gnetophytes in seed plants[135,156,201], the placement of sponges in metazoan[451,462], and the relationship between Amborellales and Nymphaeales in angiosperms[155,170,463]. Models that can accommodate data heterogeneities and model the evolutionary process underlying the phylogenomic discordance may resolve these debates. In spite of the multispecies coalescent model that considers incomplete lineage sorting, we further need models for introgression, hybrid speciation, and gene duplication and loss[128]. To some extent, the increased data volume has zoomed in the resolution of phylogenomics. When only limited gene sequences were available, species trees and gene trees were considered as equivalent to each other. With many gene sequences at hand, we have emphasized the differences of gene trees from species trees. Once we can obtain multiple alleles in different species, we would realize that allele trees, locus trees, and species trees are affected by different but interacted evolutionary processes[464]. Therefore, with the growing data volume, probably contributed by data from populations, the improvements of models might not only resolve the tree of life but also enlighten our further understanding of evolution itself[124,128,138,465].

Last but not least, we also need to confront the increase of phylogenomic data by novel analytical methods. How to make a good use of the increased data to gain comprehensive insights of evolution is still an open question that is worth exploring. On the one hand, understanding of the performance of current models and methods in detail could shed light on the development of new methods. For example, to mitigate problems during the gene tree – species reconciliation in Chapter 3, we rearranged branches in the ML gene trees to get tree topologies that offer an acceptable compromise to both reconciliation criterion and tree inference criterion. It has been suggested that the rearranged tree topologies that are accepted by the approach are often nearly ML topologies, so the approach could obtain solid reconciliation results as well as remain statistical meanings for the reconciled trees[342]. In future, the implementation of phylogenetic simulations on both trees and sequences would provide more insights on understanding the interactions of different evolutionary processes[464]. On the other hand, further integration of phylogenomic data with other genomic data could be another feasible option. The combination needs novel ideas and methods to organize the expanded phylogenomic data. It could benefit both evolutionary biology and genomics. For instance, integrating phylogenomic data with synteny information in plant genomes has given explicitly evolutionary trajectories for gene and gene families, which provide new hypotheses for further evolutionary and functional studies[466,467]. Sorting phylogenomic data into functional networks is also able to uncover new patterns in the evolution of functional modules[468,469].

In conclusion, the period has begun to fade off, that we need to use fast-speed phylogenomic approaches to chase the rush of generated sequences. In the coming future, although even more sequence data are in order, appropriate phylogenomic analysis with critical species sampling, well-performed sequencings, sophisticated models, along with high-performance computing is required to gain detailed insights on the process of adaptation and diversification of plants by studying the evolution of plant genomes.

Appendices

A. Academic CV

Personal information

Name	Zhen Li
Address	Burggravenlaan 32 bus 207, 9000 Gent, Belgium
E-mail	lizhen.cmb@gmail.com
Phone	+32 488 29 45 19
Date of birth	23/11/1985
Place of birth	Urumqi, Xinjiang, China

Education

2012.9-2018.4	**Doctoral researcher** Ghent University / VIB – Gent, Belgium Thesis: *The study of plant genome evolution by means of phylogenomics* Award for Outstanding Self-financed Students Abroad, China Scholarship Council
2008.9-2011.6	**Master of Science, Bioinformatics** Beijing Normal University – Beijing, China Thesis: *Effects of recent whole-genome duplication on adaptive expansion of gene families in flowering plants*
2004.9-2008.6	**Bachelor of Science, Biological science** Beijing Normal University – Beijing, China Award for Outstanding Undergraduate Student, National Institute of Biological Sciences (NIBS)

Publications

*contributed equally

1. **Li, Z.***, Zhang, Z*., Yan, P., Huang, S, Fei, Z., Lin, K. (2011) RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* 12(1): 540.

2. **Li, Z.***, Defoort, J.*, Tasdighian, S., Maere, S., Van de Peer, Y., De Smet, R. (2016) Gene duplicability of core genes is highly consistent across all angiosperms. *The Plant Cell* 28(2):326-44.

3. Kerchev, P.I., Waszczak, C., Lewandowska, A., Willems, A., Shapiguzov, A., **Li, Z.**, Alseekh, S., Mühlenbock, P., Hoebrichts, F., Huang, J., Van Der Kelen, K., Kangasjarvi, J., Fernie, A., De Smet, R., Van de Peer, Y., Messens, W., Van

Breusegem, F. (2016) Lack of GLYCOLATE OXIDASE 1, but not GLYCOLATE OXIDASE 2, attenuates the photorespiratory phenotype of CATALASE2-deficient Arabidopsis. *Plant Physiology* 171(3): 1704-1719.

4. De La Torre, A.R., **Li, Z.**, Van de Peer, Y., Ingvarsson, P.K. (2017) Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biology and Evolution* 34 (6): 1363-1377.

5. **Li, Z.**, De La Torre, A. R., Sterck, L., Canovas, F. M., Avila, C., Sierra, I.M., Cabezas, J. A., Cervera, M. T., Ingvarsson, P. K., Van de Peer, Y. (2017) Single-copy genes as molecular markers for phylogenomic studies in seed plants. *Genome Biology and Evolution* 9 (5): 1130-1147.

6. Cañas, R., **Li, Z.**, Pascual, M., Castro-Rodríguez, V., Avila, C. Sterck, L. Van de Peer, Canovas, F. (2017) The gene expression landscape of pine seedling tissues. *The Plant Journal* 91(6):1064-1087.

7. Zhang, G-Q.* , Liu, K-W.*, **Li, Z.***, Lohaus, R.*, Hsiao, Y-Y.*, Yoshida, K., Niu, S-C., Fujiwara, S., Lin, Y-C., Zhang, Y-Q., Xu, Q. , Mitsuda, N., Chen, L., Wang, M. , Liu, G-H. , Pecoraro L., Huang H-X., Xiao X-J. , Lin M., Wu X-Y., Wu W-L. , Chen Y-Y., Chang, S-B., Wang, Z. , Sakamoto, S., Ohme-Takagi, M., Yagi, M., Wang, J-Y., Zeng, S-J., Shen, C-Y., Yeh, C-M., Tsai W-C., Luo Y-B., Van der Peer, Y., Liu, Z-J. (2017) The *Apostasia* genome and the evolution of orchids. *Nature* 549(7672): 379-383.

8. Unver, T., Wu, Z., Sterck, L., Turktas, M. , Lohaus, R., **Li, Z**., Yang, M., He, L., Deng, T., Escalante, F., Llorens, C., Molina, F., Parmaksiz, I., Dundar, E., Eldem, V., Xie, F., Zhang, B., Ipek, A., Uranbey, S., Erayman, M., Ilhan, E., Badad, O., Ghazal, H., Lightfoot, D., Kasarla, P., Colantonio, V., Zararsiz, G., Tombuloglu, H., Mete, N., Cetin, O., Yang, H., Gao, Q., Dorado, G., Van de Peer, Y. (2017) Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of National Academy of Sciences* 114(44): E9413-E9422.

9. Tasdighian, S., Van Bel, M., **Li, Z.**, Van de Peer, Y., Carretero-Paulet, L., Maere, S. (2017) Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *The Plant Cell* 29(11): 2766-2785.

10. De Smet, R.*, Sabaghian, E.*, **Li, Z.***, Saeys, Y., Van de Peer, Y. (2017) Coordinated functional divergence of genes after genome duplication in *Arabidopsis thaliana*. *The Plant Cell* 29(11): 2786-2800.

11. Causier, B., **Li, Z.**, De Smet, R., Lloyd, J., Van de Peer, Y., Davies, B. (2017) Conservation of nonsense mediated mRNA decay complex components throughout eukaryotic evolutionary history. *Scientific Reports* 7(1): 16692

Conferences

2017	**Plant Genome Evolution 2017 (PGEV2017)** Stiges, Spain Poster: *Single-copy genes as molecular markers for phylogenomic studies in seed plants*
2017	**XIX International Botanical Congress (IBC 2017)** Shenzhen, China Poster: *Gene duplicability of core genes is highly consistent across all angiosperms*
2016	**Society for Molecular Biology and Evolution 2016 (SMBE 2016)** Gold Coast, Australia Poster: *Gene duplicability of core genes is highly consistent across all angiosperms*
2015	**Society for Molecular Biology and Evolution 2015 (SMBE 2015)** Vienna, Austria Poster: *Single-copy genes as molecular markers for phylogenomic studies in seed plants*
2014	**Benelux Bioinformatics Conference 2014 (BBC2014)** Luxemburg Poster**:** *Universal exome probes for sequence capture genotyping in conifers*

Trainings

2017	Career Guidance for PhDs and Postdocs
2016	Effective Graphical Displays
2016	Effective Scientific Communication
2016	N2N Multidisciplinary Seminar Series on Bioinformatics
2015	Microbial Evolution: Theory, Simulation and Experiment

B. Abstracts and contributions to other scientific publications

B.1. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants

Amanda R. De La Torre, Zhen Li, Yves Van de Peer, Pär K Ingvarsson

Molecular Biology and Evolution **34**(6): 1363-1377 (2016).

Abstract

The majority of variation in rates of molecular evolution among seed plants remains both unexplored and unexplained. Although some attention has been given to flowering plants, reports of molecular evolutionary rates for their sister plant clade (gymnosperms) are scarce, and to our knowledge differences in molecular evolution among seed plant clades have never been tested in a phylogenetic framework. Angiosperms and gymnosperms differ in a number of features, of which contrasting reproductive biology, life spans, and population sizes are the most prominent. The highly conserved morphology of gymnosperms evidenced by similarity of extant species to fossil records and the high levels of macrosynteny at the genomic level have led scientists to believe that gymnosperms are slow-evolving plants, although some studies have offered contradictory results. Here, we used 31,968 nucleotide sites obtained from orthologous genes across a wide taxonomic sampling that includes representatives of most conifers, cycads, ginkgo, and many angiosperms with a sequenced genome. Our results suggest that angiosperms and gymnosperms differ considerably in their rates of molecular evolution per unit time, with gymnosperm rates being, on average, seven times lower than angiosperm species. Longer generation times and larger genome sizes are some of the factors explaining the slow rates of molecular evolution found in gymnosperms. In contrast to their slow rates of molecular evolution, gymnosperms possess higher substitution rate ratios than angiosperm taxa. Finally, our study suggests stronger and more efficient purifying and diversifying selection in gymnosperm than in angiosperm species, probably in relation to larger effective population sizes.

Author contributions

I identified the single-copy genes used in this study and inferred the phylogeny of seed plants based on the alignment partitioned by a Bayesian mixture model. Both were done under supervision and with significant contributions of Lieven Sterck and Yves Van de Peer.

B.2. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity

Setareh Tasdighian, Michiel Van Bel, Zhen Li, Yves Van de Peer, Lorenzo Carretero-Paulet, Steven Maere

The Plant Cell **29**(11): 2766-2785 (2017).

Abstract

In several organisms, particular functional categories of genes, such as regulatory and complex-forming genes, are preferentially retained after whole-genome multiplications but rarely duplicate through small-scale duplication, a pattern referred to as reciprocal retention. This peculiar duplication behavior is hypothesized to stem from constraints on the dosage balance between the genes concerned and their interaction context. However, the evidence for a relationship between reciprocal retention and dosage balance sensitivity remains fragmentary. Here, we identified which gene families are most strongly reciprocally retained in the angiosperm lineage and studied their functional and evolutionary characteristics. Reciprocally retained gene families exhibit stronger sequence divergence constraints and lower rates of functional and expression divergence than other gene families, suggesting that dosage balance sensitivity is a general characteristic of reciprocally retained genes. Gene families functioning in regulatory and signaling processes were much more strongly represented at the top of the reciprocal retention ranking than those functioning in multiprotein complexes, suggesting that regulatory imbalances may lead to stronger fitness effects than classical stoichiometric protein complex imbalances. Finally, reciprocally retained duplicates are often subject to dosage balance constraints for prolonged evolutionary times, which may have repercussions for the ease with which genome multiplications can engender evolutionary innovation.

Author contributions

I performed and compared several phylogenetic analyses based on different alignments and substitution models. It was done under supervision of Riet De Smet and Yves Van de Peer.

B.3. Coordinated functional divergence of genes after genome duplication in *Arabidopsis thaliana*

Riet De Smet*, Ehsan Sabaghian*, Zhen Li*, Yvan Saeys, Yves Van de Peer

*contributed equally

The Plant Cell **29**(11): 2786-2800 (2017).

Abstract

Gene and genome duplications have been rampant during the evolution of flowering plants. Unlike small-scale gene duplications, whole-genome duplications (WGDs) copy entire pathways or networks, and as such create the unique situation in which such duplicated pathways or networks could evolve novel functionality through the coordinated sub- or neo-functionalization of its constituent genes. Here, we describe a remarkable case of coordinated gene expression divergence following WGDs in Arabidopsis thaliana. We identified a set of 92 homoeologous gene pairs that all show a similar pattern of tissue-specific gene expression divergence following WGD, with one homoeolog showing predominant expression in aerial tissues and the other homoeolog showing biased expression in tip-growth tissues. We provide evidence that this pattern of gene expression divergence seems to involve genes with a role in cell polarity and that likely function in the maintenance of cell-wall integrity. Following WGD, many of these duplicated genes evolved separate functions through subfunctionalization in growth/development and stress response. Uncoupling these processes through genome duplications likely provided important adaptions with respect to growth and morphogenesis and defense against biotic and abiotic stress.

Author contributions

I performed the phylogenomic analysis of gene families and revised the research article during peer review. Both were done under supervision and with significant contributions of Riet De Smet and Yves Van de Peer.

B.4. Genome of wild olive and the evolution of oil biosynthesis

Turgay Unvera*, Zhangyan Wu*, Lieven Sterck, Mine Turktas, Rolf Lohaus, Zhen Li, Ming Yang, Lijuan He, Tianquan Deng, Francisco Javier Escalante, Carlos Llorens, Francisco J. Roig, Iskender Parmaksiz, Ekrem Dundar, Fuliang Xie, Baohong Zhang, Arif Ipek, Serkan Uranbey, Mustafa Erayman, Emre Ilhan, Oussama Badad, Hassan Ghazal, David A. Lightfoot, Pavan Kasarla, Vincent Colantonio, Huseyin Tombuloglu, Pilar Hernandez, Nurengin Mete, Oznur Cetin, Marc Van Montagu, Huanming Yang, Qiang Gao, Gabriel Dorado, Yves Van de Peer

*contributed equally

Proceedings of the National Academy of Sciences **114**(44): E9413-E9422 (2017).

Abstract

Here we present the genome sequence and annotation of the wild olive tree (*Olea europaea* var. *sylvestris*), called oleaster, which is considered an ancestor of cultivated olive trees. More than 50,000 protein-coding genes were predicted, a majority of which could be anchored to 23 pseudochromosomes obtained through a newly constructed genetic map. The oleaster genome contains signatures of two Oleaceae lineage-specific paleopolyploidy events, dated at ~28 and ~59 Mya. These events contributed to the expansion and neofunctionalization of genes and gene families that play important roles in oil biosynthesis. The functional divergence of oil biosynthesis pathway genes, such as *FAD2*, *SACPD, EAR*, and *ACPTE*, following duplication, has been responsible for the differential accumulation of oleic and linoleic acids produced in olive compared with sesame, a closely related oil crop. Duplicated oleaster *FAD2* genes are regulated by an siRNA derived from a transposable element-rich region, leading to suppressed levels of *FAD2* gene expression. Additionally, neofunctionalization of members of the *SACPD* gene family has led to increased expression of *SACPD2*, *3*, *5*, and *7*, consequently resulting in an increased desaturation of steric acid. Taken together, decreased *FAD2* expression and increased *SACPD* expression likely explain the accumulation of exceptionally high levels of oleic acid in olive. The oleaster genome thus provides important insights into the evolution of oil biosynthesis and will be a valuable resource for oil crop genomics.

Author contributions

I developed and implemented an approach to test the placement of duplication event identified by gene tree – species tree reconciliation. All were done under supervision and with significant contributions of Rolf Lohaus and Yves Van de Peer.

B.5. Lack of GLYCOLATE OXIDASE1, but not GLYCOLATE OXIDASE2, attenuates the photorespiratory phenotype of CATALASE2-deficient Arabidopsis

Pavel Kerchev*, Cezary Waszczak*, Aleksandra Lewandowska, Patrick Willems, Alexey Shapiguzov, Zhen Li, Saleh Alseekh, Per Mühlenbock, Frank A. Hoeberichts, Jingjing Huang, Katrien Van Der Kelen, Jaakko Kangasjärvi, Alisdair R. Fernie, Riet De Smet, Yves Van de Peer, Joris Messens, Frank Van Breusegem

*contributed equally

Plant Physiology **171**(3): 1704-1719 (2016).

Abstract

The genes coding for the core metabolic enzymes of the photorespiratory pathway that allows plants with C3-type photosynthesis to survive in an oxygen-rich atmosphere, have been largely discovered in genetic screens aimed to isolate mutants that are unviable under ambient air. As an exception, glycolate oxidase (GOX) mutants with a photorespiratory phenotype have not been described yet in C3 species. Using Arabidopsis (*Arabidopsis thaliana*) mutants lacking the peroxisomal *CATALASE2* (*cat2-2*) that display stunted growth and cell death lesions under ambient air, we isolated a second-site loss-of-function mutation in *GLYCOLATE OXIDASE1* (*GOX1*) that attenuated the photorespiratory phenotype of *cat2-2*. Interestingly, knocking out the nearly identical *GOX2* in the *cat2-2* background did not affect the photorespiratory phenotype, indicating that GOX1 and GOX2 play distinct metabolic roles. We further investigated their individual functions in single *gox1-1* and *gox2-1* mutants and revealed that their phenotypes can be modulated by environmental conditions that increase the metabolic flux through the photorespiratory pathway. High light negatively affected the photosynthetic performance and growth of both *gox1-1* and *gox2-1* mutants, but the negative consequences of severe photorespiration were more pronounced in the absence of GOX1, which was accompanied with lesser ability to process glycolate. Taken together, our results point toward divergent functions of the two photorespiratory GOX isoforms in Arabidopsis and contribute to a better understanding of the photorespiratory pathway.

Author contributions

I performed the syntenic and phylogenetic analyses of the GOX gene families, and estimated selection pressures on GOX1 and GOX2. I also wrote the corresponding part in the results. All were done under supervision and with significant contributions of Riet De Smet and Yves Van de Peer.

B.6. The gene expression landscape of pine seedling tissues

Rafael A. Cañas, Zhen Li, M. Belén Pascual, Vanessa Castro-Rodríguez, Concepción Ávila, Lieven Sterck, Yves Van de Peer, Francisco M. Cánovas

The Plant Journal **91**(6): 1064-1087 (2017).

Abstract

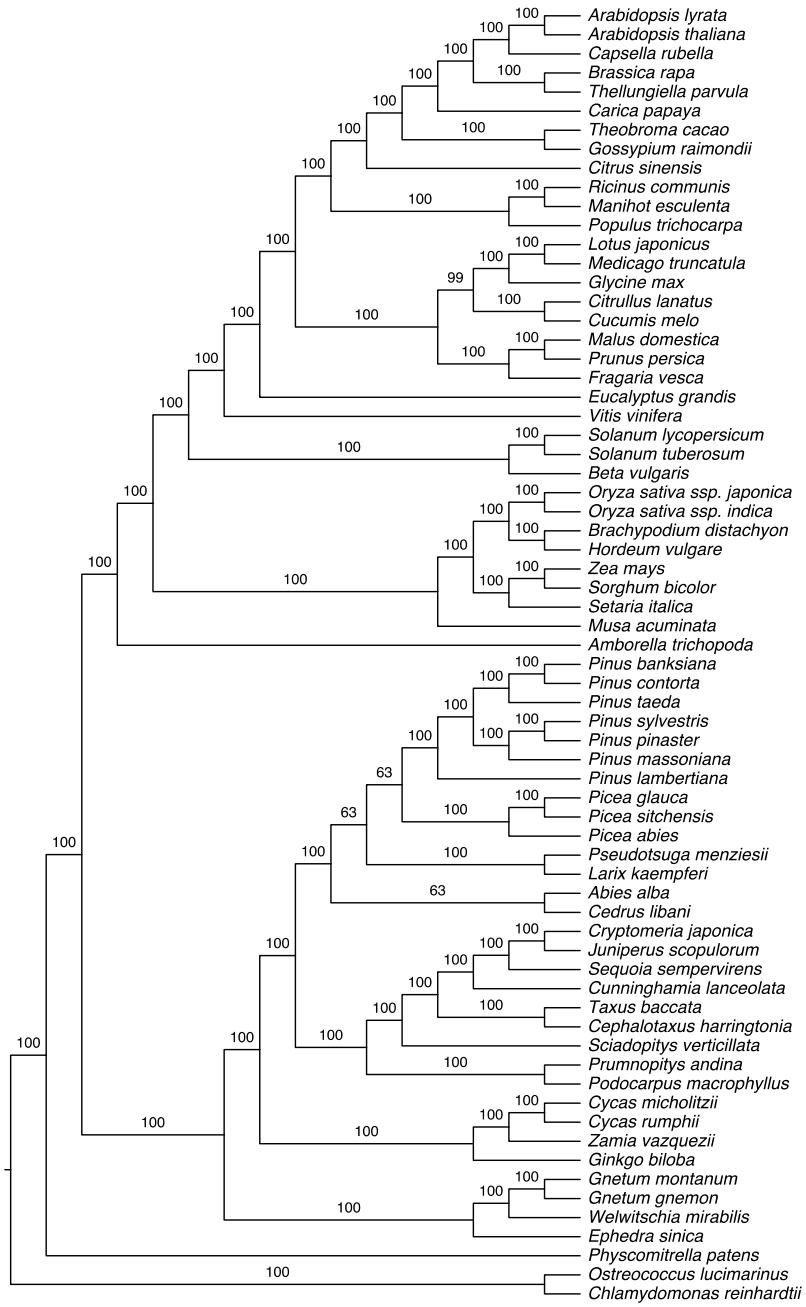
Conifers dominate vast regions of the Northern hemisphere. They are the main source of raw materials for timber industry as well as a wide range of biomaterials. Despite their inherent difficulties as experimental models for classical plant biology research, the technological advances in genomics research are enabling fundamental studies on these plants. The use of laser capture microdissection followed by transcriptomic analysis is a powerful tool for unravelling the molecular and functional organization of conifer tissues and specialized cells. In the present work, 14 different tissues from 1-month-old maritime pine (*Pinus pinaster*) seedlings have been isolated and their transcriptomes analyzed. The results increased the sequence information and number of full-length transcripts from a previous reference transcriptome and added 39 841 new transcripts. In total, 2376 transcripts were ubiquitously expressed in all of the examined tissues. These transcripts could be considered the core 'housekeeping genes' in pine. The genes have been clustered in function to their expression profiles. This analysis reduced the number of profiles to 38, most of these defined by their expression in a unique tissue that is much higher than in the other tissues. The expression and localization data are accessible at ConGenIE.org (http://v22.popgenie.org/microdisection/). This study presents an overview of the gene expression distribution in different pine tissues, specifically highlighting the relationships between tissue gene expression and function. This transcriptome atlas is a valuable resource for functional genomics research in conifers.

Author contribution

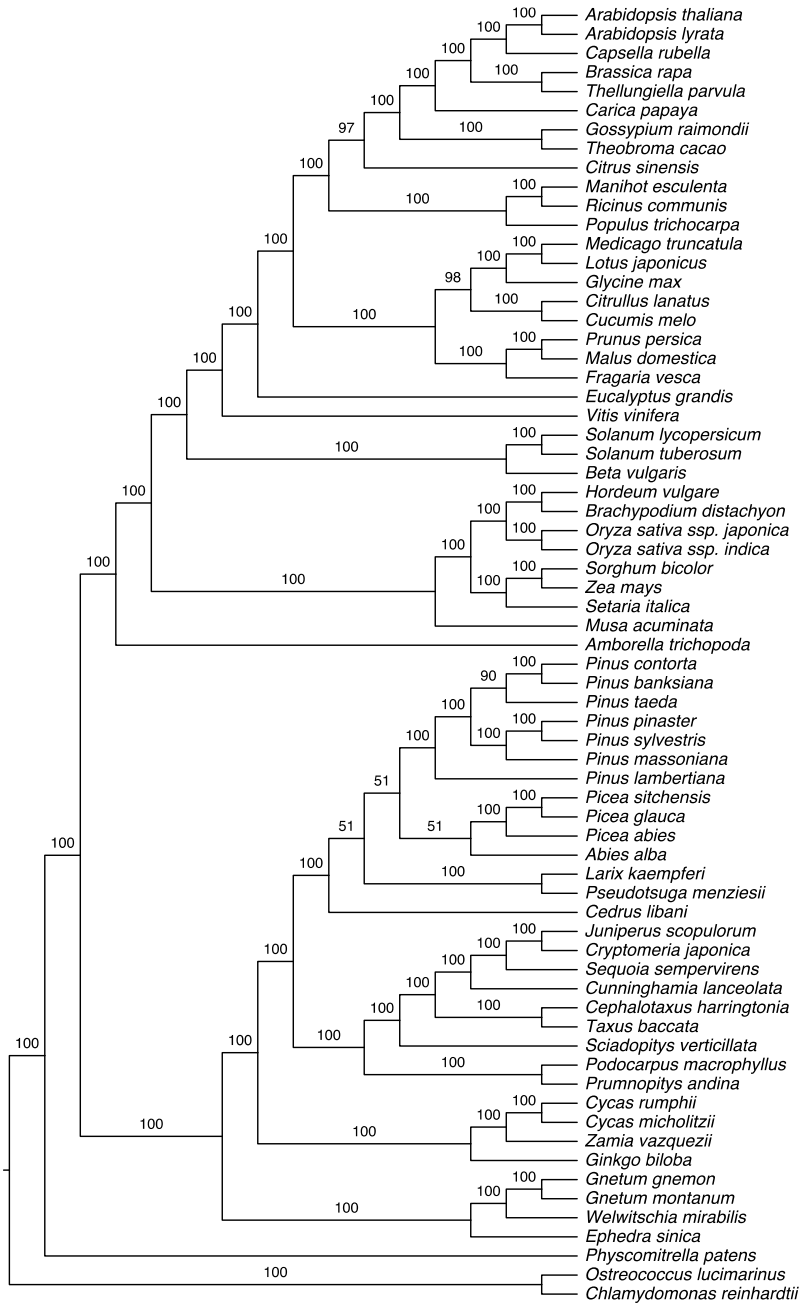
I assembled the transcriptomes of *Pinus pinaster*. It was done under supervision and with significant contributions of Lieven Sterck and Yves Van de Peer.

C. Supplementary information – Single-copy genes as molecular markers for phylogenomic studies in seed plants

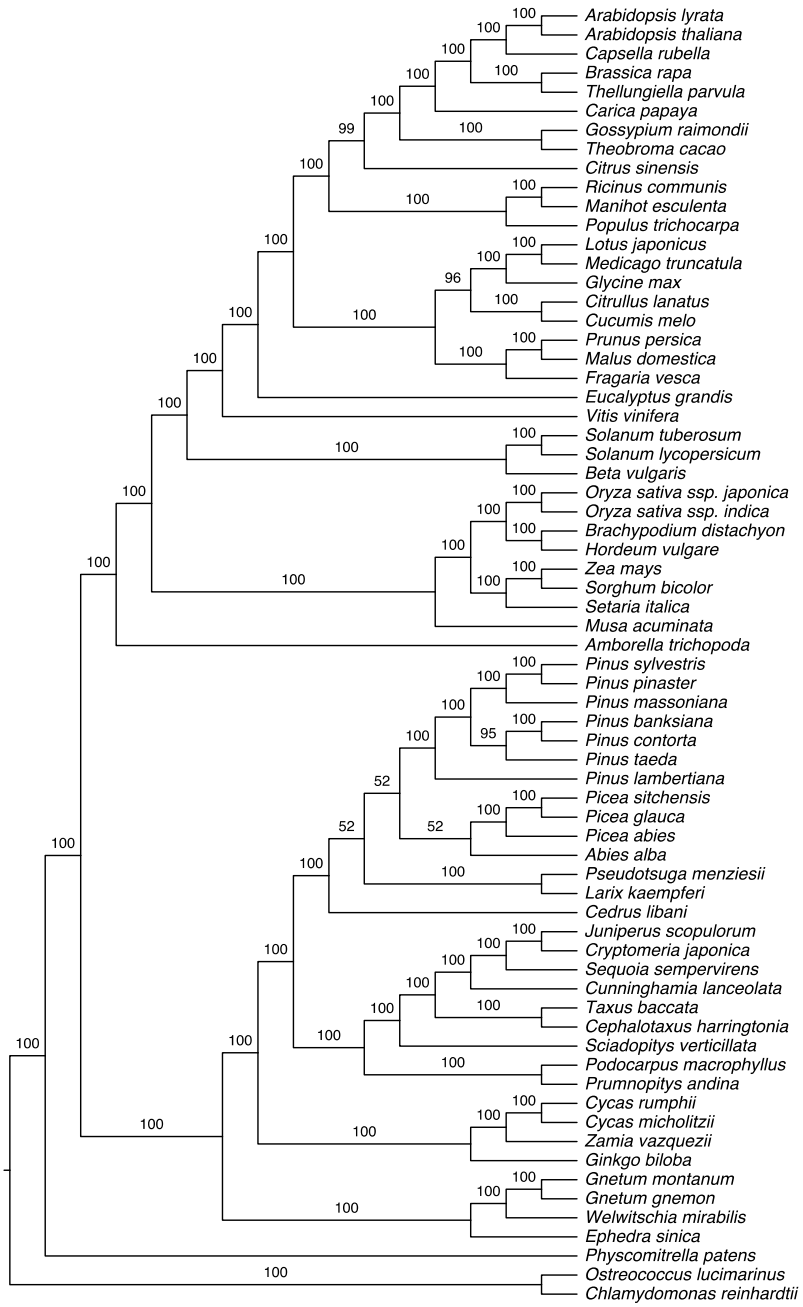
C.1. Supplementary Figures



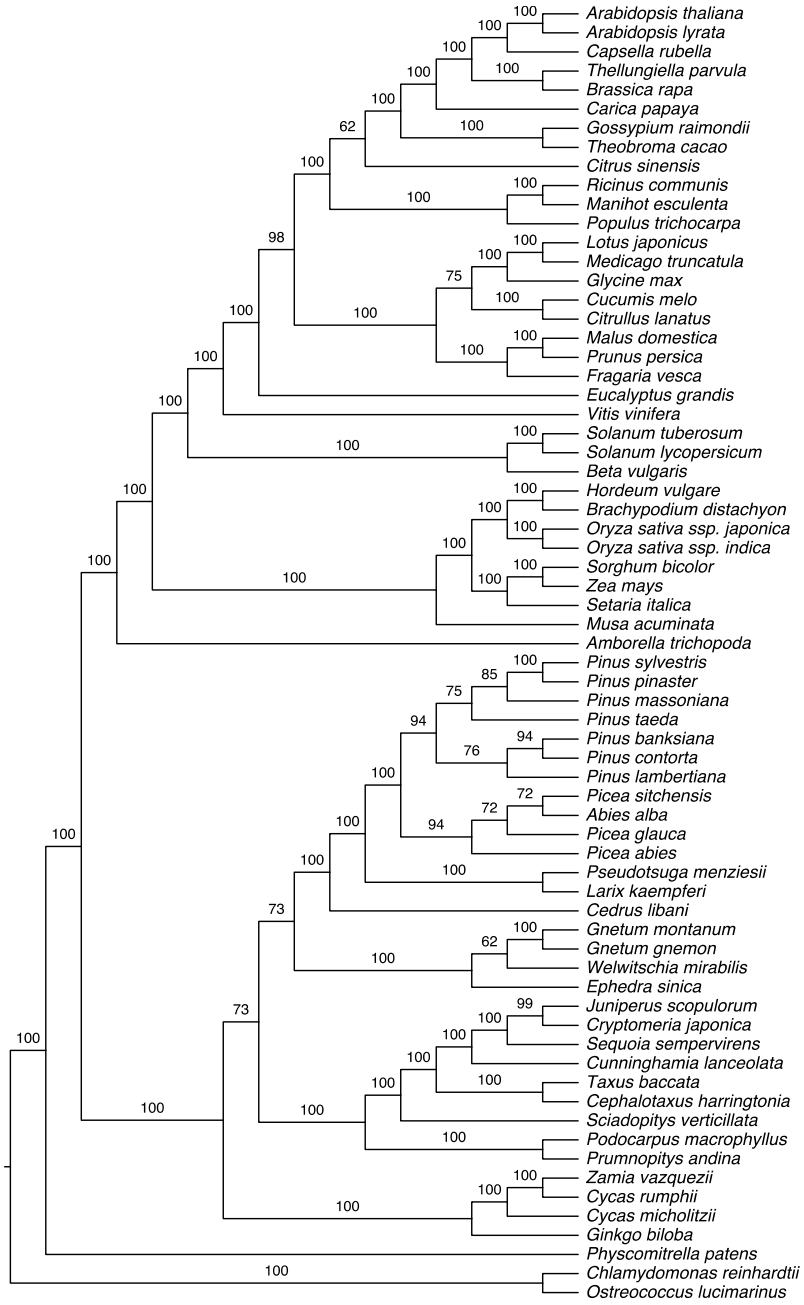
Supplementary Figure C-1. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions with one partition. Numbers on branches represent bootstrap values.



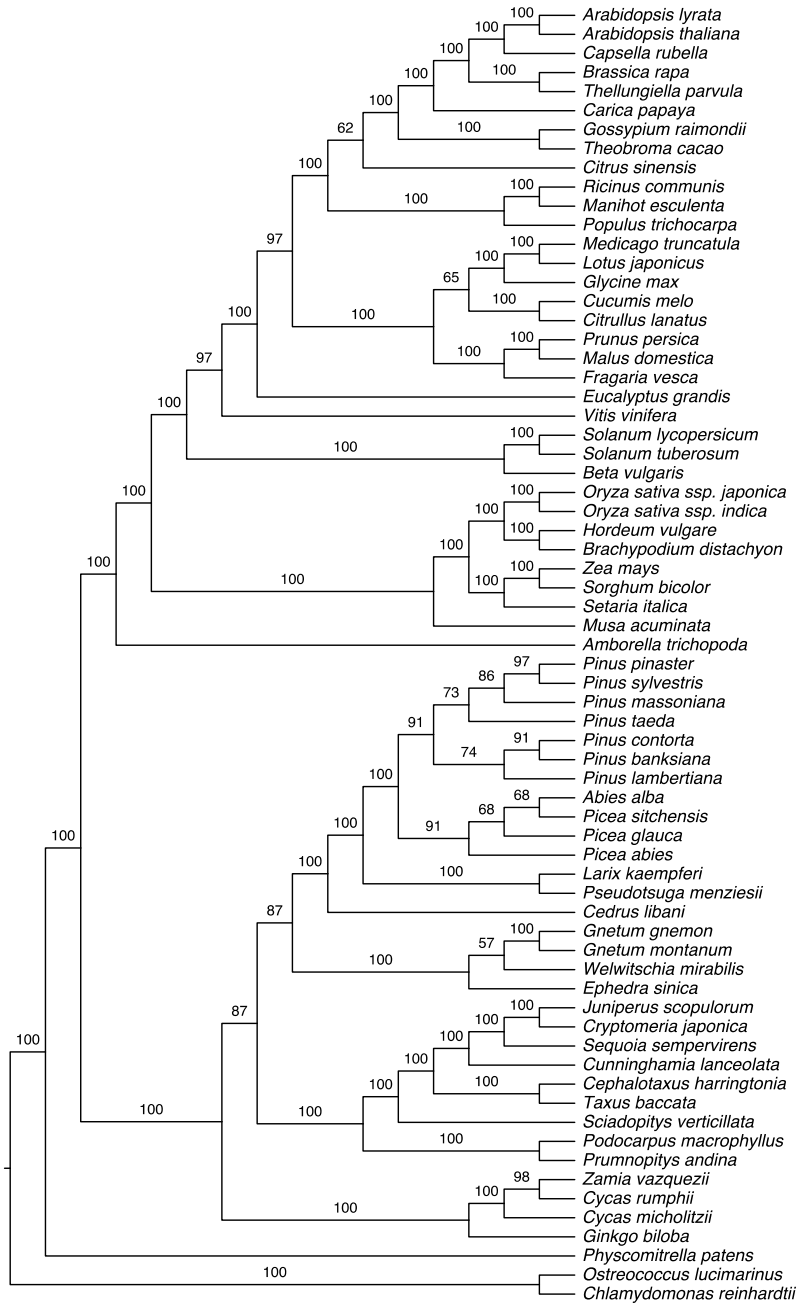
Supplementary Figure C-2. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions, but with 1st and 2nd codon partitions as one partition and 3rd codon partition as another. Numbers on branches represent bootstrap values.



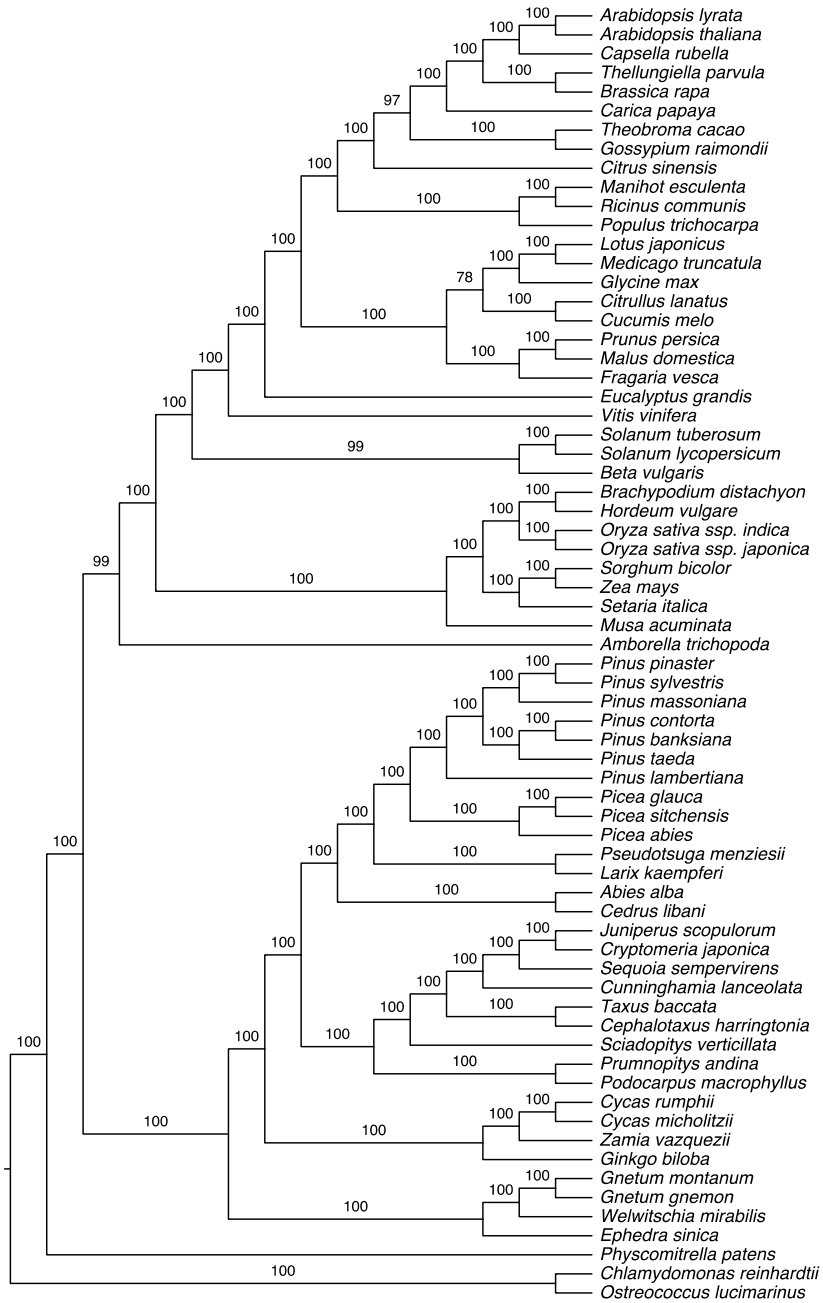
Supplementary Figure C-3. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions but with three partitions, *i.e.*, one for each codon position. Numbers on branches represent bootstrap values.



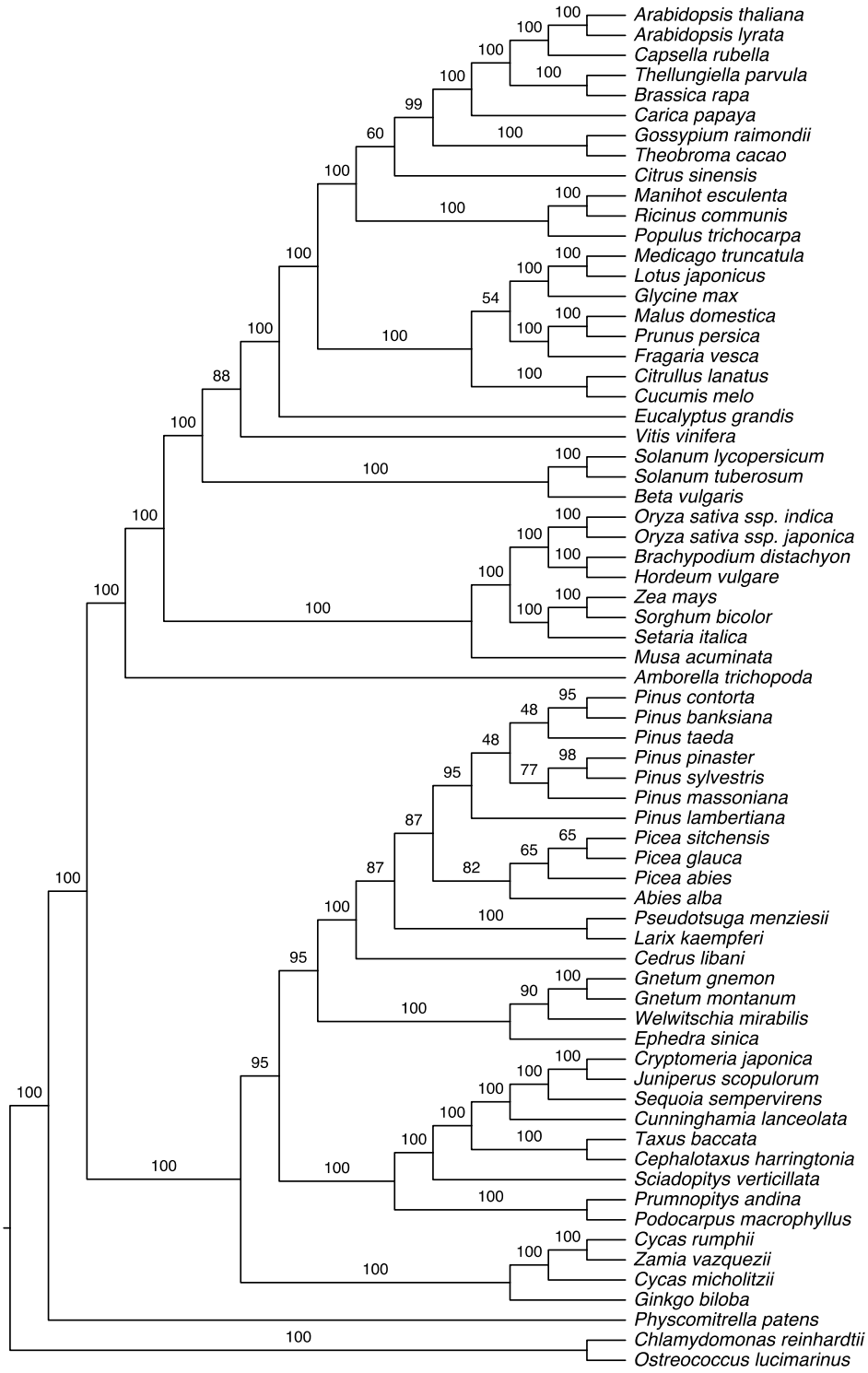
Supplementary Figure C-4. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants with 3rd codon positions removed and with position 1 and 2 as one partition. Numbers on branches represent bootstrap values.



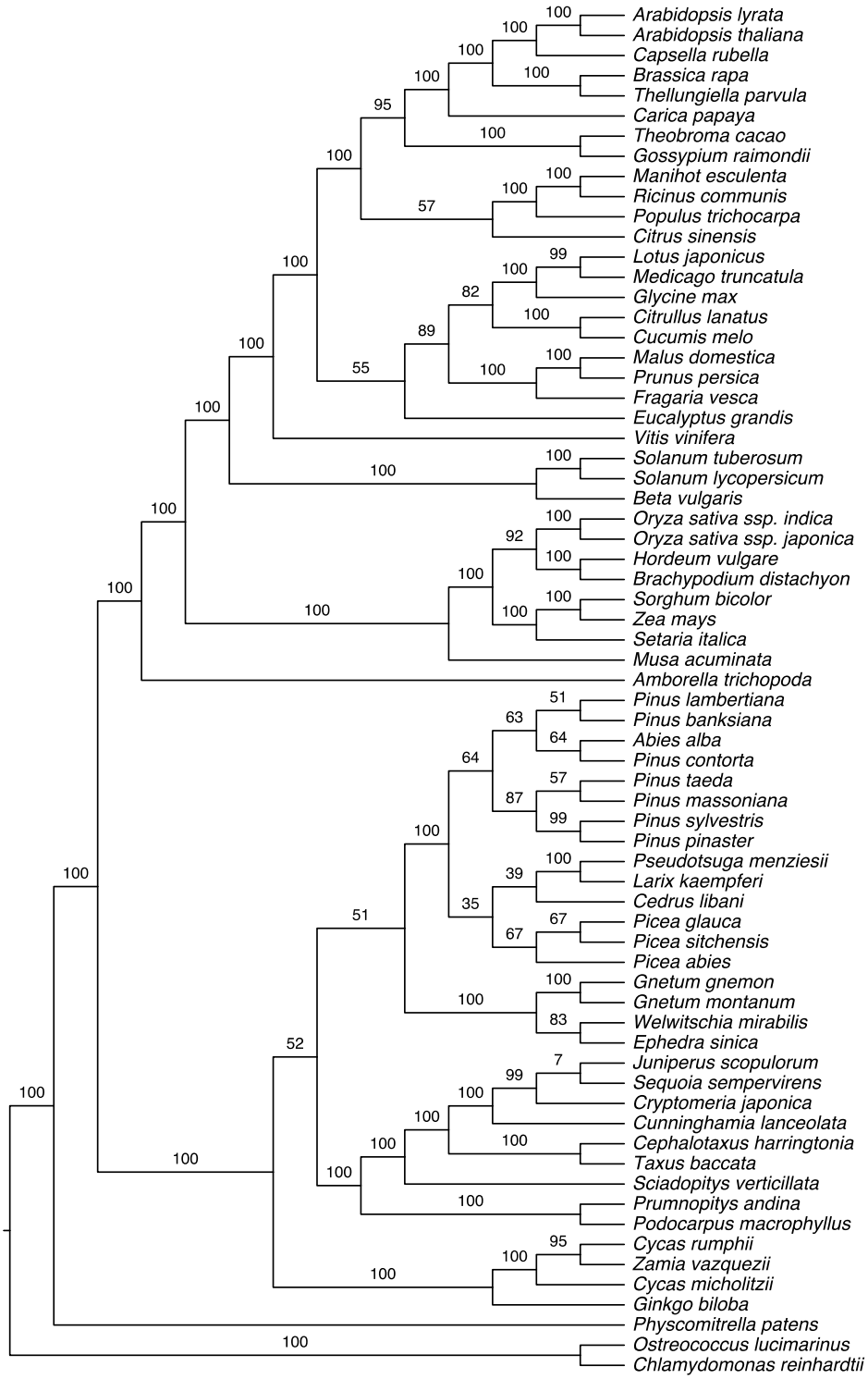
Supplementary Figure C-5. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants with 3rd codon positions removed and with position 1 and 2 used as separate partitions. Numbers on branches represent bootstrap values.



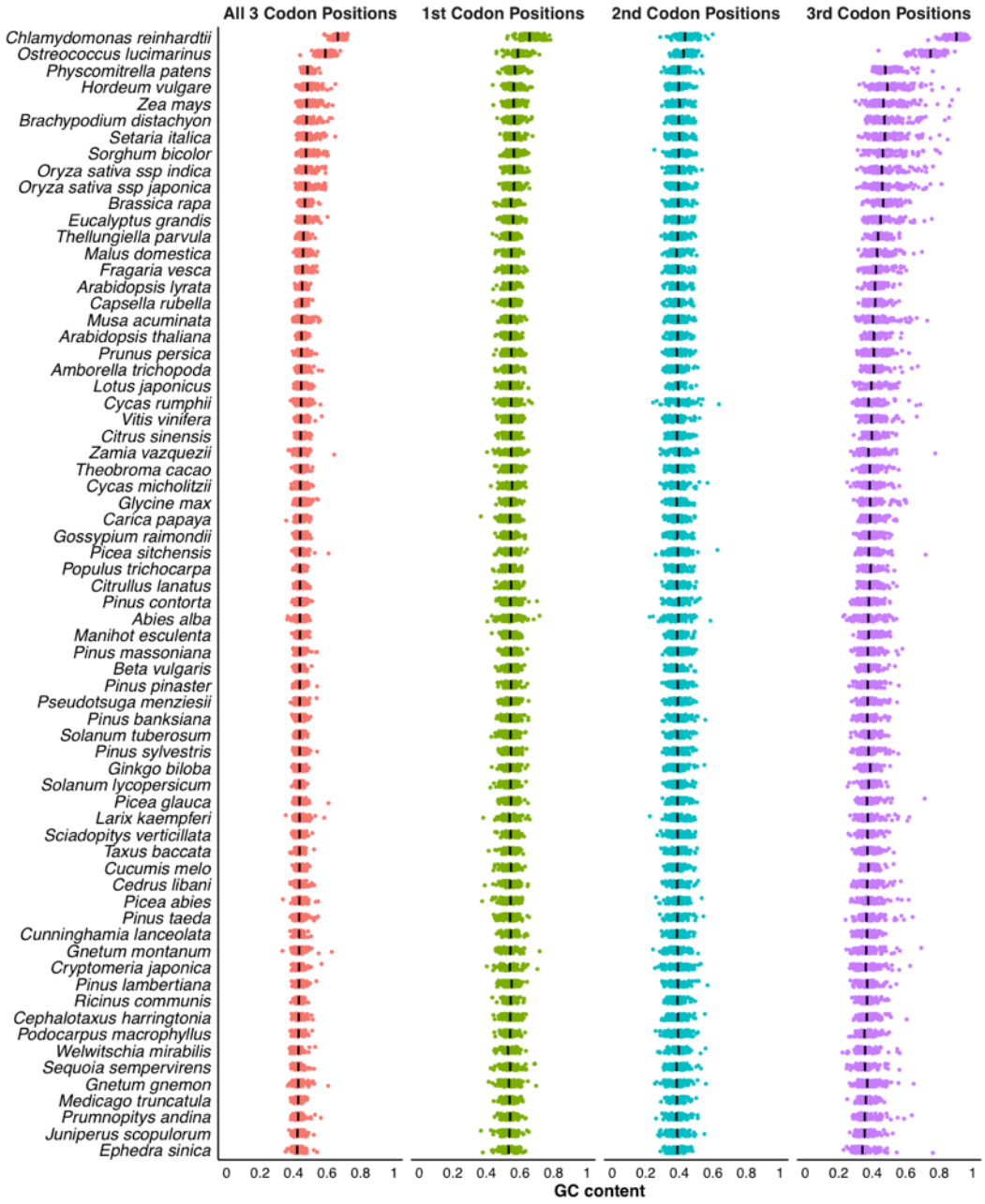
Supplementary Figure C-6. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants only with 3rd codon positions. Numbers on branches represent bootstrap values.



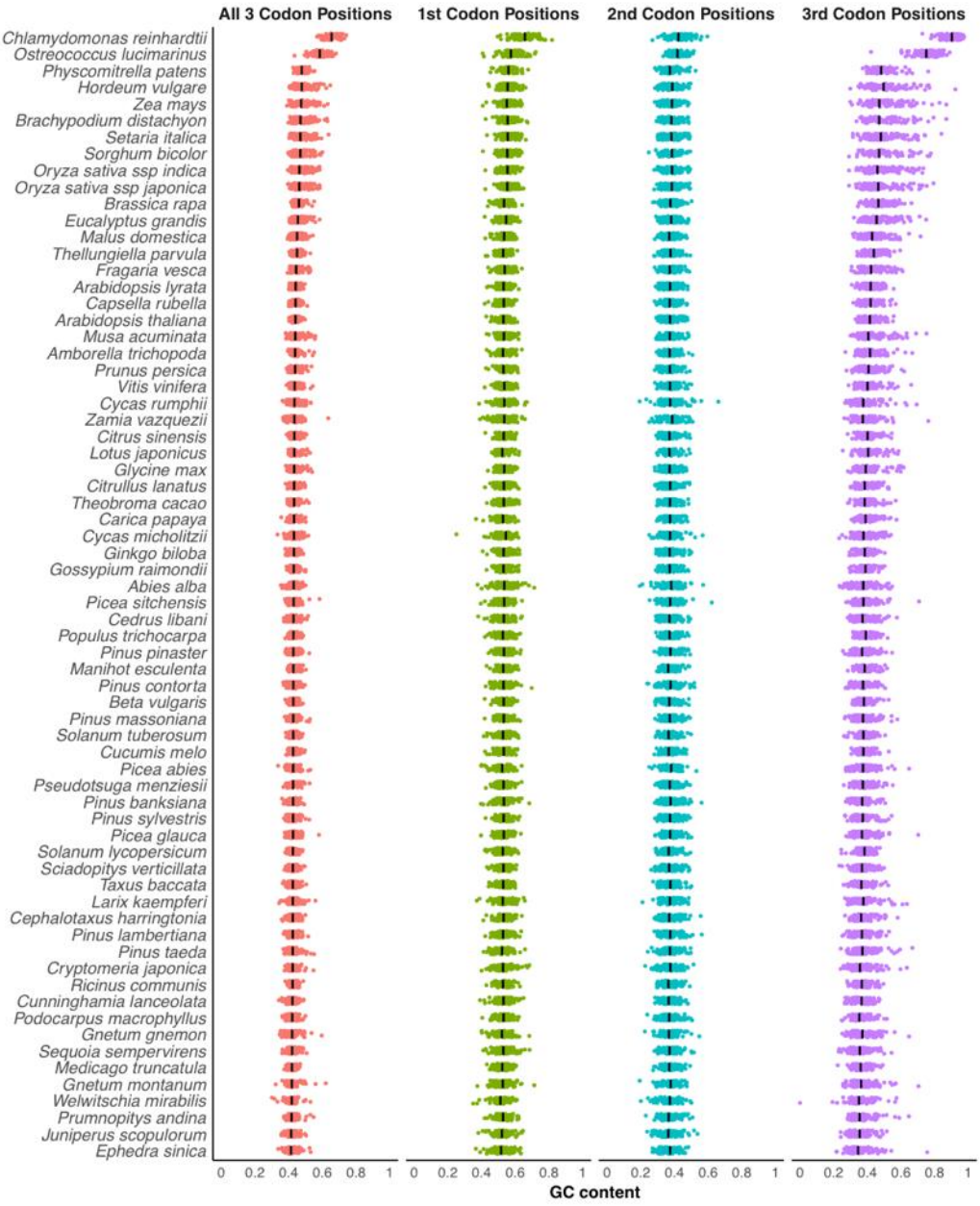
Supplementary Figure C-7. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants only with 1st codon positions. Numbers on branches represent bootstrap values.



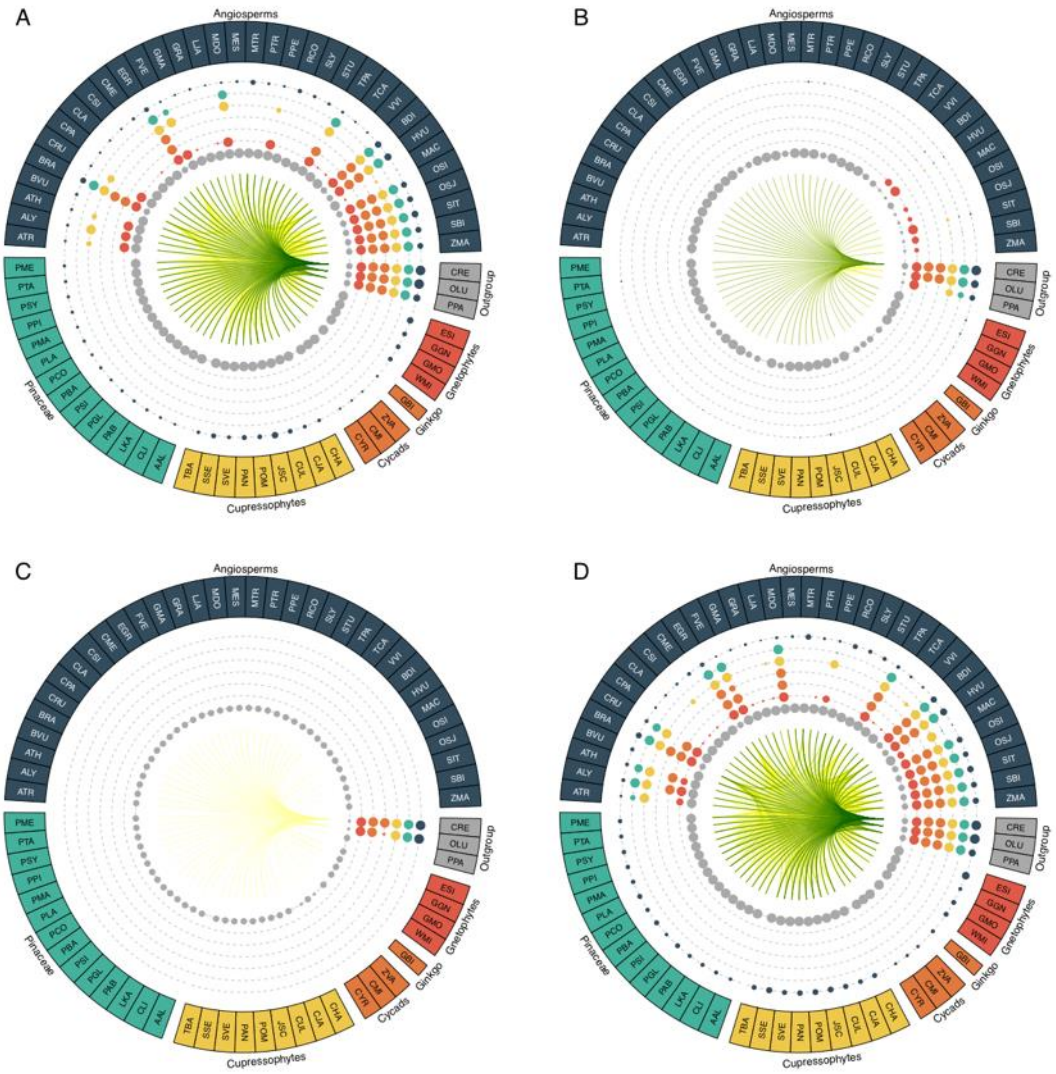
Supplementary Figure C-8. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants only with 2nd codon positions. Numbers on branches represent bootstrap values.



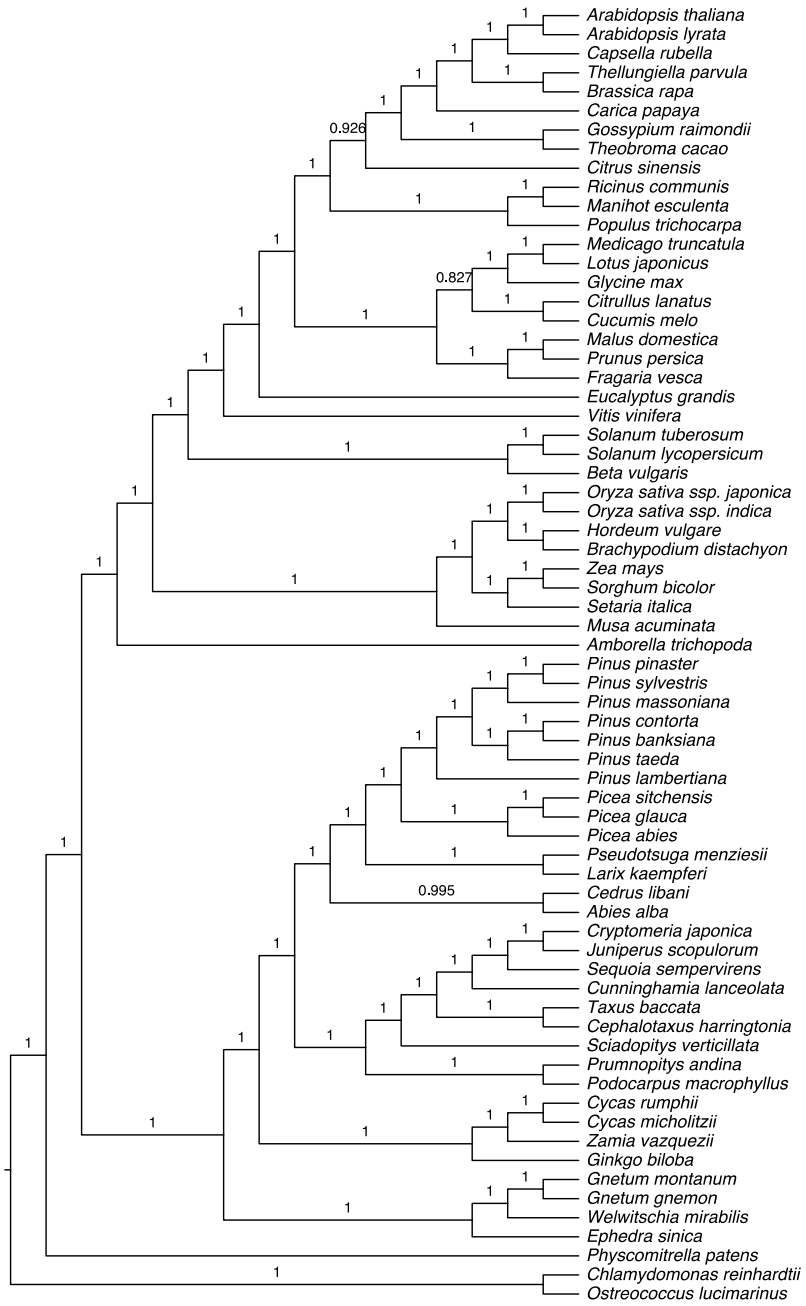
Supplementary Figure C-9. GC content of the 106 phylogenetic markers at 1st, 2nd, and 3rd codon positions for the species studied.



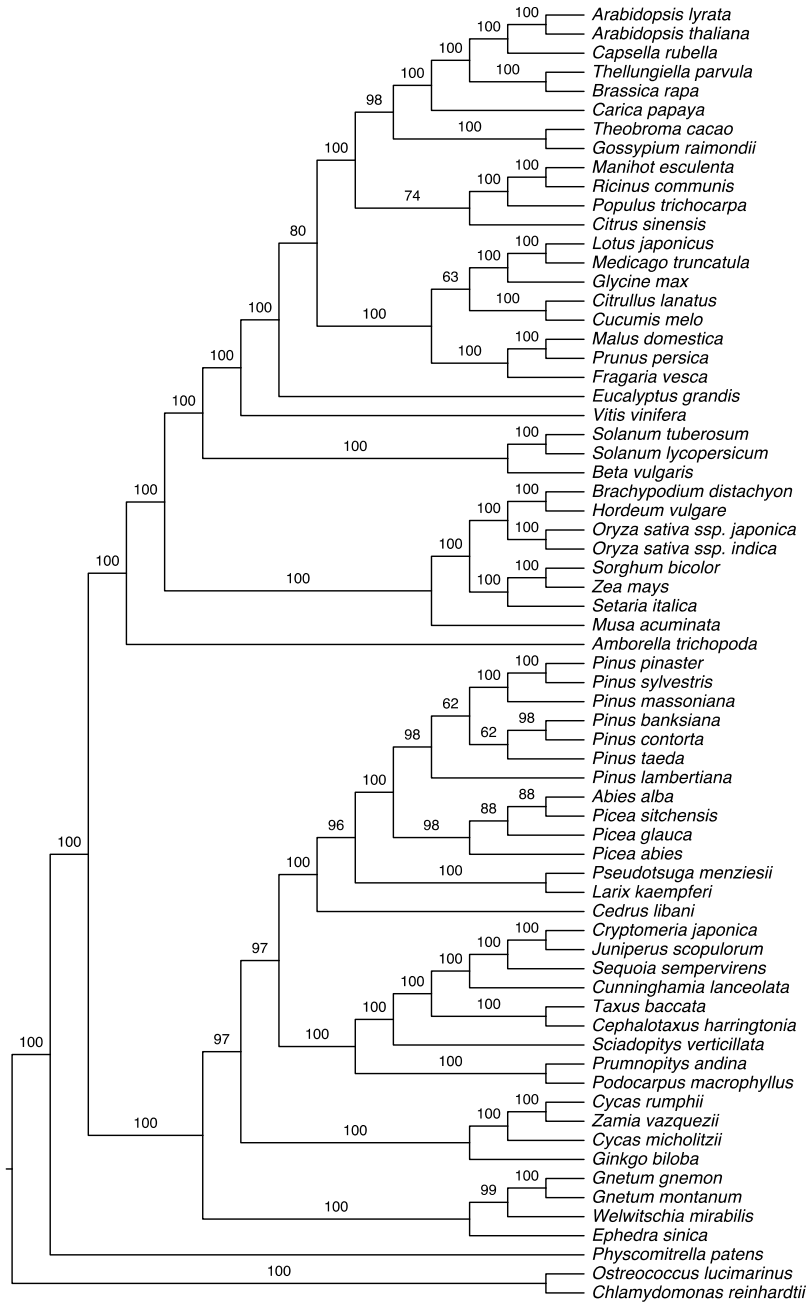
Supplementary Figure C-10. GC content of the 106 phylogenetic markers at 1st, 2nd, and 3rd codon positions for the species studied after removing sites that encode the same amino acids.



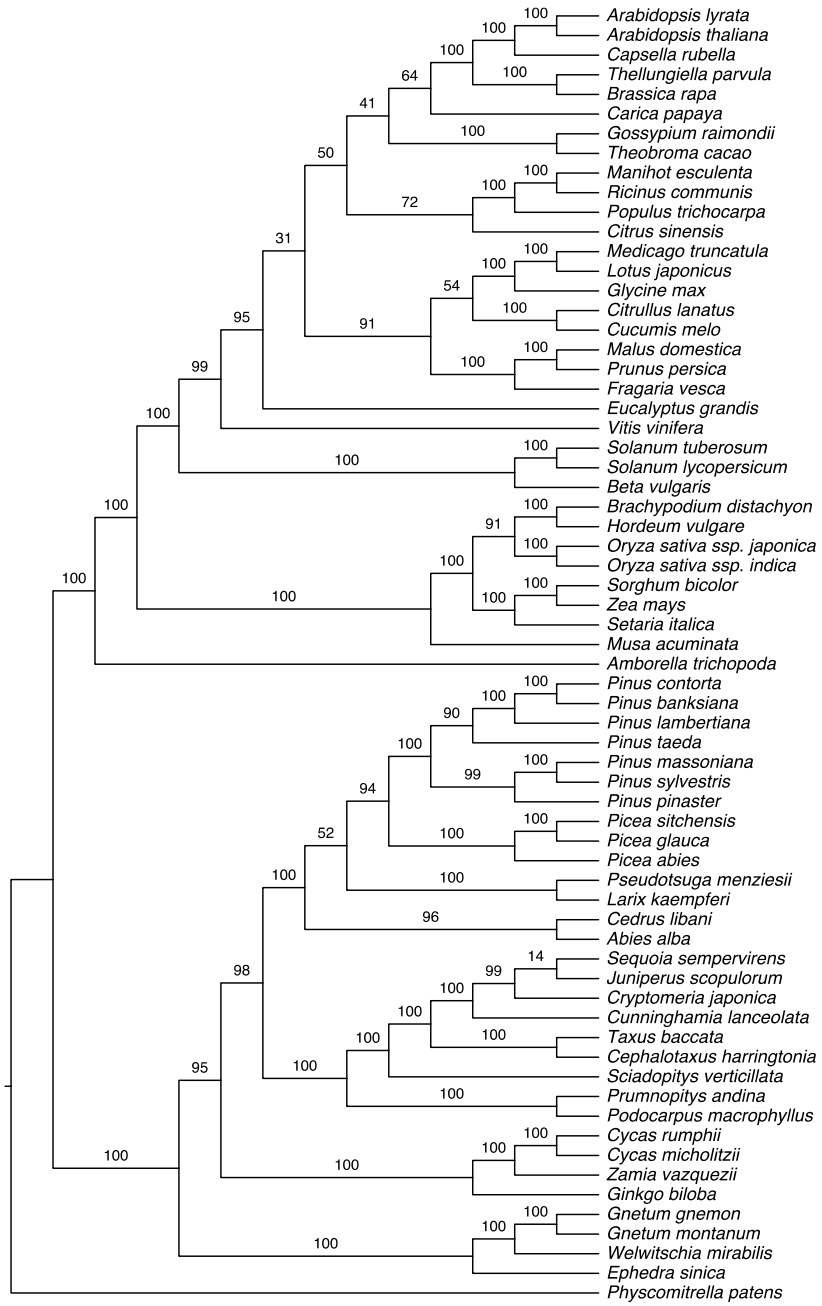
Supplementary Figure C-11. Comparison of GC content in the concatenated alignment (A) and at each codon position (B, C, and D) from 106 genes in 68 species after removing sites that encode the same amino acids. Dot size correlates with the number of species in each lineage (group) that have a significantly different GC% (Wilcox test, $P < 1 \times 10^{-3}$) with the species compared with (colors of dots correspond to the compared lineages). Lines connecting any two species represent significant difference in GC content, with most significant in green and weakest in yellow (1×10^{-3}). The full names for the species can be found in Supplementary Table 3.



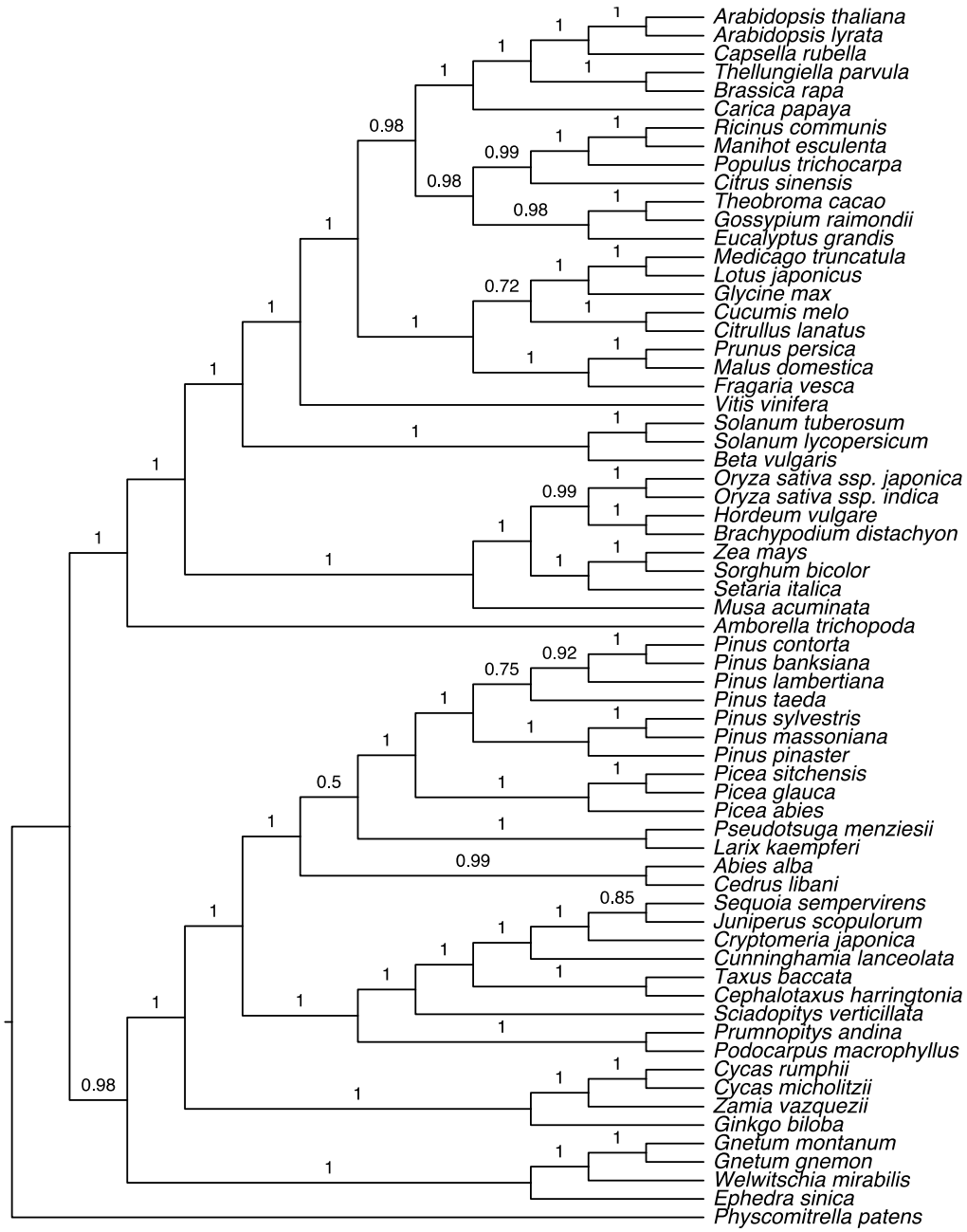
Supplementary Figure C-12. Maximum likelihood tree inferred from a concatenated codon alignment of 106 single-copy genes using the Goldman and Yang (GY) model. Numbers on branches represent support values from the SH-like approximate likelihood-ratio test.



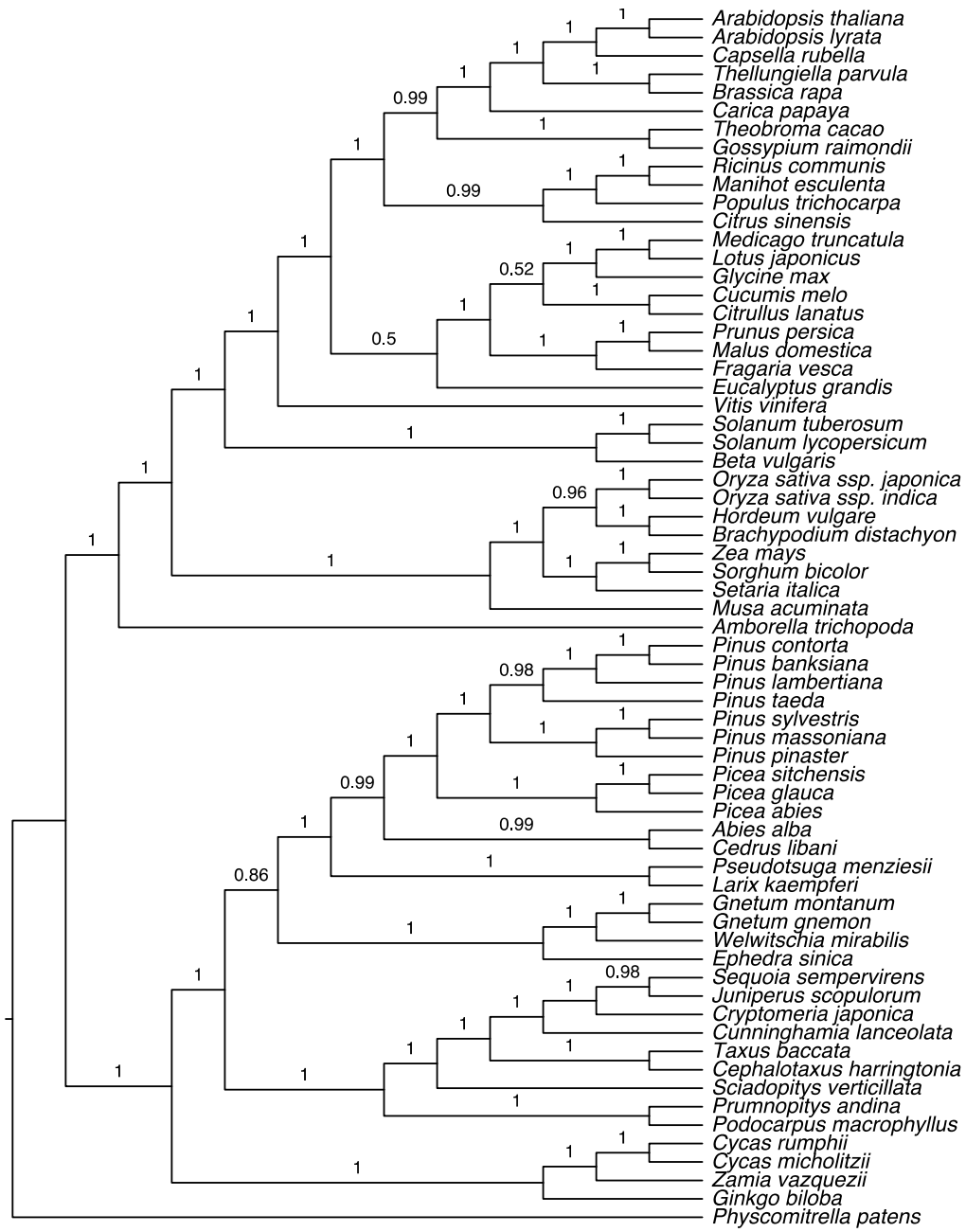
Supplementary Figure C-13. Maximum likelihood tree inferred from a concatenated amino acid alignment of 106 single-copy genes using the JTT+I+GAMMA+F model. Numbers on branches represent bootstrap values.



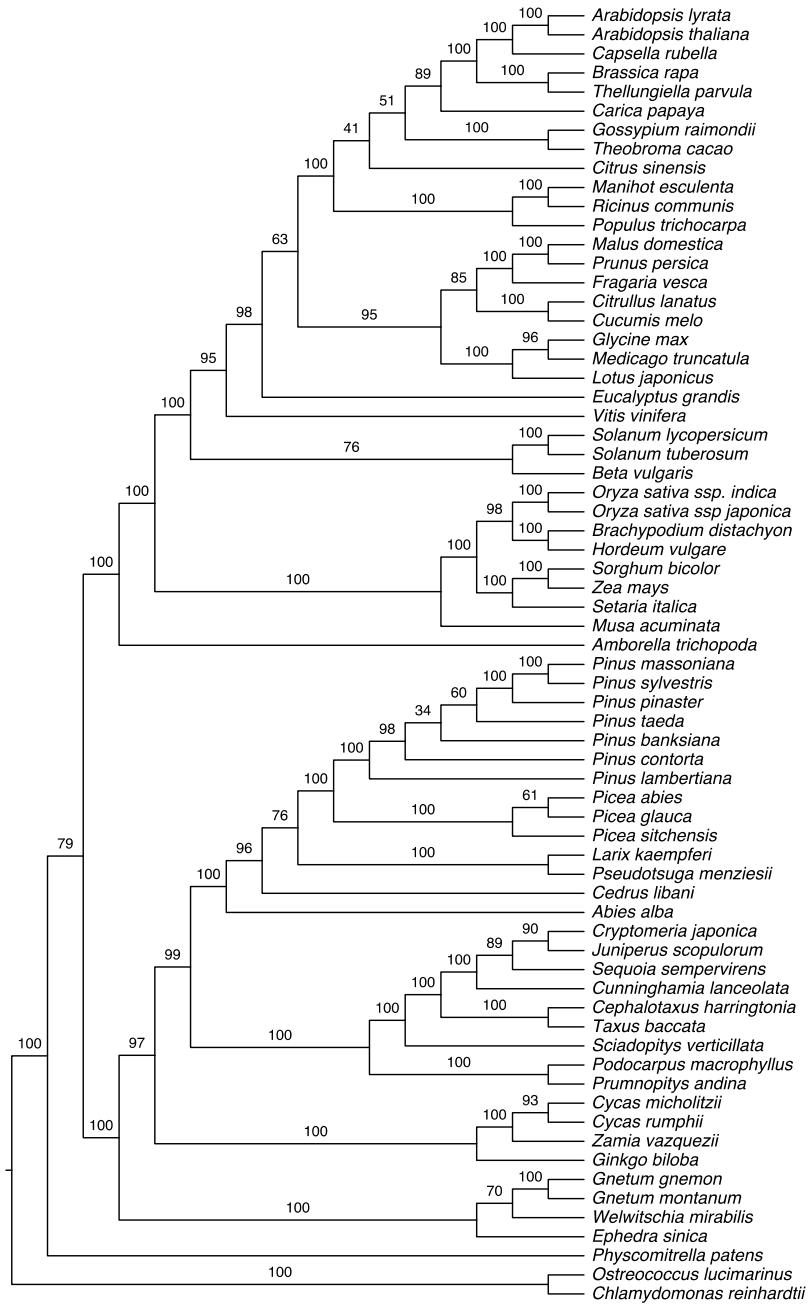
Supplementary Figure C-14. Maximum likelihood tree inferred from a reduced concatenated amino acid alignment of 106 single-copy genes using the JTT+I+GAMMA+F model. Numbers on branches represent bootstrap values.



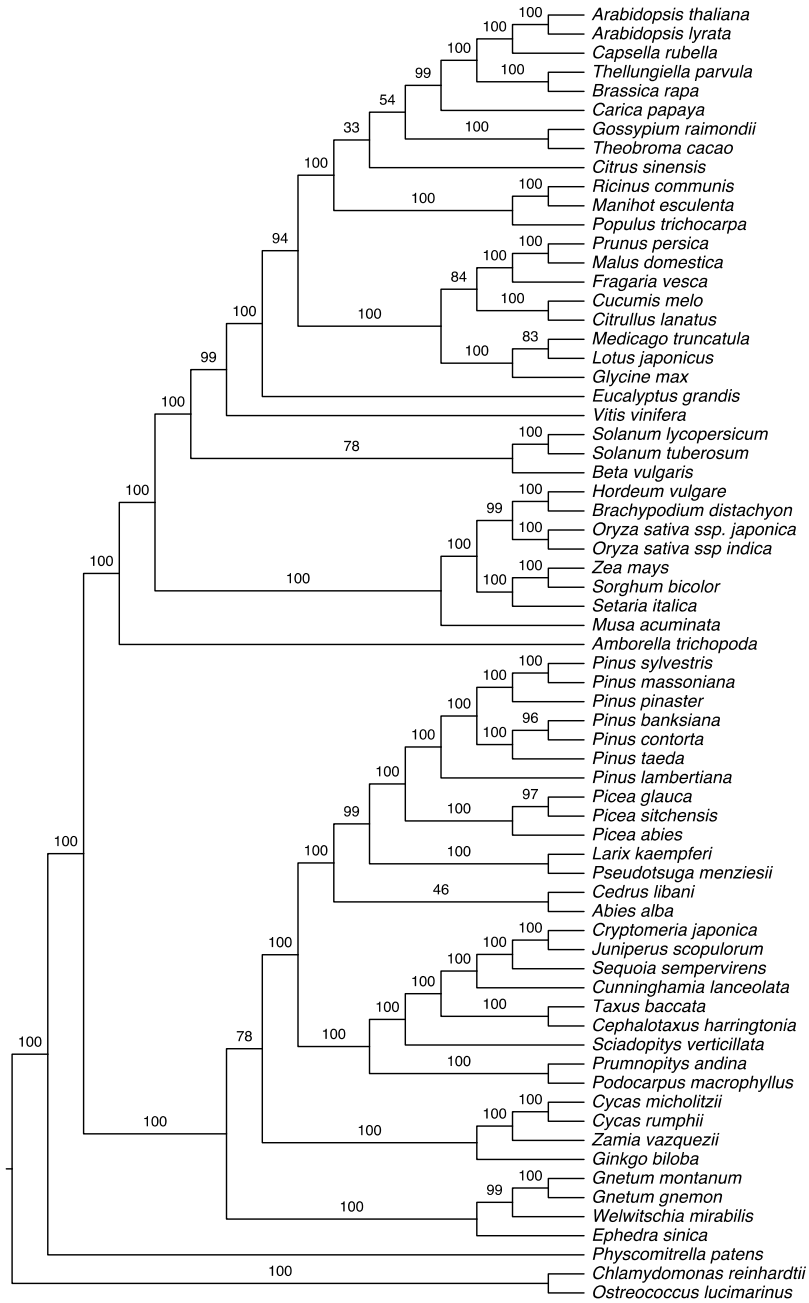
Supplementary Figure C-15. Bayesian phylogenetic tree based on the reduced amino acid concatenation of 106 single-copy genes under CAT model. Numbers indicate posterior probabilities.



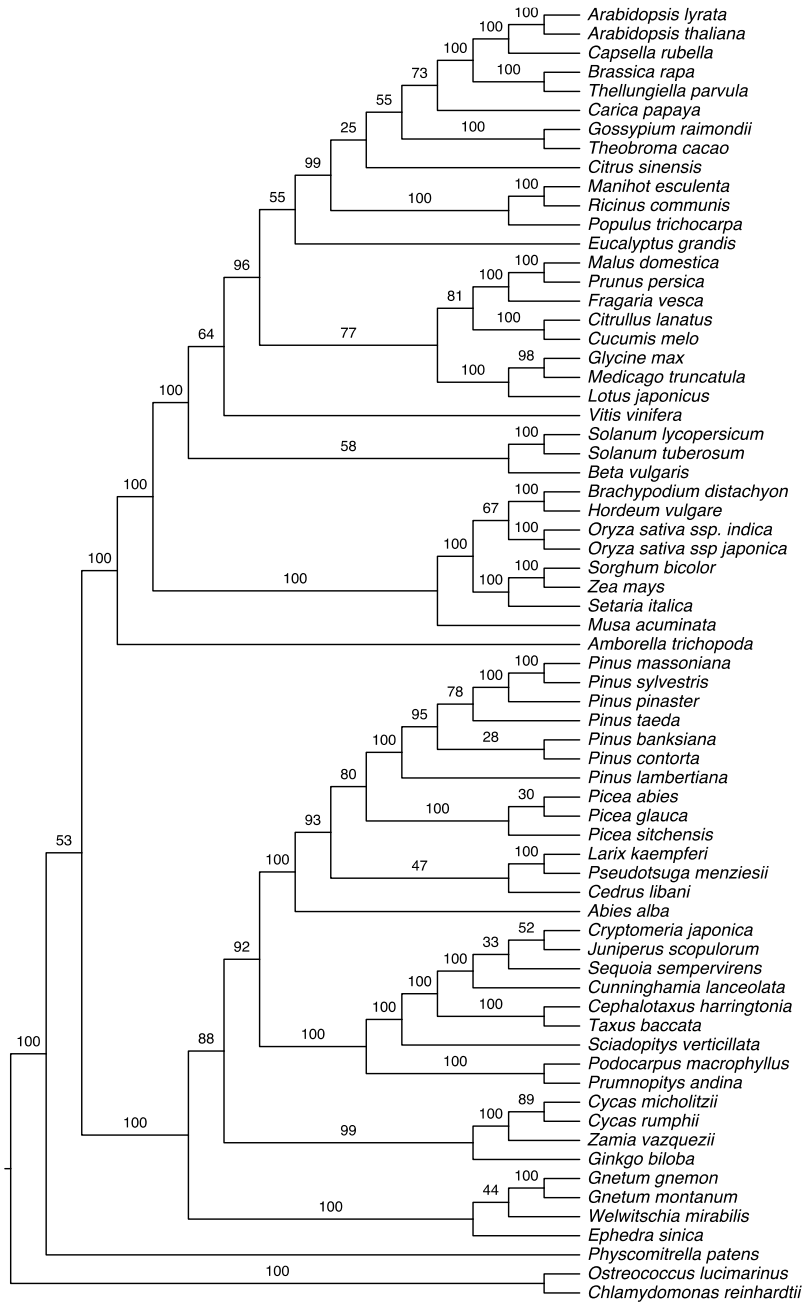
Supplementary Figure C-16. Bayesian phylogenetic tree based on the reduced amino acid concatenation of 106 single-copy genes under CAT+GTR model. Numbers indicate posterior probabilities.



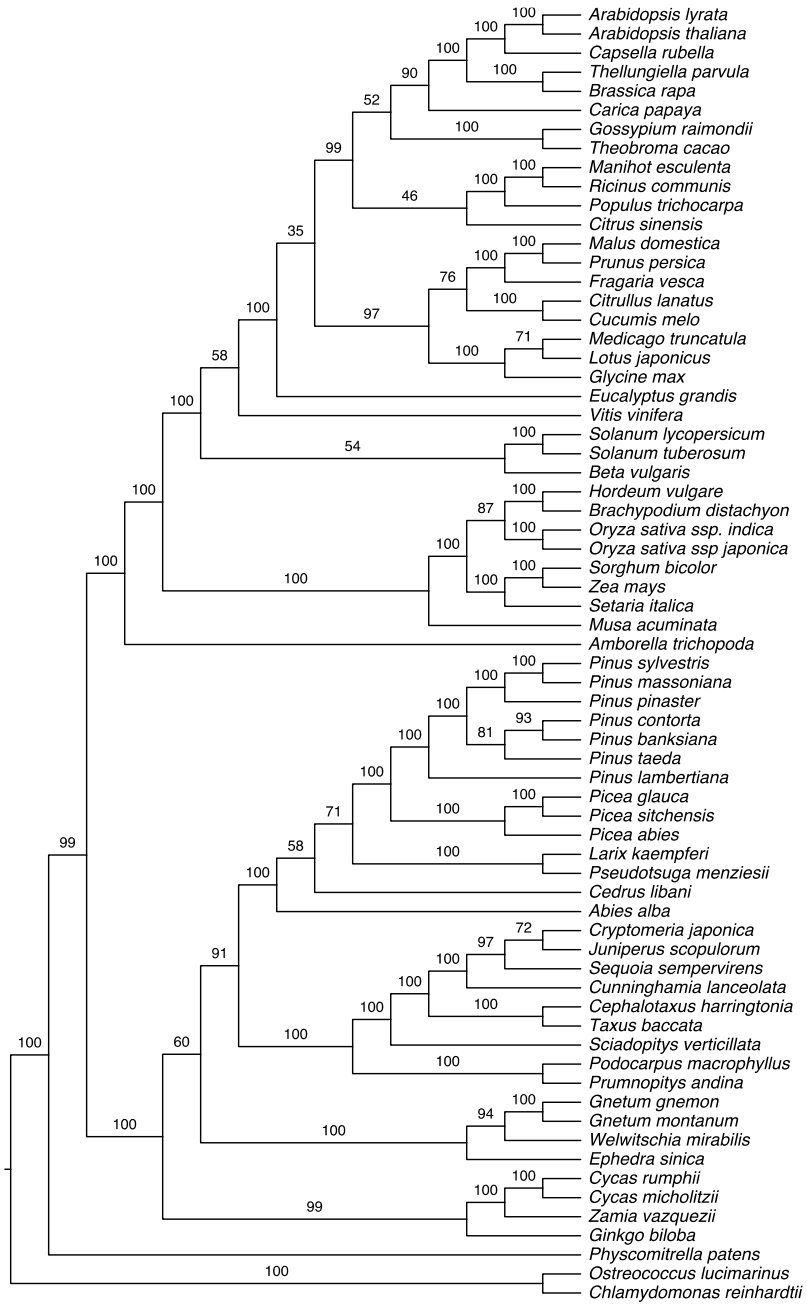
Supplementary Figure C-17. Coalescent based tree inferred from gene trees of 106 single-copy genes by STAR. Numbers on branches represent bootstrap values.



Supplementary Figure C-18. Coalescent based tree inferred from gene trees of 106 single-copy genes by ASTRA-II. Numbers on branches represent bootstrap values.

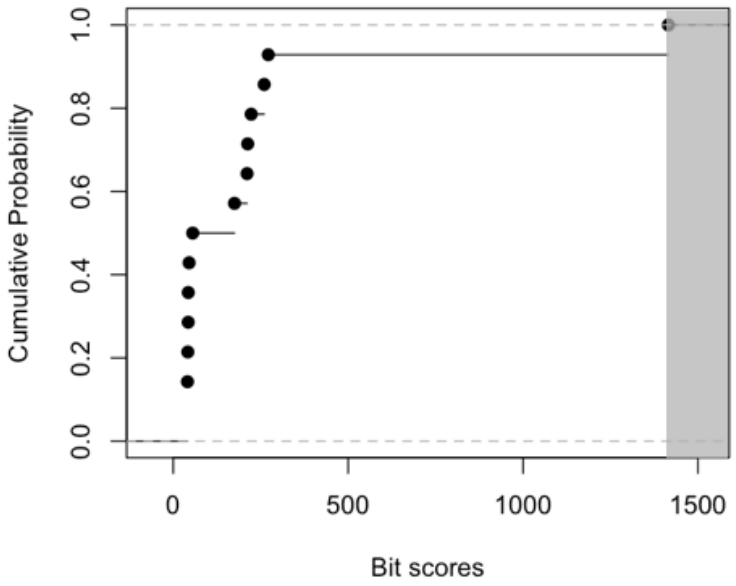


Supplementary Figure C-19. Coalescent based tree inferred from gene trees of 106 single-copy genes by STAR. Gene trees were built without considering 3rd codon positions. Numbers on branches represent bootstrap values.

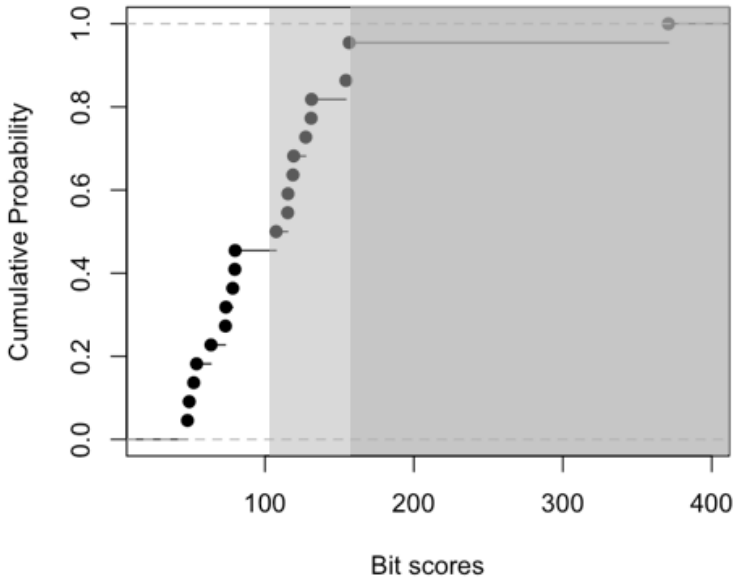


Supplementary Figure C-20. Coalescent based tree inferred from gene trees of 106 single-copy genes by ASTRAL-II. Gene trees were built without considering 3rd codon positions. Numbers on branches represent bootstrap values.

A



B



Supplementary Figure C-21. Example showing the process to assign genes to a gene family accounting for 95% of the cumulative probability of bit-scores. The dark grey blocks in both A and B show regions with over 95% cumulative probability. Dots (*i.e.*, hits) falling in the dark grey region are allocated to the gene family. In B, the light grey block denotes a region where hits have similar *E* values (Δ*E* value < 1×10^{20}) to the hit with the smallest Bit-score at the 95% border.

C.2. Supplementary Tables

Supplementary Table C-1. Integrated transcriptomes for gymnosperms from public databases

Species	TreeGenes*	PlantGDB**	oneKP	# Transcripts	# Proteins
Abies alba	25,419	-	-	20,796	17,279
Cedrus libani	-	-	70,595	65,491	37,730
Cephalotaxus harringtonia	13,997	-	-	13,927	16,578
Cryptomeria japonica	347	24,299	-	21,922	19,688
Cunninghamia lanceolata	62,140	-	68,622	70,761	48,057
Cycas micholitzii	-	-	54,202	51,130	28,901
Cycas rumphii	-	10,901	22,908	21,844	12,619
Ephedra sinica	-	-	57,134	51,088	26,873
Ginkgo biloba	-	10,210	48,343	45,631	30,404
Gnetum gnemon	10,221	6,193	-	12,994	14,889
Gnetum montanum	-	-	70,316	65,123	32,549
Juniperus scopulorum	-	-	61,299	57,857	35,004
Larix kaempferi	57,484	-	-	53,274	40,692
Pinus banksiana	16	13,040	-	12,036	13,421
Pinus contorta	32	13,570	-	12,605	14,984
Pinus lambertiana	19,509	-	-	19,292	20,561
Pinus massoniana	69,738	-	-	55,525	46,423
Podocarpus macrophyllus	12,266	-	-	12,209	14,296
Prumnopitys andina	-	-	45,616	42,269	30,894
Pseudotsuga menziesii	393,638	9,857	-	246,282	149,717
Sciadopitys verticillata	11,955	-	51,723	40,310	30,035
Sequoia sempervirens	11,517	-	-	11,462	13,461
Taxus baccata	10,554	-	54,249	44,059	32,062
Welwitschia mirabilis	-	-	6,606	6,261	6,052
Zamia vazquezii	-	7,657	50,336	33,979	24,619

* TreeGenes includes ESTs, cDNAs, TSAs, and 454 assemblies

***Cycas rumphii*, *Ginkgo biloba*, *Gnetum gnemon*, and *Zamia vazquezii* are directly from PlantGDB

Supplementary Table C-2. The short name and clade of species used in the current study.

Species	Short name	Clade
Cephalotaxus harringtonia	CHA	Cupressophytes
Cryptomeria japonica	CJA	Cupressophytes
Cunninghamia lanceolata	CUL	Cupressophytes
Juniperus scopulorum	JSC	Cupressophytes
Podocarpus macrophyllus	POM	Cupressophytes
Prumnopitys andina	PAN	Cupressophytes
Sciadopitys verticillata	SVE	Cupressophytes
Sequoia sempervirens	SSE	Cupressophytes
Taxus baccata	TBA	Cupressophytes
Zamia vazquezii	ZVA	Cycads
Cycas micholitzii	CMI	Cycads
Cycas rumphii	CYR	Cycads
Amborella trichopoda	ATR	Angiosperms
Arabidopsis lyrata	ALY	Angiosperms
Arabidopsis thaliana	ATH	Angiosperms
Beta vulgaris	BVU	Angiosperms
Brassica rapa	BRA	Angiosperms
Capsella rubella	CRU	Angiosperms
Carica papaya	CPA	Angiosperms
Citrullus lanatus	CLA	Angiosperms
Citrus sinensis	CSI	Angiosperms
Cucumis melo	CME	Angiosperms
Eucalyptus grandis	EGR	Angiosperms
Fragaria vesca	FVE	Angiosperms
Glycine max	GMA	Angiosperms
Gossypium raimondii	GRA	Angiosperms
Lotus japonicus	LJA	Angiosperms
Malus domestica	MDO	Angiosperms
Manihot esculenta	MES	Angiosperms
Medicago truncatula	MTR	Angiosperms
Populus trichocarpa	PTR	Angiosperms
Prunus persica	PPE	Angiosperms
Ricinus communis	RCO	Angiosperms
Solanum lycopersicum	SLY	Angiosperms
Solanum tuberosum	STU	Angiosperms
Thellungiella parvula	TPA	Angiosperms
Theobroma cacao	TCA	Angiosperms
Vitis vinifera	VVI	Angiosperms
Ginkgo biloba	GBI	Ginkgo
Ephedra sinica	ESI	Gnetophytes
Gnetum gnemon	GGN	Gnetophytes
Gnetum montanum	GMO	Gnetophytes
Welwitschia mirabilis	WMI	Gnetophytes
Brachypodium distachyon	BDI	Angiosperms
Hordeum vulgare	HVU	Angiosperms
Musa acuminata	MAC	Angiosperms
Oryza sativa ssp indica	OSI	Angiosperms
Oryza sativa ssp japonica	OSJ	Angiosperms
Setaria italica	SIT	Angiosperms
Sorghum bicolor	SBI	Angiosperms
Zea mays	ZMA	Angiosperms
Chlamydomonas reinhardtii	CRE	Outgroup
Ostreococcus lucimarinus	OLU	Outgroup

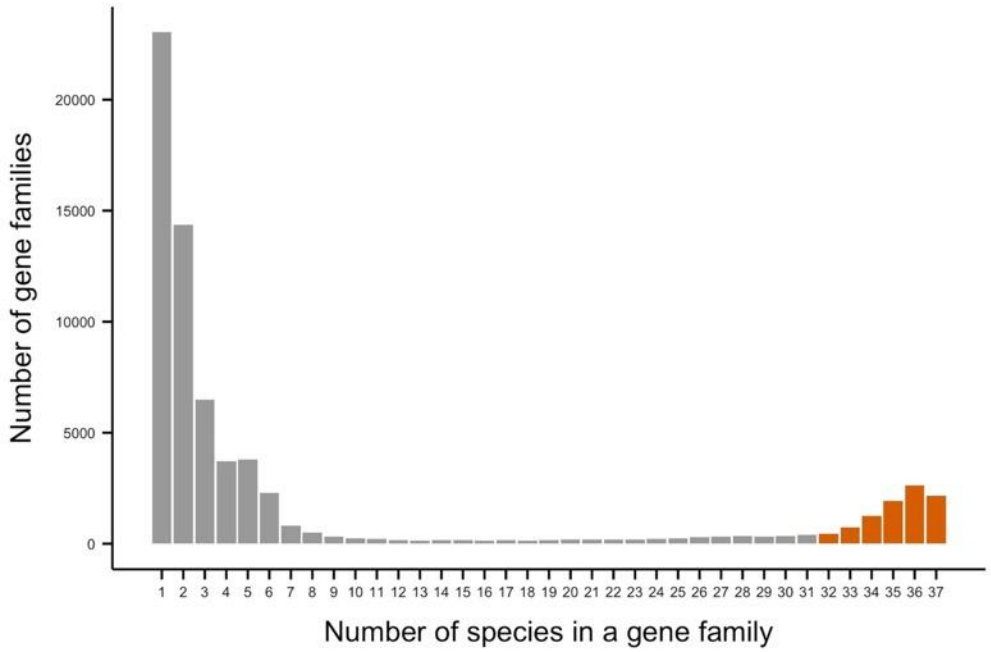
Species	Short name	Clade
Physcomitrella patens	PPA	Outgroup
Abies alba	AAL	Pinaceae
Cedrus libani	CLI	Pinaceae
Larix kaempferi	LKA	Pinaceae
Picea abies	PAB	Pinaceae
Picea glauca	PGL	Pinaceae
Picea sitchensis	PSI	Pinaceae
Pinus banksiana	PBA	Pinaceae
Pinus contorta	PCO	Pinaceae
Pinus lambertiana	PLA	Pinaceae
Pinus massoniana	PMA	Pinaceae
Pinus pinaster	PPI	Pinaceae
Pinus sylvestris	PSY	Pinaceae
Pinus taeda	PTA	Pinaceae
Pseudotsuga menziesii	PME	Pinaceae

Supplementary Table C-3. Statistics of the sequencing reads in *Pinus pinaster* and *Pinus sylvestris*

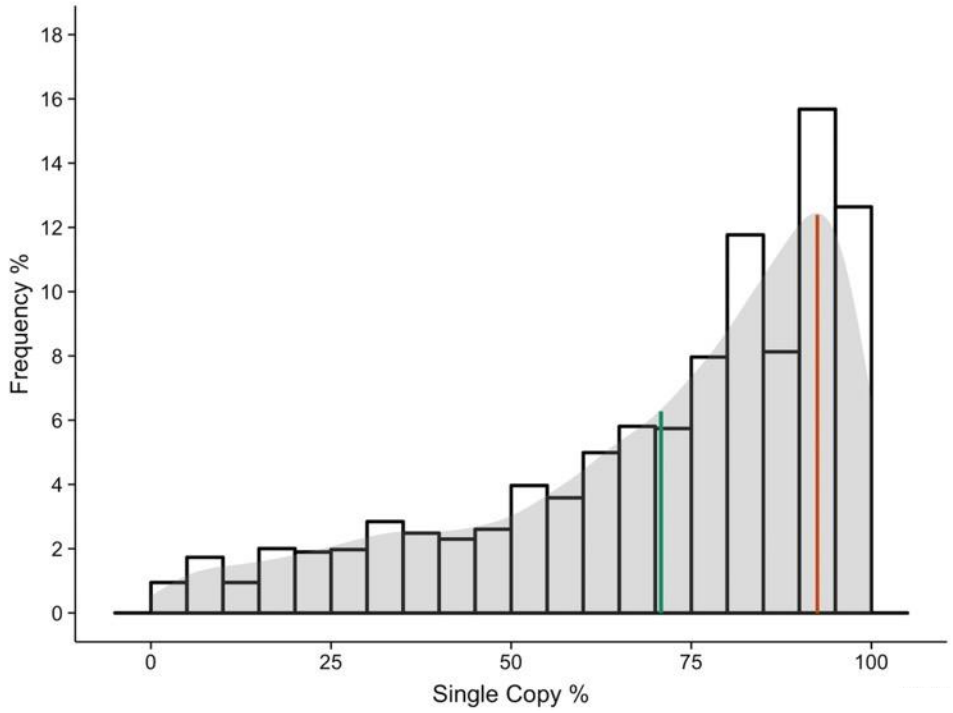
	Tissue	Raw Data	# Reads for Assembly	# Bases for Assembly	% Reads for Assembly
Pinus pinaster	Shoot Apical Meristem	548,328	489,417	193,513,935	89.26%
	Shoot Apical Meristem	511,488	450,986	246,681,796	88.17%
	Cortex Root	595,575	388,906	144,683,636	65.30%
	Cortex Hypocotyl	552,098	445,605	238,646,186	80.71%
	Pith Hypocotyl	238,995	177,643	34,516,172	74.33%
	Pith Hypocotyl	141,112	110,838	31,040,865	78.55%
	Vascular Developing Root	596,858	511,873	351,624,842	85.76%
	Cortex Developing Root	466,708	380,969	251,267,660	81.63%
	Root Apical Meristem	475,279	422,455	153,747,617	88.89%
	Root Apical Meristem	605,083	535,391	338,062,907	88.48%
	Vascular Root	297,094	173,094	78,592,365	58.26%
	Vascular Root	448,120	262,767	129,947,819	58.64%
	Vascular Cotyledon	726,853	663,607	424,191,066	91.30%
	Mesophyll Cotyledon	776,974	720,010	460,528,808	92.67%
	Pith Hypocotyl	631,059	573,629	353,808,071	90.90%
	Vascular New Needle	747,050	691,860	472,097,944	92.61%
	Vascular Hypocotyl	678,337	607,026	363,679,917	89.49%
	Developing Needle	747,508	702,247	460,245,151	93.95%
	Mesophyll New Needle	682,094	601,778	328,943,737	88.23%
	Vascular Root	712,551	673,643	335,482,084	94.54%
	PPIN_454_Reads_Reg1.RL6.sff	493,631	394,947	151,121,677	80.01%
	PPIN_454_Reads_Reg2.RL6.sff	724,677	562,481	212,858,345	77.62%
	Total	**12,397,472**	**10,541,172**	**5,755,282,600**	**85.03%**
Pinus sylvestris	Early Embryo (E1)	603,508	539,840	250,762,894	89.45%
	Cleavage (E2)	625,195	573,611	279,768,670	91.75%
	Dominant Embryo (E3DO)	749,430	711,685	506,198,526	94.96%
	Megagametophyte (E3SU)	745,590	708,247	499,030,847	94.99%
	Dominant Embryo (E4)	817,722	780,954	552,917,241	95.50%
	Subordinate Embryos (M1)	758,018	725,345	512,694,678	95.69%
	Megagametophyte (M2)	712,754	671,451	465,102,991	94.21%
	Megagametophyte (M3)	800,707	768,722	521,997,230	96.01%
	Megagametophyte (M4)	789,571	754,048	516,959,353	95.50%
	Total	**6,602,495**	**6,233,903**	**4,105,432,430**	**94.42%**

D. Supplementary information – Gene duplicability of core genes is highly consistent across all angiosperms

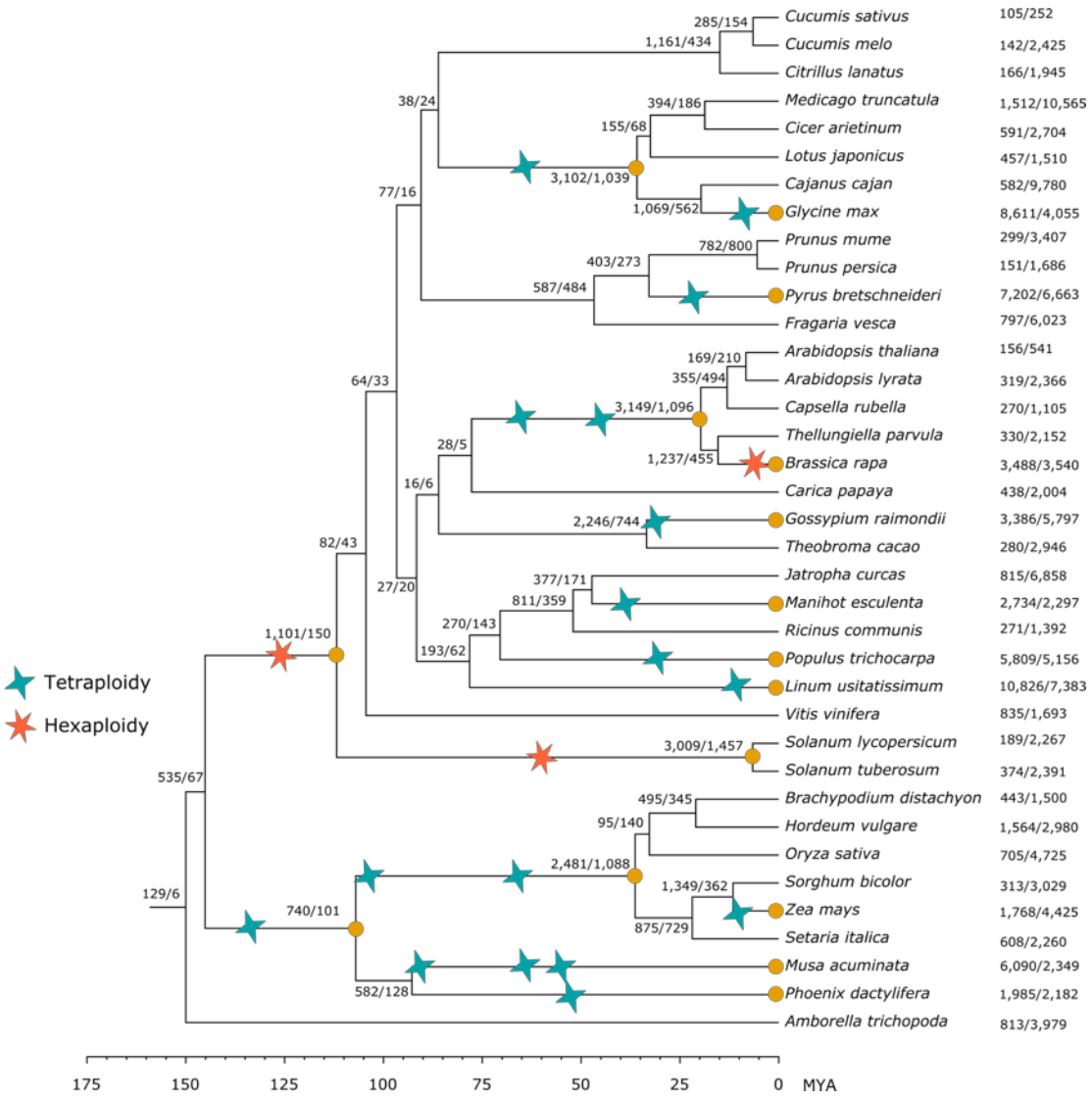
D.1. Supplementary Figures



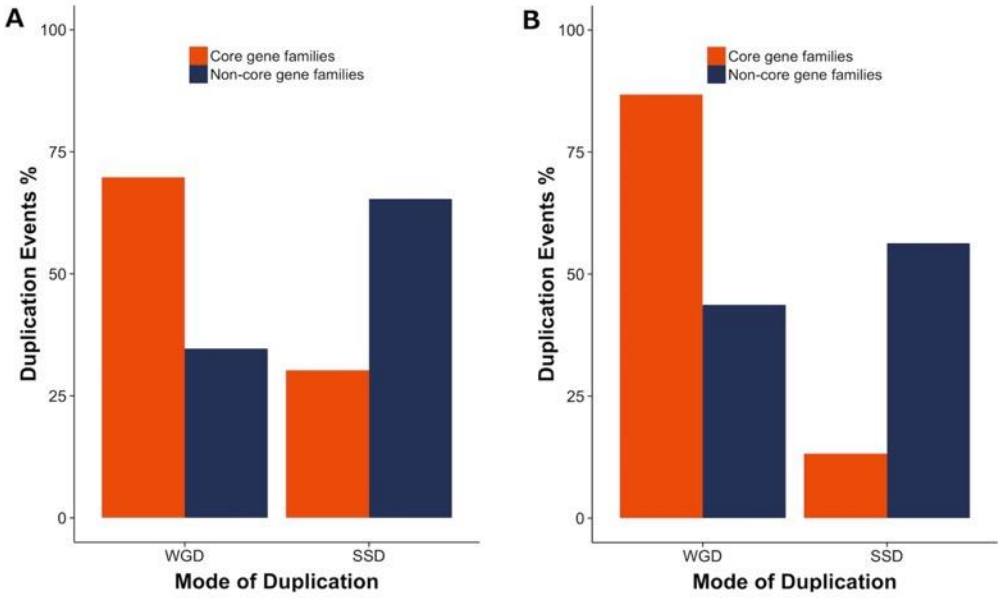
Supplementary Figure D-1. Motivation for the 32 out of 37 species cut-off to define core gene families. To distinguish core from non-core gene families we assessed the distribution of the number of species in each gene family based on all 69,542 gene families obtained by reconciliation. This distribution is U-shaped, suggesting a large number of gene families that are species- or lineage-specific (left side of the distribution) and also an excess of gene families present in the large majority of angiosperm species (right side of the distribution). Based on this distribution we decided to consider all gene families containing genes from at least 32 species as being 'core gene families'. As such we account for a limited number of putative missing orthologs from core gene families due to for instance errors in genome annotation, gene family construction errors or the presence of incomplete genomes.



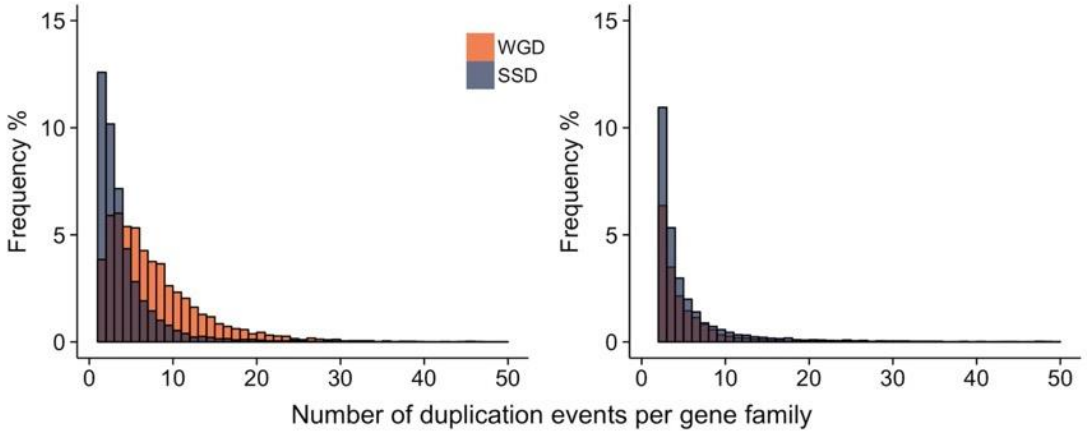
Supplementary Figure D-2. The distribution of Single-Copy Percentages (SCPs) for all core gene families, with SCPs calculated upon removing the highly duplicated genomes of *Glycine max*, *Linum usitatissimum*, *Brassica rapa*, and *Zea mays*. This distribution has a mode of 92% and a mean of 70.8%.



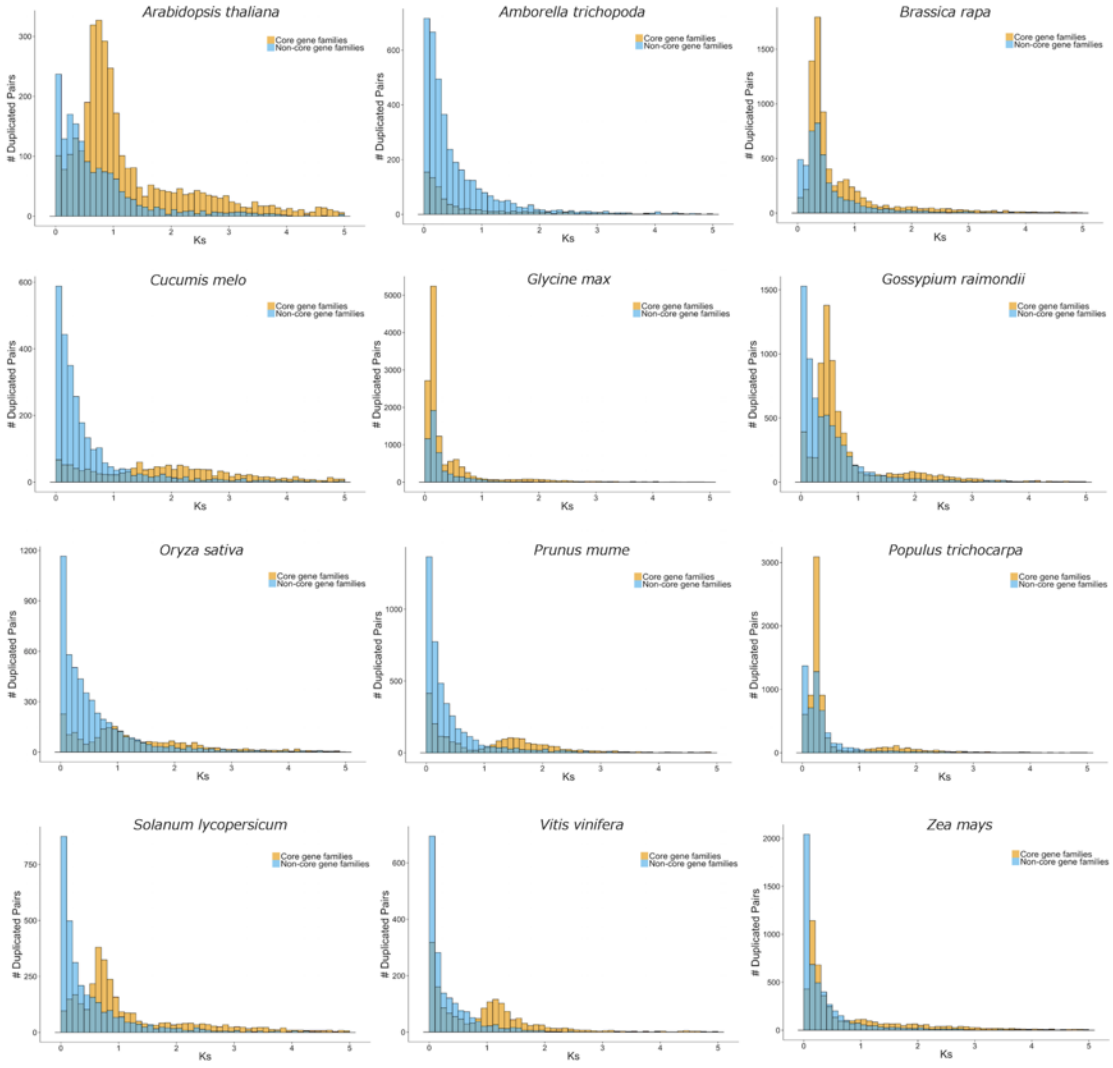
Supplementary Figure D-3. Classification of species tree nodes as SSD or WGD. On the species tree, nodes with WGDs on their parent branches were considered as WGD nodes (orange dots), while the rest of the nodes were considered as SSD nodes. Next to each node are the number of duplication events predicted by gene tree-species tree reconciliation for both core and non-core gene families (core/non-core). There are in total 93,942 predicted duplication events in core gene families and 140,786 duplication events in non-core gene families.



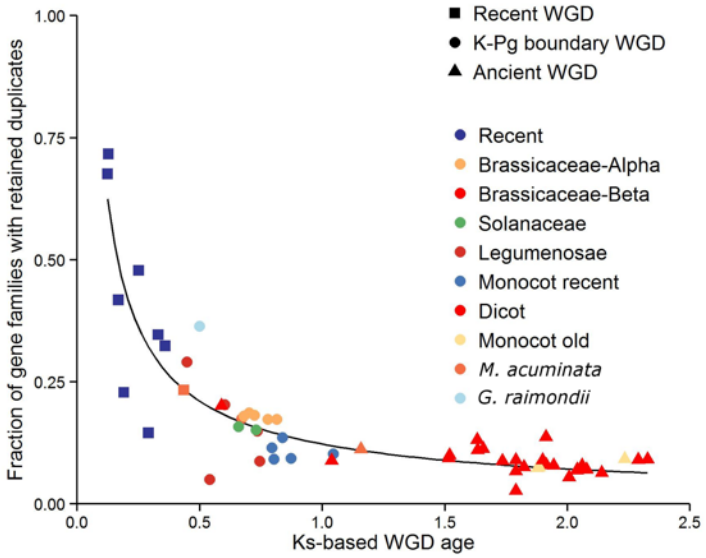
Supplementary Figure D-4. Core gene families mainly duplicate through WGD. Bar plots represent the fraction of duplication events, summed over all gene families, attributed to WGD or SSD in core and non-core gene families. Panel (A) represents results obtained from all nodes in the species tree in (Supplementary Figure D-2) and shows that for core genes families, as compared to non-core gene families, the presence of duplicates seems to be biased towards WGD-associated gene duplication ($p < 2.2\times10^{-16}$, Fisher's exact test). In panel (B) we assessed the possibility that these observations might be caused by an overrepresentation of WGD-associated nodes in the species tree for core gene families as opposed to non-core gene families: since core gene families cover by definition a larger number of species, some of the more ancient WGD events that are shared by many species will only be represented by core gene families. Hence, we repeated this analysis by only considering nodes from the species tree that are also ubiquitously present in non-core gene families (top 10 of the nodes) and came to the same conclusion ($p < 2.2\times10^{-16}$, Fisher's exact test).



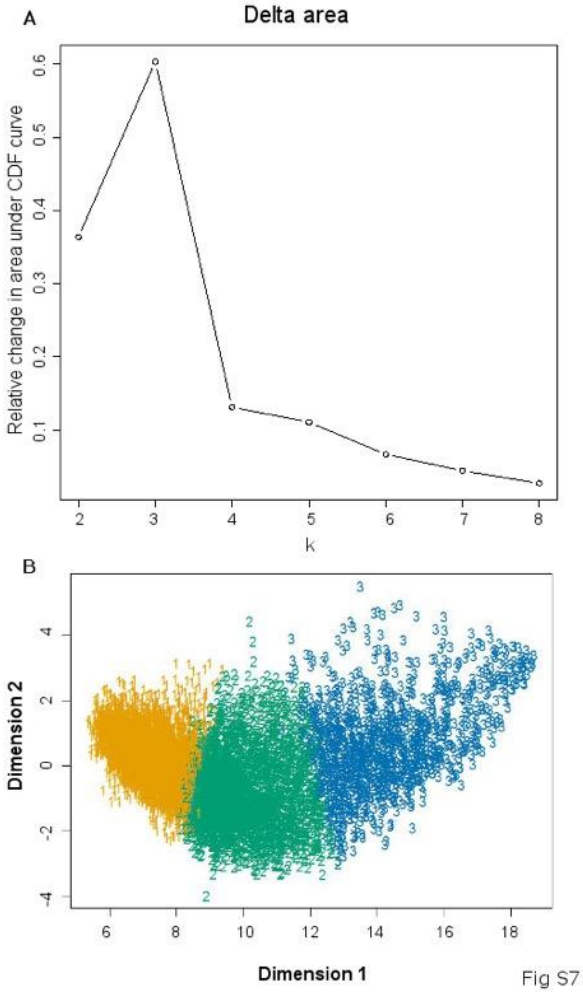
Supplementary Figure D-5. Comparison of the number of duplications for core and non-core gene families at WGD and SSD nodes on a gene family base (only illustrating gene families with no more than 50 duplications). (A) The number of WGD and SSD duplications per gene family for core gene families. There are significantly more nodes associated with WGD derived duplications than SSD derived duplications ($p < 2.2 \times 10^{-16}$, Wilcoxon-rank-sum test). (B) The number of WGD and SSD duplication per gene family for non-core gene families. Here the number of WGD derived duplications is not significantly larger than those of SSD derived duplications ($p = 1$, Wilcoxon-rank-sum test). Predicted duplication events were obtained by gene tree - species tree reconciliation (see Materials and Methods).



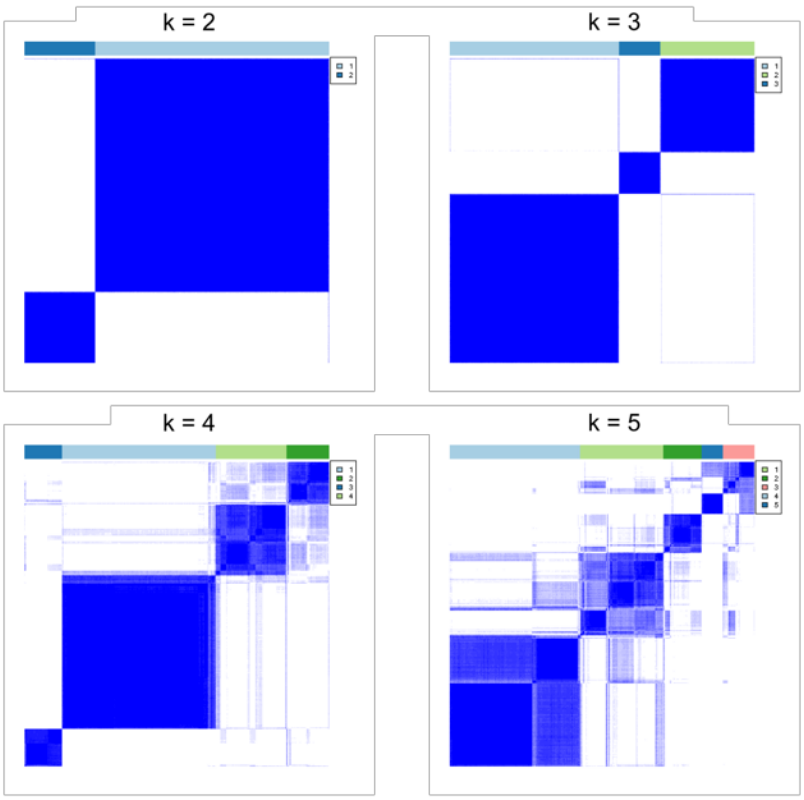
Supplementary Figure D-6. K_S-distributions of duplicated pairs from core and non-core gene families in 12 species, *i.e.*, *Arabidopsis thaliana*, *Amborella trichopoda*, *Brassica rapa*, *Cucumis melo*, *Glycine max*, *Gossypium raimondii*, *Oryza sativa*, *Prunus mume*, *Populus trichocarpa*, *Solanum lycopersicum*, *Vitis vinifera*, and *Zea mays*.



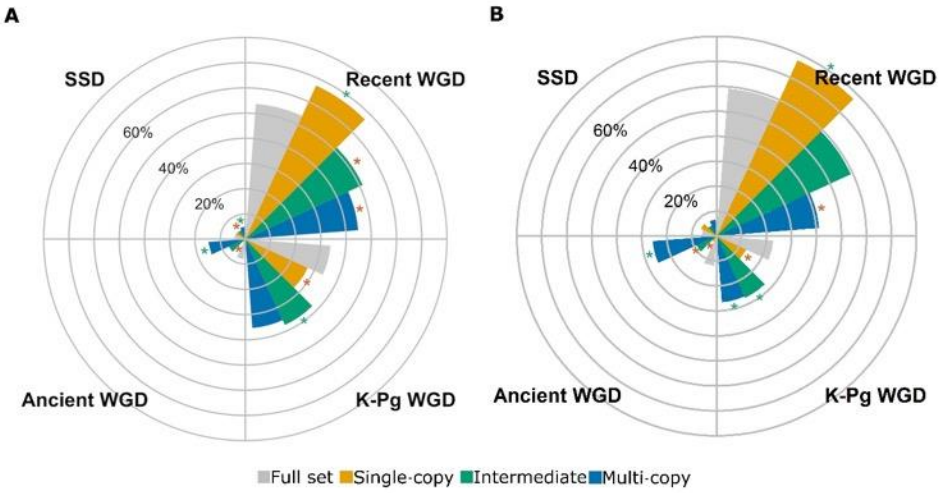
Supplementary Figure D-7. Duplicate gene retention in function of time since WGD. Each dot represents the fraction of core gene families with retained duplicates following a specific WGD (y-axis), as a function of WGD age, expressed in K_S-units (x-axis). The timing of the WGD events and the particular gene families that retained duplicates following a specific WGD event were inferred by fitting Gaussian mixture models to K_S-age distributions for all 37 species separately (see Materials and Methods). This figure is related to Figure 3-3, but here all WGD peak callings were included. Since the Dicot and Brassicaceae-Beta peaks can not be distinguished from each other they are denoted by the same color. Additional information on all the peaks is provided in the Supplementary Table D-2.



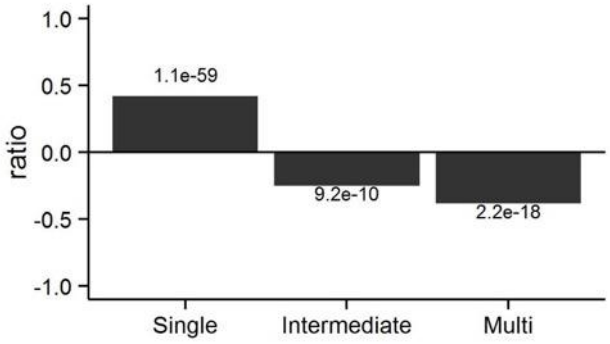
Supplementary Figure D-8. Criteria that we used to choose the optimal number of clusters for *k*-means clustering of the copy-number matrix. (A) We used the Delta Area Plot from the ConsensusClusterPlus R-package to select the optimal number of clusters. The results of 1000 clustering runs, each time on subsampled matrices, are summarized into a consensus matrix, whose values represent the proportion of clustering runs in which two items (*i.e.*, gene families) are grouped together. Hence, values in this matrix are between 0 and 1 (*i.e.*, always clustered together). The Delta Area Plot assesses the 'cleanness' of this consensus matrix: if all clustering runs agree on the same solution than this matrix only consists of 0's and 1's (bimodal distribution). To determine the optimal numbers of clusters the largest changes in these consensus values are detected by calculating the change in the area under the Cumulative Distribution of consensus values for increasing cluster number[301]. The 'Delta area' represents this change, with *k* corresponding to cluster number. (B) Corresponding multidimensional scaling plot of the copy-number matrix, with data points colored according to cluster membership.



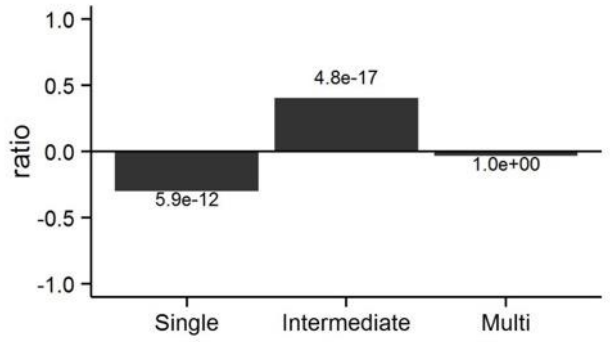
Supplementary Figure D-9. Consensus matrices obtained for different number of clusters *k.* The consensus matrix represents the number of times that two gene families belonged to the same cluster over 1,000 clustering runs of the subsampled copy-number matrix. The values within this matrix range from 0 (gene families were never grouped into the same cluster; white in this figure) to 1 (gene families were always grouped into the same cluster; blue in this figure). Here results are shown for *k* = 2-5 clusters. Color bars on top of the visualized consensus matrix indicate cluster assignments.



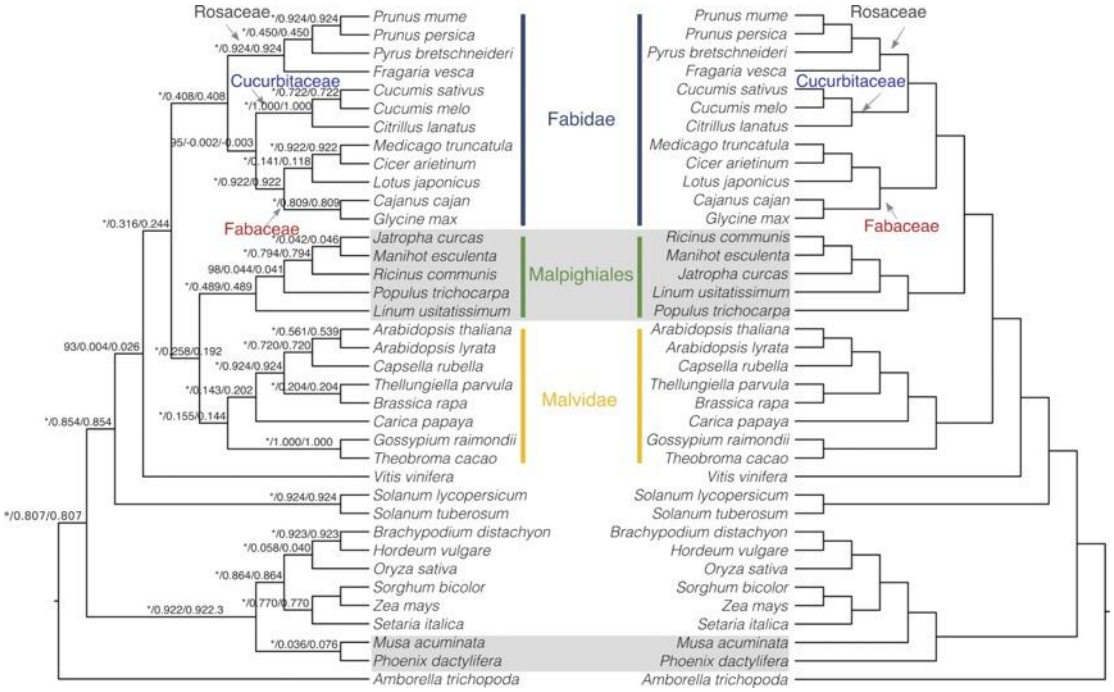
Supplementary Figure D-10. Polar diagrams depicting the fraction of duplication events in each gene family group belonging to either the ,000 clustering runs of the subsampled copy-number matrix. lasses. (A) Represents predictions of duplication timing for all core gene families, obtained by using gene tree and species tree reconciliation. This Figure is the same as Figure 3-5B. In contrast to GMM (see panel B), which provides estimates of the ages of the duplication events for each species separately, here estimates of the duplication age is based on a gene family basis and hence no averaging over species is necessary. To obtain the bar plots we normalized the absolute counts of duplication events for each node in the species tree with the number of nodes in the species tree of that duplication class, correcting for the fact that there are for instance more nodes associated to the panel duplication class. Significance values are indicated by asterisks (green = overrepresentation, red = underrepresentation) and were calculated based on the absolute counts of predicted duplications of each class. The predictions of duplication timing for all core gene families are based on GMM of K_S-based species-specific age distributions. We classified each duplicate pair to a certain duplication class depending on the K_S-peak it belonged to (see Supplementary Table D-1). The bars in the Figures represent averages, obtained from averaging over the number of duplications assigned to a certain class for all species. Statistical significant over- and underrepresentations were calculated based on the Wilcoxon-rank-sum test and are denoted by asterisks.



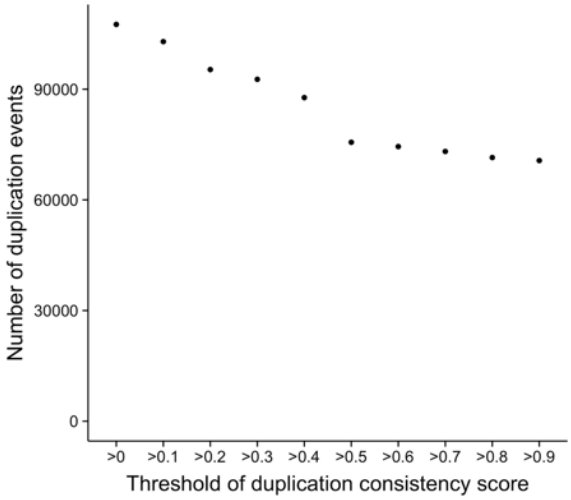
Supplementary Figure D-11. Over- and underrepresentation of an independent set of 2,090 nuclear-encoded chloroplast-targeted genes obtained from The Chloroplast Function Database [303]. The y-axis represents over- (positive values) or under- (negative values) representation of these chloroplast genes in the three different functional groups as compared to the full set. In specific, to obtain the values on the y-axis we calculated the ratio of the proportion of group genes (*i.e.*, 'Single', 'Intermediate' or 'Multi') that are chloroplast genes to the proportion of genes in the full set that are chloroplast genes. Positive values for overrepresentation (ratio > 1) and negative values for underrepresentation (ratio < 1) were obtained by subtracting one from the above described ratio. P-values as obtained by Fisher's exact test with Bonferroni multiple-testing correction are indicated on the bars.



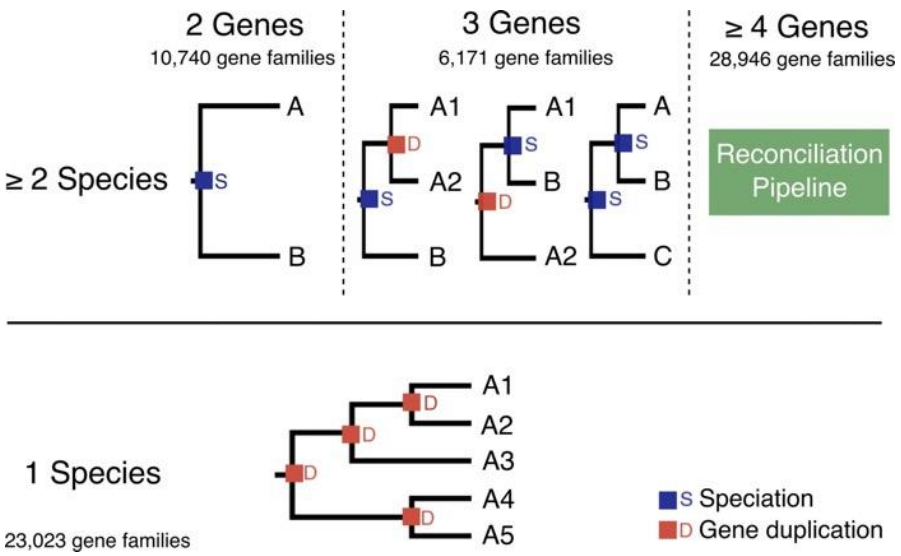
Supplementary Figure D-12. Over- and underrepresentation of an independent set of 1,795 putative transcription factors, obtained from Perez-Rodriguez *et al.*[470]. The y-axis represents over- (positive values) or under- (negative values) representations for transcription factor genes in the three different functional groups as compared to the full set. In specific, to obtain the values on the y-axis we calculated the ratio of the proportion of group genes (*i.e.*, 'Single', 'Intermediate' or 'Multi') that are transcription factors to the proportion of genes in the full set that are transcription factors. Positive values for overrepresentation (ratio > 1) and negative values for underrepresentation (ratio < 1) were obtained by subtracting one from the above described ratio. *P* values as obtained by Fisher's exact test with Bonferroni multiple-testing correction are indicated on the bars.



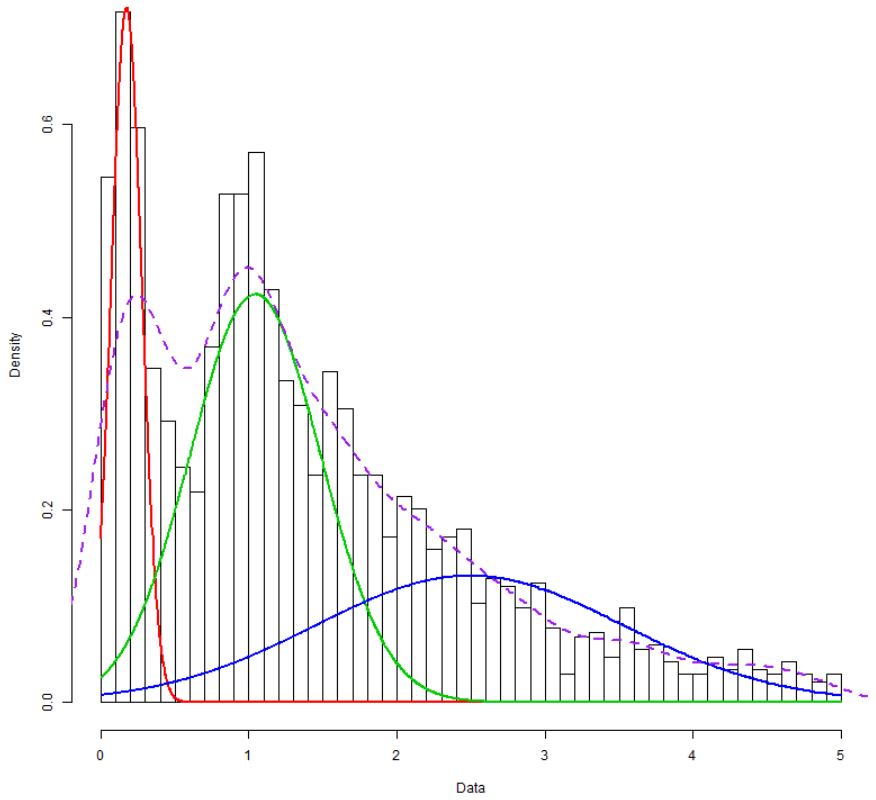
Supplementary Figure D-13. Conflicting clades between the species tree used in this paper and which we inferred from 107 core gene families (left) and the APGIII tree (right). The here obtained species tree is largely consistent with the APGIIItree[337], yet there are some conflicts. The incongruence between the positions of the Malpighiales clade in trees constructed from nuclear genes versus chloroplast genes have long been recognized, and is thought to be caused by introgressive hybridization in the ancestral lineages of Fabidae and Malvidae[233]. Moreover, due to rapid diversification at the mid-Cretaceous, the relationships within Malpighiales are hard to determine[433]. The close to zero values of IC and ICA suggest incongruence of the gene trees and the species tree on the branch leading to *Populus trichocarpa* and on the branch leading to *Jartropha curcas* and *Manihot esculenta*. Similarly, the monophyletic group consisting of Cucurbitaceae and Fabaceae is also only supported by half of the gene families used to reconstruct the species tree.



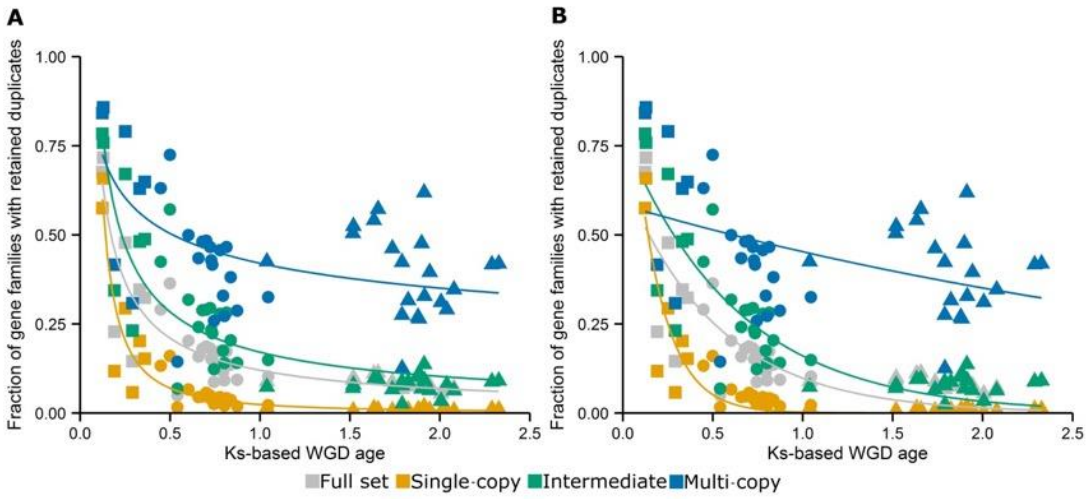
Supplementary Figure D-14. The change in the total number of predicted duplication events in core gene families in function of the threshold on the duplication consistency score. The predicted number of duplication events stays relatively stable for duplication consistency score thresholds up until 0.5, yet shows a drop for duplication consistency scores larger than 0.5. The large reduction at 0.5 can be explained by the large number of nodes in the species tree that only encompass two species and hence the large effect of an increase in the duplication consistency score threshold from 0.4 to 0.5 on the total number of duplication events: e.g. ((ath,aly)ath) will not make the cut of a duplication consistency score > 0.5.



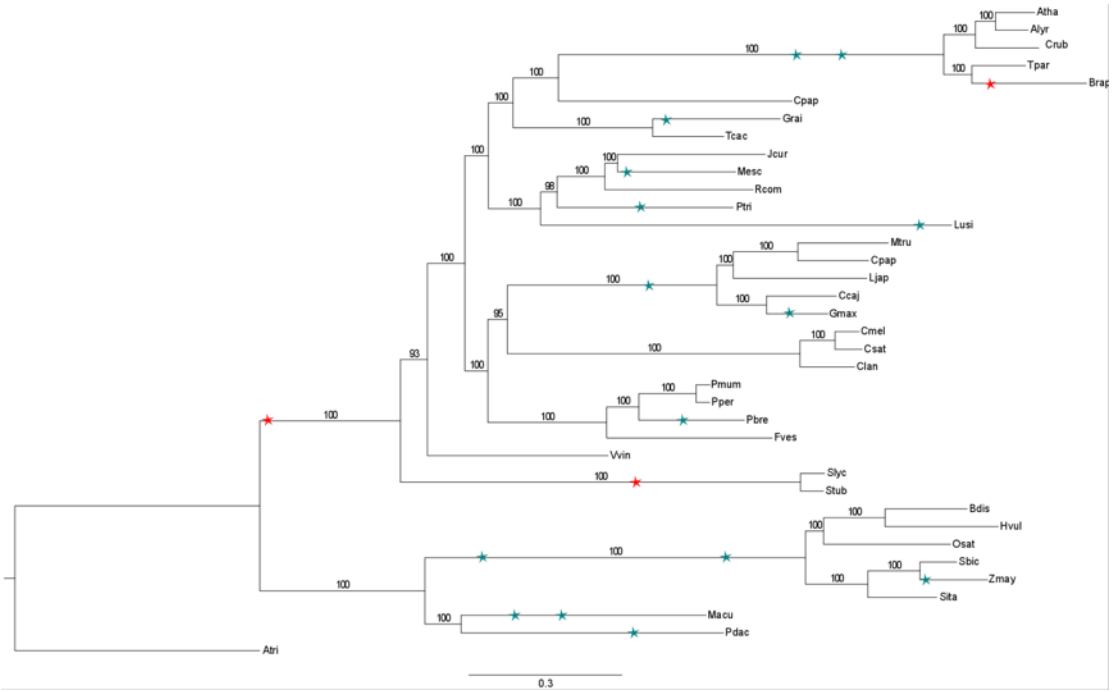
Supplementary Figure D-15. Explanation of how duplications were inferred for gene families with at least two species but no more than three genes or gene families that are only present in one species. For gene families with two genes in two species (10,740 gene families), the node connecting both genes is assumed to be a speciation node. For gene families with three genes (6,171 gene families), we mid-point rerooted the gene tree and distinguished between three possible scenarios. If the three genes come from two species, the duplication occurred either in one species or in the common ancestor of the two species, depending on the topology of the gene tree. If the three genes come from three species, we assume that no duplications have occurred in the history of the gene family (most parsimonious scenario). For gene families that only cover one species (23,023) but with two genes or more, e.g. five genes in the figure, we mid-point rerooted the gene tree and considered all nodes in the tree to be duplication nodes. For the remaining 28,946 gene families with at least four genes (including all core gene families) duplications were inferred using the reconciliation pipelines as described in Materials and Methods.



Supplementary Figure D-16. Gaussian mixture models were fit to the K_S-distribution of each species. Peaks were considered solid if they had a good visual fit with the density line (dashed purple line) and the K_S histogram and had a μ lower than 3. Flat peaks, e.g. peaks which span the whole K_S distribution, where also removed. The annotation of the peaks was done using known literature[168]. The figure shows the K_S distribution for *Sorghum bicolor*. The red and green peaks have a good fit to the density line whereas the flat blue peak shows no correspondence to density line and spans the whole K_S distribution.



Supplementary Figure D-17. Comparison of (A) power-law fit and (B) exponential fit to the data obtained from the Gaussian Mixture Modeling of K_S-based age distributions. The power-law shows consistently a better fit than the exponential, as assessed by Chi-squared Goodness-Of-Fit test (see Supplementary Table D-3).



Supplementary Figure D-18. Mapping of the whole-genome duplications and triplications on the species tree as obtained by the approach outlined in 'Dating whole-genome duplications' and as used for the simulations of gene family evolution according to the stochastic gene birth-death null model.

D.2. Supplementary Tables

Supplementary Table D-1. Comparison of the numbers of interacting protein pairs in each group to those obtained from randomized networks.

	Number of PPIs within group	Average number of PPIs within group for 1000 randomized networks	Z-score	*P* value enrichment of PPI vs random (one-sided test)	*P* value with multiple-testing correction (Bonferroni)
Full	15949	15949			
Single-copy	2550	2813.012	-1.005	0.84	1
Intermediate	2277	1740.331	2.710	0.0034	0.010
Multi-copy	1034	990.558	0.322	0.374	1

Supplementary Table D-2. Description of all identified peaks inferred from the K_S-based age distributions.

Species	*k*	μ	*σ*	λ	L_bound	H_bound	Annotation	WGD type	Included
Alyr1	4	0.095	0.086	0.131	0.000	0.289	SSD	SSD	NO
Alyr2	4	0.723	0.258	0.579	0.289	1.199	BRAalpha	KT	YES
Alyr3	4	2.038	0.720	0.227	1.199	2.970	BRABeta	OLD	NO
Alyr4	4	3.848	0.631	0.063	2.970	5.000	HighKS	HighKS	NO
Atha1	4	0.178	0.122	0.088	0.000	0.411	SSD	SSD	NO
Atha2	4	0.778	0.243	0.574	0.411	1.231	BRAalpha	KT	YES
Atha3	4	2.059	0.783	0.286	1.231	3.185	BRABeta	OLD	NO
Atha4	4	4.083	0.533	0.052	3.185	5.000	HighKS	HighKS	NO
Bdis1	4	0.182	0.108	0.144	0.000	0.400	SSD	SSD	NO
Bdis2	4	0.802	0.263	0.374	0.400	1.240	MON1	KT	YES
Bdis3	4	1.878	0.613	0.383	1.240	2.762	MON2	OLD	YES
Bdis4	4	3.688	0.671	0.100	2.762	5.000	HighKS	HighKS	NO
Brap1	3	0.331	0.082	0.513	0.000	0.479	REC	REC	YES
Brap2	3	0.701	0.340	0.334	0.479	1.292	BRAalpha	KT	YES
Brap3	3	2.220	1.025	0.153	1.292	5.000	BRABeta	OLD	NO
Cari1	4	0.047	0.039	0.118	0.000	0.155	SSD	SSD	NO
Cari2	4	0.735	0.316	0.543	0.155	1.273	LEG	KT	YES
Cari3	4	2.078	0.725	0.277	1.273	3.064	DIC	OLD	YES
Cari4	4	3.945	0.581	0.063	3.064	5.000	HighKS	HighKS	NO
Ccaj1	4	0.032	0.037	0.100	0.000	0.138	SSD	SSD	NO
Ccaj2	4	0.602	0.214	0.569	0.138	1.009	LEG	KT	YES
Ccaj3	4	1.789	0.679	0.279	1.009	2.794	DIC	OLD	YES
Ccaj4	4	3.746	0.617	0.052	2.794	5.000	HighKS	HighKS	NO
Clan1	3	0.2643	0.1755	0.2239	0.0000	0.6731	SSD	SSD	NO
Clan2	3	1.8231	0.7083	0.6317	0.6731	2.7961	DIC	OLD	YES
Clan3	3	3.7459	0.6738	0.1444	2.7961	5.0000	HighKS	HighKS	NO
Cmel1	3	0.2786	0.2019	0.1872	0.0000	0.7310	SSD	SSD	NO
Cmel2	3	1.9139	0.7743	0.6712	0.7310	2.9552	DIC	OLD	YES
Cmel3	3	3.8984	0.6355	0.1416	2.9552	5.0000	HighKS	HighKS	NO
Cpap1	3	0.249	0.202	0.306	0.000	0.765	SSD	SSD	NO
Cpap2	3	2.006	0.595	0.602	0.765	2.995	DIC	OLD	YES
Cpap3	3	3.897	0.517	0.092	2.995	5.000	HighKS	HighKS	NO
Crub1	4	0.124	0.075	0.070	0.000	0.308	SSD	SSD	NO
Crub2	4	0.814	0.273	0.593	0.308	1.289	BRAalpha	KT	YES
Crub3	4	2.039	0.724	0.263	1.289	3.027	BRABeta	OLD	NO

Supplementary Table D-2. Description of all identified peaks inferred from the K_S-based age distributions.

Species	*k*	μ	*σ*	λ	L_bound	H_bound	Annotation	WGD type	Included
Crub4	4	3.907	0.580	0.075	3.027	5.000	HighKS	HighKS	NO
Csat1	3	0.318	0.216	0.192	0.000	0.777	SSD	SSD	NO
Csat2	3	1.789	0.680	0.580	0.777	2.596	DIC	OLD	YES
Csat3	3	3.425	0.773	0.228	2.596	5.000	HighKS	HighKS	NO
Fves1	3	0.334	0.222	0.365	0.000	0.791	SSD	SSD	NO
Fves2	3	1.735	0.658	0.552	0.791	2.631	DIC	OLD	YES
Fves3	3	3.543	0.685	0.083	2.631	5.000	HighKS	HighKS	NO
Gmax1	3	0.124	0.044	0.622	0.000	0.216	REC	REC	YES
Gmax2	3	0.448	0.208	0.261	0.216	0.872	LEG	KT	YES
Gmax3	3	1.868	0.967	0.117	0.872	5.000	DIC	OLD	NO
Grai1	3	0.048	0.037	0.058	0.000	0.149	SSD	SSD	NO
Grai2	3	0.499	0.166	0.703	0.149	0.858	KT	KT	YES
Grai3	3	1.912	0.964	0.239	0.858	5.000	DIC	OLD	YES
Hvul1	3	0.011	0.010	0.115	0.000	0.042	SSD	SSD	NO
Hvul2	3	0.639	0.416	0.487	0.042	1.312	MON1	KT	NO
Hvul3	3	2.217	1.092	0.398	1.312	5.000	MON2	OLD	NO
Jcur1	3	0.120	0.116	0.274	0.000	0.432	SSD	SSD	NO
Jcur2	3	1.943	0.831	0.669	0.432	3.377	DIC	OLD	YES
Jcur3	3	4.271	0.432	0.057	3.377	5.000	HighKS	HighKS	NO
Ljap1	4	0.051	0.058	0.144	0.000	0.180	SSD	SSD	NO
Ljap2	4	0.541	0.268	0.490	0.180	1.018	LEG	KT	YES
Ljap3	4	1.790	0.655	0.252	1.018	2.634	DIC	OLD	YES
Ljap4	4	3.491	0.682	0.114	2.634	5.000	HighKS	HighKS	NO
Lusi1	3	0.128	0.056	0.726	0.000	0.249	REC	REC	YES
Lusi2	3	0.588	0.303	0.190	0.249	1.163	DIC	OLD	NO
Lusi3	3	2.265	1.025	0.084	1.163	5.000	HighKS	HighKS	NO
Macu1	5	0.075	0.039	0.021	0.000	0.198	SSD	SSD	NO
Macu2	5	0.435	0.081	0.326	0.198	0.556	MAC	KT	NO
Macu3	5	0.672	0.211	0.398	0.556	0.937	MAC	KT	NO
Macu4	5	1.158	0.398	0.220	0.937	1.782	MAC	OLD	NO
Macu5	5	2.538	1.049	0.036	1.782	5.000	HighKS	HighKS	NO
Mesc1	4	0.071	0.040	0.044	0.000	0.171	SSD	SSD	NO
Mesc2	4	0.359	0.086	0.671	0.171	0.580	REC	REC	YES
Mesc3	4	1.633	0.664	0.251	0.580	2.667	DIC	OLD	YES
Mesc4	4	3.717	0.681	0.034	2.667	5.000	HighKS	HighKS	NO
Mtru1	3	0.159	0.122	0.342	0.000	0.379	SSD	SSD	NO
Mtru2	3	0.744	0.324	0.414	0.379	1.330	LEG	KT	YES
Mtru3	3	2.338	1.063	0.244	1.330	5.000	DIC	OLD	NO
Osat1	4	0.143	0.114	0.197	0.000	0.396	SSD	SSD	NO
Osat2	4	0.873	0.266	0.356	0.396	1.300	MON1	KT	YES
Osat3	4	1.884	0.598	0.365	1.300	2.829	MON2	OLD	YES
Osat4	4	3.779	0.602	0.082	2.829	5.000	HighKS	HighKS	NO
Pbre1	3	0.010	0.010	0.290	0.000	0.038	SSD	SSD	NO
Pbre2	3	0.168	0.071	0.550	0.038	0.353	REC	REC	NO
Pbre3	3	1.564	0.950	0.160	0.353	5.000	DIC	OLD	NO
Pdac1	3	0.291	0.078	0.548	0.100	0.440	REC	KT	YES
Pdac2	3	0.706	0.375	0.394	0.440	1.354	?	?	NO
Pdac3	3	2.350	1.130	0.057	1.354	5.000	?	?	NO
Pmum1	3	0.167	0.150	0.418	0.000	0.534	SSD	SSD	NO
Pmum2	3	1.516	0.522	0.488	0.577	2.185	DIC	OLD	YES
Pmum3	3	2.813	0.957	0.094	2.162	5.000	HighKS	HighKS	NO
Pper1	3	0.194	0.153	0.391	0.000	0.571	SSD	SSD	NO
Pper2	3	1.519	0.488	0.519	0.571	2.189	DIC	OLD	YES
Pper3	3	2.894	0.946	0.089	2.189	5.000	HighKS	HighKS	NO
Ptri1	3	0.028	0.020	0.072	0.000	0.085	SSD	SSD	NO
Ptri2	3	0.251	0.067	0.719	0.085	0.428	REC	REC	YES
Ptri3	3	1.632	0.940	0.209	0.428	5.000	DIC	OLD	NO
Rcom1	3	0.278	0.197	0.186	0.000	0.736	SSD	SSD	NO
Rcom2	3	1.898	0.685	0.741	0.736	3.130	DIC	OLD	YES

Supplementary Table D-2. Description of all identified peaks inferred from the K_S-based age distributions.

Species	*k*	μ	*σ*	λ	L_bound	H_bound	Annotation	WGD type	Included
Rcom3	3	4.087	0.483	0.073	3.130	5.000	HighKS	HighKS	NO
Sbic1	3	0.175	0.103	0.187	0.000	0.406	SSD	SSD	NO
Sbic2	3	1.045	0.442	0.469	0.406	1.711	MON1	KT	YES
Sbic3	3	2.490	1.045	0.344	1.711	5.000	MON2	OLD	NO
Sita1	3	0.079	0.062	0.126	0.000	0.231	SSD	SSD	NO
Sita2	3	0.837	0.398	0.490	0.231	1.461	MON1	KT	YES
Sita3	3	2.233	1.027	0.384	1.461	5.000	MON2	OLD	YESB
Slyc1	3	0.184	0.094	0.125	0.000	0.375	SSD	SSD	NO
Slyc2	3	0.729	0.228	0.541	0.375	1.197	SOL	KT	YES
Slyc3	3	2.327	1.068	0.334	1.197	5.000	DIC	OLD	YES
Stub1	3	0.118	0.085	0.212	0.000	0.300	SSD	SSD	NO
Stub2	3	0.658	0.223	0.501	0.300	1.121	SOL	KT	YES
Stub3	3	2.289	1.071	0.286	1.121	5.000	DIC	OLD	YES
Tcac1	3	0.128	0.061	0.142	0.000	0.311	SSD	SSD	NO
Tcac2	3	1.656	0.663	0.787	0.311	2.802	DIC	OLD	YES
Tcac3	3	3.874	0.600	0.071	2.802	5.000	HighKS	HighKS	NO
Tpar1	3	0.680	0.356	0.707	0.000	1.309	BRAalpha	KT	YES
Tpar2	3	2.140	0.555	0.211	1.309	2.959	BRABeta	OLD	NO
Tpar3	3	3.835	0.632	0.082	2.959	5.000	HighKS	HighKS	NO
Vvin1	3	0.088	0.067	0.292	0.000	0.258	SSD	SSD	NO
Vvin2	3	1.038	0.494	0.611	0.258	1.767	DIC	OLD	YES
Vvin3	3	2.608	1.089	0.097	1.767	5.000	HighKS	HighKS	NO
Zmay1	3	0.191	0.104	0.532	0.000	0.392	REC	REC	YES
Zmay2	3	0.795	0.394	0.226	0.392	1.426	MON1	KT	YES
Zmay3	3	2.248	1.036	0.242	1.426	5.000	MON2	OLD	NO

Each row in the table represents one peak: *k* denotes the number of components that were fitted; μ, σ, and λ are the obtained parameters for fitted GMMs; L_bound and U_bound represent respectively the lower- and upperbound K_S-values associated with each peak; Annotation represents the annotation of the peak based on data from Vanneste *et al.*[168]; WGD types is the classification of the peak as either 'SSD', 'Recent' (REC), 'K-Pg Boundary' (KT), 'Ancient' (OLD) or 'HighKS' if they had μ-values exceeding 3.5; 'Included' indicates whether we used the peak data to create Figure 3.3 and Figure 3.5B.

Supplementary Table D-3. Comparison of the power-law and the exponential fit.

	χ^2-goodness-of-fit (p-value)	
	Power-law	Exponential
Full	0.76795 (p =1)	5.072 (p=1)
Single-copy	0.52465 (p = 1)	477.6 (p < 2.2 e-16)
Intermediate	1.3838 (p = 1)	2.0733 (p = 1)
Multi-copy	1.8271 (p = 1)	2.1274 (p = 1)

E. Supplementary information – The *Apostasia* genome and the evolution of orchids

E.1. Supplementary Notes

E.1-1 Evolution of gene family sizes

We determined the expansion and contraction of orthologous gene families using CAFÉ 2.2[376]. One hundred and thirteen gene families were expanded in the lineage leading to the orchids, whereas 1,047 families became smaller (Figure 4-2). Five hundred and twenty-two gene families were expanded in *A. shenzhenica* (six by a significant margin), compared to 557 (five by a significant margin) in *P. equestris* and 872 (34 by a significant margin) in *D. catenatum*. At the same time, 1,661 (four by a significant margin) gene families became smaller in *A. shenzhenica* compared to 1,384 (27 by a significant margin) in *P. equestris* and 703 (one by a significant margin) in *D. catenatum* (Supplementary Tables E-15–18).

E.1-2 Orchidaceae-specific gene families

A four-way comparison of Orchidaceae (*A. shenzhenica*, *D. catenatum*, and *P. equestris*), dicots (*A. thaliana*, *P. trichocarpa*, and *V. vinifera*), Poales (*A. comosus*, *B. distachyon*, *O. sativa*, and *S. bicolor*), and a group of *M. acuminata* and *P. dactylifera* found 10,377 gene families to be shared by all taxa (Figure 4-3). In addition, 474 gene families present in all three orchid genomes are found to be unique to Orchidaceae, suggesting that orchids have fewer unique gene families than the three dicots (522), but much more than the four Poales (180).

For the *A. shenzhenica*-specific gene families, we conducted a Gene ontology (GO) / Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis via enrichment pipeline (https://sourceforge.net/projects/enrichmentpipeline/) and found enrichment of the GO Terms 'RNA-directed DNA polymerase activity', 'RNA-dependent DNA replication' (Supplementary Table E-19). The Orchidaceae-specific gene families were found to be enriched for GO terms 'Cysteine-type peptidase activity', 'O-methyltransferase activity' and the KEGG pathways 'Flavone and flavonol biosynthesis', 'Stilbenoid, diarylheptanoid and gingerol biosynthesis' (Supplementary Tables E-20 and E-21). The monocot-specific gene families were enriched for the GO terms 'two-component response regulator activity', 'solute antiporter activity', 'two-component signal transduction system', 'signal transducer activity' and 'molecular transducer activity', 'hydrogen ion transmembrane transporter activity' and the KEGG pathways 'Plant hormone signal transduction' and 'RNA polymerase' (Supplementary Tables E-22 and E-23).

O-methyltransferases can collectively mono- or polymethylate a great number of plant natural products. The methylation of plant natural products can alter their

solubility and intracellular compartmentalization and can increase their antimicrobial activity[471]. For instance, amongst the genes with O-methyltransferase activity, Ash015798 is highly expressed in stem and seed, while significant expression of Ash015796 is detected in floral buds, stems, leaves and seeds (Supplementary Figure E-27). Flavonoids and related phenolic compounds have essential functions in protecting plants against pathogens. The evolution of flavonoids has enabled vascular plants to cope with pathogen attacks and damaging UV light[472]. Ash006087, a gene involved in flavone and flavonol biosynthesis, and predominantly expressed in stems and roots, might play a protective role in vegetative tissues (Supplementary Figure E-28). Stilbenes are a small family of plant secondary metabolites that are found in a limited number of plant species, such as *Vitis sylvestris*[473]. Stilbenes are important phytoalexins that accumulate in response to various biotic and abiotic stresses, and increasing evidence suggests a positive correlation between the stilbene content of plants and disease resistance[474]. The two tandem duplicates, Ash006321 and Ash006322, which are involved in the biosynthesis of stilbenoids, diarylheptanoids, and gingerol, are both highly expressed in flower buds, stems and seeds, and likely have important functions against stresses (Supplementary Figure E-29).

In addition, the GO-term 'cysteine-type peptidase activity' seem to be specifically enriched (Supplementary Table E-20). Cysteine proteases are an important class of enzymes implicated in both developmental and defense-related programmed cell death and other biological processes in plants[6]. The Arabidopsis papain-like cysteine protease CEP1 for instance is involved in tapetal programmed cell death and pollen development[475]. In addition, in Arabidopsis, cysteine proteases such as aleurain-like proteases, cathepsin B-like proteases, and vacuolar processing enzymes are correlated with the remobilization of seed storage proteins during seed germination[476]. Among A. shenzhenica cysteine-type peptidase genes, significant expression of Ash003370 could be detected in the flower bud, stem, leaf, and seed (Supplementary Figure E-30). Gene Ash003370 thus likely plays important roles in the reproductive and vegetative development of A. shenzhenica, but further studies are necessary to unravel its precise functional role.

E.1-3 Whole-genome duplication

We believe the large number of anchor points mapped on the Apostasioideae stem branch to be due to phylogenetic discordance as a result of the probably very short time interval between the shared WGD event and the divergence of Orchidaceae. Such phylogenetic discordance could theoretically be due to incomplete lineage sorting (or other difficulties in tree inference when short branches are involved due to short time intervals between successive phylogenetic events, or, alternatively, due to homeologous recombination). A possible scenario for a specific gene (tree) is illustrated in Supplementary Figure E-31. A gene in the common ancestor of all extant orchids was polymorphic before the shared orchid-specific WGD, *i.e.*, it had two

diverging alleles, G and G'. Then both alleles got duplicated during the WGD, resulting in two paralogous loci with a total of four alleles that were retained in the early orchid polyploid. If the ancestral orchid speciation event followed relatively soon after the WGD event, all four alleles may have been retained in the two diverging orchid species, which gave rise to the ancestors of the current Apostasioideae (which includes *A. shenzhenica*) and the ancestors of the rest of the orchid families including the lineage eventually leading to Epidendroideae (which includes *P. equestris* and *D. catenatum*), respectively. Over time, if both paralogous loci are retained in the current orchids and if we assume, for simplicity, that alleles eventually became fixed at the paralogous loci, *i.e.*, two of the four alleles got lost owing to genetic drift or selection, then only two of the alleles were retained at the paralogous loci forming an anchor pair. As a result, the coalescence points of such anchor pairs in the current species depend on which of the alleles were finally fixed at the paralogous loci in a species. There are in total 36 possible combinations of any two fixed alleles from each species (in a species/gene tree with two species), but only three possible coalescence points on branches in the species-level phylogenetic tree, *i.e.*, the stem branch of Orchidaceae or one of the two early-diverging branches of Orchidaceae after the ancestral speciation event. We elaborate on four examples in Supplementary Figure E-31. If all the retained anchors are all from one ancestral allele (Supplementary Figure E-31A, all from allele G), the coalescence points of the two anchor pairs are both the duplication (*i.e.*, WGD) event and they fall onto the orchid stem branch. If one ancestral species kept anchors from different alleles (Supplementary Figure E-31B and C, species A and E, respectively, one from allele G and the other from allele G') and the other ancestral species kept anchors from the same allele (Supplementary Figure E-31B and C, species E and A, respectively, all from allele G), then the coalescence point of one of the anchor pairs is the duplication event (e.g., of E_1 and E_2 in Supplementary Figure E-31B) and the coalescence point of the other anchor pair is actually the divergence event of the two ancestral alleles (e.g., of A_1 and A'_2 in Supplementary Figure E-31B). Therefore, both coalescence points map onto the orchid stem branch but seemingly indicate different phylogenetic events. In the fourth example, if one ancestral species retained both anchors from one allele (Supplementary Figure E-31D, species A, both from allele G') while the other ancestral species retained both anchors from the other allele (Supplementary Figure E-31D, species E, both from allele G), then although the two coalescence points reflect the duplications of each of the two alleles (during shared orchid WGD event) they actually erroneously map onto the two orchid subbranches.

To substantiate our hypothesis on discordance in anchor-pair mapping due to incomplete lineage sorting, we additionally built gene families as described in Materials and Methods ('Identification of WGD events in *A. shenzhenica* and phylogenomic analyses'), but now only using the three orchid genomes, plus *Asparagus officinalis* and *Amborella trichopoda*. Phylogenetic trees of 2,573 gene families were reconstructed and rooted for gene families that had at least five genes,

with at least one gene from *A. trichopoda*, *A. officinalis*, and *A. shenzhenica* plus at least one gene from either *P. equestris* or *D. catenatum* and at least one duplicate pair in at least one of the three orchid species. For computational reasons, gene families with more than 200 genes were removed from further analysis. We traversed this full set of the 2,573 rooted gene trees to explore tree topologies related to duplication events in orchids based on all paralogues and not just the limited set of anchor pairs (Supplementary Figure E-32).

We first extracted subtree(s) from each node for which the node itself included genes only from the orchids and the sister clade had genes from *A. officinalis* and/or *A. trichopoda*. Subtrees that did not have genes from at least *A. shenzhenica* and one other orchid plus a pair of paralogues were removed. Based on these criteria, we obtained 2,085 subtrees from 1,864 rooted gene trees. The subtrees were explored to look for expected topologies supporting the corresponding topologies under the four different scenarios described in Supplementary Figure E-31. The expected gene tree topologies, including those with gene loss and/or incomplete sampling taken into account (see below), are illustrated in Supplementary Figure E-31E. The bootstrap values on the branches leading to nodes that were used to distinguish different topologies, *i.e.*, whether the nodes represented speciation or duplication events, were used to evaluate the support for such events with a cutoff of greater than or equal to 50%. We found 167, 39, 45, and 72 gene trees that showed subtrees with the expected topologies under the respective four scenarios as depicted in Supplementary Figures E-31A–D. The first three topologies thus support a (whole-genome) duplication event that occurred before the divergence of Orchidaceae, while the gene trees of the last topology could erroneously support two duplication events, each occurring on one of the two subbranches of Orchidaceae (Supplementary Figure E-31E). Due to extensive gene loss (the most common fate for duplicated genes) and/or incomplete sampling, the resulting gene trees may have a second set of four prevalent topologies consisting of only three retained genes (*i.e.*, one lost or non-sampled gene). Two of these topologies support a coalescence point on the stem branch of Orchidaceae (we found 548 and 296 such gene trees), one topology could erroneously support a coalescence point on the subbranch of Orchidaceae leading to Epidendroideae (we found 429 such gene trees), and the fourth topology could erroneously support a coalescence point on the stem branch of Apostasioideae (we found 267 such gene trees) (Supplementary Figure E-31E). Considering all these scenarios together, we found that there were fewer gene (sub)trees with duplicated genes from *A. shenzhenica* than gene (sub)trees with duplicated genes from Epidendroideae, indicating massive gene loss in *A. shenzhenica*. In total, 1,450 gene families with 1,620 subtrees were considered among the 1,864 gene families with 2,085 subtrees.

E.1-4 MIKC-type genes and the evolution of the pollinium*

MIKC*-type genes are one of the major regulators of the male gametophytic developmental programme. These genes were shown to have a conserved function in both *Arabidopsis* and rice pollen development[383]. MIKC*-type genes are present in all major groups of land plants, including bryophytes, lycophytes, ferns, gymnosperms, and angiosperms[384]. In seed plants, two different monophyletic groups of MIKC*-type genes could be identified, the P- and S-subclades[384]. The former, however, is absent in all orchids except *A. shenzhenica* (Supplementary Figure E-16).

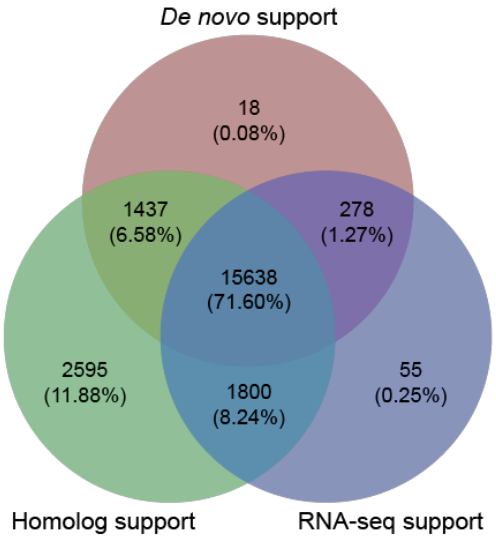
In rice, there are two S-subclade MIKC* genes, *MADS62* and *MADS63*, and one P-subclade *MADS68* gene. All three rice genes are specifically expressed late in pollen development[383], which resembles the expression of *AGL66*, *AGL104*, and *AGL30* in *Arabidopsis*[477]. This suggests that a pollen-specific expression pattern was already established in the most recent common ancestor of monocots and eudicots. The single knockdown or knockout lines, respectively, of the S-subclade *MADS62* and *MADS63* in rice did not show a mutant phenotype, but lines in which both S-subclade genes were affected showed severe defects in pollen maturation and germination[383]. This indicates that the two S-subclade genes of *MADS62* and *MADS63* act redundantly in pollen development, as do their homologs *AGL66* and *AGL104* in *Arabidopsis*[478,479].

In *Arabidopsis* no complete knockout or a knockdown of P-subclade genes were achieved. The down-regulation of the sole rice P-subclade gene *MADS68* in RNAi transgenic lines resulted in defects in pollen maturation and germination[479]. Taken together, both S- and P-subclade MIKC*-type genes confer an indispensable and highly conserved function in pollen maturation and germination in the monocot rice as well as in the eudicot *Arabidopsis*.

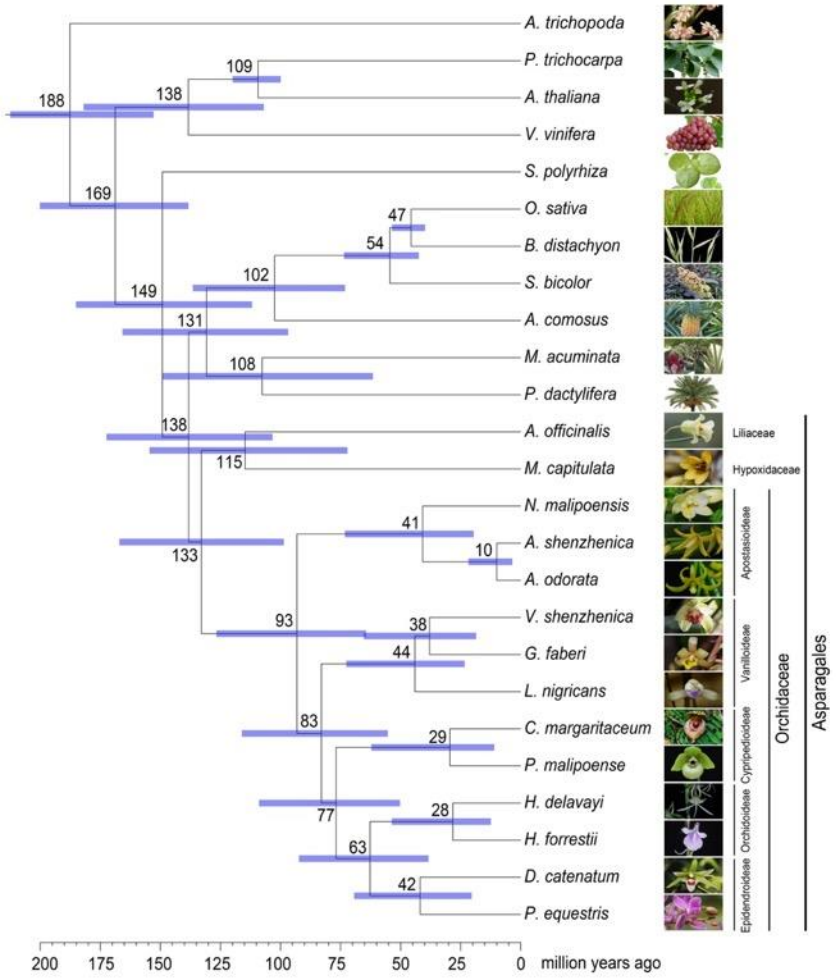
Therefore, it is possible that the loss of the P-subclade members of MIKC*-type genes is related to the evolution of the pollinium. During pollen development, a pollen mother cell undergoes meiosis to produce four haploid microspores, which are first packaged in common callose. Subsequently, due to the decomposition of the tetrad callose wall, the microspores are separated and form the pollen grains[480]. In orchids, with the exception of *Apostasia* and *Neuwiedia*, the pollen callose wall is not decomposed, leading to the formation of a tetrad of pollen grains as a unit, and these 'sticky balls' together form the pollinium[373,480]. Because the P-subclade genes have been lost in most orchids but not in *A. shenzhenica* and because their presence seems to correlate with the presence of a pollinium, we propose that the P-subclade genes provide the ability to decompose the callose wall and that their loss leads to the production of pollen that aggregate into pollinia (Figure 4-9A and C), which however will need to be confirmed with experimental data. The formation of pollinia has been an important transition during the evolution of orchids. This transition, combined with the evolution and formation of the flower lip and gynostemium, has proven very

useful for efficient insect pollination and likely lead to the emergence of reproductive barriers[194].

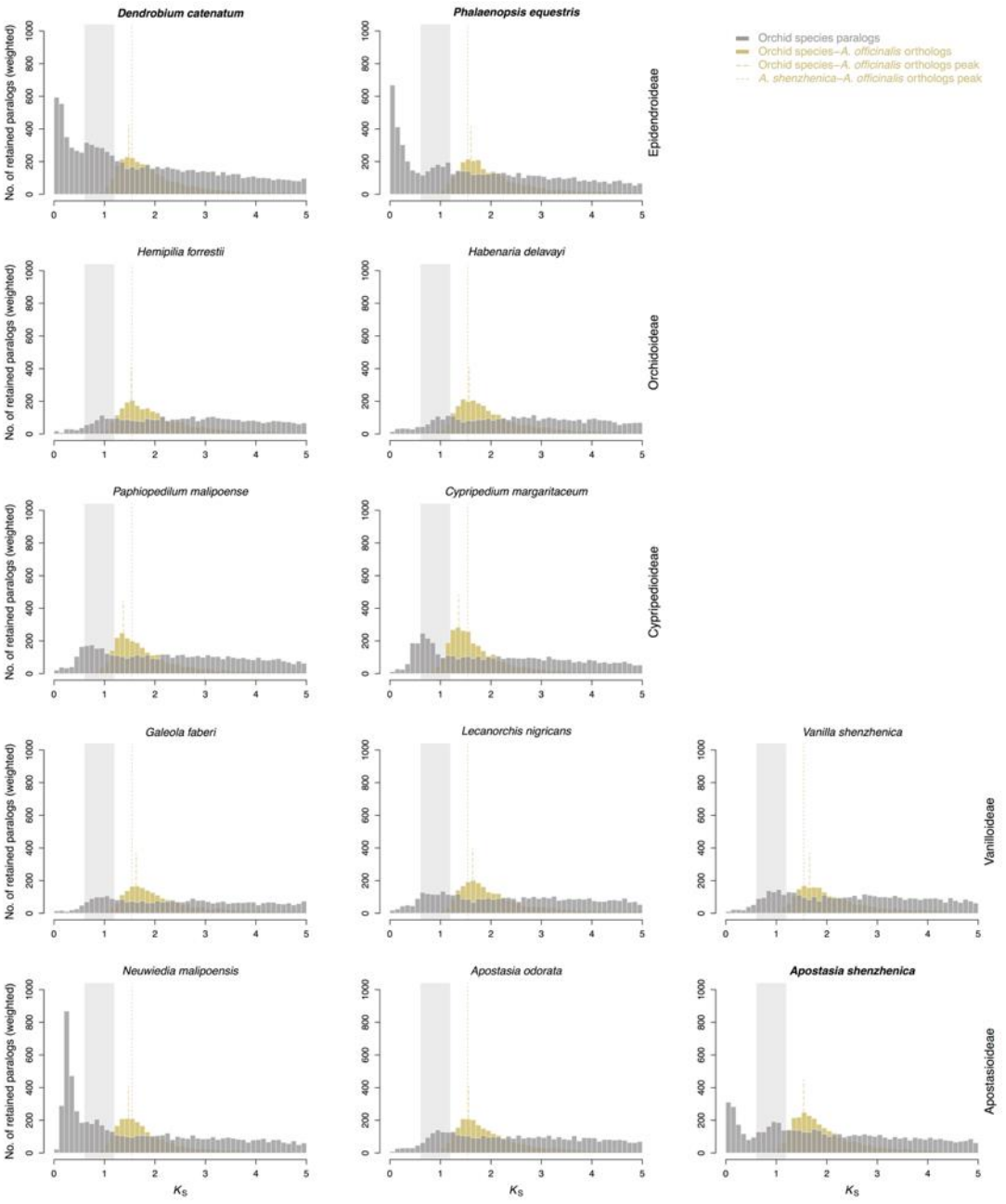
E.2. Supplementary Figures



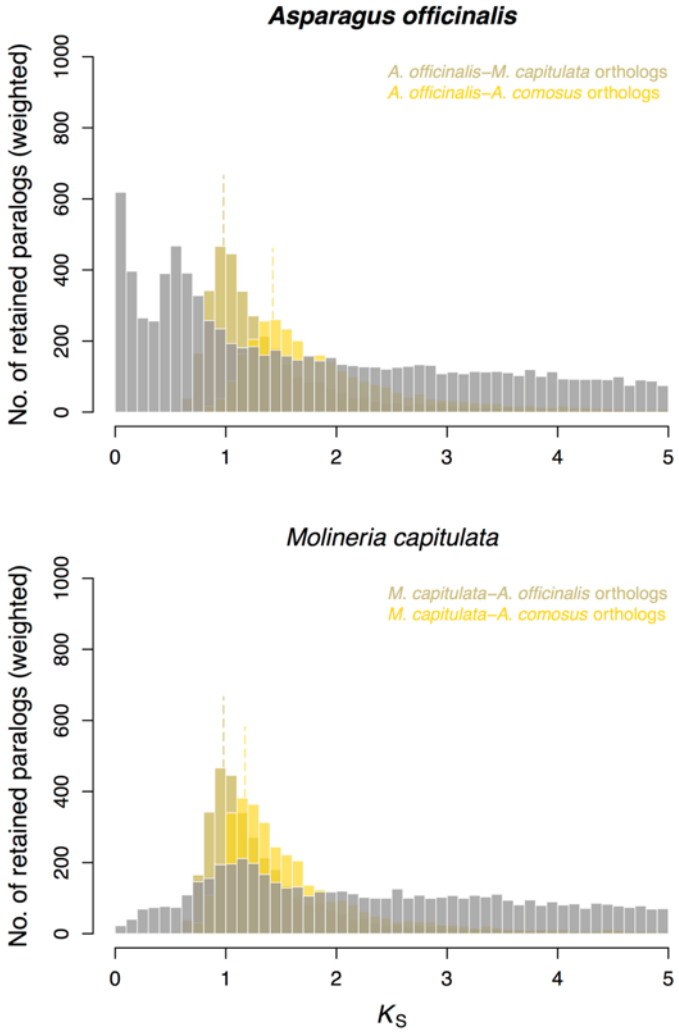
Supplementary Figure E-1. Evidence for gene annotation of *A. shenzhenica*. The respective support of final gene sets by the three methods (*De novo* prediction, Homology searching and RNA-seq mapping) are shown.



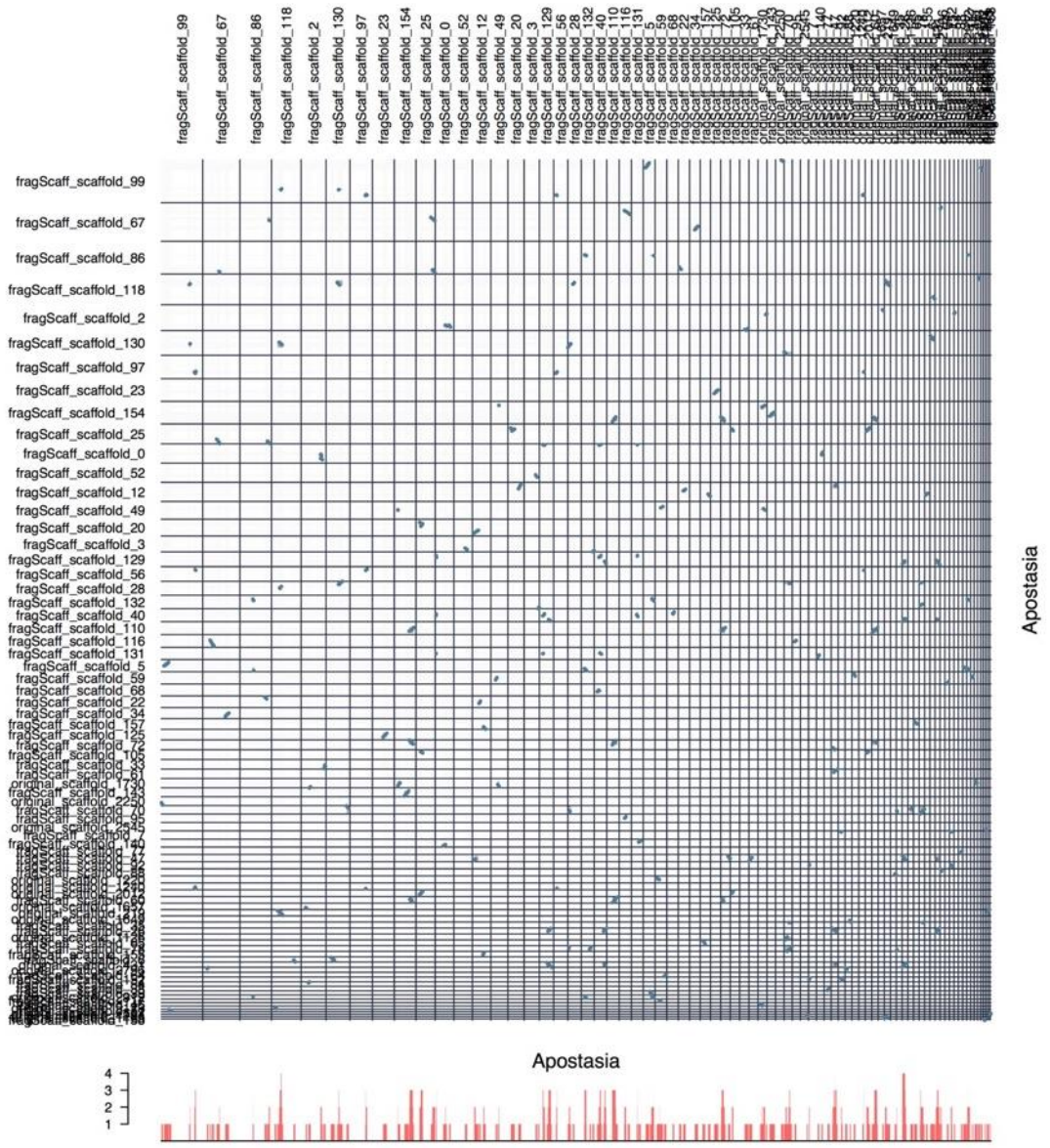
Supplementary Figure E-2. Phylogenetic tree showing the topology and divergence times for 15 genomes (*A. trichopoda*, *P. trichocarpa*, *A. thaliana*, *V. vinifera*, *S. polyrhiza*, *O. sativa*, *B. distachyon*, *S. bicolor*, *A. comosus*, *M. acuminata*, *P. dactylifera*, *A. officinalis*, *A. shenzhenica*, *P. equestris* and *D. catenatum*) and 10 transcriptomes (*A. odorata, C. margaritaceum, G. faberi, H. delavayi, H. forrestii, L. nigricans, M. capitulata, N. malipoensis, P. malipoense, V. shenzhenica*).The unigenes of the transcriptomes of the 10 'transcriptome' species were aligned to the 439 single-copy gene families of the 15 'genome' species. Hundred thirty-two single-copy gene families for the 25 species could thus be identified, which were used to construct a phylogenetic tree based on the PhyML software[99] with the GTR+Γ model, while divergence times (indicated by light blue bars at the internodes) were predicted by MCMCTREE[345]. The range of the bars indicates the 95% confidence interval of the divergence times.



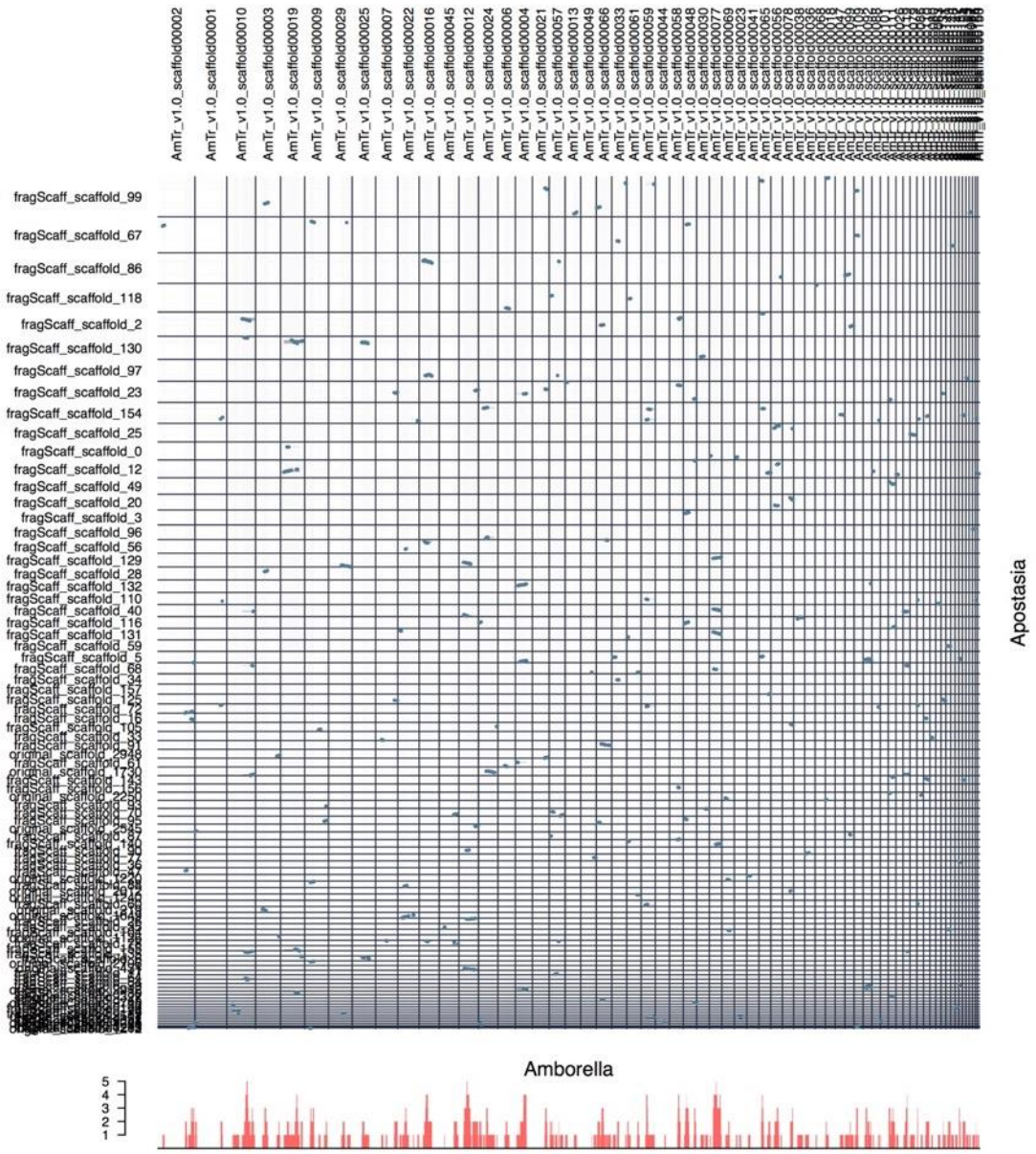
Supplementary Figure E-3. Distribution of synonymous substitutions per synonymous site (K_S) of the whole paranome for three orchid genomes and nine orchid transcriptomes. K_S distributions of paralogues are shown in gray. The light gray rectangle in the background of each plot highlights the K_S range from 0.6–1.2 in which putative WGD peaks can be identified for all 12 orchid species shown. K_S distributions of one-to-one orthologues between each orchid species and *A. officinalis*, representing their time of divergence, are shown in yellow, with long-dashed lines indicating the peak (based on KDE) of the distributions. The K_S value of the peak of the *A. shenzhenica–A. officinalis* one-to-one orthologue distribution is shown for comparison as a dashed line in each plot, indicating potential differences in substitution rates compared to *A. shenzhenica*. The three orchids with genomes are in bold.



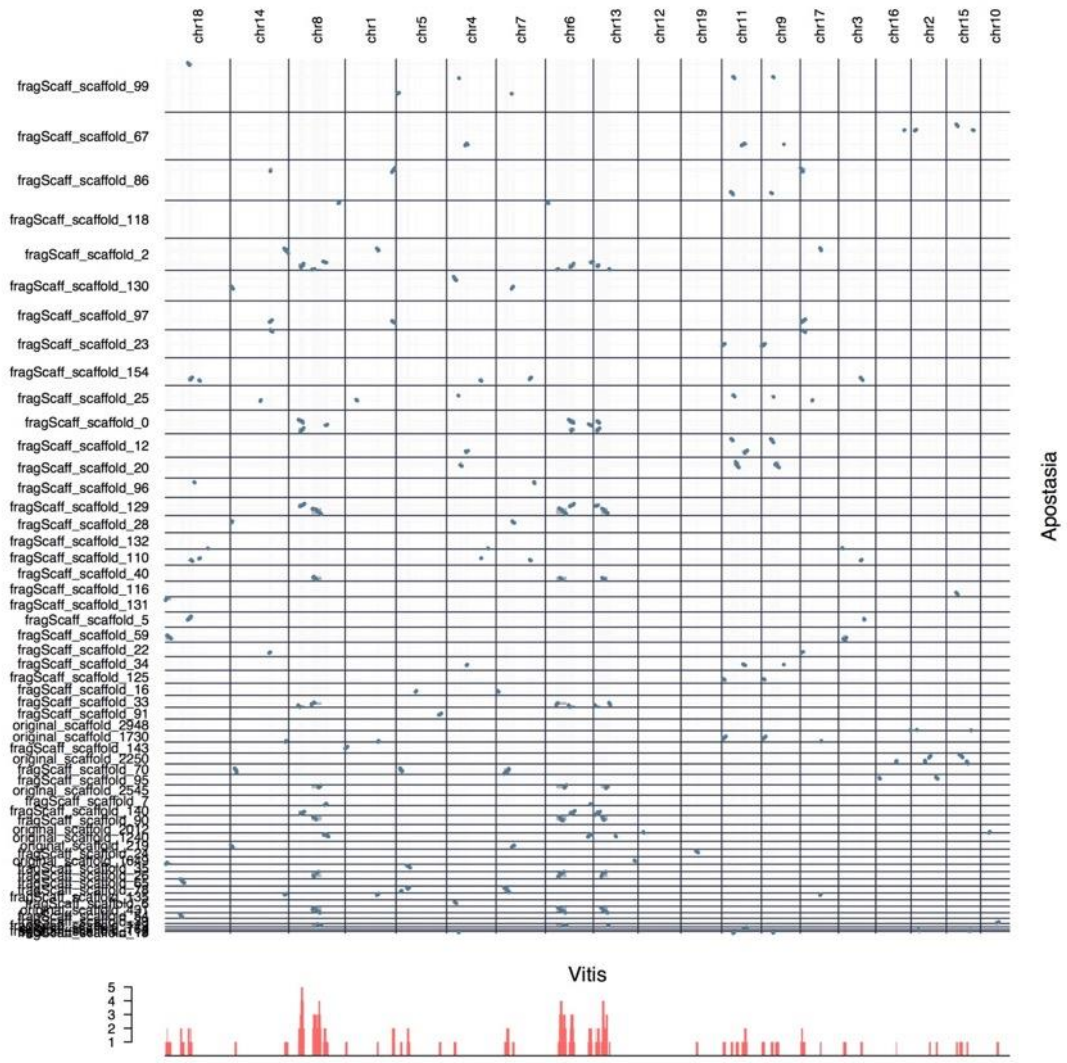
Supplementary Figure E-4. Distribution of synonymous substitutions per synonymous site (K_S) of the whole paranome for the *A. officinalis* genome and the *M. capitulata* transcriptome. K_S distributions of paralogues are shown in gray. K_S distributions of one-to-one orthologues between *A. officinalis* and *M. capitulata*, and between each of these and *A. comosus*, representing their respective times of divergence, are shown in light brown and bright yellow, respectively, with long-dashed lines indicating the peaks (based on KDE) of the distributions. The putative WGD evident from the K_S peak in *A. officinalis* does not seem to be shared with *M. capitulata*. The K_S peak in *M. capitulata* likely represents a signature of the ancient monocot τ WGD event[98,185], shared with *A. comosus* and all other commelinids as well as Orchidaceae.



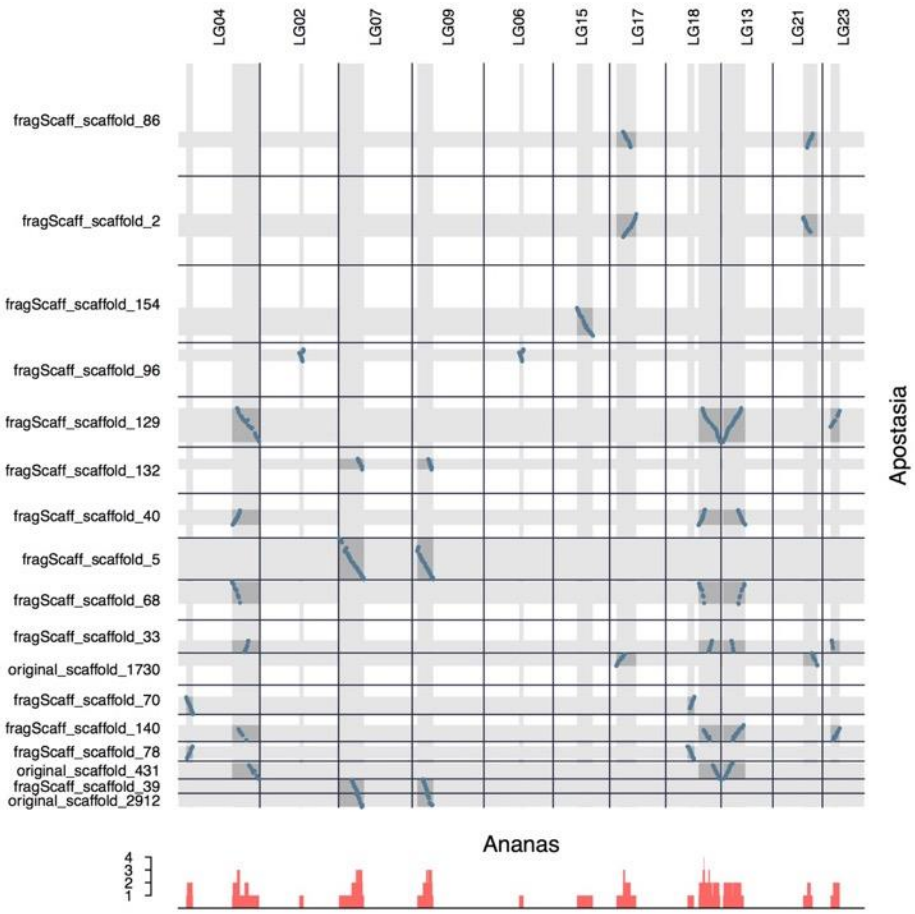
Supplementary Figure E-5. 'Synteny' dot plot of the self-comparison of *A. shenzhenica*. Only co-linear segments with at least 5 anchor pairs are shown. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Materials and Methods).



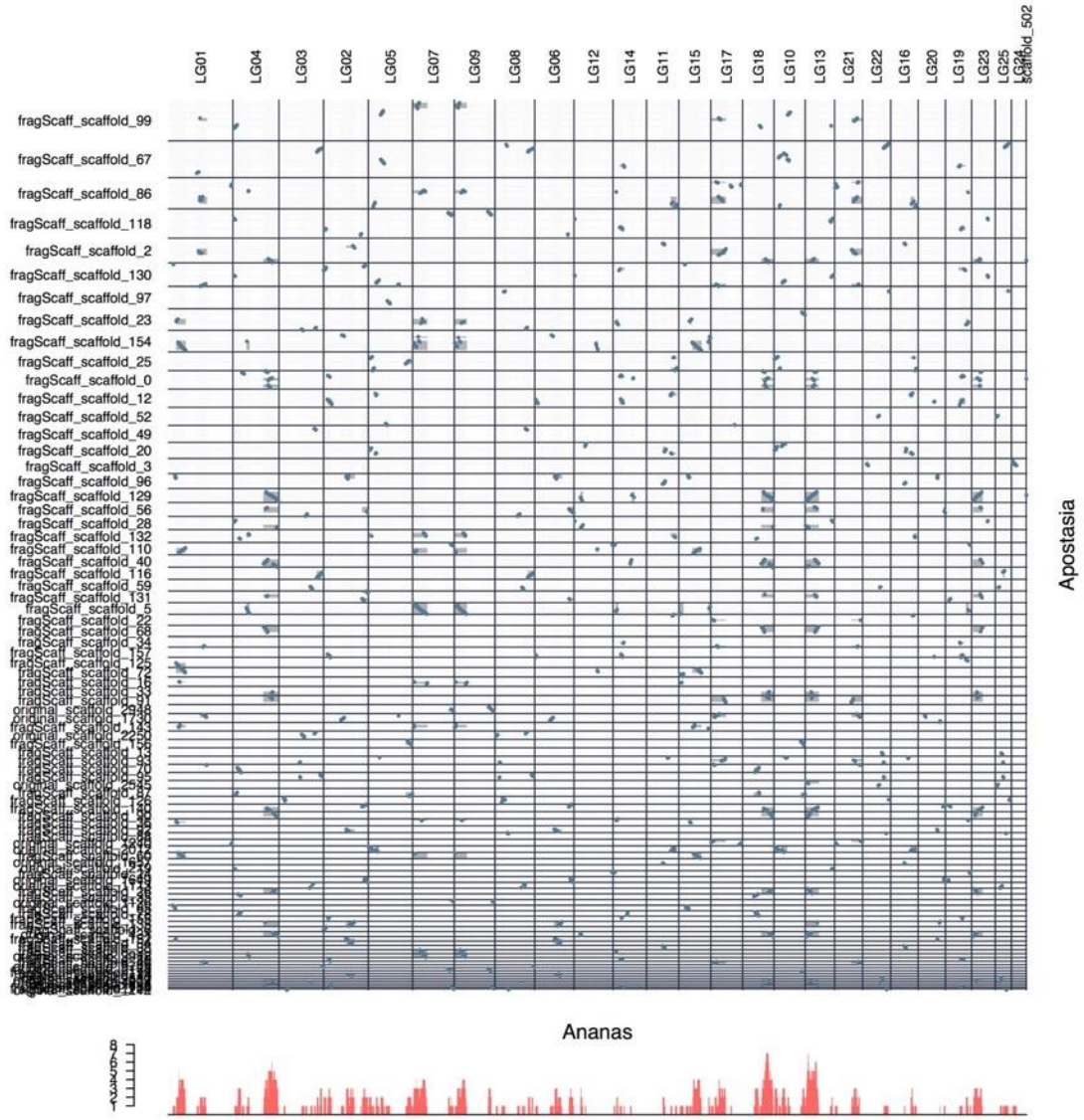
Supplementary Figure E-6. 'Synteny' dot plot of the comparison between *A. shenzhenica* and *A. trichopoda*. Only co-linear segments with at least 5 anchor pairs are shown. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Materials and Methods).



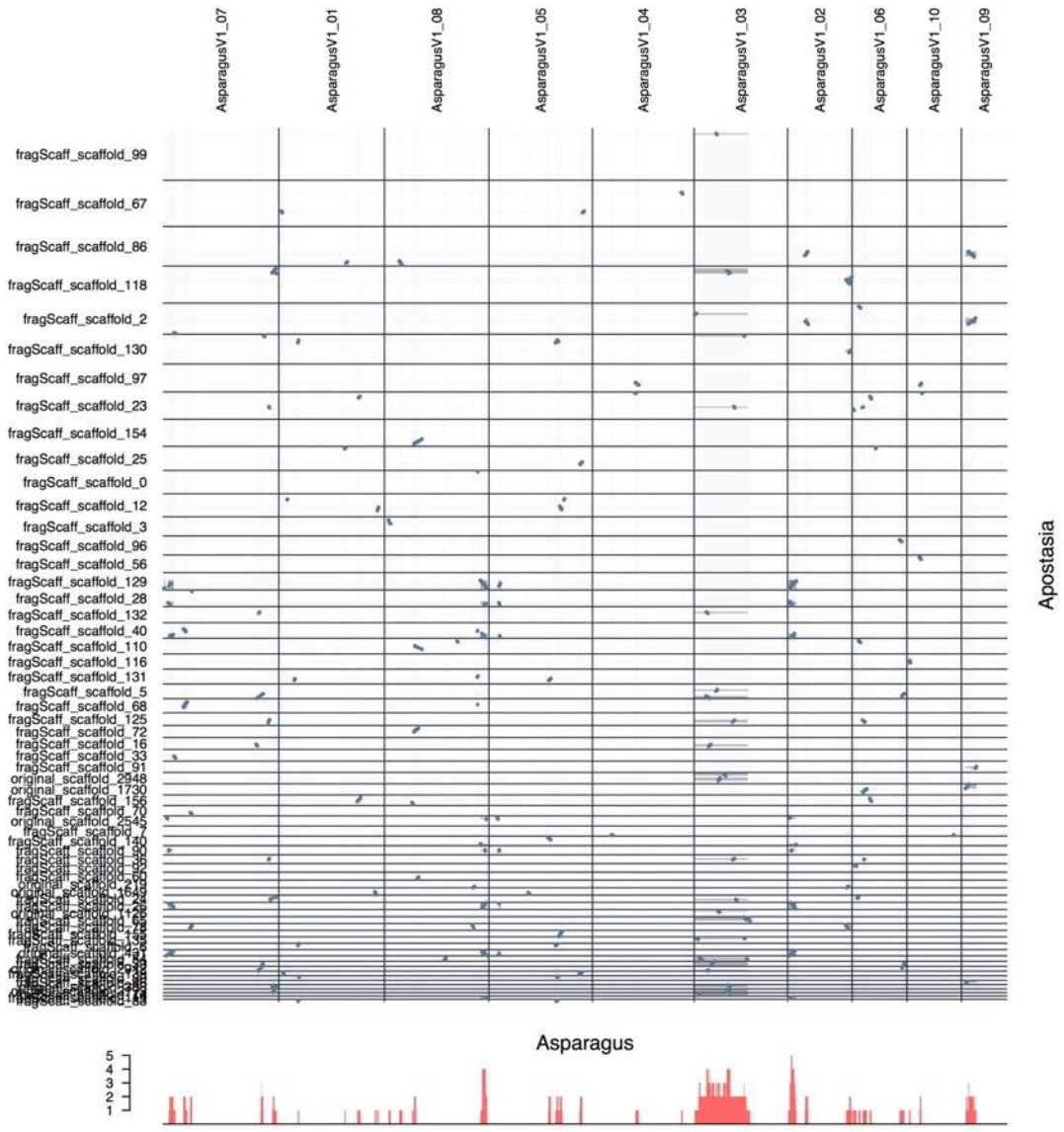
Supplementary Figure E-7. 'Synteny' dot plot of the comparison between *A. shenzhenica* and *V. vinifera*. Only co-linear segments with at least 10 anchor pairs are shown. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Materials and Methods). Only connected co-linear segments with at least 10 anchor pairs were used to calculate the duplication depths.



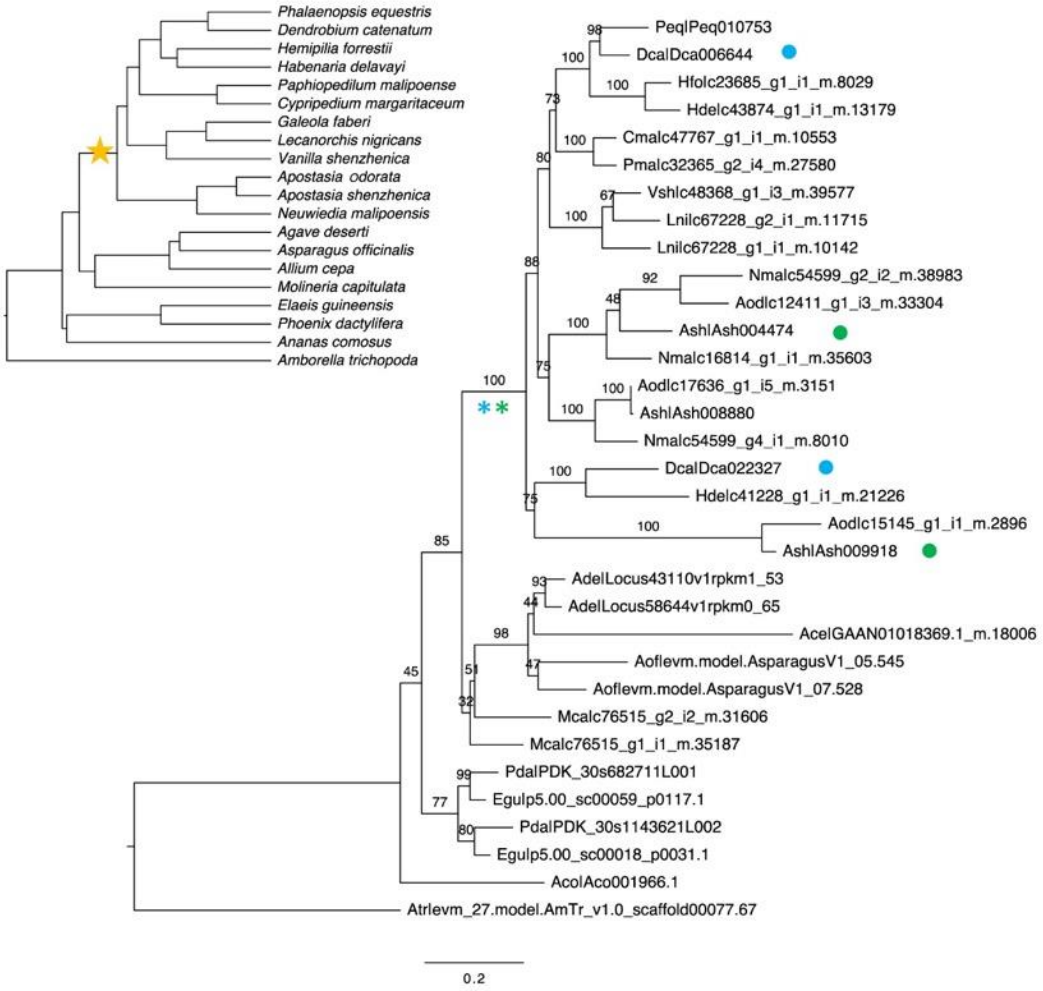
Supplementary Figure E-8. 'Synteny' dot plot of the comparison between *A. shenzhenica* and *A. comosus*. Only co-linear segments with at least 30 anchor pairs are shown. The sections on each scaffold with co-linear segments between are shown in grey. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Materials and Methods). Only connected co-linear segments with at least 15 anchor pairs were used to calculate the duplication depths.



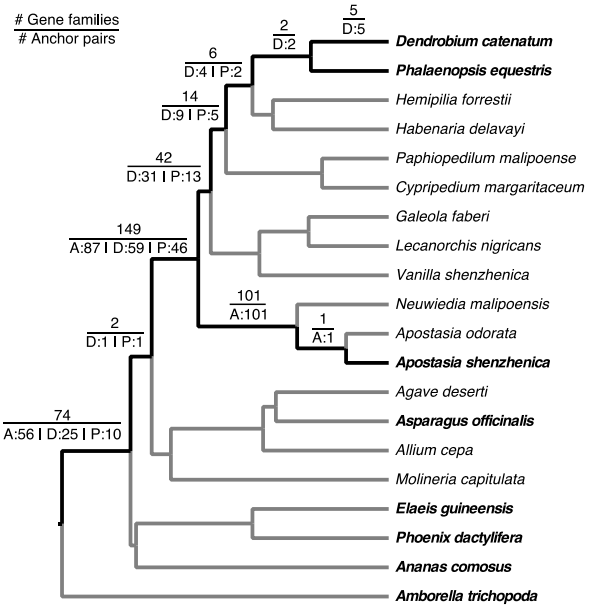
Supplementary Figure E-9. 'Synteny' dot plot of the comparison between *A. shenzhenica* and *A. comosus*. Only co-linear segments with at least 10 anchor pairs are shown. The sections on each scaffold with co-linear segments between are shown in grey. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Materials and Methods). Only connected co-linear segments with at least 10 anchor pairs were used to calculate the duplication depths.



Supplementary Figure E-10. 'Synteny' dot plot of the comparison between *A. shenzhenica* and *A. officinalis*. Only co-linear segments with at least 10 anchor pairs are shown. The sections on each scaffold with co-linear segments between are shown in grey. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Materials and Methods). Only connected co-linear segments with at least 10 anchor pairs were used to calculate the duplication depths.

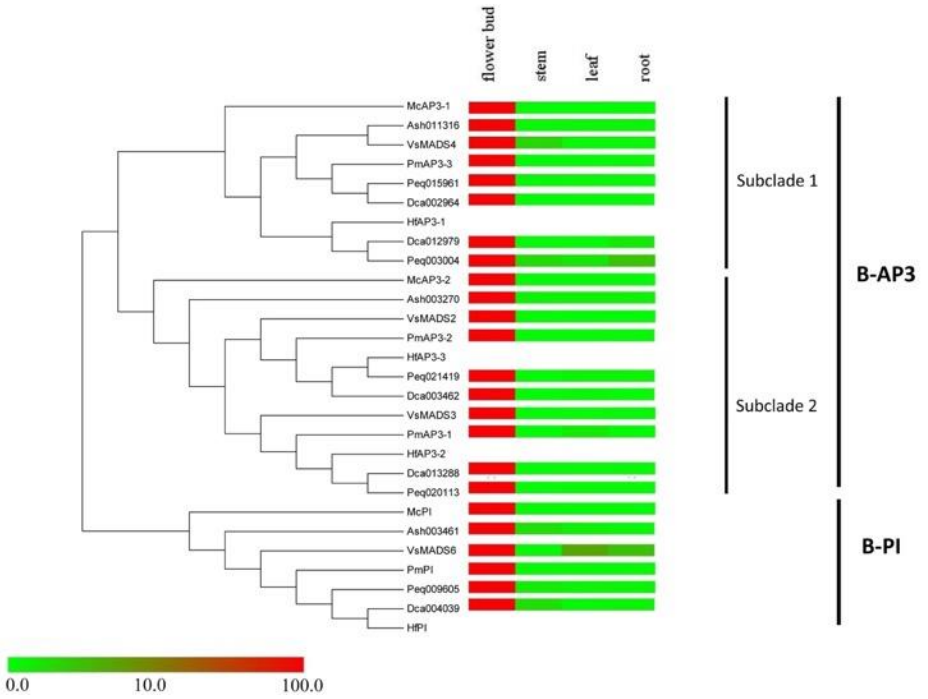


Supplementary Figure E-11. Example of coalescence points of anchor pairs mapped onto the species phylogeny. The blue and green dots in the gene tree (right) denote pairs of duplicated anchor genes from *D. catenatum* and *A. shenzhenica*, respectively. Both anchor pairs coalesce onto the same node (blue and green asterisks), which corresponds to a putative WGD event on the stem branch of the orchids in the species tree (left, yellow stars). Although two anchor pairs were found here, we count the gene family only once as support for a shared WGD event in orchids in Figure 4-8 and Supplementary Figure E-12 because the anchor pairs are from different species and support the same single event. The numbers on the gene tree branches are bootstrap values (%). For gene names, the first three letters denote species: Ace, *A. cepa*; Aco, *A. comosus*; Ade, *A. deserti*; Aod, *A. odorata*; Aof, *A. officinalis*; Ash, *A. shenzhenica*; Atr, *A. trichopoda*; Cma, *C. margaritaceum*; Dca, *D. catenatum*; Egu, *E. guineensis*; Gfa: *G. faberi*; Hde, *H. delavayi*; Hfo, *H. forrestii*; Lni, *L. nigricans*; Mca, *M. capitulata*; Nma, *N. malipoensis*; Pda, *P. dactylifera*; Peq, *P. equestris*; Pma, *P. malipoense*; Vsh, *V. shenzhenica*.

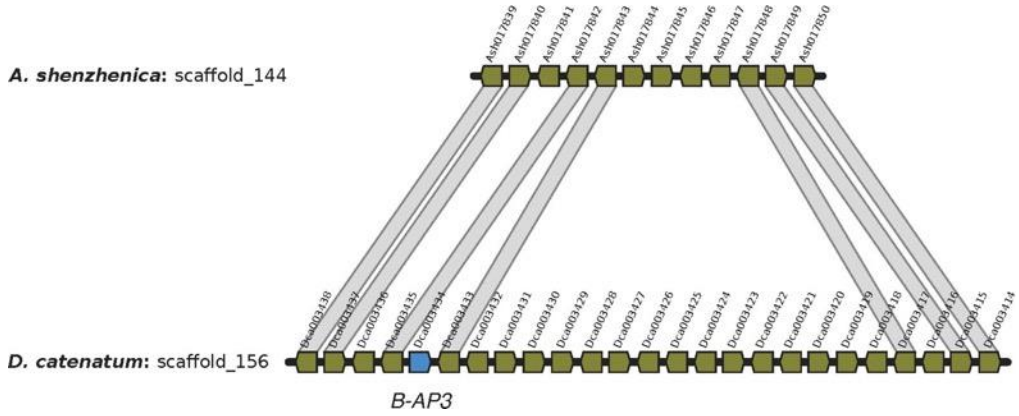


Supplementary Figure E-12. Phylogenomic analysis of orchid WGD events. The numbers on the branches of the species tree indicate the number of gene families that had one or more anchor pairs (duplicated genes found in co-linear regions) from at least one of the three orchids with genomes that coalesced on the respective branch (top), as well as the individual contributions of anchor pairs from the three orchids (bottom; A: *A. shenzhenica*; D: *D. catenatum*; and P: *P. equestris*). All the duplication events have bootstrap values over 50%.

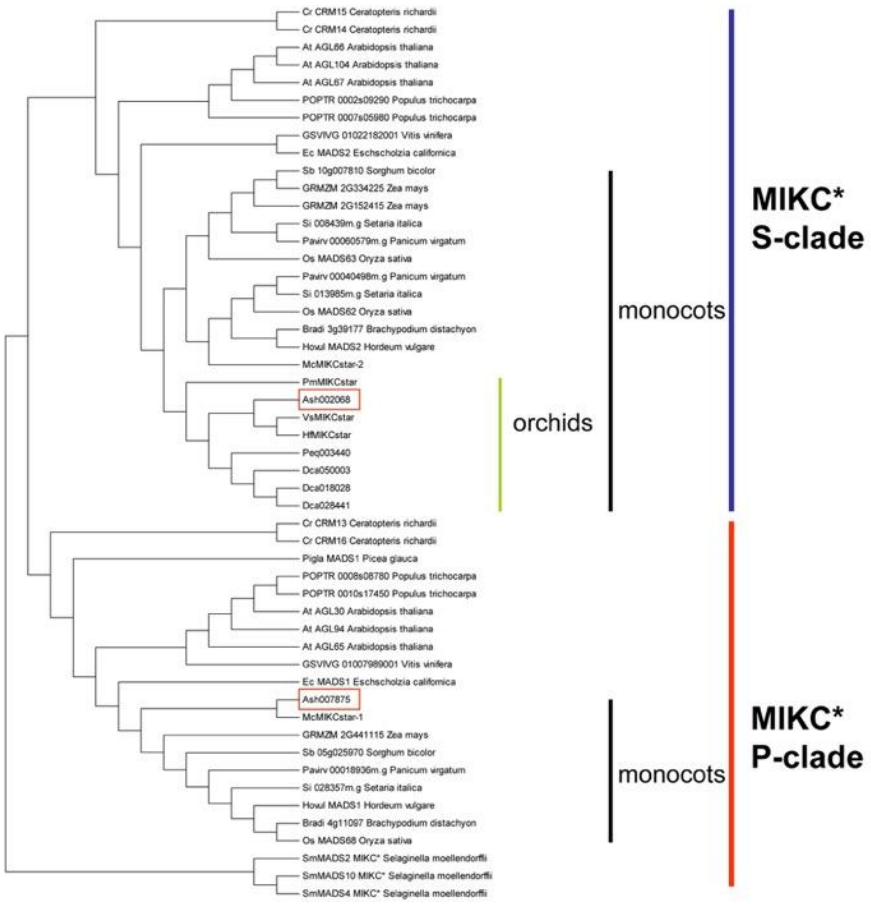
Supplementary Figure E-13. The MADS-box genes involved in orchid morphological evolution. MADS-box genes in *A. shenzhenica*, which contains 36 putative functioning MADS-box genes belonging to 15 subfamilies.



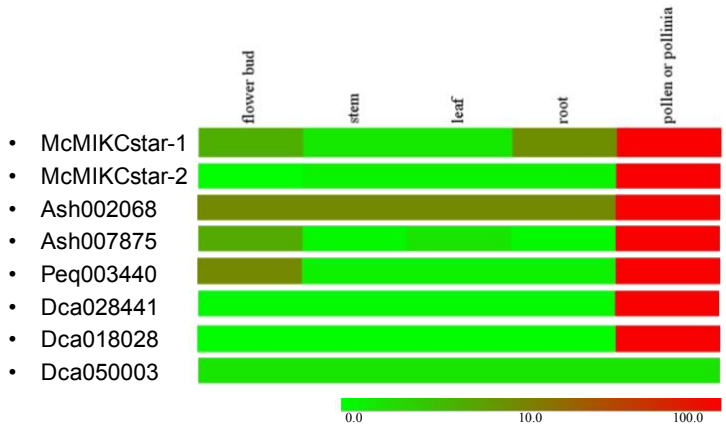
Supplementary Figure E-14. Phylogenetic and expression analysis of orchid B-AP3 genes. Ash: *A. shenzhenica*; Dca: *D. catenatum*; Hf: *H. forrestii*; Mc: *M. capitulata*; Peq: *P. equestris*; Pm: *P. malipoense*; Vs: *V. shenzhenica*. Expressions of B-class genes derived from *H. forrestii* are not shown, because only a flower sample was collected from *H. forrestii*. The expression levels (FPKM value) are represented by the color bar.



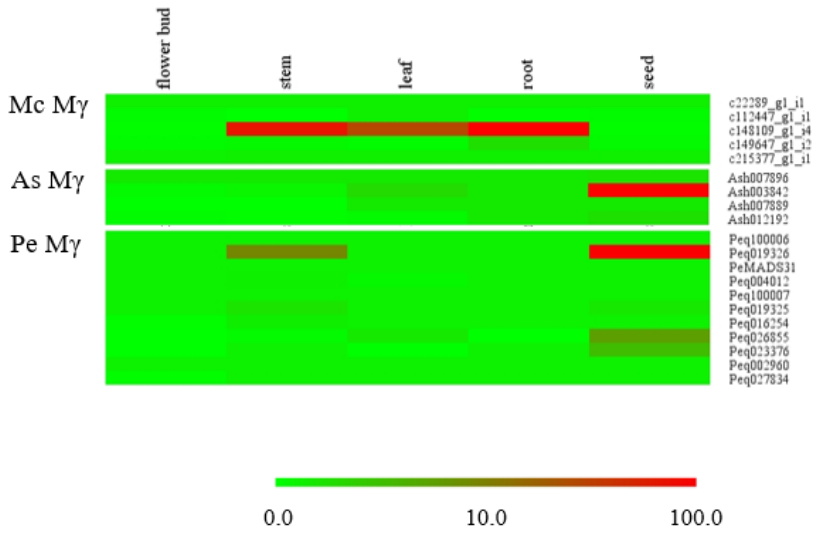
Supplementary Figure E-15. One co-linear segment between *A. shenzhenica* and *D. catenatum* showing a loss of a *B-AP3* gene in *A. shenzhenica*. The co-linearity between the two species was identified by i-ADHoRe (v3.0) using homologous pairs aligned by BLASTP (E value < 1×10^{-5} and c-score $\geqslant$ 0.5). Arrowed blocks indicate genes with orientations, and grey lines connect homologues in the co-linear region. Two genes that are on both sides of one *B-AP3* gene (blue) in *D. catenatum* have two corresponding homologues that remain in the same order in *A. shenzhenica* but without any gene in between, suggesting the B-AP3 gene got lost in *A. shenzhenica*.



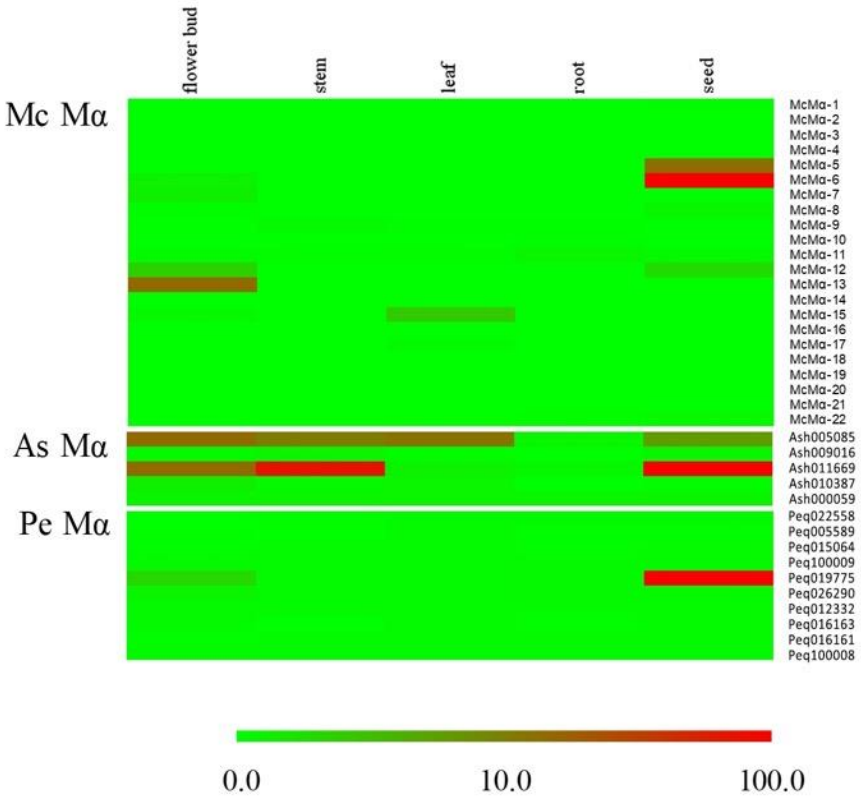
Supplementary Figure E-16. Phylogenetic Tree of MIKC*-Type Genes. The red boxes indicate MADS-box genes from *A. shenzhenica.* Ash: *A. shenzhenica*; Dca: *D. catenatum*; Hf: *H. forrestii*; Mc: *M. capitulata*; Peq: *P. equestris*; Pm: *P. malipoense*; *Vs: V. shenzhenica*. MIKC* sequences of the other species are retrieved from GenBank based on Liu *et al.*[383].



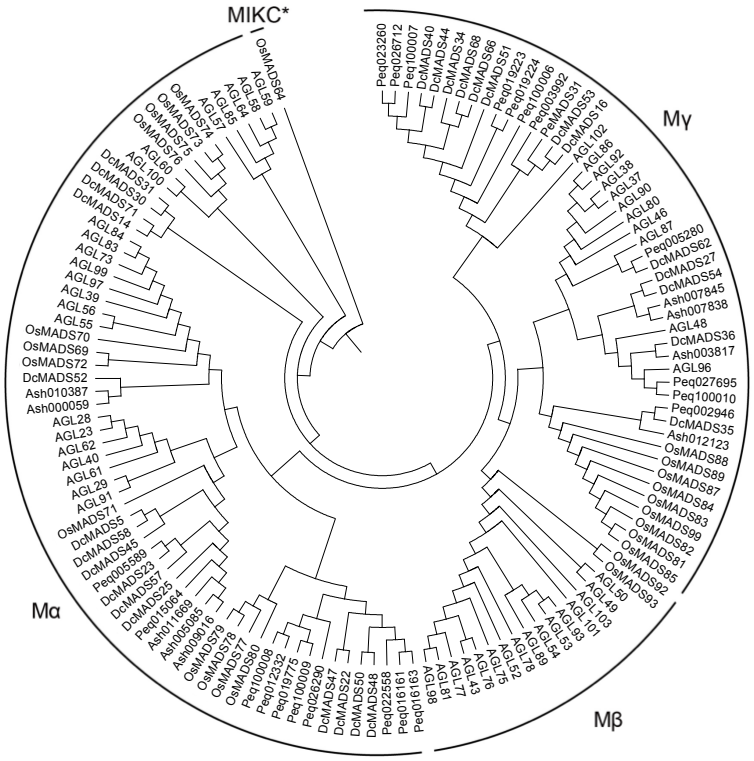
Supplementary E-17. Expression patterns of MIKC* MADS-box genes. Ash: *A. shenzhenica*; Dca: *D. catenatum*; Mc: *M. capitulata*; Peq: *P. equestris*. The expression levels (FPKM value) are represented by the color bar.



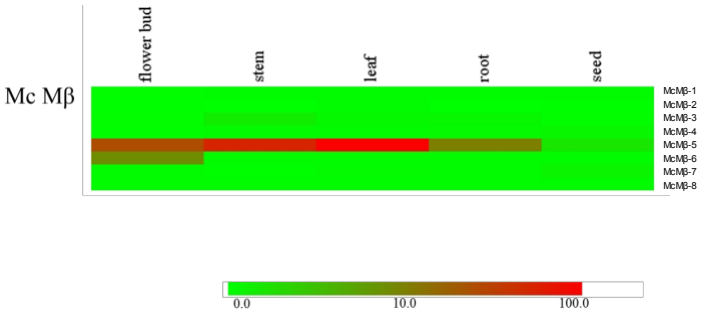
Supplementary Figure E-18. Expression of type I Mγ MADS-box genes in *M. capitulata*, *A. shenzhenica* and *P. equestris*. As: *A. shenzhenica*; Mc: *M. capitulata*; Pe: *P. equestris*. The expression levels (FPKM value) are represented by the color bar.



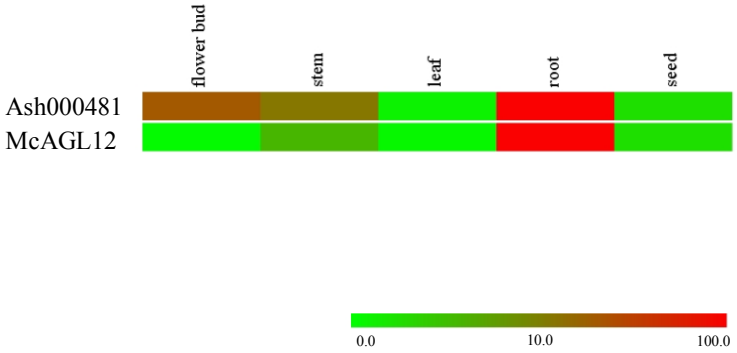
Supplementary Figure E-19. Expression of type I Mα MADS-box genes in *M. capitulata*, *A. shenzhenica* and *P. equestris*. As: *A. shenzhenica*; Mc: *M. capitulata*; Pe: *P. equestris*. The expression levels (FPKM value) are represented by the color bar.



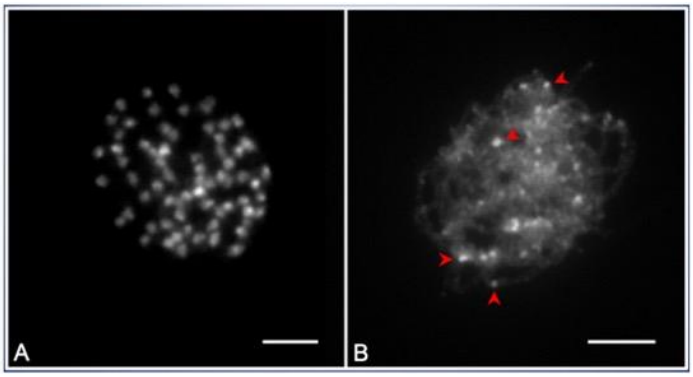
Supplementary Figure E-20. Phylogenetic tree of type I MADS-box genes. Phylogenetic analysis shows that the genomes of *A. shenzhenica*, *P. equestris* and *D. catenatum* do not contain any type I Mβ MADS–box genes (see Mβ subclade). For gene names: Ash, *A. shenzhenica*; Peq, *P. equestris*; Dc, *D. catenatum*; Os, *O. sativa*; AGL, *Agamous-like* genes from *Arabidopsis.*



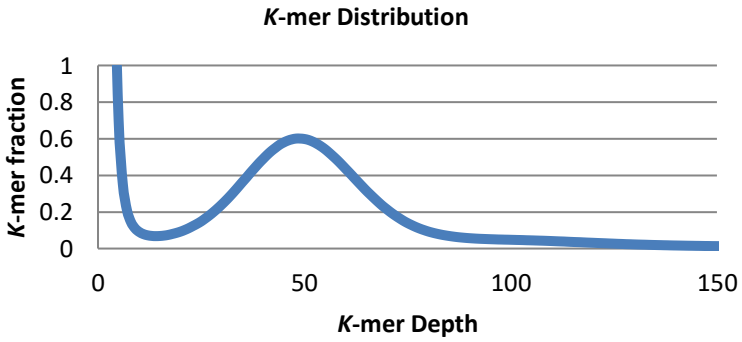
Supplementary Figure E-21. Expression of type I Mβ MADS-box genes in *M. capitulata*. The expression levels (FPKM value) are represented by the color bar.



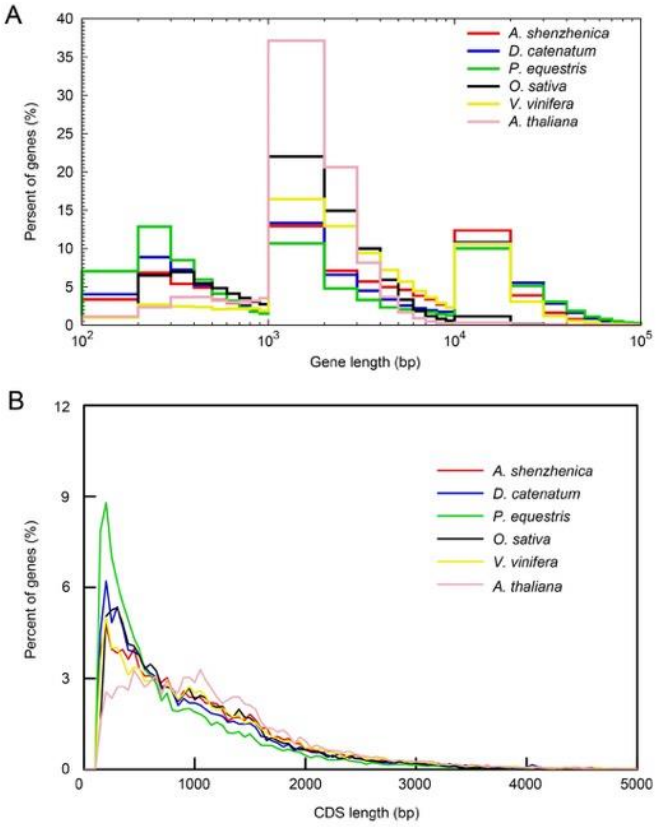
Supplementary Figure E-22. Expression level of *A. shenzhenica* and *M. capitulata AGL12* genes. *Arabidopsis AGL12* is involved in root cell differentiation and in flowering transition[388]. The expression levels (FPKM value) are represented by the color bar.



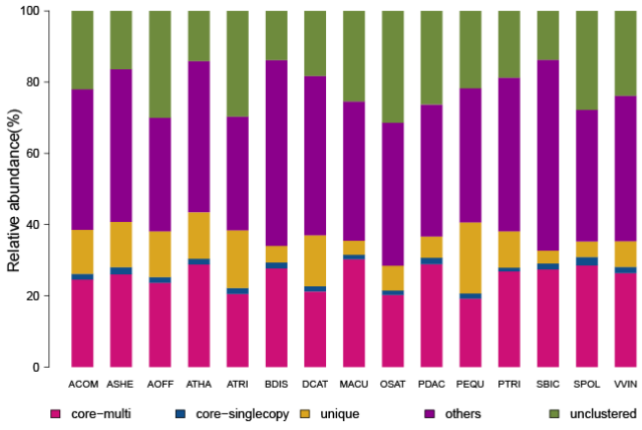
Supplementary Figure E-23. Fluorescent dye (DAPI)-stained chromosomal complements. *A. shenzhenica* presents 68 mitotic metaphase chromosomes in a root cell (A) and shows euchromatin with a few scattered heterochromatin spots (some of these are indicated by the red arrowheads) in a cell at the meiotic pachytene stage (B). The bars represent 5 µm.



Supplementary Figure E-24. *K*-mer distribution of sequencing reads. According to the distribution, we estimate that the genome size of *A. shenzhenica* is approximately 471 Mb. The analysis is based on the Illumina data.



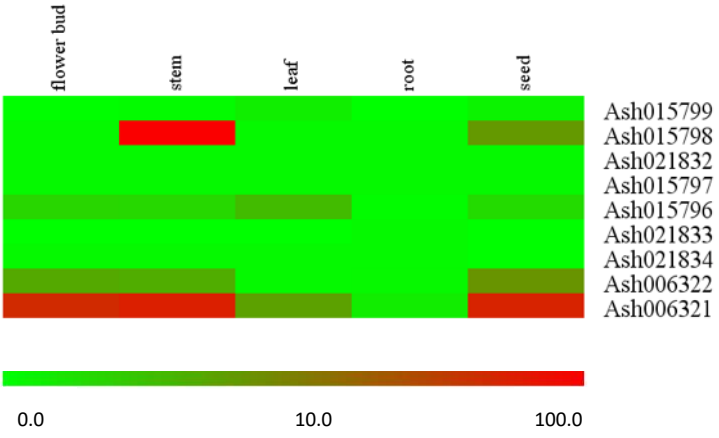
Supplementary Figure E-25. Distribution of gene length (A) and CDS length (B) for six plants.



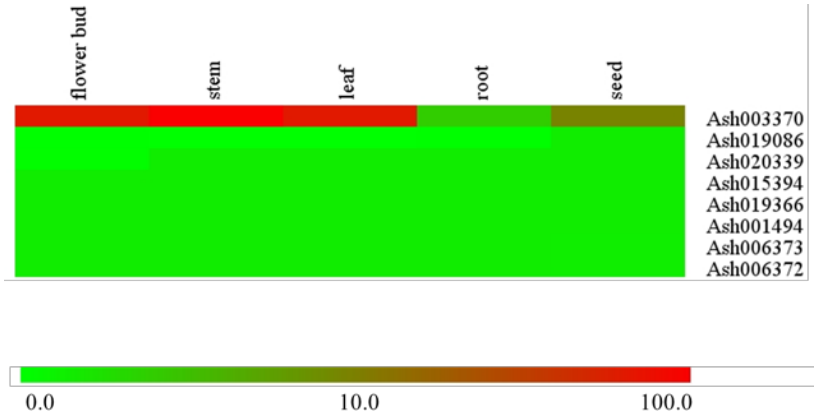
Supplementary Figure E-26. Orthologous genes found in different plant species. **Core-multi**: genes have orthologues in all other species and might have paralogues in species within one family. **Core-single copy**: genes have orthologues in all other species and no other paralogues in this species within one family. **Unique**: genes for which only one family contains genes of this species. **Other orthologues**: genes are not included in the other mentioned categories. **Unclustered genes**: genes that are unclustered into any family. ACOM, *A. comosus*; AOFF, *A. offcinalis*; ASHE, *A. shenzhenica*; ATHA, *A. thaliana*; ATRI, *A. trichopoda*; BDIS, *B. distachyon*; DCAT, *D. catenatum*; MACU, *M. acuminata*; OSAT, *O. sativa*; PDAC, *P. dactylifera*; PEQU, *P. equestris*; PTRI, *P. trichocarpa*; SBIC, *S. bicolor;* SPOL, *S. polyrrhiza;* VVIN*, V. vinifera.*

Supplementary Figure E-27. Expression of A. shenzhenica genes with O-methyltransferase activity. The expression levels (FPKM value) are represented by the color bar.

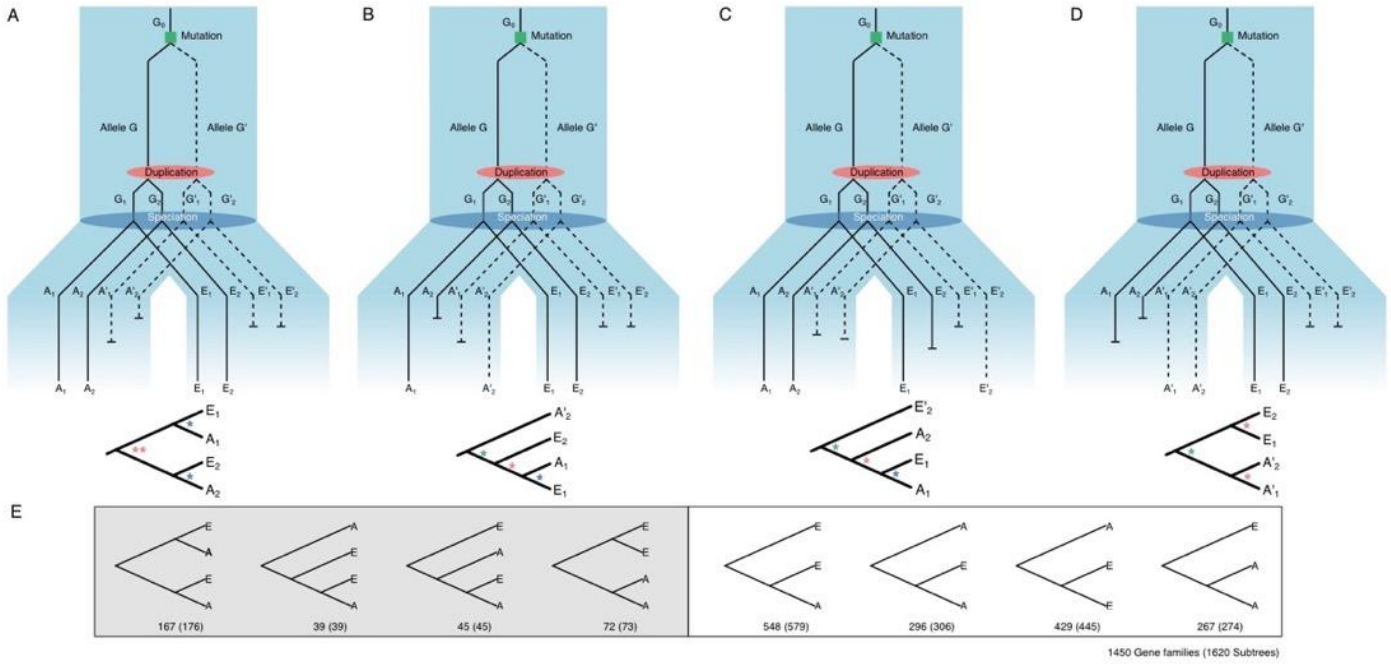
Supplementary Figure E-28. Expression of *A. shenzhenica* genes involved in flavone and flavonol biosynthesis. The expression levels (FPKM value) are represented by the color bar.



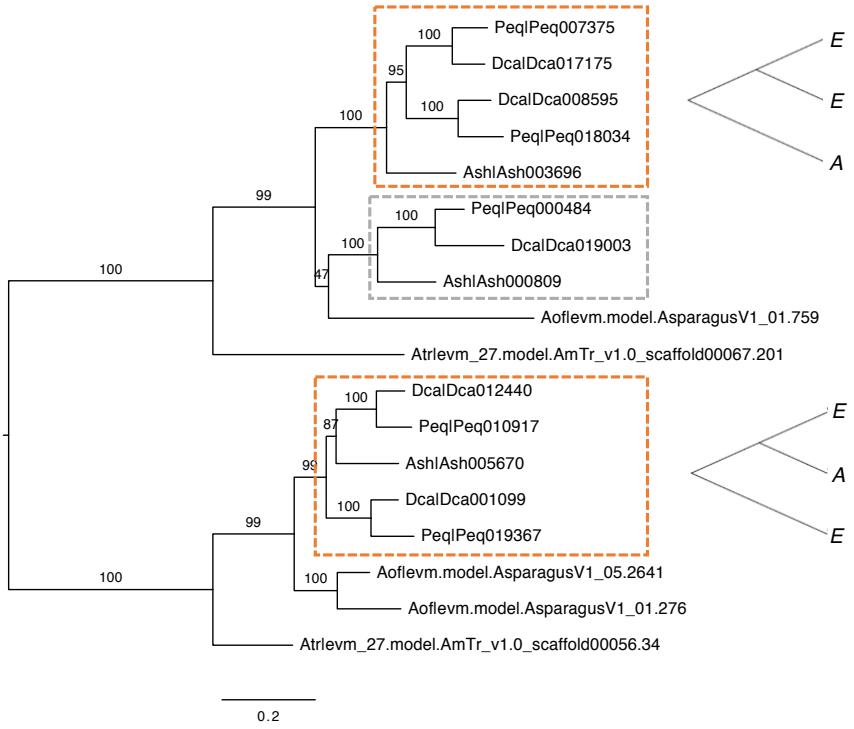
Supplementary Figure E-29. Expression of *A. shenzhenica* genes involved in biosynthesis of stilbenoids, diarylheptanoids, and gingerol. The expression levels (FPKM value) are represented by the color bar.



Supplementary Figure E-30. Expression of *A. shenzhenica* cysteine-type peptidase genes. The expression levels (FPKM value) are represented by the color bar.



Supplementary Figure E-31. Incomplete lineage sorting underlying the gene tree discordance owing to short time intervals between the orchid-specific WGD and speciation. (A)–(D) Four possible scenarios illustrating fixation of different alleles at two paralogous loci resulting from a duplication event that occurred shortly before a speciation event (see Supplementary Note E.1-3 for details). The possible resulting gene trees are shown below for each scenario with asterisks denoting possible phylogenetic events that branch the gene trees. The colors of the asterisks indicate different phylogenetic events, with blue, red and green, for speciation, duplication and mutation/allele divergence, respectively. (E) The gene trees on the left highlighted in grey are topologies that would be expected without gene loss for each scenario above; the gene trees on the right show the expected topologies including loss or incomplete sampling of one gene. The numbers below the topologies show support from phylogenomic analysis including the three orchids, plus *A. officinalis* and *A. trichopoda* based on full gene trees found and contained subtrees (in parentheses), see Supplementary Figure E-32. A, *A. shenzhenica*; E, Epidendroideae (including *D. catenatum* and *P. equestris*).



Supplementary Figure E-32. Example of the analysis of gene (sub)tree topologies. Full gene family tree with three subtrees (clades) highlighted; the highlighted subtrees each only contain genes from the three orchids (dashed rectangles). Only two of the subtrees (orange rectangles) show duplication events; one supports an independent duplication in Epidendroideae (top), and the other supports a shared duplication in orchids (bottom). The numbers on the gene tree branches are bootstrap values (%). For gene names, the first three letters denote species: Ash, *A. shenzhenica*; Atr, *A. trichopoda*; Dca, *D. catenatum*; Aof, *A. officinalis*; Peq, *P. equestris*. In the two simplified topologies, A indicates *A. shenzhenica*; E, Epidendroideae (including *D. catenatum* and *P. equestris*).

E.3. Supplementary Tables

Supplementary Table E-1. Summary of the *A. shenzhenica* genome sequencing data derived from the Illumina technology.

Insert size (bp)	Read length (bp)	Number of reads	Total data (Gb)	Sequence depth (X)
180	90	168,223,606	15.14	34.97
500	100	275,127,442	27.51	63.54
800	90	98,419,616	8.86	20.46
2000	90	91,713,060	8.25	19.06
5000	90	119,967,670	10.80	24.94
10000	90	46,631,366	4.20	9.69
20000	125	58,492,233	5.26	12.23

Supplementary Table E-2. Summary of the 3rd generation sequencing derived from the PacBio RS II.

Species	Number of bases	Number of reads	Mean read length (bp)
A. shenzhenica	5,441,238,461	1,352,628	4,023
D. catenatum	11,060,594,629	1,502,920	7,359
P. equestris	10,539,372,308	1,352,628	7,792

Supplementary Table E-3. Summary of the 10X genomics Linked-Reads sequencing derived from the Illumina technology.

Species	Read length (bp)	Raw paired reads	Raw bases	Filtered paired reads	Filtered bases
A. shenzhenica	150	369,749,121	110,924,736,300	318,763,894	95,629,168,200
D. catenatum	150	415,271,158	124,581,347,400	387,320,106	116,196,031,800
P. equestris	150	411,429,315	123,428,794,500	383,649,995	115,094,998,500

Supplementary Table E-4. Summary of the *A. shenzhenica* genome assembled by Illumina, PacBio and 10X genomics technologies.

	Scaffold		Contig	
	Length (bp)	Number	Length (bp)	Number
max_len	12,424,053		556,054	
N10	10,110,636	4	223,148	112
N20	6,237,011	8	166,933	283
N30	5,003,307	14	130,671	503
N40	3,457,059	22	103,308	780
N50	3,029,156	32	80,069	1,136
N60	2,413,737	45	63,275	1,590
N70	1,972,814	61	47,252	2,184
N80	1,402,703	82	31,086	3,022
N90	765,391	115	15,048	4,473
Total_length	348,734,287		322,901,144	
GC_rate	31.2%		33.7%	

Supplementary Table E-5. Summary of gene annotation of *A. shenzhenica*.

Gene set		Protein-coding gene number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	AUGUSTUS	26,015	8771.89	1128.15	4.68	241.14	2078.02
	GlimmerHMM	36,406	8249.20	701.52	3.39	207.17	3163.09
Homolog (exonerate)	*A. thaliana*	19,532	5099.79	904.39	4.02	225.25	1391.48
	O. sativa	21,804	4905.70	870.30	3.81	228.44	1436.23
	P. equestris	28,179	3899.98	775.81	3.35	231.88	1331.82
	S. bicolor	20,483	4829.59	881.21	3.93	223.95	1345.31
	Z. mays	20,929	4620.41	852.71	3.79	224.79	1348.81
RNA-seq (Cufflinks)		20,202	9588.04	1144.15	4.77	239.67	1471.21
CEGMA		448	11532.37	1225.80	8.46	144.82	1380.78
MAKER		23,181	7866.12	994.09	4.08	243.45	1915.22
Final set		21,841	7103.12	1099.99	4.51	244.07	1436.59

Supplementary Table E-6. The assessment of assembled genomes by BUSCO.

	A. shenzhenica				*P. equestris*				*D. catenatum*			
	Assembly		Gene set		Assembly		Gene set		Assembly		Gene set	
BUSCO	Proteins	%	Proteins	%	Proteins	%	Proteins	%	Proteins	%	Proteins	%
Complete Single-Copy	685	71.65%	575	60.15%	676	70.71%	553	57.85%	679	71.02%	550	57.53%
Complete Duplicated	210	21.97%	304	31.8%	194	20.29%	299	31.28%	205	21.44%	324	33.89%
Fragmented	20	2.09%	38	3.97%	33	3.45%	60	6.28%	29	3.03%	42	4.39%
Missing	41	4.29%	39	4.08%	53	5.54%	44	4.60%	43	4.50%	40	4.18%
Total groups	956	100%	956	100%	956	100.00%	956	100.00%	956	100.00%	956	100.00%

Supplementary Table E-7. Summary of the improved *P. equestris* and *D. catenatum* assemblies.

	P. equestris			*D. catenatum*		
	Original assembly	Add PacBio and 10X genomics	Final assembly	Original assembly	Add PacBio and 10X genomics	Final assembly
Total length (bp)	1,086,208,158	1,151,255,532	1,133,282,102	1,008,546,262	1,123,989,432	1,119,944,395
Longest scaffold (bp)	81,761,211	88,274,276	80,517,012	2,592,627	2,684,897	34,145,153
N50 of scaffold	523	499	192	723	822	213
N50 length of scaffold (bp)	359,115	408,145	1,217,477	391,462	367,633	1,055,340
Total length of contigs (bp)	1,002,400,532	1,090,349,635	1,045,027,212	955,235,028	1,092,146,538	1,060,791,339
N50 of contig	13,281	6,372	6,124	8,479	5,824	5,656
N50 length of contig (bp)	20,555	45,984	45,791	33,094	51,913	51,736
Number of protein-coding genes	29,431	-	29,545	28,910	-	29,257

Note: Final assembly is with both PacBio and 10X genomics Linked-reads Data

Supplementary Table E-8. List of 40 MADS-box genes identified in *A. shenzhenica*.

Gene ID	Name	ORF (bp)	Protein length (aa)	Type	Subfamily	Pseudogene
Ash007825	AsMADS1	1050	349	MIK*		
Ash005085	AsMADS2	714	237	Type I	Mα	
Ash011255	AsMADS3	684	227	MIKC^c	B-AP3	
Ash007845	AsMADS4	1167	388	Type I	Mγ	
Ash003440	AsMADS5	633	210	MIKC^c	B-PI	
Ash005080	AsMADS6	792	263	MIKC^c	Bs	
Ash004262	AsMADS7	822	273	MIKC^c	E	
Ash006974	AsMADS8	720	239	MIKC^c	AGL6	
AsMADS9	AsMADS9	216	71	MIKC^c	AGL6	✓
Ash015784	AsMADS10	738	245	MIKC^c	AGL6	
Ash018282	AsMADS11	657	218	MIKC^c	E	
Ash006741	AsMADS12	738	245	MIKC^c	E	
Ash002380	AsMADS13	447	148	MIKC^c	SQUA	
Ash015738	AsMADS14	630	229	MIKC^c	C/D	
Ash013457	AsMADS15	750	249	MIKC^c	SQUA	
Ash011284	AsMADS16	450	149	MIKC^c	C/D	
Ash002061	AsMADS17	1188	395	MIKC*		
Ash015066	AsMADS18	759	252	MIKC^c	ANR1	
Ash017092	AsMADS19	762	253	MIKC^c	ANR1	
Ash100002	AsMADS20	702	233	MIKC^c	C/D	
Ash002488	AsMADS21	879	292	MIKC^c	C/D	
Ash003251	AsMADS22	681	226	MIKC^c	B-AP3	
Ash007674	AsMADS23	705	234	MIKC^c	SOC	
Ash016527	AsMADS24	699	232	MIKC^c	SVP	
Ash010253	AsMADS25	486	161	MIKC^c	ANR1	
Ash017552	AsMADS26	696	231	MIKC^c	SVP	
Ash100001	AsMADS27	342	113	MIKC^c	ANR1	
Ash000481	AsMADS28	639	212	MIKC^c	AGL12	
Ash002504	AsMADS29	585	194	MIKC^c	OsMADS32	
Ash009016	AsMADS30	732	243	Type I	Mα	
Ash011669	AsMADS31	669	222	Type I	Mα	
Ash010387	AsMADS32	810	269	Type I	Mα	
Ash000059	AsMADS33	540	179	Type I	Mα	
Ash003817	AsMADS34	831	276	Type I	Mγ	
Ash007838	AsMADS35	660	219	Type I	Mγ	
AsMADS36	AsMADS36	204	67	MIKC*		✓
Ash015872	AsMADS37	648	215	MIKC^c	SOC	
AsMADS38	AsMADS38	492	163	MIKC^c	SVP	✓
Ash012123	AsMADS39	756	251	Type I	Mγ	
AsMADS40	AsMADS40	357	118	MIKC^c		✓

Supplementary Table E-9. Summary of repeat annotation of *A. shenzhenica*.

	RepBase TEs		TE Proteins		*De novo*		Combined TEs	
	Length (bp)	% Genome	Length (bp)	% of Genome	Length (bp)	% of Genome	Length (bp)	% of Genome
DNA	3,604,289	1.03	3,106,810	0.89	18,995,301	5.45	22,534,396	6.46
LINE	10,240,458	2.94	9,511,200	2.73	41,316,780	11.85	44,203,442	12.68
SINE	12,411	0.00	0	0.00	149,789	0.04	161,345	0.05
LTR	10,644,451	3.05	15,167,007	4.35	72,767,333	20.87	76,930,066	22.06
Other	5,732	0.00	0	0.00	0	0.00	5,732	0.00
Unknown	38,684	0.01	0	0.00	20,482,533	5.87	20,520,835	5.88
Total	24,555,914	7.04	27,699,462	7.94	137,241,384	39.35	146,653,786	42.05

Supplementary Table E-10. Length distribution of gene elements in twelve sequenced plants.

Species	Protein coding gene number	Average gene length (bp)	Median gene length (bp)	Average CDS length (bp)	Median CDS length (bp)	Average exon per gene	Average exon length (bp)	Median exon length (bp)	Average intron length (bp)	Median intron length (bp)	Average intergenic region length (bp)	Median intergenic region length (bp)
A. shenzhenica	21,841	6137.69	2518.00	1099.28	876.00	4.50	244.17	138.00	1438.65	403.00	9531.45	6177.00
P. equestris	29,545	8373.90	1185.00	841.32	570.00	3.56	236.05	150.00	2937.72	381.00	24903.93	12783.00
D. catenatum	29,257	7816.87	1757.00	1011.00	756.00	3.85	262.52	148.00	2387.03	329.00	22639.53	10503.50
Z. mays	39,815	3361.45	1824.00	1091.32	915.00	4.51	241.77	134.00	646.06	152.00	49040.89	23226.50
S. bicolor	27,160	2942.00	2165.00	1260.98	1095.00	4.85	259.90	136.50	436.44	145.00	22363.18	5336.00
O. sativa	35,402	2177.38	1479.50	998.58	810.00	3.80	262.69	142.00	420.79	150.00	9459.07	4625.00
P. heterocycla	31,987	3099.24	2373.00	1210.22	981.00	5.28	229.03	132.00	440.95	209.00	35884.38	21907.50
A. thaliana	26,637	1909.57	1593.00	1242.78	1065.00	5.23	237.50	134.00	157.54	98.00	2569.35	1223.00
V. vinifera	25,328	6129.29	3160.00	1177.49	918.00	6.12	192.54	123.00	967.97	208.00	13026.83	5015.00
A. officinalis	27,375	6835.52	3385.00	1004.02	732.00	5.03	199.52	124.00	1446.29	361.00	34806.10	13301.00
A. comosus	27,024	4341.80	2903.00	1171.28	927.00	5.53	211.68	123.00	699.41	295.00	9286.57	4375.00
A. trichopoda	25,933	5761.95	1308.00	962.18	669.00	4.16	231.31	136.00	1519.05	397.00	20684.78	11242.00

Note: We considered the start and stop codons as the two boundaries for calculating gene length.

Supplementary Table E-11. Gene function annotation of *A. shenzhenica*.

		Number	Percent (%)
Total		21,841	
Annotated	InterPro	14,693	67.27
	GO	10,499	48.07
	KEGG	10,283	47.08
	SwissProt	11,578	53.01
	TrEMBL	18,277	83.68
	NCBI non-redundant	18,243	83.52
Unannotated		3,449	15.79

Supplementary Table E-12. Summary of ncRNA annotation of *A. shenzhenica*.

Type		Number	Average length (bp)	Total length (bp)	% of genome
miRNA		43	125.56	5,399	0.00155
tRNA		203	74.75	15,174	0.00435
rRNA	rRNA	452	162.68	73,530	0.02109
	18S	35	771.83	27,014	0.00775
	28S	26	192.96	5,017	0.00144
	5.8S	11	151.73	1,669	0.00048
	5S	380	104.82	39,830	0.01142
snRNA	snRNA	93	103.60	9,635	0.00276
	CD-box	45	98.42	4,429	0.00127
	HACA-box	0	0.00	0	0.00000
	splicing	48	108.46	5,206	0.00149
	scaRNA	0	0.00	0	0.00000

Supplementary Table E-13. Information about the transcriptomes used in this study.

Family	Subfamily	Species	Tissues used in genome annotation	Tissues used in expression analysis	Tissues used in WGD and phylogenetic analysis	Number of predicted proteins	Number of predicted proteins with plant homologs
Orchidaceae	Apostasioideae	*Apostasia shenzhenica*	Flower bud, leaf, root, seed, stem, tuber	flower bud, pollen, stem, root, leaf, seed			
		Apostasia odorata			flower bud	23,504	18,030
		Neuwiedia malipoensis			flower	25,211	23,011
	Vanilloideae	*Vanilla shenzhenica*		flower bud, pollinia, stem, root, leaf	floral bract, lip, sepal, column, leaf, node with leaf and root, aerial root	25,767	18,188
		Lecanorchis nigricans			flower	19,529	16,608
		Galeola faberi			flower	19,093	16,969
	Cypripedioideae	*Paphiopedilum malipoense*		flower bud, pollinia, stem, root, leaf	flower bud	24,360	20,631
		Cypripedium margaritaceum			flower, leaf	22,803	16,485
	Orchidoideae	*Hemipilia forrestii*		flower	flower	20,641	18,295
		Habenaria delavayi			leaf	20,260	17,842
	Epidendroideae	*Phalaenopsis equestris*		flower bud, pollinia, stem, root, leaf, seed			
		Dendrobium catenatum		flower bud, pollinia, stem, root, leaf			
Hypoxidaceae		*Molineria capitulata*		flower bud, pollen, stem, root, leaf, seed	flower bud	26,430	19,865

Supplementary Table E-14. Summary of orthologous gene families in 15 sequenced plant species.

Species	Genes	Unclustered genes	Clustered genes	Families	Unique families	Unique families genes	Common families	Common families genes	Single Copy	Average genes per family
A. comosus	27,024	5,950	21,074	13,279	936	3,346	4,120	7,079	439	1.587
A. shenzhenica	21,841	3,573	18,268	11,995	562	2,789	4,120	6,121	439	1.523
A. officinalis	27,375	8,220	19,155	12,014	901	3,521	4,120	6,920	439	1.594
A. thaliana	26,637	3,750	22,887	12,719	859	3,466	4,120	8,108	439	1.799
A. trichopoda	25,933	7,699	18,234	12,200	1,044	4,206	4,120	5,758	439	1.495
B. distachyon	26,415	3,655	22,760	15,344	421	1,240	4,120	7,748	439	1.483
D. catenatum	29,257	5,339	23,918	14,050	1,036	4,183	4,120	6,638	439	1.702
M. acuminata	34,241	8,710	25,531	12,865	538	1,359	4,120	10,792	439	1.985
O. sativa	35,402	11,106	24,296	16,352	958	2,473	4,120	7,604	439	1.486
P. dactylifera	23,890	6,281	17,609	11,011	444	1,431	4,120	7,331	439	1.599
P. equestris	29,545	6,420	23,125	13,752	1,197	5,887	4,120	6,112	439	1.682
P. trichocarpa	40,984	7,683	33,301	14,471	1,362	4,181	4,120	11,440	439	2.301
S. bicolor	27,160	3,723	23,437	15,749	361	984	4,120	7,893	439	1.488
S. polyrrhiza	18,357	5,095	13,262	10,076	264	797	4,120	5,672	439	1.316
V. vinifera	25,328	6,032	19,296	12,808	643	1,833	4,120	7,113	439	1.507

Note: Unique families = families present only in one species

Supplementary Table E-15. The GO term enrichment of *A. shenzhenica* lineage significantly contracted gene families.

GO ID	GO Term	GO Class	*P* value	Adjusted *P* value	x1	x2	n	N	GO level
GO:0005507	copper ion binding	MF	1.69E-18	9.95E-17	9	74	18	21841	7
GO:0003824	catalytic activity	MF	2.61E-12	1.54E-10	18	4970	18	21841	2
GO:0008152	metabolic process	BP	9.77E-12	5.76E-10	18	5347	18	21841	2
GO:0005488	binding	MF	1.38E-10	8.16E-09	18	6194	18	21841	2
GO:0004713	protein tyrosine kinase activity	MF	3.13E-10	1.85E-08	9	597	18	21841	7
GO:0055114	oxidation-reduction process	BP	9.01E-10	5.31E-08	9	673	18	21841	3
GO:0006468	protein phosphorylation	BP	9.12E-10	5.38E-08	9	674	18	21841	6
GO:0016491	oxidoreductase activity	MF	4.06E-09	2.39E-07	9	799	18	21841	3
GO:0046872	metal ion binding	MF	9.87E-08	5.82E-06	10	1582	18	21841	5
GO:0005524	ATP binding	MF	2.15E-07	1.27E-05	9	1265	18	21841	8
GO:0005515	protein binding	MF	1.14E-05	0.00067253	9	2031	18	21841	3
GO:0044238	primary metabolic process	BP	0.000556837	0.032853412	10	4120	18	21841	3

Note: N: total gene number; n: gene number in the list; x1: gene number with a GO term in the list; x2: gene number with a GO term in total.

Supplementary Table E-16. The GO term enrichment of *A. shenzhenica* lineage significantly expanded gene families.

GO ID	GO Term	GO Class	*P* value	Adjusted *P* value	x1	x2	n	N	GO level
GO:0006259	DNA metabolic process	BP	4.44E-40	2.93E-38	38	702	65	21841	5
GO:0003964	RNA-directed DNA polymerase activity	MF	3.37E-36	2.22E-34	29	322	65	21841	7
GO:0006278	RNA-dependent DNA replication	BP	3.37E-36	2.22E-34	29	322	65	21841	7
GO:0003723	RNA binding	MF	1.95E-28	1.28E-26	29	593	65	21841	4
GO:0015074	DNA integration	BP	3.37E-23	2.22E-21	18	177	65	21841	6
GO:0044249	cellular biosynthetic process	BP	3.61E-18	2.38E-16	32	1796	65	21841	4
GO:0016740	transferase activity	MF	2.60E-17	1.71E-15	32	1921	65	21841	3
GO:0044237	cellular metabolic process	BP	1.94E-16	1.28E-14	41	3758	65	21841	3
GO:0003676	nucleic acid binding	MF	3.22E-16	2.12E-14	30	1783	65	21841	3
GO:0044238	primary metabolic process	BP	5.37E-15	3.54E-13	41	4120	65	21841	3
GO:0008152	metabolic process	BP	2.29E-13	1.51E-11	44	5347	65	21841	2
GO:0003824	catalytic activity	MF	5.12E-08	3.38E-06	35	4970	65	21841	2
GO:0004144	diacylglycerol O-acyltransferase activity	MF	5.00E-07	3.30E-05	3	6	65	21841	8
GO:0005488	binding	MF	4.34E-06	0.000287	36	6194	65	21841	2
GO:0045017	glycerolipid biosynthetic process	BP	7.81E-05	0.005158	3	28	65	21841	5
GO:0046914	transition metal ion binding	MF	9.27E-05	0.006116	12	1110	65	21841	6
GO:0006979	response to oxidative stress	BP	0.000326	0.021548	3	45	65	21841	4
GO:0004601	peroxidase activity	MF	0.000446	0.029449	3	50	65	21841	3

Note: N: total gene number; n: gene number in the list; x1: gene number with a GO term in the list; x2: gene number with a GO term in total.

Supplementary Table E-17. The KEGG Pathway enrichment of *A. shenzhenica* lineage significantly contracted gene families.

MapID	MapTitle	*P* value	Adjusted *P* value	x	y	n	N
map00053	Ascorbate and aldarate metabolism	5.53E-17	4.42E-16	8	57	18	21841
map04626	Plant-pathogen interaction	2.09E-10	1.67E-09	8	364	18	21841
map01100	Metabolic pathways	9.33E-06	7.46E-05	9	1982	18	21841

Note: N: total gene number; n: gene number in the list; x: gene number with a KEGG term in the list; y: gene number with a KEGG term in total.

Supplementary Table E-18. The KEGG pathway enrichment of *A. shenzhenica* lineage significantly expanded gene families.

MapID	MapTitle	*P* value	Adjusted *P* value	x	y	n	N
map03440	Homologous recombination	1.23E-31	1.36E-30	20	104	65	21841
map03008	Ribosome biogenesis in eukaryotes	6.19E-13	6.81E-12	12	199	65	21841
map03018	RNA degradation	1.64E-12	1.80E-11	12	216	65	21841
map00073	Cutin, suberine and wax biosynthesis	0.000266	0.002924	3	42	65	21841
map03015	mRNA surveillance pathway	0.00097	0.010668	4	147	65	21841
map00360	Phenylalanine metabolism	0.001194	0.013139	3	70	65	21841

Note: N: total gene number; n: gene number in the list; x: gene number with a KEGG term in the list; y: gene number with a KEGG term in total.

Supplementary Table E-19. GO term enrichment of *A. shenzhenica*-specific gene families.

GO ID	GO Term	GO Class	*P* value	Adjusted *P* value	x1	x2	n	N	GO level
GO:0003964	RNA-directed DNA polymerase activity	MF	3.24E-05	0.010411	67	322	2789	21841	7
GO:0006278	RNA-dependent DNA replication	BP	3.24E-05	0.010411	67	322	2789	21841	7

Note: N: total gene number; n: gene number in the list; x1: gene number with a GO term in the list; x2: gene number with a GO term in total.

Supplementary Table E-20. GO term enrichment of Orchidaceae-specific gene families.

GO ID	GO Term	GO Class	*P* value	Adjusted *P* value	x1	x2	n	N	GO level
GO:0008171	O-methyltransferase activity	MF	3.91E-07	9.90E-05	8	28	568	21841	6
GO: 0008234	Cysteine-type peptidase activity	MF	4.45E-05	0.011261	6	26	568	21841	7

Note: N: total gene number; n: gene number in the list; x1: gene number with a GO term in the list; x2: gene number with a GO term in total.

Supplementary Table E-21. KEGG pathway enrichment of Orchidaceae-specific gene families.

MapID	MapTitle	*P* value	Adjusted *P* value	x	y	n	N
map00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	7.75E-07	3.87E-05	12	78	568	21841
map00944	Flavone and flavonol biosynthesis	5.79E-05	0.0028937	7	39	568	21841

Note: N: total gene number; n: gene number in the list; x: gene number with a KEGG term in the list; y: gene number with a KEGG term in total.

Supplementary Table E-22. GO term enrichment of monocot-specific gene families.

GO ID	GO Term	GO Class	*P* value	Adjusted *P* value	x1	x2	n	N	GO level
GO:0000156	two-component response regulator activity	MF	5.43E-06	0.000521	3	20	38	21841	4
GO:0015299	solute:hydrogen antiporter activity	MF	1.23E-05	0.00118	3	26	38	21841	8
GO:0015298	solute:cation antiporter activity	MF	1.23E-05	0.00118	3	26	38	21841	7
GO:0000160	two-component signal transduction system (phosphorelay)	BP	1.54E-05	0.001483	3	28	38	21841	4
GO:0015300	solute:solute antiporter activity	MF	1.91E-05	0.001833	3	30	38	21841	7
GO:0004871	signal transducer activity	MF	7.96E-05	0.007643	3	48	38	21841	3
GO:0060089	molecular transducer activity	MF	7.96E-05	0.007643	3	48	38	21841	2
GO:0015078	hydrogen ion transmembrane transporter activity	MF	0.000188	0.018061	3	64	38	21841	9
GO:0015297	antiporter activity	MF	0.000245	0.023559	3	70	38	21841	6
GO:0015291	secondary active transmembrane transporter activity	MF	0.000364	0.034939	3	80	38	21841	5

Note: N: total gene number; n: gene number in the list; x1: gene number with a GO term in the list; x2: gene number with a GO term in total.

Supplementary Table E-23. KEGG pathway enrichment of monocot-specific gene families.

Map ID	Map Title	P value	Adjusted P value	x	y	n	N
map04075	Plant hormone signal transduction	0.000748	0.007479	5	418	38	21841
map03020	RNA polymerase	0.004582	0.045821	2	58	38	21841

Note: N: total gene number; n: gene number in the list; x: gene number with a KEGG term in the list; y: gene number with a KEGG term in total.

Bibliography

1. Mayr, E. *The Growth of Biological Thought*. (Harvard University Press, 1982).
2. Aristotle. *The Works of Aristotle*. **3,** (Oxford Clarendon Press, 1931).
3. Gill, T. A New Translation of Aristotle's “History OF Animals”. *Science* **33,** 730–738 (1911).
4. Lovejoy, A. O. A. O. 1.-1. *The great chain of being*. (MPublishing, University of Michigan Library, 1964).
5. Linné, C. V. *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.* (Holmiae, Impensis Direct. Laurentii Salvii, 1758).
6. Baack, E. J. & Rieseberg, L. H. A genomic view of introgression and hybrid speciation. *Curr Opin Genet Dev* **17,** 513–518 (2007).
7. Ragan, M. A. Trees and networks before and after Darwin. *Biol Direct* **4,** 43–discussion 43 (2009).
8. Linné, C. V. *Philosophia botanica: in qua explicantur fundamenta botanica cum definitionibus partium, exemplis terminorum, observationibus rariorum, adjectis figuris aeneis.* (Godofr Kiesewetter, 1751).
9. Stevens, P. F. Augustin Augier's “Arbre Botanique” (1801), a Remarkable Early Botanical Representation of the Natural System. *Taxon* **32,** 203 (1983).
10. Archibald, J. D. Edward Hitchcock’s Pre-Darwinian (1840) ‘Tree of Life’. *Journal of the History of Biology* **42,** 561–592 (2009).
11. Baum, D. A. & Smith, S. D. *Tree Thinking*. (Roberts & Company, 2013).
12. Darwin, C. *On the Origin of Species*. (OUP Oxford, 2008).
13. Hossfeld, U. & Levit, G. S. Phylogeny: ‘Tree of life’ took root 150 years ago. *Nature* **540,** 38–38 (2016).
14. Hossfeld, U., Watts, E. & Levit, G. S. The First Darwinian Phylogenetic Tree of Plants. *Trends in plant science* **22,** 99–102 (2017).
15. Dayrat, B. The roots of phylogeny: how did Haeckel build his trees? *Systematic Biology* **52,** 515–527 (2003).
16. Mayr, E. & Bock, W. J. Classifications and other ordering systems. *Journal of Zoological Systematics and Evolutionary Research* **40,** 169–194 (2002).
17. Diamond, J. *Obituary: Ernst Mayr (1904-2005). Nature* **433,** 700–701 (2005).
18. Pagel, M. & Meade, A. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **363,** 3955–3964 (2008).
19. The Angiosperm Phylogeny Group. An Ordinal Classification for the Families of Flowering Plants. *Annals of the Missouri Botanical Garden* **85,** 531 (1998).
20. The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of …* (2016).
21. O'Leary, M. A. *et al.* The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **339,** 662–667 (2013).

22. Hedges, S. B. & Kumar, S. Discovering the timetree of life. *The timetree of life* (2009).
23. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* **32,** 835–845 (2015).
24. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34,** 1812–1819 (2017).
25. Pace, N. R. Time for a change. *Nature* **441,** 289–289 (2006).
26. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1,** 16048 (2016).
27. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–236 (2013).
28. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541,** 353–358 (2017).
29. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521,** 173–179 (2015).
30. Koonin, E. V. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol* **11,** 209 (2010).
31. Christenhusz, M. & Byng, J. W. The number of known plants species in the world and its annual increase. *Phytotaxa* (2016). doi:10.11646/phytotaxa.261.3.1
32. Stockey, R. A., Graham, S. W. & Crane, P. R. Introduction to the Darwin special issue: The abominable mystery1. *American Journal of Botany* **96,** 3–4 (2009).
33. Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284,** 2124–2129 (1999).
34. Koonin, E. V. Darwinian evolution in the light of genomics. *Nucleic Acids Research* **37,** 1011–1034 (2009).
35. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3,** 722–732 (2005).
36. Mallet, J. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution* **20,** 229–237 (2005).
37. Andersson, J. O. Lateral gene transfer in eukaryotes. *Cell. Mol. Life Sci.* **62,** 1182–1197 (2005).
38. Martin, W. F. Too Much Eukaryote LGT. *Bioessays* **39,** 1700115 (2017).
39. Fitch, W. M. Homology a personal view on some of the problems. **16,** 227–231 (2000).
40. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6,** 361–375 (2005).
41. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470,** 187–197 (2011).
42. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
43. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
44. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **57,** 1 (2017).
45. Chan, C. X. & Ragan, M. A. Next-generation phylogenomics. *Biol Direct* **8,** 3 (2013).
46. Green, E. D., Rubin, E. M. & Olson, M. V. The future of DNA sequencing. *Nature* **550,** 179–181 (2017).

47. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research* **8,** 163–167 (1998).
48. Wen, J., Liu, J., Ge, S., Xiang, Q.-Y. J. & Zimmer, E. A. Phylogenomic approaches to deciphering the tree of life. *Jnl of Sytematics Evolution* **53,** 369–370 (2015).
49. Genome 10K Community of Scientists. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. *Journal of Heredity* **100,** 659–674 (2009).
50. Koepfli, K.-P., Paten, B., Genome 10K Community of Scientists & O'Brien, S. J. The Genome 10K Project: a way forward. *Annu Rev Anim Biosci* **3,** 57–111 (2015).
51. Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research* **42,** D699–D704 (2013).
52. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3,** 17 (2014).
53. Sun, Y. *et al.* Fish-T1K (Transcriptomes of 1,000 Fishes) Project: large-scale transcriptome data for fish evolution studies. *Gigascience* **5,** 18 (2016).
54. Zhang, G. *et al.* Genomics: Bird sequencing project takes off. *Nature* **522,** 34–34 (2015).
55. Normile, D. Plant scientists plan massive effort to sequence 10,000 genomes. *Science* (2017). doi:10.1126/science.aan7165
56. Mcpherson, J. D. Next-generation gap. *Nat Meth* **6,** S2–5 (2009).
57. Sanger, F. *et al.* Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* **265,** 687–695 (1977).
58. Simpson, J. T. & Pop, M. The Theory and Practice of Genome Sequence Assembly. *Annual Review of Genomics and Human Genetics* **16,** 153–172 (2015).
59. Haas, B. J. & Zody, M. C. Advancing RNA-Seq analysis. *Nature biotechnology* **28,** 421–423 (2010).
60. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet* **12,** 671–682 (2011).
61. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
62. Patterson, C. Homology in classical and molecular biology. *Mol Biol Evol* **5,** 603–625 (1988).
63. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314,** 1041–1052 (2001).
64. Enright, A., Van Dongen, S. & Ouzounis, C. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30,** 1575–1584 (2002).
65. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13,** 2178–2189 (2003).
66. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
67. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
68. Sonnhammer, E. L. L. *et al.* Big data and other challenges in the quest for orthologs. *Bioinformatics* **30,** 2993–2998 (2014).
69. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16,** 157 (2015).

70. Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111,** E4859–68 (2014).
71. Ballesteros, J. A. & Hormiga, G. A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Mol Biol Evol* **33,** 2117–2134 (2016).
72. Kuzniar, A., van Ham, R. C. H. J., Pongor, S. & Leunissen, J. A. M. The quest for orthologs: finding the corresponding gene across genomes. **24,** 539–551 (2008).
73. Forslund, K. *et al.* Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx542
74. Chatzou, M. *et al.* Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics* **17,** 1009–1023 (2016).
75. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
76. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217 (2000).
77. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30,** 3059–3066 (2002).
78. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797 (2004).
79. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7,** 539–539 (2011).
80. Liu, K. *et al.* SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* **61,** 90–106 (2012).
81. Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. & Warnow, T. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324,** 1561–1564 (2009).
82. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56,** 564–577 (2007).
83. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34,** W609–12 (2006).
84. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).
85. Felsenstein, J. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Biology* **27,** 401–410 (1978).
86. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* **5,** 50 (2005).
87. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
88. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* 21–132 (Elsevier, 1969). doi:10.1016/B978-1-4832-3211-9.50009-7

89. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16,** 111–120 (1980).
90. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* **17,** 368–376 (1981).
91. Hasegawa, M., Kishino, H. & Yano, T.-A. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22,** 160–174 (1985).
92. Tavaré, S. *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences*. **17,** 57–86 (American Mathematical Society, 1986).
93. Yang, Z. *Computational Molecular Evolution*. (Oxford University Press, 2006).
94. Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nature* **514,** 550–553 (2014).
95. Luo, Y., Fu, C., Zhang, D.-Y. & Lin, K. Overlapping genes as rare genomic markers: the phylogeny of gamma-Proteobacteria as a case study. *Trends Genet* **22,** 593–596 (2006).
96. Putnam, N. H. *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317,** 86–94 (2007).
97. Tang, H., Bowers, J. E., Wang, X. & Paterson, A. H. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences* **107,** 472–477 (2010).
98. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated Syntenic and Phylogenomic Analyses Reveal an Ancient Genome Duplication in Monocots. *Plant Cell* **26,** 2792–2802 (2014).
99. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59,** 307–321 (2010).
100. Gil, M., Zanetti, M. S., Zoller, S. & Anisimova, M. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. *Mol Biol Evol* **30,** 1270–1280 (2013).
101. Kozlov, A. M., Aberer, A. J. & Stamatakis, A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31,** 2577–2579 (2015).
102. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5,** e9490 (2010).
103. Nguyen, L.-T., Schmidt, H. A., Haeseler, von, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32,** 268–274 (2015).
104. Liu, K., Linder, C. R. & Warnow, T. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS ONE* **6,** e27731 (2011).
105. Zhou, X., Shen, X., Hittinger, C. T. & Rokas, A. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol Biol Evol* (2017). doi:10.1093/molbev/msx302
106. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M. & Gabaldón, T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Research* **42,** D897–902 (2014).
107. Proost, S. *et al.* PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research* **43,** D974–81 (2015).

108. Van Bel, M. *et al.* PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research* (2017). doi:10.1093/nar/gkx1002
109. Mau, B. & Newton, M. A. Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics* **6,** 122 (1997).
110. Yang, Z. & Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol* **14,** 717–724 (1997).
111. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61,** 539–542 (2012).
112. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25,** 2286–2288 (2009).
113. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29,** 1969–1973 (2012).
114. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **10,** e1003537 (2014).
115. Pagel, M. & Meade, A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* **53,** 571–581 (2004).
116. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology* **62,** 611–615 (2013).
117. Ayres, D. L. *et al.* BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology* **61,** 170–173 (2012).
118. Kreft, Ł., Botzki, A., Coppens, F., Vandepoele, K. & Van Bel, M. PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics* **33,** 2946–2947 (2017).
119. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3,** 217–223 (2011).
120. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33,** 1635–1638 (2016).
121. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an rpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8,** 28–36 (2016).
122. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? **22,** 225–231 (2006).
123. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497,** 327–331 (2013).
124. Lemmon, E. M. & Lemmon, A. R. High-Throughput Genomic Data in Systematics and Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **44,** 99–121 (2013).
125. Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* **9,** e1000602 (2011).
126. Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L. & Tamura, K. Statistics and truth in phylogenomics. *Mol Biol Evol* **29,** 457–472 (2012).

127. Song, H., Sheffield, N. C., Cameron, S. L., Miller, K. B. & Whiting, M. F. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Systematic Entomology* **35,** 429–448 (2010).
128. Cooper, E. D. Overly simplistic substitution models obscure green plant phylogeny. *Trends in plant science* **19,** 576–582 (2014).
129. Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E. & Burleigh, J. G. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology* **14,** 23 (2014).
130. Dávalos, L. M. & Perkins, S. L. Saturation and base composition bias explain phylogenomic conflict in Plasmodium. *Genomics* **91,** 433–442 (2008).
131. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* **7 Suppl 1,** S4 (2007).
132. Lockhart, P., Steel, M. & Sullivan, J. A Tale of Two Processes. *Systematic Biology* **54,** 948–951 (2005).
133. Kolaczkowski, B. & Thornton, J. W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431,** 980–984 (2004).
134. Wu, C.-S., Wang, Y.-N., Hsu, C.-Y., Lin, C.-P. & Chaw, S.-M. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biology and Evolution* **3,** 1284–1295 (2011).
135. Zhong, B. *et al.* Systematic error in seed plant phylogenomics. *Genome Biology and Evolution* **3,** 1340–1348 (2011).
136. Lanfear, R., Welch, J. J. & Bromham, L. Watching the clock: Studying variation in rates of molecular evolution between species. *Trends in Ecology & Evolution* **25,** 495–503 (2010).
137. Crotty, S. M. *et al.* GHOST: Recovering Historical Signal from Heterotachously-evolved Sequence Alignments. *bioRxiv* 174789 (2017). doi:10.1101/174789
138. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24,** 332–340 (2009).
139. Nieto Feliner, G. *et al.* Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity* **118,** 513–516 (2017).
140. Rieseberg, L. H. & Willis, J. H. Plant speciation. *Science* **317,** 910–914 (2007).
141. Comai, L. The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6,** 836–846 (2005).
142. Lynch, M. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290,** 1151–1155 (2000).
143. Hahn, M. W., Han, M. V. & Han, S.-G. Gene family evolution across 12 Drosophila genomes. *PLoS Genet* **3,** e197 (2007).
144. Kellogg, E. A. Has the connection between polyploidy and diversification actually been tested? *Curr Opin Plant Biol* **30,** 25–32 (2016).
145. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences* **108,** 4069–4074 (2011).

146. Garsmeur, O. *et al.* Two evolutionarily distinct classes of paleopolyploidy. *Mol Biol Evol* **31,** 448–454 (2014).
147. Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. Evolution of plant genome architecture. *Genome Biol* **17,** 37 (2016).
148. Wendel, J. F., Lisch, D., Hu, G. & Mason, A. S. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr Opin Genet Dev* **49,** 1–7 (2018).
149. Zhao, M., Zhang, B., Lisch, D. & Ma, J. Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. *Plant Cell* **29,** 2974–2994 (2017).
150. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat Rev Genet* **18,** 411–424 (2017).
151. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* **15,** 150 (2015).
152. Zhang, N., Zeng, L., Shan, H. & Ma, H. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol* **195,** 923–937 (2012).
153. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* **29,** 1695–1701 (2012).
154. Mirarab, S., Bayzid, M. S., Boussau, B. & Warnow, T. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* **346,** 1250463 (2014).
155. Xi, Z., Liu, L., Rest, J. S. & Davis, C. C. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. *Systematic Biology* **63,** 919–932 (2014).
156. Xi, Z., Rest, J. S. & Davis, C. C. Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PLoS ONE* **8,** e80870 (2013).
157. Woods, S. *et al.* Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet* **9,** e1003330 (2013).
158. Makino, T., McLysaght, A. & Kawata, M. Genome-wide deserts for copy number variation in vertebrates. *Nat Commun* **4,** 2283 (2013).
159. Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A. & Ciccarelli, F. D. Low duplicability and network fragility of cancer genes. *Trends Genet* **24,** 427–430 (2008).
160. He, X. & Zhang, J. Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol* **23,** 144–151 (2006).
161. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* **60,** 433–453 (2009).
162. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* **102,** 5454–5459 (2005).
163. Blanc, G. & Wolfe, K. H. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16,** 1679–1691 (2004).
164. Tasdighian, S. *et al.* Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity. *Plant Cell* **29,** 2766–2785 (2017).

165. Edger, P. P. & Pires, J. C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17,** 699–717 (2009).
166. Duarte, J. M. *et al.* Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* **10,** 61 (2010).
167. De Smet, R. *et al.* Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences* **110,** 2898–2903 (2013).
168. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Research* **24,** 1334–1347 (2014).
169. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473,** 97–100 (2011).
170. Amborella Genome Project. The Amborella genome and the evolution of flowering plants. *Science* **342,** 1241089 (2013).
171. Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci USA* **106,** 5737–5742 (2009).
172. Lohaus, R. & Van de Peer, Y. Of dups and dinos: evolution at the K/Pg boundary. *Curr Opin Plant Biol* **30,** 62–69 (2016).
173. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* **12,** 1093–1101 (2000).
174. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20,** 3643–3646 (2004).
175. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* **10,** 725–732 (2009).
176. Proost, S. *et al.* i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research* **40,** e11–e11 (2012).
177. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40,** e49–e49 (2012).
178. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16,** 1667–1678 (2004).
179. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications from age distributions revisited. *Mol Biol Evol* **30,** 177–190 (2013).
180. Li, Z. *et al.* Early genome duplications in conifers and other seed plants. *Science Advances* **1,** e1501084 (2015).
181. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422,** 433–438 (2003).
182. Jiao, Y. *et al.* A genome triplication associated with early diversification of the core eudicots. *Genome Biol* **13,** R3 (2012).
183. Mckain, M. R. *et al.* A Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales. *Genome Biology and Evolution* **8,** 1150–1164 (2016).
184. Ruprecht, C. *et al.* Revisiting ancestral polyploidy in plants. *Science Advances* **3,** e1603195 (2017).

185. Ming, R. *et al.* The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* (2015). doi:10.1038/ng.3435
186. Marcet-Houben, M. & Gabaldón, T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLoS Biol* **13,** e1002220 (2015).
187. Thomas, G. W. C., Ather, S. H. & Hahn, M. W. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Systematic Biology* **66,** 1007–1018 (2017).
188. Cai, J. *et al.* The genome sequence of the orchid Phalaenopsis equestris. *Nat Genet* (2014). doi:10.1038/ng.3149
189. Zhang, G.-Q. *et al.* The Dendrobium catenatum Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci Rep* **6,** 19029 (2016).
190. Ramírez, S. R., Gravendeel, B., Singer, R. B., Marshall, C. R. & Pierce, N. E. Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. *Nature* **448,** 1042–1045 (2007).
191. Gustafsson, A. L. S., Verola, C. F. & Antonelli, A. Reassessing the temporal evolution of orchids with new fossils and a Bayesian relaxed clock, with implications for the diversification of the rare South American genus Hoffmannseggella (Orchidaceae: Epidendroideae). *BMC Evolutionary Biology* **10,** 177 (2010).
192. Chen, S., Kim, D.-K., Chase, M. W. & Kim, J.-H. Networks in a large-scale phylogenetic analysis: reconstructing evolutionary history of Asparagales (Lilianae) based on four plastid genes. *PLoS ONE* **8,** e59472 (2013).
193. Chomicki, G. *et al.* The velamen protects photosynthetic orchid roots against UV-B damage, and a large dated phylogeny implies multiple gains and losses of this function during the Cenozoic. *New Phytol* **205,** 1330–1341 (2015).
194. Givnish, T. J. *et al.* Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proc Biol Sci* **282,** 20151553 (2015).
195. Fiz-Palacios, O., Schneider, H., Heinrichs, J. & Savolainen, V. Diversification of land plants: insights from a family-level phylogenetic analysis. *BMC Evolutionary Biology* **11,** 341 (2011).
196. Chase, M. W. & Reveal, J. L. A phylogenetic classification of the land plants to accompany APG III. *Bot J Linn Soc* **161,** 122–127 (2009).
197. Chaw, S. M., Parkinson, C. L., Cheng, Y., Vincent, T. M. & Palmer, J. D. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc Natl Acad Sci USA* **97,** 4086–4091 (2000).
198. Wang, X.-Q. & Ran, J.-H. Evolution and biogeography of gymnosperms. *Mol. Phylogenet. Evol.* **75,** 24–40 (2014).
199. Haston, E., Richardson, J. E., Stevens, P. F., Chase, M. W. & Harris, D. J. The Linear Angiosperm Phylogeny Group (LAPG) III: a linear sequence of the families in APG III. *Bot J Linn Soc* **161,** 128–131 (2009).
200. Lee, E. K. *et al.* A functional phylogenomic view of the seed plants. *PLoS Genet* **7,** e1002411 (2011).
201. Zhong, B., Yonezawa, T., Zhong, Y. & Hasegawa, M. The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol Biol Evol* **27,** 2855–2863 (2010).

202. Lu, Y., Ran, J.-H., Guo, D.-M., Yang, Z.-Y. & Wang, X.-Q. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PLoS ONE* **9,** e107679 (2014).
203. Doyle, J. A. Phylogeny of vascular plants. *Annual Review of Ecology and Systematics* (1998). doi:10.2307/221719
204. Burleigh, J. G. & Mathews, S. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American Journal of Botany* **91,** 1599–1613 (2004).
205. Cibrián-Jaramillo, A. *et al.* Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution. *Genome Biology and Evolution* **2,** 225–239 (2010).
206. Ran, J.-H., Gao, H. & Wang, X.-Q. Fast evolution of the retroprocessed mitochondrial rps3 gene in Conifer II and further evidence for the phylogeny of gymnosperms. *Mol. Phylogenet. Evol.* **54,** 136–149 (2010).
207. Burleigh, J. G., Barbazuk, W. B., Davis, J. M., Morse, A. M. & Soltis, P. S. Exploring Diversification and Genome Size Evolution in Extant Gymnosperms through Phylogenetic Synthesis. *Journal of Botany* **2012,** 1–6 (2012).
208. Mathews, S. Phylogenetic relationships among seed plants: Persistent questions and the limits of molecular data. *American Journal of Botany* **96,** 228–236 (2009).
209. Wu, C.-S., Chaw, S.-M. & Huang, Y.-Y. Chloroplast phylogenomics indicates that Ginkgo biloba is sister to cycads. *Genome Biology and Evolution* **5,** 243–254 (2013).
210. Zwickl, D. J. & Hillis, D. M. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* **51,** 588–598 (2002).
211. Salas-Leiva, D. E. *et al.* Conserved genetic regions across angiosperms as tools to develop single-copy nuclear markers in gymnosperms: an example using cycads. *Mol Ecol Resour* **14,** 831–845 (2014).
212. Levin, R. A., Whelan, A. & Miller, J. S. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in Lycium (Solanaceae). *Mol. Phylogenet. Evol.* **53,** 881–890 (2009).
213. Zeng, L. *et al.* Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat Commun* **5,** 4956 (2014).
214. Wu, F., Mueller, L. A., Crouzillat, D., Pétiard, V. & Tanksley, S. D. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics* **174,** 1407–1420 (2006).
215. Li, Z. *et al.* Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28,** 326–344 (2016).
216. La Torre, De, A. R. *et al.* Insights into conifer giga-genomes. *Plant Physiol* **166,** 1724–1732 (2014).
217. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497,** 579–584 (2013).
218. Neale, D. B. *et al.* Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol* **15,** R59 (2014).
219. Birol, I. *et al.* Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29,** 1492–1497 (2013).

220. Warren, R. L. *et al.* Improved white spruce (Picea glauca) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J* **83,** 189–212 (2015).
221. La Torre, De, A. R., Lin, Y.-C., Van de Peer, Y. & Ingvarsson, P. K. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in picea gene families. *Genome Biology and Evolution* **7,** 1002–1015 (2015).
222. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7,** e1002195 (2011).
223. Velasco, R. *et al.* The genome of the domesticated apple (Malus × domestica Borkh.). *Nat Genet* **42,** 833–839 (2010).
224. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463,** 178–183 (2010).
225. Wang, X. *et al.* The genome of the mesopolyploid crop species Brassica rapa. *Nat Genet* **43,** 1035–1039 (2011).
226. Wang, K. *et al.* The draft genome of a diploid cotton Gossypium raimondii. *Nat Genet* **44,** 1098–1103 (2012).
227. Rensing, S. A. *et al.* The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* **319,** 64–69 (2008).
228. Tuskan, G. A. *et al.* The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
229. Myburg, A. A. *et al.* The genome of Eucalyptus grandis. *Nature* **509,** 356–362 (2014).
230. Waterhouse, R. M., Zdobnov, E. M. & Kriventseva, E. V. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biology and Evolution* **3,** 75–86 (2011).
231. De Torre-Bárcena, J. E. *et al.* The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. *PLoS ONE* **4,** e5764 (2009).
232. Zhu, X.-Y. *et al.* Mitochondrial matR sequences help to resolve deep phylogenetic relationships in rosids. *BMC Evolutionary Biology* **7,** 217 (2007).
233. Sun, M. *et al.* Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Mol. Phylogenet. Evol.* **83C,** 156–166 (2014).
234. Palmé, A. E., Pyhäjärvi, T., Wachowiak, W. & Savolainen, O. Selection on nuclear genes in a Pinus phylogeny. *Mol Biol Evol* **26,** 893–905 (2009).
235. Gernandt, D. S., López, G. G., García, S. O. & Liston, A. Phylogeny and classification of Pinus. *Taxon* **54,** 29–42 (2005).
236. Lin, C.-P., Huang, J.-P., Wu, C.-S., Hsu, C.-Y. & Chaw, S.-M. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. *Genome Biology and Evolution* **2,** 504–517 (2010).
237. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51,** 492–508 (2002).
238. Seo, T.-K. & Kishino, H. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. *Systematic Biology* **57,** 367–377 (2008).
239. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics*. (Oxford University Press, USA, 2000).

240. Xia, X., Xie, Z., Salemi, M., Chen, L. & Wang, Y. An index of substitution saturation and its application. *Mol. Phylogenet. Evol.* **26,** 1–7 (2003).
241. Ren, F., Tanaka, H. & Yang, Z. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Systematic Biology* **54,** 808–818 (2005).
242. Liu, L., Yu, L., Pearl, D. K. & Edwards, S. V. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* **58,** 468–477 (2009).
243. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31,** i44–52 (2015).
244. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol* **31,** 1261–1271 (2014).
245. Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Mol Biol Evol* **33,** 1606–1617 (2016).
246. Crisp, M. D. & Cook, L. G. Cenozoic extinctions account for the low diversity of extant gymnosperms compared with angiosperms. *New Phytol* **192,** 997–1009 (2011).
247. Nagalingum, N. S. *et al.* Recent synchronous radiation of a living fossil. *Science* **334,** 796–799 (2011).
248. Wu, C.-S., Lai, Y.-T., Lin, C.-P., Wang, Y.-N. & Chaw, S.-M. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. *Mol. Phylogenet. Evol.* **52,** 115–124 (2009).
249. Braukmann, T. W. A., Kuzmina, M. & Stefanović, S. Loss of all plastid ndh genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. *Curr. Genet.* **55,** 323–337 (2009).
250. Ruhlman, T. A. *et al.* NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biol* **15,** 100 (2015).
251. Wakasugi, T. *et al.* Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii. *Proc Natl Acad Sci USA* **91,** 9794–9798 (1994).
252. Denton, J. F. *et al.* Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol* **10,** e1003998 (2014).
253. Rigault, P. *et al.* A white spruce gene catalog for conifer genome analyses. *Plant Physiol* **157,** 14–28 (2011).
254. Ralph, S. G. *et al.* A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis). *BMC Genomics* **9,** 484 (2008).
255. Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Rieseberg, L. H. & Aitken, S. N. Expression Divergence Is Correlated with Sequence Evolution but Not Positive Selection in Conifers. *Mol Biol Evol* **33,** 1502–1516 (2016).
256. Chen, J. *et al.* Sequencing of the needle transcriptome from Norway spruce (Picea abies Karst L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. *BMC Genomics* **13,** 589 (2012).
257. Ruttink, T. *et al.* Orthology Guided Assembly in highly heterozygous crops: creating a reference transcriptome to uncover genetic diversity in Lolium perenne. *Plant Biotechnol J* **11,** 605–617 (2013).

258. Zimin, A. *et al.* Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics* **196,** 875–890 (2014).
259. Cañas, R. A., Canales, J., Gómez-Maldonado, J., Avila, C. & Cánovas, F. M. Transcriptome analysis in maritime pine using laser capture microdissection and 454 pyrosequencing. *Tree Physiol.* **34,** 1278–1288 (2014).
260. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
261. Duvick, J. *et al.* PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research* **36,** D959–65 (2008).
262. Canales, J. *et al.* De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnol J* (2013). doi:10.1111/pbi.12136
263. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. *German conference on bioinformatics* (1999).
264. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Research* **9,** 868–877 (1999).
265. Hall, B. G. *Phylogenetic Trees Made Easy*. (Sinauer Associates Incorporated, 2004).
266. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21,** 3448–3449 (2005).
267. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164–1165 (2011).
268. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11,** 725–736 (1994).
269. Xia, X. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol* **30,** 1720–1728 (2013).
270. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).
271. Ohno, S. *Evolution by Gene Duplication*. (Springer Berlin Heidelberg, 1970). doi:10.1007/978-3-642-86659-3
272. Davis, J. C. & Petrov, D. A. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* **2,** E55 (2004).
273. Koonin, E. V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* **5,** R7 (2004).
274. Liang, H., Plazonic, K. R., Chen, J., Li, W.-H. & Fernández, A. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet* **4,** e11 (2008).
275. Makino, T., Hokamp, K. & McLysaght, A. The complex relationship of gene duplication and essentiality. *Trends Genet* **25,** 152–155 (2009).
276. Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424,** 194–197 (2003).
277. Seoighe, C. & Gehring, C. Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. **20,** 461–464 (2004).
278. Aury, J.-M. *et al.* Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444,** 171–178 (2006).
279. Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L. & Vandepoele, K. The flowering world: a tale of duplications. *Trends in plant science* **14,** 680–688 (2009).

280. Birchler, J. A., Riddle, N. C., Auger, D. L. & Veitia, R. A. Dosage balance in gene regulation: biological implications. *Trends Genet* **21,** 219–226 (2005).
281. Birchler, J. A. & Veitia, R. A. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* **19,** 395–402 (2007).
282. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences* **109,** 14746–14753 (2012).
283. Brunet, F. G. *et al.* Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* **23,** 1808–1816 (2006).
284. Huminiecki, L. & Heldin, C. H. 2R and remodeling of vertebrate signal transduction engine. *BMC Biol* **8,** 146 (2010).
285. Makino, T. & McLysaght, A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences* **107,** 9270–9274 (2010).
286. Rodgers-Melnick, E. *et al.* Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Research* **22,** 95–105 (2012).
287. Buggs, R. J. A. *et al.* Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr Biol* **22,** 248–252 (2012).
288. McGrath, C. L. & Lynch, M. in *Polyploidy and Genome Evolution* 1–20 (Springer Berlin Heidelberg, 2012). doi:10.1007/978-3-642-31442-1_1
289. Douglas, G. M. *et al.* Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. *Proceedings of the National Academy of Sciences* **112,** 2806–2811 (2015).
290. Paterson, A. H. *et al.* Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. **22,** 597–602 (2006).
291. Armisén, D., Lecharny, A. & Aubourg, S. Unique genes in plants: specificities and conserved features throughout evolution. *BMC Evolutionary Biology* **8,** 280 (2008).
292. Han, F., Peng, Y., Xu, L. & Xiao, P. Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes. *BMC Genomics* **15,** 504 (2014).
293. Barker, M. S. *et al.* Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* **25,** 2445–2455 (2008).
294. Soltis, D. E., Buggs, R. J. A., Doyle, J. J. & Soltis, P. S. What we still don't know about polyploidy. *Taxon* **59,** 1387–1403 (2010).
295. Carretero-Paulet, L. & Fares, M. A. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Mol Biol Evol* **29,** 3541–3551 (2012).
296. Conant, G. C. Comparative Genomics as a Time Machine: How Relative Gene Dosage and Metabolic Requirements Shaped the Time-dependent Resolution of Yeast Polyploidy. *Mol Biol Evol* (2014). doi:10.1093/molbev/msu250
297. Bailey, N. T. J. *The Elements of Stochastic Processes with Applications to the Natural Sciences*. (John Wiley & Sons, 1964).
298. Rabier, C.-E., Ta, T. & Ané, C. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol Biol Evol* **31,** 750–762 (2014).

299. Lloyd, A. H. *et al.* Meiotic Gene Evolution: Can You Teach a New Dog New Tricks? *Mol Biol Evol* (2014). doi:10.1093/molbev/msu119
300. Lynch, M. & Conery, J. S. The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* **3,** 35–44 (2003).
301. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52,** 91–118 (2003).
302. Smith, S. A. & Donoghue, M. J. Rates of molecular evolution are linked to life history in flowering plants. *Science* **322,** 86–89 (2008).
303. Myouga, F. *et al.* The Chloroplast Function Database II: a comprehensive collection of homozygous mutants and their phenotypic/genotypic traits for nuclear-encoded chloroplast proteins. *Plant Cell Physiol.* **54,** e2–e2 (2013).
304. Cuéllar Pérez, A. *et al.* The non-JAZ TIFY protein TIFY8 from Arabidopsis thaliana is a transcriptional repressor. *PLoS ONE* **9,** e84891 (2014).
305. Freeling, M. & Thomas, B. C. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* **16,** 805–814 (2006).
306. Lloyd, J. & Meinke, D. A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. *Plant Physiol* **158,** 1115–1129 (2012).
307. Lloyd, J. P., Seddon, A. E., Moghe, G. D., Simenc, M. C. & Shiu, S.-H. Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. *Plant Cell* **27,** 2133–2147 (2015).
308. Freeling, M. *et al.* Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Research* **18,** 1924–1937 (2008).
309. Wang, Y. *et al.* Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. *PLoS ONE* **6,** e28150 (2011).
310. Woodhouse, M. R., Tang, H. & Freeling, M. Different gene families in Arabidopsis thaliana transposed in different epochs and at different frequencies throughout the rosids. *Plant Cell* **23,** 4241–4253 (2011).
311. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449,** 54–61 (2007).
312. van Nimwegen, E. Scaling laws in the functional content of genomes. *Trends Genet* **19,** 479–484 (2003).
313. Molina, N. & van Nimwegen, E. Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends Genet* **25,** 243–247 (2009).
314. Kitano, H. Biological robustness. *Nat Rev Genet* **5,** 826–837 (2004).
315. Siegel, J. J. & Amon, A. New insights into the troubles of aneuploidy. *Annu. Rev. Cell Dev. Biol.* **28,** 189–214 (2012).
316. Bridgham, J. T., Brown, J. E., Rodríguez-Marí, A., Catchen, J. M. & Thornton, J. W. Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet* **4,** e1000191 (2008).
317. Dean, E. J., Davis, J. C., Davis, R. W. & Petrov, D. A. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet* **4,** e1000113 (2008).
318. Kaltenegger, E. & Ober, D. Paralogue Interference Affects the Dynamics after Gene Duplication. *Trends in plant science* **20,** 814–821 (2015).

319. Zhang, J. & Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat Rev Genet* (2015). doi:10.1038/nrg3950
320. Lynch, M. & Katju, V. The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20,** 544–549 (2004).
321. Bekaert, M., Edger, P. P., Pires, J. C. & Conant, G. C. Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell* **23,** 1719–1728 (2011).
322. Hahn, M. A., van Kleunen, M. & Müller-Schärer, H. Increased phenotypic plasticity to climate may have boosted the invasion success of polyploid Centaurea stoebe. *PLoS ONE* **7,** e50284 (2012).
323. Beest, te, M. *et al.* The more the better? The role of polyploidy in facilitating plant invasions. *Ann Bot* **109,** 19–45 (2012).
324. Chao, D.-Y. *et al.* Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science* **341,** 658–659 (2013).
325. Vanneste, K., Maere, S. & Van de Peer, Y. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **369,** 20130353 (2014).
326. Selmecki, A. M. *et al.* Polyploidy can drive rapid adaptation in yeast. *Nature* **519,** 349–352 (2015).
327. Sunshine, A. B. *et al.* The fitness consequences of aneuploidy are driven by condition-dependent gene effects. *PLoS Biol* **13,** e1002155 (2015).
328. Dunham, M. J. *et al.* Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* **99,** 16144–16149 (2002).
329. Gresham, D. *et al.* The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet* **4,** e1000303 (2008).
330. Yang, C. *et al.* Evolution of physiological responses to salt stress in hexaploid wheat. *Proceedings of the National Academy of Sciences* **111,** 11882–11887 (2014).
331. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134,** 341–352 (2008).
332. D'Antonio, M. & Ciccarelli, F. D. Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput Biol* **7,** e1002029 (2011).
333. Alvarez-Ponce, D. & Fares, M. A. Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network. *Genome Biology and Evolution* **4,** 1263–1274 (2012).
334. Slotte, T. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45,** 831–835 (2013).
335. Muse, S. V. & Gaut, B. S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11,** 715–724 (1994).
336. Yap, V. B., Lindsay, H., Easteal, S. & Huttley, G. Estimates of the effect of natural selection on protein-coding content. *Mol Biol Evol* **27,** 726–734 (2010).
337. Bremer, B. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* **161,** 105–121 (2009).
338. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology* **55,** 539–552 (2006).

339. Stolzer, M. *et al.* Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28,** i409–i415 (2012).
340. Hahn, M. W. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* **8,** R141 (2007).
341. Nguyen, T.-H., Ranwez, V., Berry, V. & Scornavacca, C. Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods. *PLoS ONE* **8,** e73667 (2013).
342. Wu, Y.-C., Rasmussen, M. D., Bansal, M. S. & Kellis, M. TreeFix: statistically informed gene tree error correction using species trees. *Systematic Biology* **62,** 110–120 (2013).
343. Oliver, T., Schmidt, B., Nathan, D., Clemens, R. & Maskell, D. Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics* **21,** 3431–3432 (2005).
344. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
345. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24,** 1586–1591 (2007).
346. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17,** 32–43 (2000).
347. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3,** 418–426 (1986).
348. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* **15,** 1153–1160 (2005).
349. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26,** 1572–1573 (2010).
350. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2013 update. *Nucleic Acids Research* **41,** D816–23 (2013).
351. De Bodt, S., Hollunder, J., Nelissen, H., Meulemeester, N. & Inzé, D. CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *New Phytol* **195,** 707–720 (2012).
352. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* **41,** D808–15 (2013).
353. Van Landeghem, S. *et al.* Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE* **8,** e55814 (2013).
354. Takahashi, N. *et al.* The DNA replication checkpoint aids survival of plants deficient in the novel replisome factor ETG1. *EMBO J.* **27,** 1840–1851 (2008).
355. Pauwels, L. *et al.* NINJA connects the co-repressor TOPLESS to jasmonate signalling. *Nature* **464,** 788–791 (2010).
356. Van Leene, J. *et al.* Targeted interactomics reveals a complex core cell cycle machinery in Arabidopsis thaliana. *Mol Syst Biol* **6,** 397 (2010).
357. Bassard, J.-E. *et al.* Protein-protein and protein-membrane associations in the lignin pathway. *Plant Cell* **24,** 4465–4482 (2012).
358. Domenichini, S. *et al.* Evidence for a role of Arabidopsis CDT1 proteins in gametophyte development and maintenance of genome integrity. *Plant Cell* **24,** 2779–2791 (2012).

359. Eloy, N. B. *et al.* SAMBA, a plant-specific anaphase-promoting complex/cyclosome regulator is involved in early development and A-type cyclin stabilization. *Proceedings of the National Academy of Sciences* **109,** 13853–13858 (2012).
360. Antoni, R. *et al.* PYRABACTIN RESISTANCE1-LIKE8 plays an important role for the regulation of abscisic acid signaling in root. *Plant Physiol* **161,** 931–941 (2013).
361. Cromer, L. *et al.* Centromeric cohesion is protected twice at meiosis, by SHUGOSHINs at anaphase I and by PATRONUS at interkinesis. *Curr Biol* **23,** 2090–2099 (2013).
362. Di Rubbo, S. *et al.* The clathrin adaptor complex AP-2 mediates endocytosis of brassinosteroid insensitive1 in Arabidopsis. *Plant Cell* **25,** 2986–2997 (2013).
363. Heijde, M. *et al.* Constitutively active UVR8 photoreceptor variant in Arabidopsis. *Proceedings of the National Academy of Sciences* **110,** 20326–20331 (2013).
364. Spinner, L. *et al.* A protein phosphatase 2A complex spatially controls plant cell division. *Nat Commun* **4,** 1863 (2013).
365. Fonseca, S. *et al.* bHLH003, bHLH013 and bHLH017 are new targets of JAZ repressors negatively regulating JA responses. *PLoS ONE* **9,** e86182 (2014).
366. Gadeyne, A. *et al.* The TPLATE adaptor complex drives clathrin-mediated endocytosis in plants. *Cell* **156,** 691–704 (2014).
367. Vercruyssen, L. *et al.* ANGUSTIFOLIA3 binds to SWI/SNF chromatin remodeling complexes to regulate transcription during Arabidopsis leaf development. *Plant Cell* **26,** 210–229 (2014).
368. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.* **57,** 289–300 (1995).
369. Roberts, D. L. & Dixon, K. W. Orchids. *Curr Biol* **18,** R325–9 (2008).
370. Givnish, T. J. *et al.* Orchid historical biogeography, diversification, Antarctica and the paradox of orchid dispersal. *Journal of Biogeography* **43,** 1905–1916 (2016).
371. Chen, L.-J. & Liu, Z.-J. *Apostasia shenzhenica, a new species of Apostasioideae (Orchidaceae) from China*. **29,** 38–41 (Plant Science Journal, 2011).
372. Kocyan, A., Qiu, Y. L., Endress, P. K. & Conti, E. A phylogenetic analysis of Apostasioideae (Orchidaceae) based on ITS, trnL-F and matK sequences. *Plant Syst. Evol.* **247,** 203–213 (2004).
373. Dressler, R. L. *Phylogeny and Classification of the Orchid Family*. (Cambridge University Press, 1993).
374. Kocyan, A. & Endress, P. K. Floral Structure and Development of Apostasiaand Neuwiedia(Apostasioideae) and their Relationships to Other Orchidaceae. *International Journal of Plant Sciences* **162,** 847–867 (2001).
375. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
376. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22,** 1269–1271 (2006).
377. Schranz, M. E., Mohammadin, S. & Edger, P. P. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr Opin Plant Biol* **15,** 147–153 (2012).
378. Tank, D. C. *et al.* Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol* **207,** 454–467 (2015).

379. Tsai, W.-C., Kuoh, C.-S., Chuang, M.-H., Chen, W.-H. & Chen, H.-H. Four DEF-like MADS box genes displayed distinct floral morphogenetic roles in Phalaenopsis orchid. *Plant Cell Physiol.* **45,** 831–844 (2004).
380. Pan, Z.-J. *et al.* Flower development of Phalaenopsis orchid involves functionally divergent SEPALLATA-like genes. *New Phytol* **202,** 1024–1042 (2014).
381. Mondragón-Palomino, M. & Theissen, G. Why are orchid flowers so diverse? Reduction of evolutionary constraints by paralogues of class B floral homeotic genes. *Ann Bot* **104,** 583–594 (2009).
382. Johnson, S. D. & Edwards, T. J. The structure and function of orchid pollinaria. *Plant Syst. Evol.* **222,** 243–269 (2000).
383. Liu, Y. *et al.* Functional conservation of MIKC*-Type MADS box genes in Arabidopsis and rice pollen maturation. *Plant Cell* **25,** 1288–1303 (2013).
384. Kwantes, M., Liebsch, D. & Verelst, W. How MIKC* MADS-box genes originated and evidence for their conserved function throughout the evolution of vascular plant gametophytes. *Mol Biol Evol* **29,** 293–302 (2012).
385. Masiero, S., Colombo, L., Grini, P. E., Schnittger, A. & Kater, M. M. The emerging importance of type I MADS box transcription factors for plant reproduction. *Plant Cell* **23,** 865–872 (2011).
386. Zotz, G. & Winkler, U. Aerial roots of epiphytic orchids: the velamen radicum and its role in water and nutrient uptake. *Oecologia* **171,** 733–741 (2013).
387. Gravendeel, B., Smithson, A., Slik, F. J. W. & Schuiteman, A. Epiphytism and pollinator specialization: drivers for orchid diversity? *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **359,** 1523–1535 (2004).
388. Tapia-López, R. *et al.* An AGAMOUS-related MADS-box gene, XAL1 (AGL12), regulates root meristem cell proliferation and flowering transition in Arabidopsis. *Plant Physiol* **146,** 1182–1192 (2008).
389. Ibarra-Laclette, E. *et al.* Architecture and evolution of a minute plant genome. *Nature* **498,** 94–98 (2013).
390. Zhang, H. & Forde, B. G. An Arabidopsis MADS box gene that controls nutrient-induced changes in root architecture. *Science* **279,** 407–409 (1998).
391. Leseberg, C. H., Li, A., Kang, H., Duvall, M. & Mao, L. Genome-wide analysis of the MADS-box gene family in Populus trichocarpa. *Gene* **378,** 84–94 (2006).
392. Arora, R. *et al.* MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics* **8,** 242 (2007).
393. Jersáková, J. *et al.* Genome size variation in Orchidaceae subfamily Apostasioideae: filling the phylogenetic gap. *Bot J Linn Soc* **172,** 95–105 (2013).
394. Leitch, I. J. *et al.* Genome size diversity in orchids: consequences and evolution. *Ann Bot* **104,** 469–481 (2009).
395. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2,** 231–239 (1988).
396. Butler, J. *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Research* **18,** 810–820 (2008).
397. English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE* **7,** e47768 (2012).
398. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9,** e112963 (2014).

399. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology* **34,** 303–311 (2016).
400. Adey, A. *et al.* In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Research* **24,** 2041–2049 (2014).
401. Mostovoy, Y. *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Meth* **13,** 587–590 (2016).
402. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6,** 11 (2015).
403. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).
404. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27,** 757–763 (2011).
405. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20,** 2878–2879 (2004).
406. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).
407. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
408. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7,** 562–578 (2012).
409. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).
410. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **27,** 29–34 (1999).
411. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **28,** 45–48 (2000).
412. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31,** 365–370 (2003).
413. Zdobnov, E. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
414. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25,** 955–964 (1997).
415. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25,** 1335–1337 (2009).
416. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
417. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8,** 1494–1512 (2013).
418. Gross, S. M. *et al.* De novo transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana. *BMC Genomics* **14,** 563 (2013).
419. Duangjit, J., Bohanec, B., Chan, A. P., Town, C. D. & Havey, M. J. Transcriptome sequencing to produce SNP-based genetic maps of onion. *Theor. Appl. Genet.* **126,** 2093–2101 (2013).

420. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).
421. International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463,** 763–768 (2010).
422. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
423. Magallón, S., Hilu, K. W. & Quandt, D. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *American Journal of Botany* **100,** 556–573 (2013).
424. Fostier, J. *et al.* A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **27,** 749–756 (2011).
425. Ostlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research* **38,** D196–D203 (2010).
426. D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488,** 213–217 (2012).
427. Al-Dous, E. K. *et al.* De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). *Nature biotechnology* **29,** 521–527 (2011).
428. Wang, W. *et al.* The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat Commun* **5,** 3311 (2014).
429. Olsen, J. L. *et al.* The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature* **530,** 331–335 (2016).
430. Gandolfo, M., Nixon, K. & Crepet, W. A new fossil flower from the Turonian of New Jersey: Dressiantha bicarpellata gen. et sp. nov. (Capparales). *American Journal of Botany* **85,** 964 (1998).
431. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **107,** 18724–18728 (2010).
432. Crepet, W. & Nixon, K. Fossil Clusiaceae from the late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. *American Journal of Botany* **85,** 1122 (1998).
433. Xi, Z. *et al.* Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences* **109,** 17519–17524 (2012).
434. JANSSEN, T. & Bremer, K. The age of major monocot groups inferred from 800+rbcL sequences. *Bot J Linn Soc* **146,** 385–398 (2004).
435. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proceedings of the National Academy of Sciences* **107,** 5897–5902 (2010).
436. Clarke, J. T., Warnock, R. C. M. & Donoghue, P. C. J. Establishing a time-scale for plant evolution. *New Phytol* **192,** 266–301 (2011).
437. Heled, J. & Drummond, A. J. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology* **61,** 138–149 (2012).
438. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research* **43,** D257–60 (2015).
439. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28,** 2731–2739 (2011).

440. Eisen, J. A. & Fraser, C. M. Phylogenomics: intersection of evolution and genomics. *Science* **300,** 1706–1707 (2003).
441. Dobzhansky, T. Nothing in Biology Makes Sense except in the Light of Evolution. *The American Biology Teacher* **35,** 125–129 (1973).
442. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164,** 1645–1656 (2003).
443. Wall, J. D. Estimating ancestral population sizes and divergence times. *Genetics* **163,** 395–404 (2003).
444. Lemmon, A. R., Brown, J. M., Stanger-Hall, K. & Lemmon, E. M. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology* **58,** 130–145 (2009).
445. Roure, B., Baurain, D. & Philippe, H. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol* **30,** 197–214 (2013).
446. Tang, H. *et al.* An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics* **15,** 312 (2014).
447. Sato, S. *et al.* Genome structure of the legume, Lotus japonicus. *DNA Res.* **15,** 227–239 (2008).
448. Posada, D. Phylogenetic models of molecular evolution: next-generation data, fit, and performance. *J Mol Evol* **76,** 351–352 (2013).
449. Gadagkar, S. R., Rosenberg, M. S. & Kumar, S. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **304B,** 64–74 (2005).
450. Cummings, M. P. *et al.* Comparing bootstrap and posterior probability values in the four-taxon case. *Systematic Biology* **52,** 477–487 (2003).
451. Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Publishing Group* **1,** 0126 (2017).
452. Copetti, D. *et al.* Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proceedings of the National Academy of Sciences* **114,** 12003–12008 (2017).
453. Bergsten, J. A review of long-branch attraction. *Cladistics* **21,** 163–193 (2005).
454. Poe, S. & Swofford, D. L. Taxon sampling revisited. *Nature* **398,** 299–300 (1999).
455. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat Genet* **48,** 1077–1082 (2016).
456. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol Biol Evol* **28,** 2239–2252 (2011).
457. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol* **32,** 244–257 (2015).
458. Gatesy, J., DeSalle, R. & Wahlberg, N. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Systematic Biology* **56,** 355–363 (2007).
459. Veeckman, E., Ruttink, T. & Vandepoele, K. Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. *Plant Cell* **28,** 1759–1768 (2016).

460. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evolutionary Biology* **7 Suppl 1,** S2 (2007).
461. Arenas, M. Trends in substitution models of molecular evolution. *Front Genet* **6,** 319 (2015).
462. Simion, P. *et al.* A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr Biol* **27,** 958–967 (2017).
463. Simmons, M. P. & Gatesy, J. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Mol. Phylogenet. Evol.* **91,** 98–122 (2015).
464. Mallo, D., De Oliveira Martins, L. & Posada, D. SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology* **65,** 334–344 (2016).
465. DeWitt, W. S., Mesin, L., Victora, G. D., Minin, V. N. & Matsen, F. A. Using genotype abundance to improve phylogenetic inference. *Mol Biol Evol* **338,** 67 (2018).
466. Zhao, T. & Schranz, M. E. Network approaches for plant phylogenomic synteny analysis. *Curr Opin Plant Biol* **36,** 129–134 (2017).
467. Zhao, T. *et al.* Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation. *Plant Cell* **29,** tpc.00312.2017–1292 (2017).
468. Ruprecht, C. *et al.* Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *Plant J* **90,** 447–465 (2017).
469. De Smet, R., Sabaghian, E., Li, Z., Saeys, Y. & Van de Peer, Y. Coordinated Functional Divergence of Genes after Genome Duplication in Arabidopsis thaliana. *Plant Cell* **29,** tpc.00531.2017–2800 (2017).
470. Pérez-Rodríguez, P. *et al.* PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research* **38,** D822–7 (2010).
471. Ibrahim, R. K. *et al.* Enzymology and compartmentation of polymethylated flavonol glucosides in chrysosplenium americanum. *Phytochemistry* **26,** 1237–1245 (1987).
472. Mouradov, A. & Spangenberg, G. Flavonoids: a metabolic network mediating plants adaptation to their real estate. *Front Plant Sci* **5,** 666 (2014).
473. Duan, D. *et al.* Genetic diversity of stilbene metabolism in Vitis sylvestris. *J. Exp. Bot.* **66,** 3243–3257 (2015).
474. Schnee, S., Viret, O. & Gindro, K. Role of stilbenes in the resistance of grapevine to powdery mildew. *Physiological and Molecular Plant Pathology* **72,** 128–133 (2008).
475. Zhang, D. *et al.* The cysteine protease CEP1, a key executor involved in tapetal programmed cell death, regulates pollen development in Arabidopsis. *Plant Cell* **26,** 2939–2961 (2014).
476. Lu, H. *et al.* Subfamily-Specific Fluorescent Probes for Cysteine Proteases Display Dynamic Protease Activities during Seed Germination. *Plant Physiol* **168,** 1462–1475 (2015).
477. Honys, D. & Twell, D. Transcriptome analysis of haploid male gametophyte development in Arabidopsis. *Genome Biol* **5,** R85 (2004).
478. Verelst, W., Saedler, H. & Münster, T. MIKC* MADS-protein complexes bind motifs enriched in the proximal region of late pollen-specific Arabidopsis promoters. *Plant Physiol* **143,** 447–460 (2007).

479. Adamczyk, B. J. & Fernandez, D. E. MIKC* MADS domain heterodimers are required for pollen maturation and tube growth in Arabidopsis. *Plant Physiol* **149,** 1713–1723 (2009).
480. Wolter, M. & Schill, R. Ontogenie von Pollen, Massulae und Pollinien bei den Orchideen. (1986).