The evolutionary significance of gene and genome duplications

Kevin Vanneste

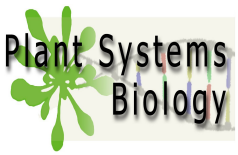
Promoter: Prof. Dr. Yves Van de Peer
Co-Promoter: Prof. Dr. Ir. Steven Maere

Ghent University (UGent) / Flanders Institute for Biotechnology (VIB)
Faculty of Sciences
UGent Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology
Bioinformatics and Systems Biology Research Division

This research was made possible by an Aspirant Fellowship from the Fund for Scientific Research Flanders (FWO-Vlaanderen).

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor (PhD) in Sciences: Bioinformatics.

Academic year: 2013-2014



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 (co-promoter)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Olivier De Clerck*

Faculty of Sciences, Department of Biology, Ghent University

Prof. Dr. Koen Geuten*

Faculty of Sciences, Department of Biology, KU Leuven

Prof. Dr. Eric Schranz*

Faculty of Science, Biosystematics Group, Wageningen University

Dr. Guy Baele*

Faculty of Medicine, Department of Microbiology and Immunology, KU Leuven

** Member of the reading commission*

Acknowledgements

“I don’t know half of you half as well as I should like,
and I like less than half of you half as well as you deserve.”

Bilbo Baggins (culinary expert, recreational dragon slayer),

“The Fellowship of the Ring”

This applies to, but is not limited to, my two promoters (for taking the time to guide this work - thanks for the patience, I can only imagine how much it must have taken at times), the members of my PhD jury (for taking the time to evaluate this work - thanks for the positive feedback), the FWO (for taking a huge leap of faith - thanks for the funding), my colleagues (for proving that choosing the right lab can make a difference - thanks for the professional guidance, the overall positive atmosphere, and perhaps most importantly, making the inner nerd in me much stronger), and the IT staff (for providing essential support - thanks for forgiving me when I crashed a server here and there).
Also my family (for learning me at a young age that life was going to be an uphill battle - thanks for the wisdom), my heroes (for giving me the perseverance to fight that battle - thanks for the courage), my friends (for being there and forgiving my personal shortcomings as a friend while embracing the strong points, whether I’ve known you for several years or just a few months - thanks for the heart), and my girlfriend Lieze (for you being you - thanks for showing the road).

Contents

List of figures

List of tables

Abbreviations

K_N	number of non-synonymous substitutions per non-synonymous site
K_S	number of synonymous substitutions per synonymous site
A	Adenine
AA	Amino Acid
AIC	Akaike Information Criterion
APG	Angiosperm Phylogeny Group
BEB	Bayes Empirical Bayes
BLAST	Basic Local Alignment Search Tool
C	Cytosine
cDNA	Complementary DNA
CDS	CoDing Sequence
CI	Confidence Interval
DDC	Duplication-Degeneration-Complementation
DNA	DeoxyriboNucleic Acid
ENCODE	Encyclopedia of DNA Elements
ESS	Effective Sample Size
EST	Expressed Sequence Tag
G	Guanine
GA	Genetic Algorithm
Gb	Giga-base
GPU	Graphics Processing Unit
GTR	General Time Reversible
HPC	High-Performance Computing
IAD	Innovation-Amplification-Divergence
JTT	Jones-Taylor-Thornton
K-Pg	Cretaceous-Paleogene
k_{cat}	catalytic constant
K_m	Michaelis dissociation constant
KDE	Kernel Density Estimation
KT	Cretaceous-Tertiary

LBA Long Branch Attraction

LG Le-Gascuel

lnL likelihood

Mb Mega-base

MCMC Markov Chain Monte Carlo

ML Maximum Likelihood

MLE Maximum Likelihood Estimation

mRNA messenger-RNA

MSA Multiple Sequence Alignment

mya million years ago

NGS Next Generation Sequencing

ORF Open Reading Frame

RGL Reciprocal Gene Loss

RNA RiboNucleic Acid

RRT Relative Rate Test

SNP Single Nucleotide Polymorphism

SSD Small-Scale Duplication

T Thymine

U Uracil

UCED UnCorrelated Exponential Distribution

UCLD UnCorrelated Lognormal Distribution

WAG Whelan And Goldman

WGD Whole Genome Duplication

Glossary

Allopolyploid: Polyploids that result from the merger of different species.

Angiosperm: A plant whose ovules are enclosed in an ovary (flowering plant).

Autopolyploid: Polyploids that result from the merger of the same species.

Collinear: The conservation of both gene content and order within homologous regions.

Cytotype: Refers to the chromosomal factor of one individual compared to another (e.g., haploid versus diploid).

Developmental plasticity: A single genotype's ability to alter its developmental processes and phenotypic outcomes in response to different environmental conditions.

Diploid: A cell or an organism consisting of two sets of chromosomes.

Eudicots: A group of flowering plants whose seeds typically contain two embryonic leaves.

Evolutionary spandrel: A trait that originated as the byproduct of constraints on the development of other traits and typically receives some secondary functionality that can be mistaken for primary functionality in the absence of knowledge of the constraints that gave raise to the spandrel.

Fractionation: The process of gene loss from homeologous genomic regions after a whole genome duplication.

Gametophyte: The sexual and usually haploid phase in the life cycle of plants with alternating generations that produces the gametes from which the zygote and sporophyte arises.

Genetic drift: Change in the frequency of alleles in a population between different generations due to stochastic events related to population structure and size.

Haploid: A cell or an organism consisting of a single set of unpaired chromosomes.

Heterosis: The greater fitness of a hybrid individual carrying different alleles of genes relative to either of the two corresponding homozygous parents. Also called hybrid vigour.

Homeolog: A gene created by a whole genome duplication.

Homolog: A gene related to a second gene by descent from a common ancestral DNA sequence.

Hybrid: An offspring resulting from the cross between parents of different species.

Macro-evolution: Major evolutionary change, especially with regard to the evolution of whole taxonomic groups over long periods of time.

Mendelian genetics: A set of theories that attempts to explain inheritance and biological diversity according to the tenets of Gregor Mendel regarding the transmission of genetic characters from parent organisms to their offspring.

Micro-evolution: Evolutionary change within a species or small group of organisms, especially over a short period.

Minority cytotype disadvantage: A frequency-dependent reproductive disadvantage in polyploids caused by the fact that ineffective matings with the diploid progenitor majority cytotype result in a net loss of reduced $2n$ gametes that are not available to form new polyploids.

Monocots: A group of flowering plants whose seeds typically contain one embryonic leaf.

Non-synonymous substitution: A substitution in a codon that changes the amino acid that the codon codes for.

Ohnolog: A gene created by a whole genome duplication.

Ortholog: A gene created by a speciation event.

Paralog: A gene created by a duplication event.

Polyploid: A cell or an organism having more than twice the haploid number of chromosomes.

Pre-adaptation: An adaptation which serves a different purpose from the one for which it evolved.

Pseudogene: A defective segment of DNA that resembles a gene but cannot be transcribed.

Punctuated equilibrium: A theory that postulates that evolution proceeds by long periods of relative stasis interspersed with short periods of drastic changes where many species become extinct and new species emerge.

Saltational process: A process whereby the changes between different generations of a population are much more sudden and pronounced than can be explained by selection on standing genetic variation.

Segregation load: Reduction in fitness caused by the inability of a sexually producing population to be composed entirely of heterozygotes even when these genotypes are the most fit.

Specificity constant: A measure for the efficiency of an enzyme.

Sporophyte: The dominant asexual and usually diploid phase in the life cycle of plants with alternating generations that produces the spores from which the gametophyte arises.

Synapomorphy: A characteristic present in an ancestral species and shared exclusively (in more or less modified form) by its evolutionary descendants.

Synonymous substitution: A substitution in a codon that does not change the amino acid that the codon codes for. Also called a silent substitution.

Syntenic: The conservation of gene content within homologous regions.

Transgressive segregation: The formation of extreme phenotypes that are observed in segregating hybrid populations when compared with parental lines.

Chapter 1

Introduction and research goals

“Most people who travel look only at what they are directed to look at.
Great is the power of the guidebook maker, however ignorant.”

John Muir (Scottish-American naturalist),

“Travels in Alaska”

For the author contributions, see page 30.

1.1 The structure of DNA and the central dogma

1.1.1 The molecule of life

DNA (deoxyribonucleic acid), the molecule that carries the blueprint of life, had its structure first described in 1953[1]. A concise overview is presented in figure 1.1. DNA resides in the cell nucleus and consists out of a sequence of four different nitrogenous bases (nucleobases): adenine (A), cytosine (C), guanine (G), and thymine (T). These nucleobases are positioned sequentially on two long anti-parallel polymer strands in a double helix configuration where hydrogen bonds link complementary bases on opposite strands: A with T and C with G. T and C are pyrimidines (heterocyclic aromatic compounds that consist out of a pyrimidine ring), while A and G are purines (heterocyclic aromatic compounds that have a imidazole ring fused to the pyrimidine ring), so that a purine is always paired with a pyrimidine.

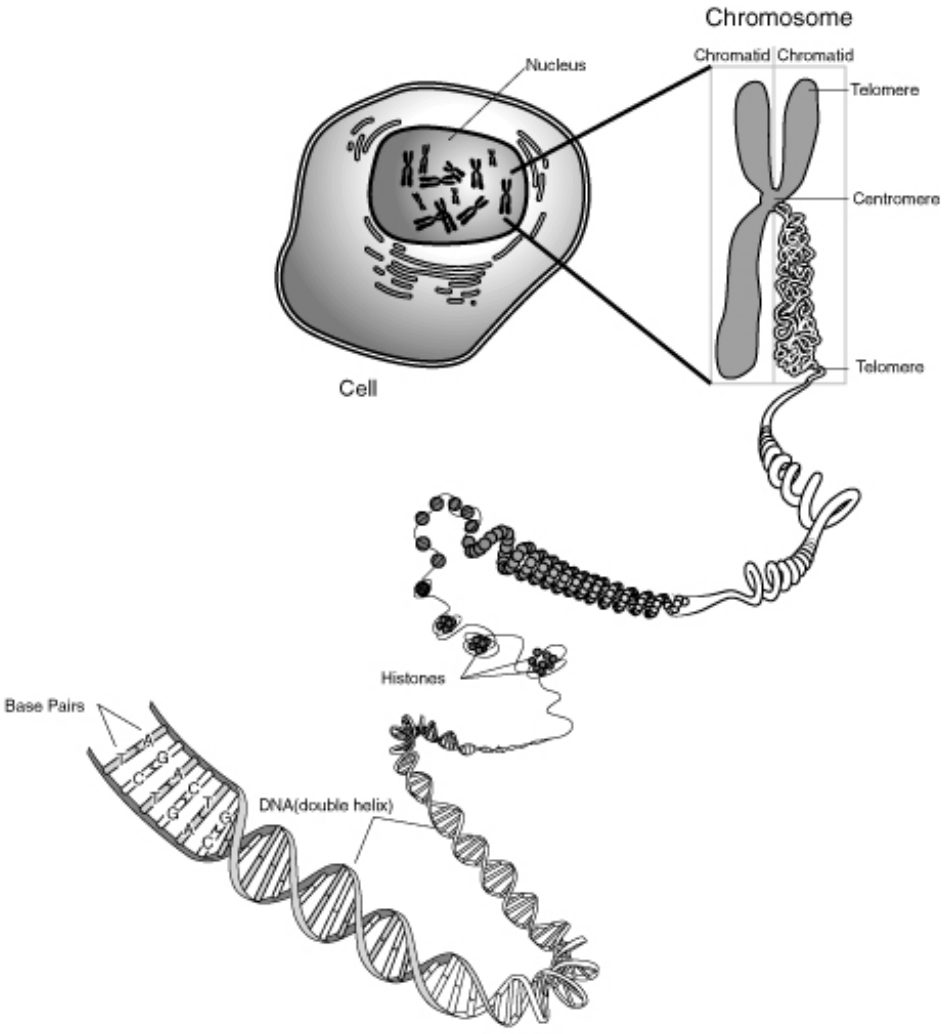


Figure 1.1: The structure of chromatin. The structure of chromatin is depicted, from the four nucleobases that make up the double helix to the dense chromosomal structures during metaphase. Picture from Wikimedia Commons.

The DNA molecule does however not appear in this naked state but instead is configured in nucleosomes, which consist out of this double helix wrapped around a histone protein complex, so that

the total resembles a 'beads-on-a-string' configuration (see figure 1.1). Entry and exit of the DNA onto a histone complex is taken care off by a linker histone (H1), which together with the nucleosome forms the chromatosome. Through addition of H1, this beads-on-a-string configuration coils into a 30 nm diameter helical structure, which is known as the 30 nm filament or simply the chromatin. The chromatin is distributed over several separate molecules, the chromosomes. Several copies of the chromosomes can be present, but most eukaryotes have a double set of chromosomes, also referred to as a diploid state. The dense chromosome structure as depicted in figure 1.1 is only visible in the metaphase of cell division, where each chromosome is duplicated to provide a copy to each daughter cell. Such metaphase chromosomes consist out of two chromatids (each one copy of the duplicated chromosome) that are joined at the centromere, while the terminal arms are called the telomeres[2].

1.1.2 The central dogma of molecular biology

The central dogma of molecular biology was formulated not long after the description of the DNA structure, and explains the flow of sequence information from DNA to proteins[3,4]. This is concisely illustrated in figure 1.2. DNA information is transferred to the next generation of cells (or individuals) by the process of replication. The hydrogen bonds between both strands are broken after which each strand serves as a template for the creation of its own new anti-parallel strand by DNA polymerases to ensure that each DNA copy contains exactly the same information. The information contained within the genes, functional units of hereditary information within the DNA sequence that exert certain functionality, is transferred outside the cell nucleus via an intermediate information transfer molecule, the RNA (ribonucleic acid). RNA is generated from DNA by the process of transcription. RNA polymerases temporarily break the hydrogen bonds between both DNA strands, after which either strand can serve as a template for the creation of a single-stranded RNA molecule. This RNA molecule thus carries exactly the same sequence information as contained within the DNA, with the exception that the nucleobase uracil (U) replaces T. The RNA molecule can migrate to the cell cytoplasm where its sequence information is transferred to proteins by the process of translation. Ribosomes produce proteins based on the RNA sequence information because every triplet of three bases within the RNA, also referred to as a codon, corresponds to one amino acid that is built into the protein sequence. The codon table for most diploid eukaryotic species is depicted in figure 1.3. Proteins are responsible for a wide variety of functionality in the cell, ranging from enzymatic activity to cellular signalling and structural roles. The central dogma thus explains how to blueprint contained within the DNA leads to functionality of the cell[5].

1.1.3 The central dogma has expired

If the view presented above seems simplistic, that is because it largely is. Numerous elaborations to the central dogma have been described[6]. Both extensive post-transcriptional and post-translational changes take place. RNA molecules are typically not transferred to the cytoplasm immediately after transcription, but first undergo extensive changes that produce messenger-RNA (mRNA). RNA-splicing removes the bases of the RNA that will not be translated into amino acids because they are part of the DNA intragenic regions (introns) so that only the expressed regions (exons) are retained. Differential splicing is an intrinsic property of biological eukaryotic systems that allows biological regulation by producing alternative

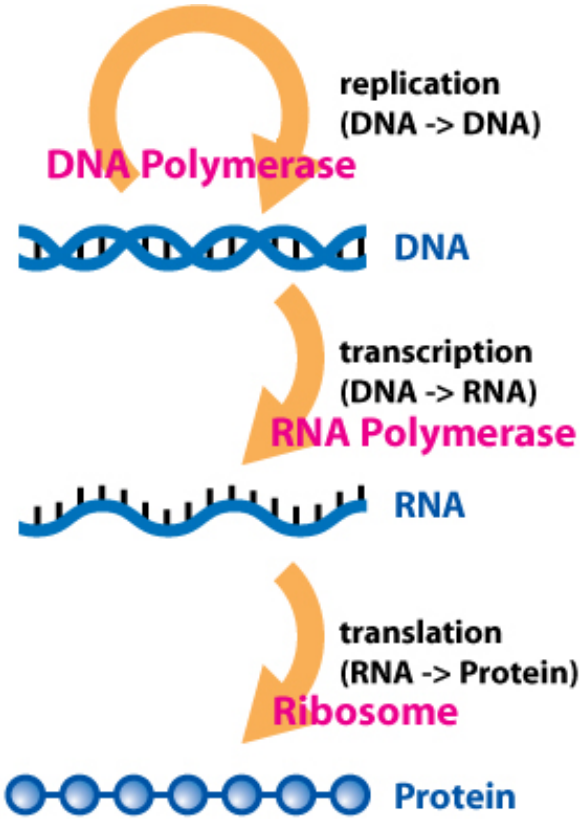


Figure 1.2: The central dogma. The unidirectional information flow from DNA to proteins is depicted, along with the molecules and processes by which this is mediated. Picture from Wikimedia Commons.

transcripts with modified functionality based on the same DNA sequence[7]. RNA editing can even allow to alter the sequence of the RNA transcript itself[8]. Proteins themselves can undergo a particularly diverse set of modifications after translation, including amongst others, phosphorylation, acetylation, methylation, and proteotic cleavage, which all can modify their precise functionality[9].

Apart from these elaborations, numerous contradictions to the unidirectional flow of sequence information and functionality of their carrier molecules have also been described[6]. Non-coding DNA also contains much biologically meaningful information that can enhance or repress transcription[10]. Chromatin states do not influence DNA sequence information but can control access to it, so that these 'epigenetic' marks can have a profound impact on which parts of the DNA can be transcribed[11]. Reverse transcriptases allow to copy RNA into DNA[12]. The role of RNAs is not limited to information transmission, since many small RNA molecules play an important role in cellular regulation, such as micro-RNAs and small inhibitory RNAs that can control chromatin structure, transcription, and translation[13]. RNAs can also have catalytic functions analogous to those of proteins[11].

Lastly, the continuity of DNA sequence information over different generations in the central dogma is a strong oversimplification because many small mistakes can happen during replication, including substitutions, insertions, and deletions of bases[14]. Nevertheless, the central dogma provides a good starting point for this dissertation as most analyses typically focus on the evolution of protein-coding genes, without having to touch upon many of the elaborations and contradictions mentioned above.

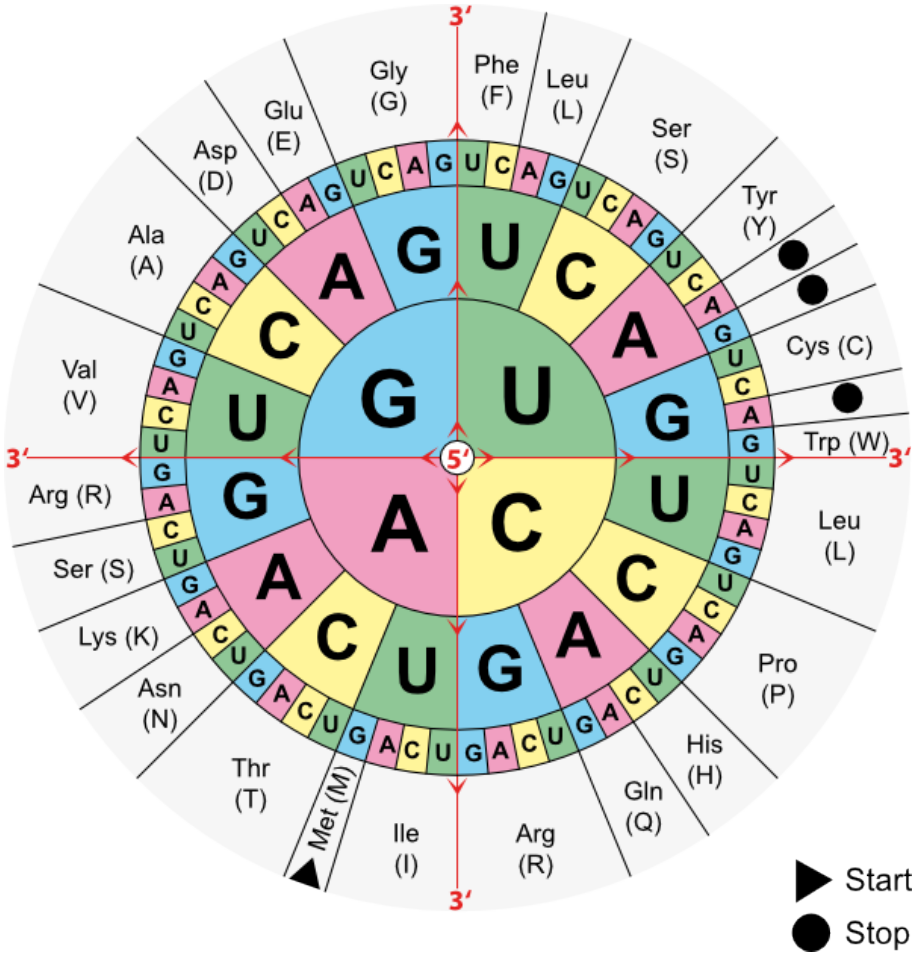


Figure 1.3: The genetic code. The general codon table that applies to most eukaryotes is illustrated. Each triplet of bases (a codon) codes for a particular amino acid. The circle should be read starting from the centre outwards, and demonstrates that the genetic code is redundant, i.e., several codons code for the same amino acid. Start and stop codons, which initiate and terminate transcription respectively, are also indicated. Picture from Wikimedia Commons.

1.2 Nothing makes sense except in the light of evolution

The above title refers to the title of a famous essay from 1973 by the renowned evolutionary biologist Theodosius Dobzhansky[15]. Although this quote has been used very widely in journal articles, dissertations, and powerpoint presentations at conferences and symposia to the extent I would almost be inclined not to use it myself, I do so for one simple reason, namely the validity of its statement. Whether considering biodiversity (e.g., taxonomy etc.), ecology (e.g., species competition etc.), or functional biology (e.g., physiology and molecular biology), evolution represents the underlying theme that can explain much of the observations done in those respective fields, and also represents the backbone I wish to use for this dissertation. This is why this section contains a concise introduction into some major theories that have dominated evolutionary thinking since the paradigm shift introduced by Darwin in the 19th century. This overview is by no means supposed to be a history lesson (or even very complete for that matter), as it would be rather difficult to summarize all the work that has been produced in just a few paragraphs, but rather serves to introduce some important evolutionary principles, many of which will return in later chapters.

1.2.1 Darwinism

In his seminal book "On the origin of species"[16], Charles Darwin introduced many of the conceptual pillars that have dominated evolutionary thinking ever since. Whole books have been written about his work, but two major points deserve attention[17]. First, Darwin proposed that the immense biodiversity witnessed on earth can be explained by evolutionary change through 'descent with modification'. Influenced by Lyell's work in the "Principles of Geology"[18], where the latter described gradual changes over extended periods of time as explaining current geological phenomena, Darwin recognized that small gradual changes between subsequent generations of a population on a micro-evolutionary scale lead to the observed patterns of species biodiversity on a macro-evolutionary scale. Hence, micro-evolutionary changes result in macro-evolutionary phenomena. This proved very controversial in a time period dominated by a theologically inspired *scala naturae*, wherein an order of complexity between different species was recognized but the steps between the ladder considered as static without any possibility of change[19].

Second, Darwin also provided the primary mechanism that explains this evolutionary change, namely 'natural selection', which ensures that in every generation of new individuals only the most fit (i.e., most adapted to their environment) survive to adulthood and reproduce. Because unfit individuals cannot survive and reproduce, their unfavourable characteristics are slowly purged from the population. On the other hand, selection for certain favourable traits over different generations increases the frequency of these traits in the population (i.e., more individuals possess them) and can also make them more expressed (i.e., the traits themselves become more pronounced). Darwin placed an important emphasis on the gradualism of this process, namely many small favourable changes accumulate slowly over subsequent generations (e.g., small phenotypic changes in a population), so that over long periods of time they give rise to more drastic changes (e.g., the origin of a new reproductively isolated species). The mechanism of natural selection thus explains how small micro-evolutionary changes gradually result in macro-evolutionary phenomena. Although Darwin regarded natural selection as the most important mechanism, he did acknowledge that other more elusive mechanisms probably are also at play[20].

1.2.2 Neo-Darwinism and the modern synthesis

Darwin did however not know how traits were passed on to the next generation, but instead assumed a model of blended inheritance, wherein all traits of the two parents were blended into the offspring. The beginning of the 20th century saw the rediscovery of Mendelian genetics by prominent early century geneticists such as Correns, de Vries, and von Tschermak. Early Mendelian genetics were however considered incompatible with Darwin's work because it was difficult to explain how a single mutation that invokes a favourable change could spread to a whole population instead of being just blended out. A seminal paper by Ronald Fisher[21] demonstrated however by means of mathematical modelling that natural selection acting on a whole population obeying Mendelian inheritance was compatible with both ideas and therefore reconciled both theories, ushering in the era of population genetics, often referred to as neo-Darwinism.

Neo-Darwinism gave rise to the modern synthesis in the middle of the 20th century, fuelled to a large extent by the development of the central dogma (see 1.1.2). As both the molecule of life and its structure became known, they led to a small revolution in evolutionary thinking because it was thought

the mechanism by which gradual change takes place, as advocated by Darwin, was finally understood. The central dogma provided a universal platform on which descent with modification could be understood for all species, also referred to as the 'unity of life' by Dobzhanksy[15]. Small point mutations and larger changes such as deletions and insertions of short DNA stretches were viewed as rare errors in DNA replication that represent however a constant source of genetic variation, and consequently phenotypic variation, in the population. Beneficial mutations that lead to phenotypes more fit under the environment are indirectly selected for through the direct action of natural selection on the phenotype. Because such individuals are more fit and hence better able to survive and produce offspring, beneficial mutations (and their effect on the phenotype) can spread through the population. The subsequent accumulation of many small beneficial mutations on a micro-evolutionary scale over long periods of time eventually gives rise to larger macro-evolutionary changes. Deleterious mutations occur but are efficiently purged from the population because those individuals are not fit enough to survive until maturity and hence cannot reproduce. Note that some view a clear distinction between early 20th century neo-Darwinism and the later elaborations fuelled by the central dogma in the mid-20th century modern synthesis[17], while others mention them in the same breath without a clear distinction[22].

Because the central dogma provided an ideal atomic basis for Darwin's work, an adaptionist thinking centred largely around natural selection began to prevail[23]. Genomes were seen as well-organized libraries of hereditary information whose DNA sequence was strongly shaped by natural selection. The species was considered the durable unit of evolution, of which all aspects were seen as efficient design: a species consists of well-adapted individuals, whose well-adapted organs consist of well-adapted cells that are given form through their well-adapted DNA, whose sequence is shaped by natural selection[24]. Observations not fitting within this framework were often ascribed to trade-offs between two traits (because the two traits cannot be optimized simultaneously), whereas other mechanisms such as genetic drift were acknowledged but often conveniently forgotten (by assuming that the latter only plays a minor role in populations so small that they are likely to go extinct anyway). A counter-reaction developed as illustrated by a seminal paper by Stephen Gould and Richard Lewontin[20], who emphasized that phyletic and developmental constraints may lead to 'evolutionary spandrels' that exist only because of those constraints. Such evolutionary spandrels may then perhaps acquire some secondary functionality because they exist anyway, which may be mistaken for primary well-adapted functionality in the absence of knowledge about the constraints that gave rise to the trait in the first place. The term spandrel derives from an analogy with San Marco's Cathedral in Venice (Italy), where spandrels are the triangular shapes that result primarily as an architectural constraint of fitting a dome upon rounded arches (see figure 1.4). Afterwards, these spandrels received secondary aesthetic functionality by filling them with grandiose biblical scenarios. An example of an evolutionary spandrel are the 'male-mimicking' genitalia of the female spotted hyena[25]. It seems unlikely that their 'mock penis', which is basically an enlargement of the clitoris and birth canal, originated through direct selection for such a trait because it has many adverse effects (birth needs to happen through this very small canal leading to a high death rate of both cubs and mothers, whereas the sight of spotted hyenas mating is also not for the faint of heart). Rather, it appears much more likely that this trait originated as a by-product of selection for female dominance and larger size in this species through enhanced testosterone production. Note that for this particular example, 'secondary

aesthetic functionality' is hard to discern, which is perhaps why the justification of using the analogy of the spandrels of San Marco's Cathedral in biology remains controversial to this day[26].

Figure 1.4: Illustration of a spandrel. Close-up of one of the spandrels of San Marco's Cathedral in Venice (Italy). The spandrel results as an architectural by-product from fitting a dome upon rounded arches, and was later filled with biblical scenarios. Picture from Wikimedia Commons.

1.2.3 An extended evolutionary synthesis(?)

By the end of the 20th century, some aspects of the selectionist-based view of the modern synthesis were called into question. On a micro-evolutionary scale, the notion of a well-organized DNA sequence was challenged by several discoveries. The early discovery of transposons in maize by Barbara McClintock indicated the existence of mal-adaptive parasitic DNA elements that were able to escape selection[27]. Especially the neutral theory developed by Motoo Kimura[28] became a prominent player. Based on the degeneracy of the genetic code (see figure 1.3), it was recognized that many mutations can occur in codons that do not change the amino acid (i.e., synonymous mutations), and hence are free from natural selection. The latter remains heavily discussed to this day, as for instance codon usage bias may indicate adaptive evolution of the genetic code itself, hinting synonymous mutations may not be as neutral as once thought[29]. The nearly-neutral theory, an extension of the neutral theory that incorporates population size and was developed by Kimura's student Tomoka Ohta[30], demonstrated the importance of genetic drift in smaller populations. On the one hand, slightly deleterious mutations that do not drastically affect the phenotype can spread relatively easily through the population by chance. Beneficial mutations on the other hand require a substantial selective advantage to overcome genetic drift so that many slightly beneficial mutations never reach fixation. The completion of the human genome sequencing project demonstrated the power of genetic drift, as a whopping ~98-99% appeared to consist out of junk DNA[31].

Other genome sequencing projects similarly indicated that much of the genome structure was not due to efficient selection, because many genomes appeared littered with abundant introns and mobile genetic elements that emerged passively in response to population structure[32]. More sophisticated statistical tests were developed[33] that indicated the signature of selection may be more abundant in genomes than expected based on their large proportion of junk DNA[34]. Such results were viewed as a strong argument for selectionist thinking by some[35], but called into question by others[36].

Because much of the countermovement against adaptionist thinking based on natural selection was centred around the neutral theory and genetic drift, this has led for many biologists to a certain dichotomy in their thinking as considering evolution either 'neutral or adaptive', which this dissertation will also conveniently adhere to. However, neutral evolution is not the only non-selectionist process that can explain micro-evolutionary phenomena. In a much applauded book by Mary Jane West-Eberhard[37], she made a very strong case for developmental plasticity as an intrinsic property of biological systems that can lead to micro-evolutionary changes, which later may be consolidated by rapid genetic changes. In this regard, it is also important to note that many of the most prominent opponents of selectionist thinking did not disregard natural selection, but rather insisted much allelic and phenotypic variation exists that is shaped through other processes[23]. The discussion between 'neutral and adaptive thinking' is still very much alive today. For instance, recent efforts by the ENCODE project, which aims to construct an encyclopaedia of all DNA elements in the human genome[38], indicated an estimated 80% of human DNA is in fact functional. This even led some to argue that there is no such thing as junk DNA[39]. These claims were however swiftly called into question[40].

On a macro-evolutionary scale, the discontinuity of the fossil record where gradual series are often not observed, which traditionally is explained through the imperfection of the fossil record due to the fact that proper conditions for fossilization are the exception rather than the rule, was re-evaluated in light of the punctuated equilibrium theory[41]. The latter describes the discontinuity of the fossil record as a saltational process characterized by sudden changes wherein periods of extremely rapid speciation are alternated with long periods of relative stasis. In other words, the gradualism of the evolutionary process as advocated by Darwin is put into question by postulating that major macro-evolutionary phenomena can originate very rapidly. Micro-evolution is thus essentially decoupled from macro-evolution in this view, in contrast to the Darwinian notion that long periods of micro-evolution lead to macro-evolution. The theory of punctuated equilibrium revisited some ideas first postulated by Richard Goldschmidt[42], who viewed macro-evolution as punctuated speciation events through the creation of 'hopeful monsters'. He hypothesized that the latter could be created by systematic genomic mutations that affect the whole genome, or developmental macro-mutations wherein a small mutation in a developmental gene has drastic consequences on the overall phenotype. Most of these changes would result in hopeless monsters, but once in a while, a successful monster could arise and give rise to a completely new evolutionary lineage. Goldschmidt did thus not only decouple micro-evolution from macro-evolution, he in effect proposed micro-evolutionary processes can only rarely result in macro-evolutionary phenomena, for which he was largely ridiculed at the time[43]. His hopeful monsters were re-evaluated in light of the punctuated equilibrium theory as a potential saltational origin for the essential features of key adaptions, also referred to as pre-adaptations, after which these features can be fine-tuned by rapid genetic changes[23].

Analysis of branching patterns in both animal and plant lineages suggests that major features of their complexity did not arise in a gradual way[44]. Goldschmidt's view on developmental macromutations has received support from the study of homeotic genes that have a major impact on the specification of animal body segment[45] and plant organ identity[46]. In animals for instance, the dorsal shell of turtles consists out of modified ribs with a shoulder girdle inside the rib cage for which no gradual series are available in the fossil record, which is remarkable due to their high chance of fossilization. Detailed analysis suggests that this can largely be attributed towards changes in the expression of a few Hox-genes during development, much in line with a saltational origin[47]. In plants, flower development is under the tight control of only a few homeotic flowering genes, in which changes can also have drastic effects on floral architecture, such as for instance the sudden appearance of the female inflorescence in maize (the "ear")[48]. It has been suggested that plants are ideal candidates for such saltational events because of their vegetative modular additive growth that can be changed more dramatically[49]. Nevertheless, punctuated equilibrium and hopeful monsters remain extremely controversial to this day[50].

Some of the above controversies led to the view by some that the modern synthesis lacked the capacity to adequately explain many of the upcoming and blooming research fields by the beginning of the 21st century. A meeting of 16 prominent evolutionary scientists convened in Altenberg (Austria) in July 2008 to discuss the possibility for an extended evolutionary synthesis that can better incorporate topics that are more difficult to reconcile with the modern synthesis such as evolvability, phenotypic plasticity, epigenetic inheritance, junk DNA, and self-organizing systems[17]. Other evolutionary scientists however strongly contest this notion and emphasize that many of these topics can easily be reconciled with fundamental aspects already present within the modern synthesis[51]. A thorough discussion about the merits of an extended evolutionary synthesis falls however most definitely outside of the scope of this dissertation. Luckily, despite much of the discussion mentioned above, evolutionary biologists remain strongly united by their view that "nothing makes sense except in the light of evolution", although it may be quite difficult to make sense of evolution itself.

1.3 Darwin's abominable mystery

1.3.1 The mystery

'Darwin's abominable mystery' is a term often encountered in evolutionary plant biology, and has been used in a wide variety of contexts: the phylogenetic relationships between and within different clades of flowering plants, the angiosperm fossil record, the angiosperm ancestor, and the evolution of the flower. However, the mystery refers strictly speaking only to the very rapid rise and origin of most major extant angiosperm clades in the mid to late Cretaceous according to the fossil record, as communicated in one of Darwin's letters to his contemporary scientist Joseph Datlon Hooker[19]. For an overview of the geological timescale, see figure 1.5 (the mid-Cretaceous corresponds roughly to 105 million years ago (mya)). This was of great interest to Darwin, not as much driven by his great passion for plant biology, but rather by his realization that the angiosperms represented the single largest threat to his theory of gradual descent with modification. He even stated explicitly on page 189 in "On the origin of species"[16]:

"If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down." Angiosperms are one of the most successful higher eukaryotic clades with over 350,000 estimated extant species of flowering plants[52]. The lack of fossil evidence for the existence of angiosperms in the early Cretaceous, in combination with the fact that many mid-Cretaceous fossils resemble extant taxa, could be interpreted as an incredibly rapid radiation of the angiosperms in line with a saltational origin, which conflicted directly with Darwin's beliefs on the gradualism of evolution[19]. Darwin attributed the sudden and extremely rapid rise of the angiosperms therefore largely to the incompleteness of the fossil record[16], although he later also considered other possibilities such as an undiscovered archipelago where angiosperms evolved for a very long time before rapidly spreading to most other land masses, or a co-evolutionary event with insects that hastened their diversification[19]. Darwin's reluctance to accept a saltational origin for species has been suggested by some to be responsible to a large extent for the reluctance and controversy around hopeful monsters and the punctuated equilibrium theory in the modern synthesis[23].

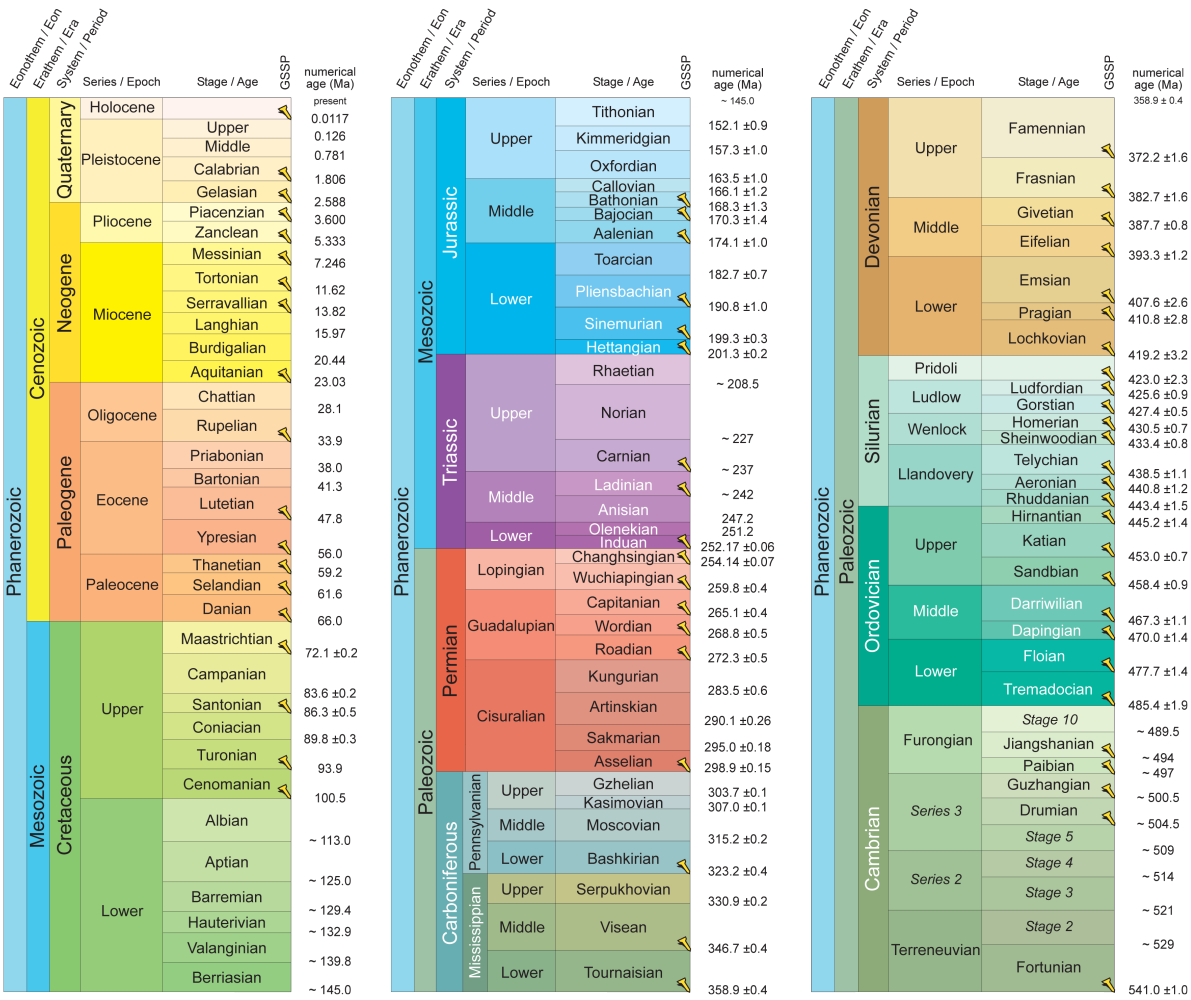


Figure 1.5: The geological time scale. Overview of the geological time scale in the Phanerozoic with all eras, periods, epochs, and stages, together with their age in million years. GSSP refers to Global Boundary Stratotype Section and Point, which is an internationally agreed upon reference point on a stratigraphic section that defines the lower boundary of a stage. The term Tertiary is now officially deprecated, but is still often used in literature to denote the combined Paleogene and Neogene periods. Picture adapted from the International Commission on Stratigraphy, version of January 2013.

1.3.2 The mystery re-visited

There has been significant progress in our understanding of the plant fossil record since the time of Darwin. Advances in the study of leaf fossils, which traditionally were considered to have low systematic value because of their developmental plasticity, indicate a large bloom in angiosperm biodiversity in the mid-Cretaceous. The oldest leaf fossils sharing primitive angiosperm synapomorphies originate from the Aptian (113.0-125.0 mya)[53]. Studies on fossil flowers and fruits, which have high systematic value, similarly indicate an enormous angiosperm biodiversity in the mid-Cretaceous, while older fossils are largely absent[54]. In particular the study of fossil pollen, which is of good systematic value in the absence of other megafossils, has contributed to our understanding of the angiosperm origin. Fossil pollen sharing several primitive angiosperm synapomorphies have been described dating back to the Hauterivian (129.4-132.9 mya) or Valanginian (132.9-139.8 mya)[55], while fossil pollen sharing eudicot synapomorphies appear at the Barremian-Aptian boundary (~125 mya) at several localities that are geographically widespread[56]. Reports of older angiosperm fossils, such as *Archaefructus* from the late Jurassic[57], turned out to be due to radiometric dating errors[58], while possible angiosperm-like fossil pollen from the late-Triassic lack enough deterministic characters to be placed confidently within the angiosperms[59]. Progress in the plant fossil record has thus pushed back the stem of the angiosperms towards ~130 mya, situated in the early Cretaceous, although the rapid radiation of the angiosperm crown group during the mid-Cretaceous remains firmly established. The pattern of angiosperm appearance in the early to mid-Cretaceous shows no strong relationship to the separation of the supercontinent Pangea into its two daughters Laurasia (containing present-day North America and Eurasia) and Gondwana (containing present-day South America, Africa, Madagascar, Australia, Antarctica, the Arabian Peninsula, and India). Although separation of the latter by the Tethyan Ocean was already well underway by the early Cretaceous, the mid-Cretaceous angiosperm radiation seems to have been little affected by the oceanic barrier that the Tethyan Ocean represented. The earliest mid-Cretaceous fossils are associated with the equatorial regions, which were presumably much hotter than present-day equatorial climates, and then show a pattern of polewards dispersal, with high-latitude climates also being much hotter than their present-day equivalents[60].

The rise in molecular dating studies over the last decennia has largely corroborated these fossil results. A diverse series of large-scale dating studies, using different species, methodologies, and fossil information, agree on the rapid radiation of most crown group angiosperms in the mid-Cretaceous[61–67]. More specialized dating studies focusing on particular angiosperm clades also agree on these time estimates, including the fabidae[68], malvidae[69], asterids[70], and most monocot clades[71]. These dating studies are however typically less congruent on their estimates for the stem of the angiosperms, which vary widely from as early as 140 mya in line with an early-Cretaceous origin[63,72], to older than 200 mya[64,65], and all values in between[66].

Both fossil evidence and molecular data thus agree that the mid-Cretaceous represents a period of "layer upon layer of rapid radiation"[63] for the angiosperms, but also indicate that the stem of the angiosperms most likely was already present (long) before the mid-Cretaceous. In a sense, the origin of the angiosperms in Darwin's abominable mystery has thus been resolved (i.e., they were most likely already present long before that time), but their very rapid rise remains very much an abominable mystery

(i.e., the extremely rapid mid-Cretaceous angiosperm radiation is still not in line with gradualism). Many theories exist that try to reconcile the lack of angiosperm evidence before the mid-Cretaceous, such as for instance an origin in isolated freshwater lake-related wetlands from where other habitats were later quickly invaded[73]. Nevertheless, the mid-Cretaceous angiosperm radiation remains a remarkable enigma that is concentrated, in geological terms, on a very short period[74]. Because of this, the mid-Cretaceous angiosperm radiation might perhaps be considered as one of the best examples of a truly saltational event[48], although the mere notion of this remains vividly debated[50]. Figure 1.6 provides a concise overview of angiosperm diversification through time, illustrating the relationships between some of the major plant clades that originated during the Cretaceous.

1.3.3 The Cretaceous-Paleogene extinction event

Since the mid-Cretaceous radiation discussed above, the Cretaceous-Paleogene (K-Pg) extinction event arguably had the largest impact on angiosperm evolution. Note that before the term “Tertiary” was officially deprecated, this event was known as the Cretaceous-Tertiary (KT) extinction, a term which still can be encountered very often. This event constitutes the most recent of the five major mass extinctions recorded in the Phanerozoic eon[78], in which an estimated ∼75% of all species became extinct during a relatively small time period[79]. Despite the well-established narrow timeframe at 66.0 mya for the K-Pg boundary itself, several factors probably contributed to this extinction event for an extended period of time before and after this boundary, such as increased volcanism, greenhouse warming, and in particular a bolide impact near Chicxulub (Mexico)[80]. Recent evidence indicates that this cataclysmic impact led to high levels of infrared radiation in the earth’s higher atmosphere, resulting in worldwide firestorms that set whole ecosystems ablaze, which would have killed off most organisms that could not seek shelter[81]. Nevertheless, the impact of this event on angiosperm evolution was underestimated for a long time because a remarkably large fraction of plant families have survived past the K-Pg boundary[82]. This is in contrast to more obvious changes in the animals, where several large animal lineages went completely extinct, the textbook example being the non-avian dinosaurs[83]. More recent evidence learned however that global dust clouds blocking sunlight and photosynthesis for years after the impact event, in combination with an unstable changing environment for a prolonged period, were especially problematic for stationary plant communities, as evidenced by the extinction of about one-third to three-fifths of plant species[84] and global deforestation[85]. The abundance and dominance structure of angiosperm communities was severely disrupted over a period of several million years after the K-Pg extinction event, and although representatives from most angiosperm families survived, the new angiosperm communities and species that came to dominate during the Cenozoic were drastically different from those in the Cretaceous[82].

After the K-Pg extinction event, angiosperm biodiversity increased throughout the Cenozoic to its present-day observed levels, which is also illustrated concisely on figure 1.6. Angiosperm diversification in the Cenozoic was most likely influenced very heavily by continental drift characteristics, especially so for Gondwana. Land mass movements of Laurasia had less influence on plant diversification characteristics because the Bering Strait often connected both North America and Eurasia, providing Holarctic biota and active routes for dispersal. Gondwana was however mostly characterized by continuing separation of its

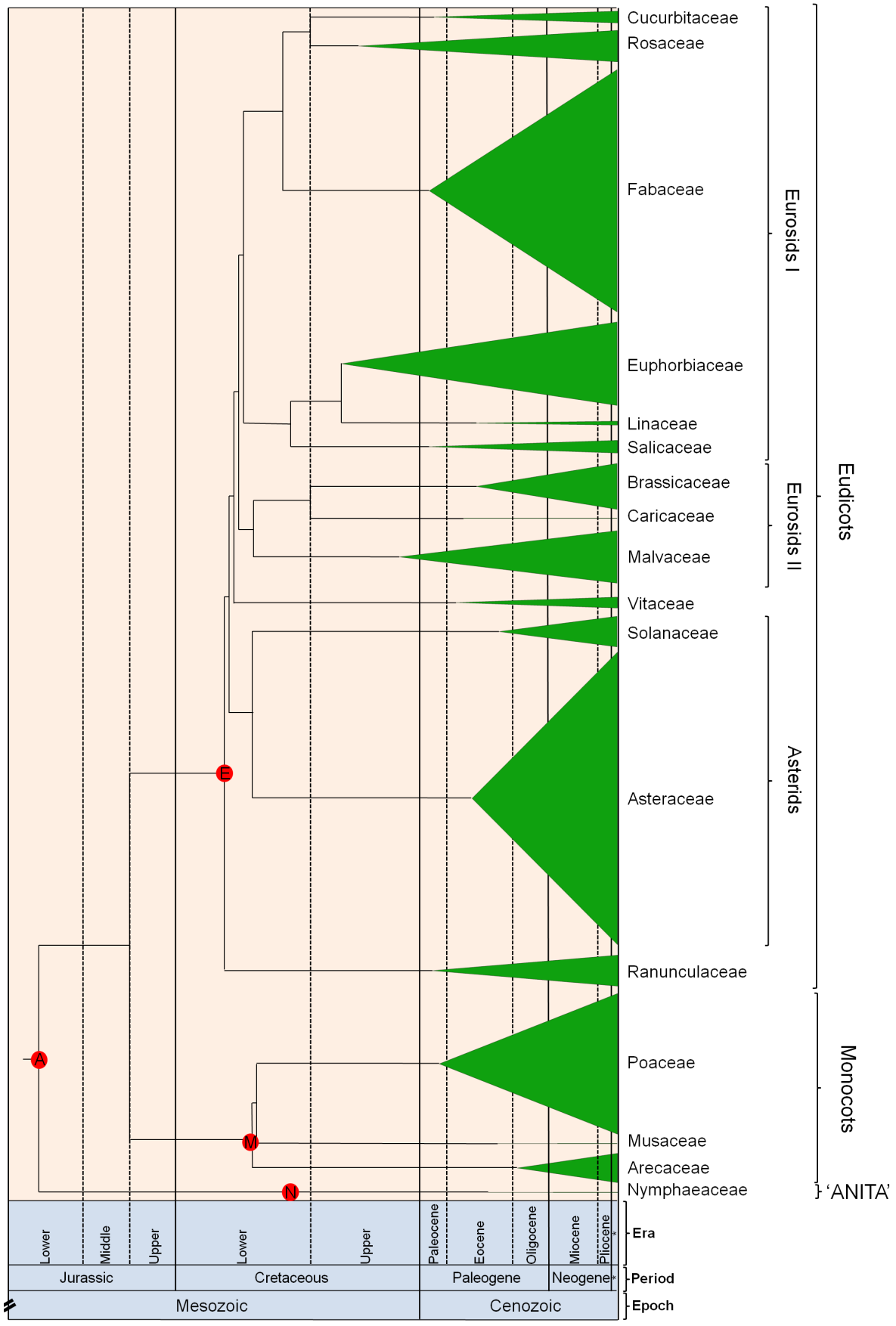

Cucurbitaceae
Rosaceae
Fabaceae
Euphorbiaceae
Linaceae
Salicaceae
Brassicaceae
Caricaceae
Malvaceae
Vitaceae
Solanaceae
Asteraceae
Ranunculaceae
Poaceae
Musaceae
Arecaceae
Nymphaeaceae
Eurosids I
Eurosids II
Asterids
Eudicots
Monocots
'ANITA'
A
E
M
N
Lower
Middle
Upper
Lower
Upper
Paleocene
Eocene
Oligocene
Miocene
Pliocene
Era
Jurassic
Cretaceous
Paleogene
Neogene
Period
Mesozoic
Cenozoic
Epoch

Figure 1.6: Concise overview of angiosperm diversification through time. The relationship between some major angiosperm clades is illustrated on a simplified geological time scale. The full geological time scale in the Phanerozoic is illustrated in figure 1.5 (the asterisk represents the Quaternary period that consists out of the Holocene and Pleistocene epochs). Only angiosperm families that currently have a representative species with a full genome sequence available (or transcriptome assembly - see chapter 4 for overview) are indicated to provide a clear image on their relationships. Names for higher-level clades are indicated on the right of the figure. Age estimates are based on detailed molecular dating studies within the Eurosids I[68,75], Eurosids II[69], monocots[71], and a recent large-scale molecular dating study within the angiosperms[63] for divergence events not described by the former. Green triangles represent the diversification of crown groups, and tips are on scale with the total number of extant species for each family according to estimates from the Missouri Botanical Garden available at `www.mobot.org`. Diversification for each family between the origin of the crown group and the present-day number of species is indicated as a monotonous increase because estimates for rates of diversification are still putative[61], and it remains difficult to detect rate shifts within the angiosperms[76]. The red circles represent critical fossil evidence on the earliest existence for major clades (and are consequently not on scale according to the chronogram): (**A**) fossil pollen sharing several primitive angiosperm synapomorphies from the Hauterivian (129.4-132.9 mya) or Valanginian (132.9-139.8 mya) for the total group of angiosperms[55]; (**N**) a fossil flower from the Nympaeales from the Late Aptian-Early Albian (~113 mya) for the 'ANITA' clade, the latter being a set of the earliest diverging lineages of extant angiosperms[77]; (**E**) fossil pollen sharing several eudicot synapomorphies from the Barremian-Aptian boundary (~125 mya) for the total group of eudicots[56]; and (**M**) fossil flowers from the *Araceae* from the Late Aptian-Early Albian (~113 mya) for the total group of monocots[66].

continental fragments (South America, Africa, Madagascar, Australia, Antarctica, the Arabian Peninsula, and India), which led to greater isolation and independent evolution of endemic biota after the K-Pg boundary. This active isolation most likely spurred plant diversification. Progressive shift later during the Cenozoic northwards of South America, India, and Australia led to regional climatic shifts and adaptive radiations, while the progressive closure of South America and India with North America and Eurasia, respectively, enabled mixing of northern and southern floras[86]. Assessment of ancient angiosperm diversification rates is however not straightforward. There are for instance many examples of particular plant families that expanded very strongly shortly after the K-Pg extinction event, including some very large and particularly successful present-day ones. These include the Orchidaceae[87], Brassicaceae[69], Fabaceae[88], Poaceae[89], and Piperaceae[90]. However, all these examples are anecdotal in nature and do not offer a profound insight into whether angiosperm diversity in general also radiated more strongly right after the K-Pg mass extinction, or rather just experienced a gradual increase throughout the Cenozoic. A thorough understanding of angiosperm diversification rates since the K-Pg boundary is complicated by the fact that analysis of diversification based on present-day biodiversity (which can 'easily' be counted), does not properly account for the total number of species that went extinct (which have to be both properly fossilized and discovered), and therefore may be biased to overestimate diversification rates[91]. In absence of the body of overwhelming evidence for the mid-Cretaceous angiosperm radiation, more sophisticated tools are required that can explicitly deal with this. Such tools have only recently received a boost in attention and development[92] so that few studies are yet available. At least one such study demonstrated that net angiosperm diversification increased markedly in the warm beginning of the Cenozoic (~66-54 mya), before decreasing in the cooler middle and end of the Cenozoic[76]. The latter study was based on a family-level phylogenetic analysis, which may not provide adequate resolution for diversification rates at the genus level[74]. However, coupled with evidence on the individual plant families listed above, this indicates that angiosperms may have experienced a moderate radiation not long after the K-Pg extinction event in which plant community structure recovered, but more research will be required to properly confirm this.

1.4 Gene duplication

Genes evolve through several small errors during replication, but can sometimes also be duplicated in their entirety during this process. The importance of gene duplication in evolution was first emphasized by Susumu Ohno in his seminal book “Evolution by gene duplication”[93]. Ohno proposed that the action of mutation on individual gene loci alone cannot explain the evolution of novel and/or expanded functionality, because this requires the creation of new gene loci with previously non-existent functions. Rather, the duplication of existing gene loci allows for the creation of new ‘raw’ genetic material that can be used to evolve novel and/or expanded functionality. Ohno thus postulated that natural selection merely modified while redundancy created. Genome sequencing in the past decennia has demonstrated that all prokaryotic and eukaryotic genomes are indeed characterized by high numbers of duplicated genes originating from continuous small-scale duplications (SSDs)[94]. Genes created by duplication are also referred to as paralogs, in contrast to genes created by speciation events, which are referred to as orthologs (both share however a common ancestor and are therefore homologs).

There are several molecular mechanisms that can lead to SSDs. The first mechanism is unequal crossing over, which leads to the creation of a new gene copy very close to the original gene, also referred to as a tandem duplicate[95]. The second mechanism is duplicative transposition through non-allelic homologous recombination or non-homologous end joining, which leads to the creation of a gene copy that can be located very far from the original gene on the same or a different chromosome[96]. The third mechanism is retrotransposition through the reverse transcription of mRNA into cDNA that is inserted back into the genome. Such duplicated copies can also be dispersed over the complete genome and are recognizable by their lack of introns. They are however often non-functional because they lack the necessary regulatory sequences that were not included in the cDNA[97]. The fourth mechanism is large-scale duplication, including polyploidy, and will be the subject of the next section (see 1.5). A thorough in-depth overview of these exact molecular mechanisms is out of the scope of this dissertation and can be found in Li[5]. Rather, we will focus on the three main scenarios for evolutionary innovation through gene duplication as envisaged by Ohno, which have stood the test of time remarkably well, together with some of their more complex derived models that have been formulated[98].

1.4.1 Gene conservation

In the first scenario, the duplicated gene copy is kept because it allows to maintain the original gene function, which is therefore known as gene conservation (see figure 1.7)[99]. It is thought to be especially important for initial duplicate retention[100]. Note that immediately after gene duplication, the distinction between the ‘original’ and ‘new’ gene copy is of course largely semantic, but still useful for conceptualizing their evolutionary fates. There are two models that explain gene conservation. The first model focuses on the functional redundancy provided by the duplicate, which serves as a buffer against deleterious mutations in the original gene. It has been proposed that this only plays a minor role[101]. The second model focuses on dosage amplification, which entails the duplicate is kept because it is beneficial to provide more of the original gene product. This probably plays a more important role and many such examples have been described, a typical example being the highly duplicated ribosomal RNA that is

required for quick translation during growth and development[102]. Both the functional redundancy and dosage amplification models of gene conservation predict that the new and original gene copy will remain highly identical in their sequence.

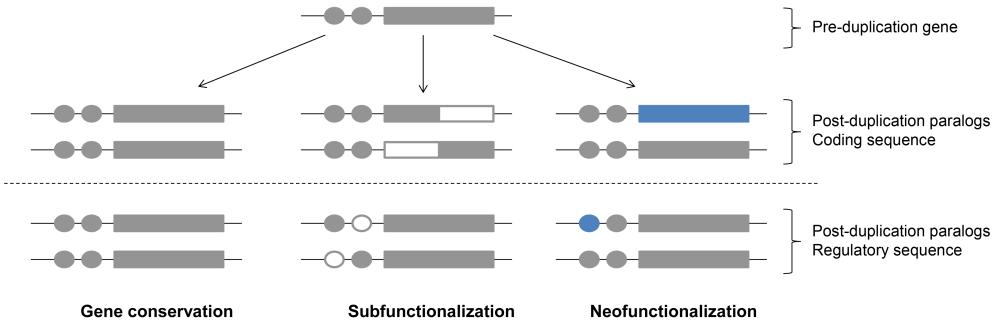


Figure 1.7: Illustration of the three major outcomes after gene duplication. The grey circles and bar represent two regulatory sequences and the coding sequence of the ancestral gene, respectively. The arrows indicate the three major fates of the ancestral gene after duplication. In gene conservation, both the regulatory and coding sequences remain the same. In subfunctionalization, pre-duplication functionality is divided among the daughter paralogs, which can be both on the level of the coding and/or regulatory sequence as indicated by the respective loss of grey colour, so that both copies are required for proper functioning. In neofunctionalization, one of the duplicates acquires new functionality by changes in the regulatory and/or coding sequence, as indicated by the gain of blue colour.

1.4.2 Subfunctionalization

In the second scenario, the duplicated copy is kept because the ancestral functionality is divided over the two daughter copies, a process which is known as subfunctionalization (see figure 1.7)[103]. Several theoretical models exist that make different assumptions about the functionality of the original gene, and the role of adaptive and neutral processes in the evolutionary trajectories of the post-duplication gene copies. In the segregation avoidance model, both gene copies are kept to avoid the segregation load[104]. The latter entails that it can be advantageous to keep a gene locus in a heterozygous condition, but a fraction of descendants will always have a homozygous configuration and hence lose this advantage. If one of the alleles is however duplicated into a new gene copy, a permanent heterozygous state can be attained. This model resembles dosage amplification to some extent, except for the fact that the two duplicates will differ in some critical residues in their sequence.

The “Duplication-Degeneration-Complementation” (DDC) model is the most widely known and unambiguous example of subfunctionalization. It assumes that both gene loci undergo complementary degenerative (i.e., non-adaptive) mutations after duplication so that the ancestral functionality is divided over both daughter copies, which therefore both are required for proper functioning[103]. In the qualitative version of DDC, the original gene locus has two (or more) functions that are divided over both daughter copies. In the quantitative version of DDC, the original gene locus has only one function that is post-duplication carried out by both daughter copies. A typical example of the latter is expression efficiency, which can quickly deteriorate by degenerative complementary mutations in the regulatory sequences of both copies, so that both need to be retained to attain pre-duplication expression levels.

The specialization and gene sharing models are co-option models that are very similar and only differ in their definition of what makes up the ancestral functionality. In the specialization model, the pre-duplication gene has one function that post-duplication is refined and optimized among the daughter paralogs expressed in several tissues or developmental stages (e.g., each paralog becomes more efficient in a specific tissue)[105]. In the gene sharing model, the pre-duplication gene has two or more functions that cannot be optimized jointly because beneficial mutations for one function adversely affect the other[106]. Both models thus assume that the pre-duplication gene has an adaptive conflict wherein one aspect cannot be optimized without negatively affecting the other. Both models also assume that duplication followed by positive selection allows to optimize the new function(s) of both daughter paralogs. Both models differ however in their definition of the original function, which can sometimes be very difficult to assess. Additionally, it can be argued that resolving such an adaptive conflict in fact represents a form of neofunctionalization (see below)[98].

1.4.3 Neofunctionalization

In the third scenario, the duplicated copy evolves a completely novel function, also referred to as neofunctionalization (see figure 1.7). This is regarded as one of the hallmark mechanisms of Ohno, although the exact prevalence of neofunctionalization is still very controversial and it is often regarded as a rather rare phenomenon[98]. In the Dykhuizen-Hartl model, it is assumed that mutations accumulate by neutral genetic drift in the duplicated copy, which is freed from purifying selection because its function is maintained by the original copy. This degenerate copy can however by chance gain a novel function during its random walk, for instance because its gene product suddenly becomes functional due to a drastic environmental shift[107,108]. In the adaptive model, the duplicated copy also goes through a random walk, but it is assumed that the novel functionality is attained through adaptive mutations that are positively selected for[109].

Two particular gene sharing models demonstrate the difficulty in distinguishing between the outlined fates of gene duplicates. The “Escape from Adaptive Conflict” (EAC) model describes gene sharing rather as a subfunctionalization mechanism[110,111]. The pre-duplication gene has two conflicting subfunctions that are independently optimized in either paralogous daughter gene by positive selection. The related “Innovation, Amplification, and Divergence” (IAD) model however describes gene sharing rather as a neofunctionalization mechanism[112,113]. A minor activity arises in the pre-duplication gene, and increased requirement for this (minor) activity is first met by gene amplification (e.g., through formation of tandem arrays). After this, adaptive mutations lead to divergence and specialization of some of the duplicated copies. Both models are gene sharing models because the pre-duplication gene has more than one function and positive selection afterwards drives the evolution of both post-duplication daughter paralogs, but have small differences in how they address the post-duplication functionality. The plethora of theoretical gene models will be addressed in more detail in chapter 2, but the three major scenarios of neofunctionalization, subfunctionalization, and gene conservation, serve as a useful simplification in the meantime.

1.5 Polyploidy

1.5.1 Polyploid formation

Polyploidy is the fourth major mechanism by which genes become duplicated (see 1.4), and is defined as possessing more than two complete sets of chromosomes[114]. Whole genome duplication (WGD) is another term for polyploidy that has become more popular in the genomics era. Genes duplicated by WGD are also often referred to as homeologs or ohnologs. There are several cytological mechanisms by which WGD can occur, which will be explained with a focus on plants because polyploidy is especially abundant there (see below). In somatic polyploidy, WGD occurs in the vegetative sporophyte tissue of plants, i.e., the dominant asexual stage in their life cycle that makes up the vegetative tissues of the plant. This happens typically in wounded tissues or tumours so that the sporophyte will be of mixed ploidy[115], but can also happen in the zygote or young embryo so that the complete sporophyte is polyploid[116]. It is thought somatic polyploidy represents only a minor route towards polyploidization[114]. In polyspermy, an egg is fertilized by more than one sperm nucleus, which is well described for instance in orchids[117], but in general also considered to be only of minor importance[114]. The third mechanism involves gametic non-reduction, also referred to as meiotic nuclear restitution, and is considered the most important route towards polyploidization[114]. It is based on the formation of unreduced $2n$ gametes that did not undergo proper meiosis and therefore are diploid, which form a tetraploid plant when an unreduced egg and pollen meet[118]. Unreduced gametes can be produced through alterations in meiotic spindle morphology and orientation, defects in meiotic cell plate formation, and complete loss of the first or second meiotic division[119]. First and second division restitution refer to the formation of unreduced gametes through such errors in the first and second meiotic cell division, respectively. Empirical estimates of unreduced gamete production in plants vary widely but are relatively high[120,121], from on average 0.56% in non-hybrids to 27.52% in hybrids, the latter resulting from the cross between parents of different species[114]. Despite being considered the major route towards polyploidization, these levels are still seen as restrictive because newly formed tetraploid plants need to cope with the minority cytotype disadvantage, which is a frequency-dependent reproductive disadvantage caused by the fact that ineffective matings with the diploid progenitor majority cytotype result in a net loss of reduced $2n$ gametes that are not available to form new tetraploids (the cytotype refers to the chromosomal factor of one individual compared to another, for instance haploid versus diploid)[122]. In particular, crosses of unreduced $2n$ gametes with reduced n gametes result in triploid hybrids that are frequently less fit and more sterile, also referred to as a 'triploid block'[114,118,123]. Recent modelling approaches that account for the gametic contribution of such triploids, which despite their increased sterility still produce an excess of unreduced $3n$ gametes that can cross with reduced n gametes to form tetraploids, indicate however that this triploid stage may rather represent an intermediate step between the diploid and tetraploid, also referred to as a 'triploid bridge'[124].

Polyploids are categorized as either auto- or allopolyploid depending on their parental species[114]. Autopolyploids result from the merger of the same species. Because of this, their two subgenomes typically pair as multivalents during cell division, which often results in meiotic and mitotic abnormalities. Allopolyploids result from the merger of two different species. Their two subgenomes therefore are genetically more distant so that each subgenome pairs as bivalents and less abnormalities are present.

Segmental allopolyploids exhibit both bi- and multivalent pairing and are considered as a rare intermediate, although they may be more prevalent then originally thought[125]. In fact, genome sequencing has indicated that the traditional cytological definition may not capture all possibilities, as for instance autopolyploids resulting from two genetically very similar parents may demonstrate bivalent pairing, so that the cytological and genetic definition used for auto- and allopolyploidy may differ[126]. This also entails that autopolyploids, traditionally thought to be more rare through abnormalities during cell division[114], are more frequent than anticipated, especially since they often morphologically resemble their parental species and were therefore overlooked in classical sampling studies[127].

1.5.2 Polyploidy is especially abundant in plants

Polyploidy in general appears a more frequent phenomenon in evolution than traditionally appreciated[128]. Several ancient WGDs, referred to as paleopolyploidizations, have been uncovered in most evolutionary lineages. Examples of well-established paleopolyploidizations are illustrated in figure 1.8 and include two rounds of WGD in the vertebrate ancestor with a third one in the teleost fish lineage[129–131], three WGDs in the ciliate *Paramecium tetraurelia*[132], and one WGD in the ancestor of the hemiascomycete *Saccharomyces cerevisiae* after its divergence from the *Kluyveromyces* clade[133,134]. However, especially in the plant lineage a large number of paleopolyploidizations have been uncovered[52,128,135]. It is now commonly accepted that two whole genome duplications occurred in the ancestor of all angiosperms, so that all angiosperms are in fact paleopolyploids[136]. Furthermore, a hexaploidy event predates the origin of all core eudicots, which make up approximately 75% of extant angiosperm diversity[137–139], while traces of a WGD at the base of the monocots also suggest a WGD shared by most, if not all, monocots[140]. In addition, several more recent independent WGDs have been unveiled in many different plant lineages. As a result, the genomes of some extant plant species carry the remains of up to six successive genome duplications[141].

The number of uncovered successful paleopolyploidizations pales however in comparison with the vast amount of species that underwent a recent WGD, referred to as a neopolyploidizations. A very large number of plant species are recent polyploids[126], with an estimated 35% of all vascular plants species being neopolyploids[142]. An especially high number of invasive plant species are neopolyploids, with estimates going up to 50%[143,144]. Many neopolyploids are also found in stressful environments such as the Arctic where they can make up to 80% of all plant species in some regions[145,146]. Many of these estimates however need to be interpreted with due caution, as they can easily be subject to sampling biases because it is very difficult to adequately sample plant biodiversity given their sheer number, so that a proper large-scale systematic framework is still lacking[125]. Examples of neopolyploids in other evolutionary lineages are more anecdotal, but many examples are nevertheless known in the arthropods and lower vertebrate lineages such as amphibians, reptiles, and fish[147].

The overabundance of both neo- and paleopolyploidizations in plants compared to other evolutionary lineages is quite striking and can to some extent be attributed to some of their intrinsic characteristics that favour WGD[148]. They have indeterminate growth during their life cycle, which entails that there is a higher chance that somatic polyploidy can occur, especially so for perennial species. They also frequently exhibit traits such as the loss of self-incompatibility, which enables selfing, and the gain of apomixis, which

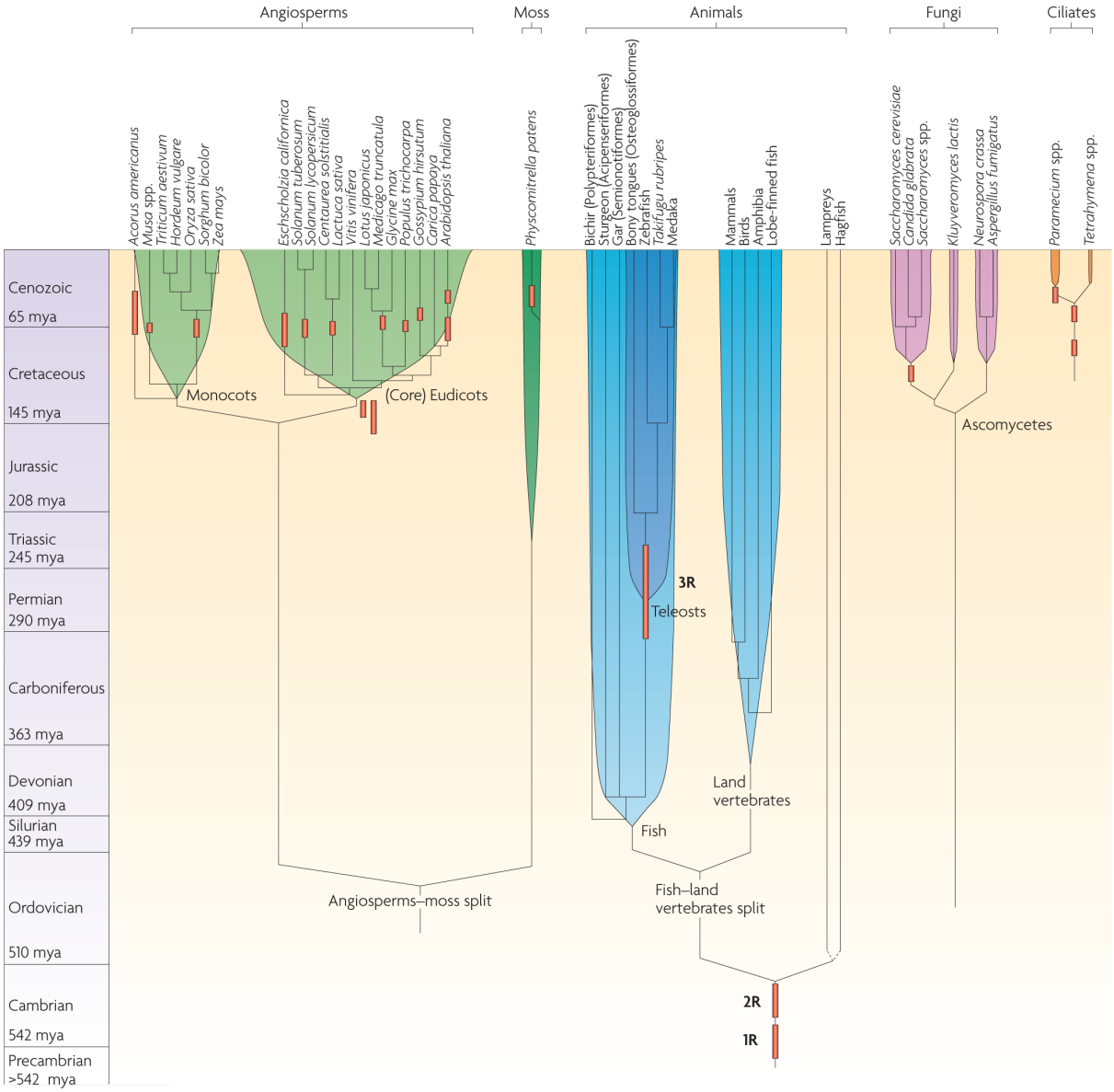


Figure 1.8: Overview of paleopolyploidizations in different evolutionary lineages. Simplified representation of different eukaryotic evolutionary lineages where well-established paleopolyploidizations have been discovered. WGDs are indicated by coloured bars that represent a rough estimate on their age. The double bar at the base of the eudicots in the angiosperms represents the eudicot-shared hexaploidy. Picture adapted from Van de Peer et al.[128].

enables asexual reproduction; and experience a weaker gene flow. These are all characteristics that can help to establish a polyploid plant population in the face of the minority cytotype disadvantage[146]. Consequently, polyploidy is also more strongly associated with asexual reproduction and hermaphroditism in animals, despite the relative paucity of such traits in animals[149].

1.5.3 The long-term fate of polyploids is heavily disputed

The prevalence of both neo- and paleopolyploidizations in several different eukaryotic lineages has been firmly established[150], but the overabundance of neopolyploids compared to the number of known paleopolyploidizations represents an interesting contradiction. One frequently encountered explanation is that neopolyploids still have to stand the test of time and may not survive in the long run[151]. In this

regard, two long-standing opposite views regard polyploidy either as an evolutionary dead end[152,153], or as a road towards evolutionary success[154].

Much research has been dedicated to this topic, especially in the plant lineage because of the high frequency of WGD occurrence there, but studies have typically found support for both scenarios. Recently formed polyploids need to cope with the minority cytotype disadvantage (see before). Although plants display some favourable traits that can mitigate this, the extent to which these traits really alleviate the minority cytotype disadvantage remains largely speculative[155], and it could be that most fit neopolyploids never get the chance to turn into established paleopolyploids because they simply could not overcome the bottleneck of finding enough suitable mating partners to establish a viable population[122]. Recently formed polyploids typically display large meiotic and mitotic abnormalities resulting in genomic instability through improper chromosome pairing, which has detrimental effects on plant fertility and fitness[156]. The study of mutant *Arabidopsis thaliana tam-1* plants that cannot enter meiosis II and therefore increase in ploidy in subsequent generations, suggests that this genomic instability is polyploidy-associated, as *tam-1* plants with higher ploidy levels experience more detrimental effects, resulting in a strong drive to go back to lower ploidy levels via genomic reductions[157]. The combination of genomic plasticity negatively affecting plant fitness and the minority cytotype disadvantage may help to explain why polyploid plant species display lower speciation rates and higher extinction rates compared to diploids, resulting in a lower net diversification rate[158].

The fact that all extant angiosperms and vertebrates are paleopolyploids[131,136] indicates however that polyploidization at the very least does not always constitute a dead end. An estimated 15% and 31% of speciation events in flowering plants and ferns, respectively, were accompanied by a ploidy increase[142]. Most recent insights explaining the evolutionary success of polyploids have focused on their duplicated genome, which simultaneously provides thousands of novel genes for evolution to tinker with. Even though the large majority of these genes are lost through pseudogenization[159], the small remaining fraction can lead to novel and/or expanded functionality through Ohno's classical models of neofunctionalization, subfunctionalization, and gene conservation (see 1.4)[93,98]. Interestingly, a large fraction of retained duplicates are most likely guarded against loss through dosage-balance constraints on the stoichiometry of whole duplicated pathways and/or macromolecular complexes[160], which includes many regulatory and developmental genes[161]. These genes are kept not because they provide an advantage, but rather because their loss could disrupt important pathways and/or macromolecular complexes and therefore would have a negative effect on the phenotype. Resolution of dosage-balance constraints over time can thus provide polyploid species with an important toolbox that can be rewired to execute novel functions[162], and allow them to cope with new ecological opportunities and/or challenges[163]. The ecological conditions that allow the initial establishment and long-term success of polyploids have been a major question in early polyploidy research for a long time, but progress in this regard has shifted somewhat to the background due to the explosion in research on their genomic composition[125]. Recently formed polyploids are traditionally considered to be good colonizers that have a large ecological tolerance, which gives them an adaptive advantage as invasive species[149,164]. Such generalizations should however once more be treated with due caution because of the paucity of large-scale systematic data and the many exceptions that can be found[125].

1.5.4 Inference of WGDs

Inference of WGDs is crucial to understanding their abundance and evolutionary role. Although neopolyploids can relatively easily be identified because their genomes are still in a tetraploid (or higher) state[165], they undergo diploidization over time to return to a diploid state[166] so that advanced computational approaches are required to successfully identify hidden paleopolyploidizations in diploid genomes[167]. There are three widely-applied methods available.

The first method is based on collinearity, i.e., the conservation of gene content and order of large duplicated segments within and between different genomes[168,169]. Within the same species, despite extensive fractionation (the loss of duplicate genes) and chromosomal rearrangements after WGD[170], several duplicated segments can typically be identified that map to each other all over the genome. Between different species, comparison of these duplicated segments with other genomes where it is well established how many WGDs occurred, can also help to establish paleopolyploid history. An example of how collinearity allows to infer WGD history is presented in figure 1.9. This method is generally quite powerful, but does rely on extensive positional information, which may be problematic for fragmented assemblies in low-coverage sequenced genomes. Collinearity is also more problematic in lineages where several subsequent WGDs occurred, because each successive WGD scrambles the positional information from older WGDs[167].

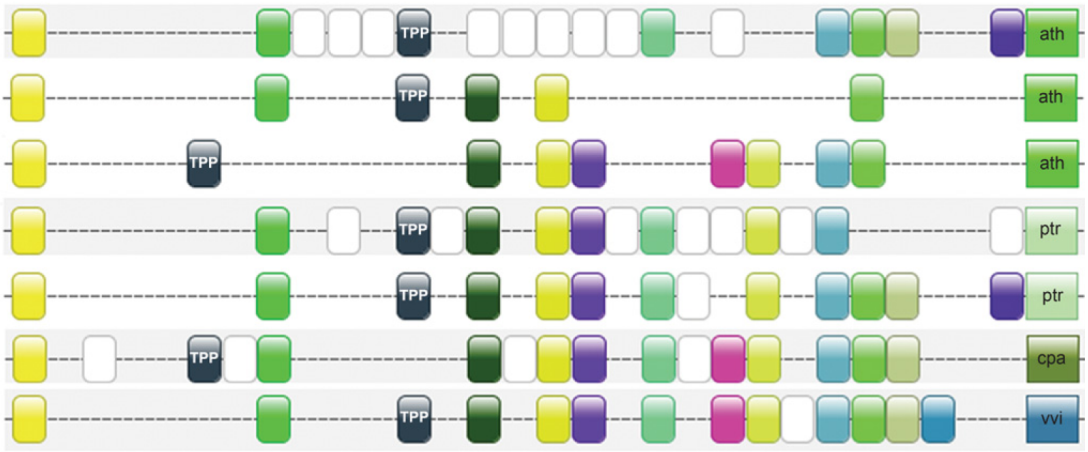


Figure 1.9: Collinearity allows to infer paleopolyploid history. Collinearity illustrated based on one of the *Arabidopsis thaliana Trehalose-6-Phosphate Phosphatase* (*TPP*) genes. The whole figure represents a multiplicon, a collection of duplicated segments that map to each other within and between different species, as identified using the PLAZA v1.0 database[171] with *Arabidopsis thaliana TPPC* as a query gene. Each line represents a duplicated segment, while the boxes represent the genes on these segments. Boxes with the same colour represent members of the same gene family, which were most likely created by a large-scale duplication event and often have preserved gene order. ath, ptr, cpa, and vvi, refer to *Arabidopsis thaliana*, *Populus trichocarpa*, *Carica papaya*, and *Vitis vinifera*, respectively. *Vitis vinifera* has a well-established WGD history with no additional polyploidizations since the eudicot hexaploidy[137]. The duplication ratio for *Vitis*:*Carica*:*Populus*:*Arabidopsis* is 1:1:2:3, indicating no WGDs since the eudicot hexaploidy in *Carica*, one WGD since the eudicot hexaploidy in *Populus*, and two WGDs since the eudicot hexaploidy in *Arabidopsis* of which one segment is lost or not identifiable anymore. Picture adapted from Vandesteene et al.[172].

The second method encompasses tree-based approaches. Phylogenetic trees of gene families within the species of interest and several other genomes are reconstructed to established how many topologies correspond to predefined duplication scenarios[173]. Tree reconciliation methods such as the NOTUNG package[174] allow to compare the individual gene family topologies with the species tree and infer the nodes in the topologies that correspond to duplication and speciation events based on

a parsimony principle. If the majority of gene family topologies contain a node that is labelled as a duplication node for all genes belonging to the same set of species, this node is considered to represent a WGD event in the history of the species tree. Since gene families are however very plastic and can expand or contract very rapidly during evolution, several WGD scenarios in the evolutionary past of the species need to be compared and statistically tested to robustly infer where exactly the WGD occurred on the species tree. This method is especially useful to evaluate very old paleopolyploidizations where collinearity information is not recognizable anymore. This method was for instance used to detect the shared paleopolyploidization among both the seed plants and angiosperms, which occurred respectively ~319 and ~192 mya[136]. Tree-based approaches are however computationally very intensive, and also require extensive sequence information from other species to build reliable topologies for evaluation.

The third method is based on paranome age distributions. These consist out of the contribution of all duplicated paralogous gene families within the same genome plotted against their age of duplication. The latter is based on the number of synonymous substitutions per synonymous site (abbreviated as K_S), which is a proxy for the age of duplicated genes because synonymous substitutions do not change the amino acid and are therefore putatively neutral (see 1.2.2) so that they accumulate changes at a constant rate[175]. Age distributions of duplicates retained from SSDs are L-shaped with many recent duplicates and fewer older duplicates. Additional peaks can be superimposed on the L-shaped background, and represent sudden bursts of new gene duplicates that were created contemporarily by large-scale duplication events such as WGDs[159]. WGDs in the evolutionary past of the species are thus recognizable by superimposed peaks on the L-shaped SSD background distribution. Age distributions are a very popular tool to detect WGDs[52,135,136,138,176–186], because they only require sequence information from the species under investigation without the requirement for positional information. They are consequently computationally also very cheap. Their main disadvantage consists out of the fact that WGD peaks cannot always be unambiguously distinguished from the SSD background, especially so for older events where many duplicates have been lost since[176]. This is why usually mixtures of normal distributions are fitted to the age distribution to elucidate real WGD peaks from smaller background deviations[135,177], often in combination with methods that identify significant peak features changes[182,187]. An example of how age distributions can help to infer paleopolyploid history is presented in figure 1.10.

1.5.5 Dating of WGDs

Once paleopolyploidizations have been identified, obtaining a reliable WGD age estimate can help to further elucidate their evolutionary role. Early approaches relied on a constant molecular clock that assumes divergence accumulates at a constant rate, so that the contribution of rate of divergence and time to the total observed divergence can be separated based on reliable fossil calibrations[188]. Early WGD age estimates therefore relied on estimates of general substitution rates in plants[189] to convert their mean divergence, for instance the location of the WGD peak in a K_S age distribution, into an absolute age estimate[135]. It has however been firmly established by now that evolution generally is not clock-like[190], because evolutionary rates are linked to life history traits such as generation time[191] and also other factors such as gene length, GC content, and codon bias[192].

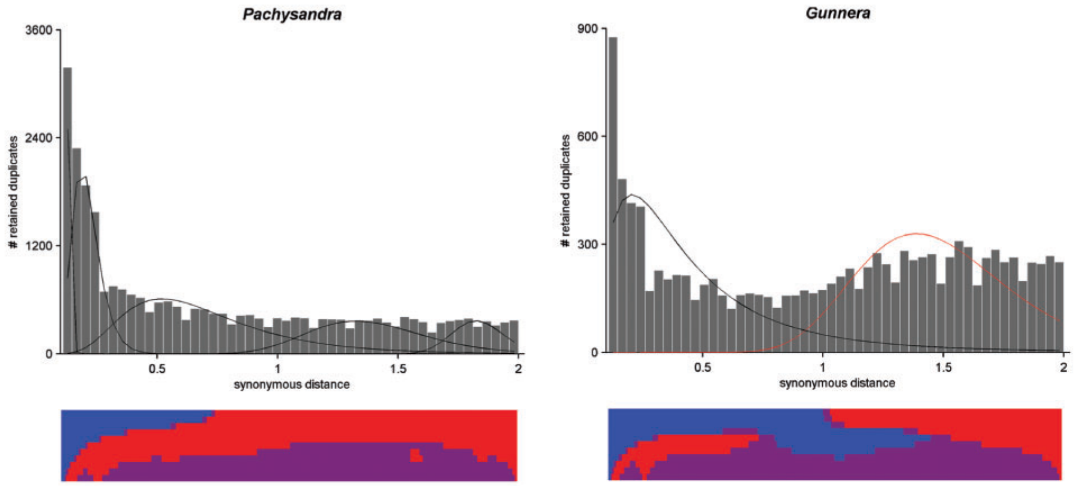


Figure 1.10: Age distributions allow to infer paleopolyploid history. The left panel illustrates the age distribution of *Pachysandra*, an early-branching eudicot genus that did not undergo any WGD since the angiosperm-shared paleopolyploidization that occurred ~192 mya. The distribution displays the L-shaped SSD background consisting of many newly created gene duplicates and fewer retained ancient duplicates. A mixture of five normal components is fitted onto the distribution in black, of which none was found to correspond to a significant WGD peak feature change as indicated by the colour scheme underneath the distribution. Blue and red colours indicate a significant increase and decrease of the first derivative in the age distribution, respectively. Apart from the change in the beginning of the distribution, corresponding to the shift from the initial L-shape into its flat tail, there are no further significant peak feature changes. The right panel illustrates the age distribution of *Gunnera*, a genus that shared the eudicot hexaploidy. A WGD peak is superimposed on the SSD background distribution around a K_S of 1.5. A mixture of two normal components is fitted onto the distribution, of which the component coloured in red corresponds to a significant WGD peak feature change as indicated by the colour scheme underneath the distribution that changes from blue to red at that location. Picture adapted from Vekemans et al.[139].

Relaxed clock methods that can deal with evolutionary rate variation are thus preferable[194]. Several relaxed clock methods have been implemented, which originally assumed an autocorrelated clock where branches that share a direct common ancestor also share similar evolutionary rates[195]. These include the popular r8s package[196] that uses a penalized likelihood method that minimizes rate changes between the different branches[197], and MCMCTREE that uses a Bayesian framework for estimating species divergence times[198].

Fawcett et al.[193] were the first to use such methods to provide a comprehensive temporal framework for all known paleopolyploidizations in plants. Remarkably, they demonstrated a tentative clustering of many paleopolyploidizations with the K-Pg mass extinction event described before (see 1.3.3). An overview of their results is presented in figure 1.11 and suggests that polyploids established around that time had a greater chance of survival[151,193], which is in line with data from teleost fishes where it was found that the teleost-specific WGD (see figure 1.8) probably alleviated the risk of extinction[199]. Explanations for enhanced polyploid establishment at the K-Pg boundary mostly focused on adaptive mechanisms that could have favoured polyploid survival over that of their diploid progenitors. Transgressive segregation, the formation of more extreme phenotypes in the polyploid population compared to their diploid parents, can lead to more phenotypic variability[200]. The latter is especially pronounced in allopolyploids that display strong hybrid vigour (heterosis) through the combination of novel allelic combinations not found in either parent. This phenotypic variability is probably enhanced by their plastic genomic background, which is characterized by extensive structural changes, expression changes, and epigenetic repatterning[201]. This genomic plasticity and phenotypic variability most often have a negative effect on polyploid fitness through chromosomal abnormalities during cell division and unstable phenotypes (see 1.5.3), resulting

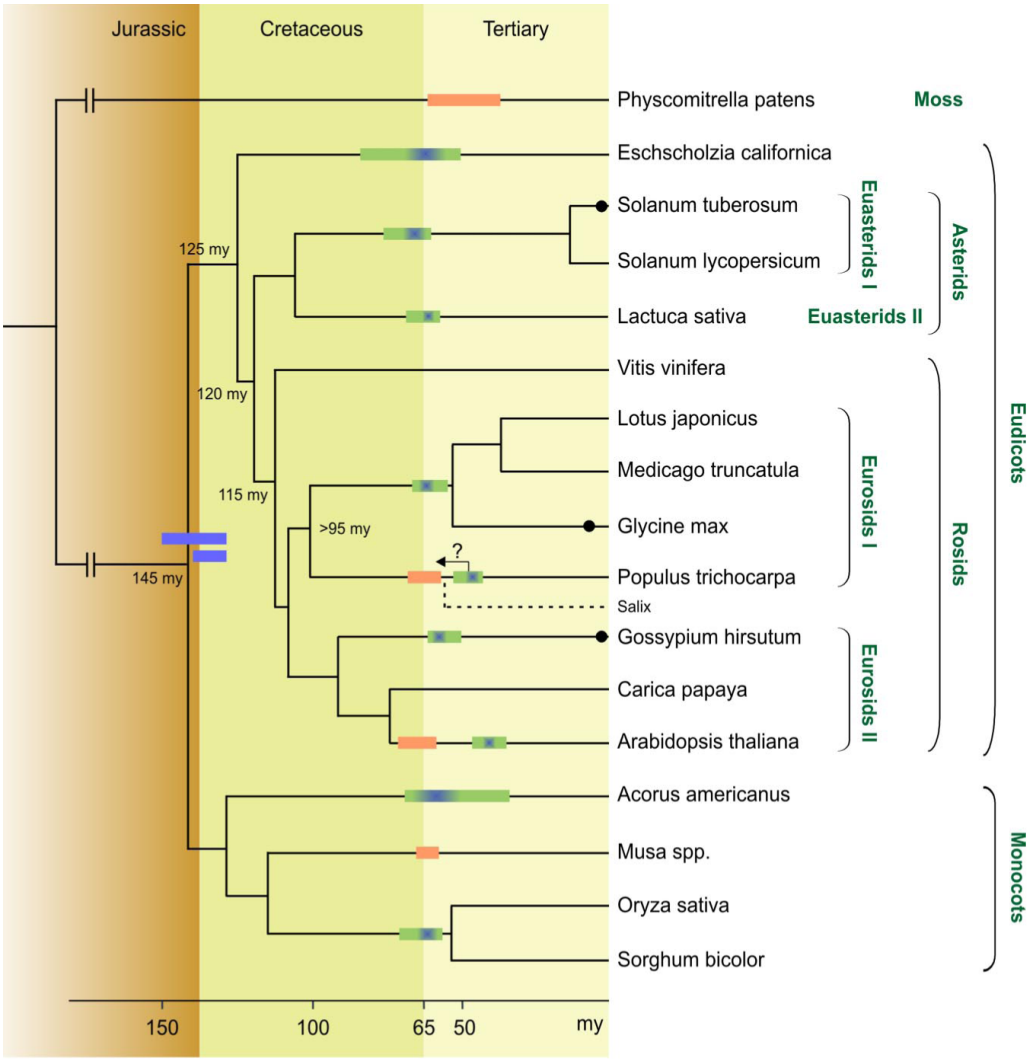


Figure 1.11: Temporal framework for paleopolyploidizations in the green plants. A simplified topology of the green plants is depicted, with WGDs indicated by green bars that denote their 95% age confidence intervals. The dark green portions of the bars are centred on their best age estimates. Orange bars are WGD age estimates from literature. Blue bars denote the hexaploidy event shared by the eudicots. The arrow and question mark for *Populus trichocarpa* indicate a suggested correction when trying to correct for its slower evolutionary rate based on sequence data from *Salix*. The black dots indicate very recent polyploidy events, which have only partially diploidized so far. Figure adapted from Fawcett et al.[193].

in outcompetition by their stable and highly specialized diploid progenitors. However, around the K-Pg boundary, polyploid genomic plasticity and phenotypic variability probably rather represented a higher adaptive potential that allowed newly formed polyploids to react more quickly to the drastically and quickly changing conditions by exploiting their potential for broad ecological tolerance as invasive colonizing species[151,193].

1.6 Research goals

1.6.1 Towards a better understanding of evolutionary models for the maintenance of gene duplicates

There is a sharp contrast between the large number of detailed theoretical models of evolution after gene duplication, on the one hand, and the lack of clear experimental evidence for the various predictions

made by these models, on the other[98]. There are two key problems to this issue. The first is the lack of knowledge about the functional properties of the ancestral pre-duplication gene, which is an important distinguishing feature among models that otherwise are very similar, for instance the specialization and gene sharing models for subfunctionalization (see 1.4.2). Since the pre-duplication genes do not longer exist, many of the events that led from the ancestral gene to the present-day duplicates remain obscure. In most studies, the activities of the pre-duplication ancestor are inferred from unduplicated present-day outgroup genes that are assumed to have retained similar functional properties, but this is only an approximation. The central hurdle to surpass involves rewinding the evolutionary record to obtain the sequence and activity of the ancestral proteins. Recent developments in sequencing and bio-informatics however now enable to reconstruct ancestral genes and proteins and characterize them in detail[202,203].

The second problem is the lack of knowledge about whether neutral of adaptive molecular processes drove evolution of the post-duplication paralogs, which can also be an important distinguishing feature amongst otherwise very similar models, for instance the adaptive and Dykhuizen-Hartl models for neofunctionalization (see 1.4.3). Resurrection of ancestral gene loci solves the first problem but not the latter. An increasingly powerful suite of tests for detecting positive selection amongst sequences[204], accounting for lineage-specific variation[205], among-site variation[206], or a combination of both[207], have been developed in the last years and allow to test for positive selection in the post-duplication paralogs. Although their use remains controversial[208,209], they provide a powerful tool when they are combined with experimental validation and adequate precautions in their interpretation are taken[210].

We have used the yeast *MALS* gene family as a model system to gain insight in the molecular mechanisms and evolutionary forces shaping the fate of duplicated genes. The *MALS* genes encode α-glucosidases that allow yeast to metabolize complex carbohydrates, and possess several key features that make them ideal to study duplicate gene evolution[211]. It is a large gene family with several recent and ancient duplication events, of which the present-day enzymes have diversified substrate specificities that can easily be measured. Furthermore, both extensive *MALS* gene sequences from many fungal genomes and a crystal structure of one of the present-day enzymes are available. High-confidence predictions of ancestral gene sequences therefore allow to assess their changing functionality and detect the adaptive or neutral molecular processes they underwent during their divergence. This is the subject of chapter 2.

1.6.2 Obtain better tools to reliably infer paleopolyploidizations

More reliable inference of paleopolyploidizations will help in better understanding both their abundance and evolutionary role. All methods used for detecting paleopolyploidizations arguably possess their strong and weak points (see 1.5.4), but the ease of use of paranome age distributions coupled with their low computational cost make them ideal for exploratory purposes. Maere et al.[178] introduced a new approach to infer WGDs based on age distributions, which uses a quantitative duplicate population dynamics model that simulates the death and birth of genes by both SSD and WGD in an age distribution. Optimization of model parameters to empirical age distributions allowed to successfully dissect the quantitative contribution of the last three WGDs that occurred during the evolutionary past of the model plant *Arabidopsis thaliana*.

Nevertheless, dissection of even older WGD events that have been confirmed through other methods remains difficult[212]. One of the main reasons for this, which also applies to the use of standard-practice mixture modelling techniques, is the use of K_S as a proxy for the age of duplicated genes. A first concern is the stochastic nature of synonymous substitutions, whereby the synonymous substitution levels of simultaneously duplicated paralogous pairs show increasing variation with time since duplication[5]. As a consequence, older WGD peaks will be progressively flattened and dispersed over the distribution until they gradually blend into the L-shaped SSD background, an effect that is exacerbated by their on-going duplicate loss[135,176,177]. A second concern are K_S saturation effects. With increasing age since duplication, paralogous pairs start to accumulate multiple substitutions per site and the evolutionary models employed for K_S estimation are unable to fully correct for this, leading to K_S estimates that are systematically lower than the real synonymous substitution levels and eventually saturate[5]. Because of this saturation effect, older gene duplicates are wrongfully lumped together at lower K_S values so that an artificial saturation peak may be generated in the age distribution, which could be mistaken for a WGD peak[177,213].

We have used a two-step approach to investigate how K_S stochasticity and saturation affect the shape of K_S-based age distributions for various species. First, we performed artificial evolution of coding sequences for different timespans that take into account species-specific genome characteristics, and afterwards re-estimated the corresponding synonymous distances to quantify K_S stochasticity and saturation. Second, we incorporated these effects in the duplicate population dynamics model introduced by Maere et al.[178] and simulated K_S-based age distributions corresponding to predefined real age distributions with and without WGDs, in order to examine how K_S stochasticity and saturation affect their shape. This is the subject of chapter 3.

1.6.3 Provide an up-to-date temporal framework for paleopolyploid abundance

Insights gained from the tentative clustering of plant paleopolyploidizations with the K-Pg boundary (see 1.5.5) demonstrate how a robust temporal framework can help to identify characteristics that contributed towards polyploid evolutionary success[193]. Nevertheless, dating of such ancient events is particularly troublesome[214], so that the proposed clustering of WGDs with the K-Pg boundary was considered an interesting hypothesis that was however burdened with some limitations due to the restricted amount of sequence data available at that time and the use of methods for sequence divergence estimation that were still under active development[215].

In particular, only six complete genome sequences and a few transcriptome assemblies were available, limiting both the taxon sampling and possibility to implement proper primary fossil calibrations. Dating was done using the penalized likelihood inference method implemented in the r8s program[196]. This software incorporates an autocorrelated relaxed clock model, which is an assumption that seems unlikely in light of the sparse taxon sampling considered[195], and violation thereof may lead to inconsistent age estimates[216]. Because few species were available, calibrations were implemented as fixed secondary point calibrations, which may lead to illusionary precision of the time estimates[217].

Recent years have seen a huge increase in plant (whole genome) sequence data becoming available[218], in addition to the development of more powerful Bayesian methods for sequence divergence

estimation[219–221], as well as more powerful high-performance computing systems that allow such intensive Bayesian algorithms to be run on a massive scale. We have therefore revisited the hypothesized link between the K-Pg mass extinction and successful WGDs, taking full advantage of these advances. All available plant genome sequences were collected to obtain a much broader coverage of the overall angiosperm phylogeny. Dating was based on the powerful Bayesian framework implemented in the BEAST package[220], using an uncorrelated relaxed clock model that assumes a lognormal distribution on evolutionary rates[219], which should be better equipped to deal with rate shifts between different branches compared to autocorrelated relaxed clocks when taxon sampling is limited[65]. Lastly, primary fossil calibrations were selected, implemented as flexible lognormal calibration priors that represent the error associated with the age of the fossil in a more realistic manner[67,222]. This is the subject of chapter 4.

1.6.4 Gain a better insight into the evolutionary significance of gene and genome duplications

A better insight into the fates of genes after duplication, combined with better tools to reliably infer WGDs that simultaneously duplicate all genes present within the genome, put into a proper temporal framework, will help to obtain a better understanding of the significance of (plant) WGDs in evolution. We incorporated the data and results gathered in this dissertation accordingly within the extensive framework of WGD that is slowly emerging as a result of the continued efforts of the broad scientific community involved in polyploidy research. In particular, we addressed the long-standing question whether WGD is an evolutionary dead end, or rather, a road towards evolutionary success. It is now well established that several successful paleopolyploidizations occurred during plant evolution[52,128], which makes it difficult to classify WGD merely as an evolutionary dead end. On the other hand, the discrepancy between the low number of successful paleopolyploidizations and the vast amount of recently formed polyploids indicates that most of these neopolyploids will most likely not stand the test of time[151], so that WGD neither can be classified solely as a road towards evolutionary success. An updated framework for the significance of WGD in evolution is therefore critically dependant upon factors that can adequately explain this enigma. This is the subject of chapter 5.

1.7 Author contributions

The content of this chapter was written by myself. It resulted from the many fruitful discussions with both my promoters and all partners I had the chance to work with during my PhD studies.

Chapter 2

Functional innovation through gene duplication

Karin Voordeckers*, Chris Brown*, Kevin Vanneste, Elisa van der Zande, Arnout Voet, Steven Maere, Kevin Verstrepen. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *Plos Biology* **10**(12):e1001446. * contributed equally

Abstract

Gene duplications are believed to facilitate evolutionary innovation. However, the mechanisms shaping the fate of duplicated genes remain heavily debated because the molecular processes and evolutionary forces involved are difficult to reconstruct. Here, we study a large family of fungal glucosidase genes that underwent several duplication events. We reconstruct all key ancestral enzymes and show that the very first preduplication enzyme was primarily active on maltose-like substrates, with trace activity for isomaltose-like sugars. Structural analysis and activity measurements on resurrected and present-day enzymes suggest that both activities cannot be fully optimized in a single enzyme. However, gene duplications repeatedly spawned daughter genes in which mutations optimized either isomaltase or maltase activity. Interestingly, similar shifts in enzyme activity were reached multiple times via different evolutionary routes. Together, our results provide a detailed picture of the molecular mechanisms that drove divergence of these duplicated enzymes and show that whereas the classic models of dosage, sub-, and neofunctionalization are helpful to conceptualize the implications of gene duplication, the three mechanisms co-occur and intertwine.

For the author contributions, see page 54.

2.1 Introduction

In a seminal book, Susumu Ohno argued that gene duplication plays an important role in evolutionary innovation[93]. He outlined three distinct fates of retained duplicates that were later formalized by others[94,98]. First, after a duplication event, one paralog may retain the ancestral function, whereas the other allele may be relieved from purifying selection, allowing it to develop a novel function (later called "neofunctionalization"). Second, different functions or regulatory patterns of an ancestral gene might be split over the different paralogs (later called "subfunctionalization"[103,223]). Third, duplication may preserve the ancestral function in both duplicates, thereby introducing redundancy and/or increasing activity of the gene ("gene dosage effect"[224]).

Recent studies have shown that duplications occur frequently during evolution, and most experts agree that many evolutionary innovations are linked to duplication[159,225–227]. A well-known example are crystallins, structural proteins that make up 60% of the protein in the lenses of vertebrate eyes. Interestingly, paralogs of many crystallins function as molecular chaperones or glycolytic enzymes. Studies suggest that on multiple occasions, an ancestral gene encoding a (structurally very stable) chaperone or enzyme was duplicated, with one paralog retaining the ancestral function and one being tuned as a lens crystallin that played a crucial role in the optimization of eyesight[228,229].

The molecular mechanisms and evolutionary forces that lead to the retention of duplicates and the development of novel functions are still heavily debated, and many different models leading to Ohno's three basic outcomes have been proposed[94,98,230,231]. Some more recent models blur the distinction between neo- and subfunctionalization[232]. Co-option models, for example, propose that a novel function does not develop entirely de novo but originates from a pre-existing minor function in the ancestor that is co-opted to a primary role in one of the postduplication paralogs[98,230]. Examples of such co-option models include the "gene sharing" or "Escape from Adaptive Conflict" (EAC) model[110,111,223,233,234] and the related "Innovation, Amplification, and Divergence" (IAD) model[112,113,235]. The IAD model describes co-option as a neofunctionalization mechanism. A novel function arises in the preduplication gene, and increased requirement for this (minor) activity is first met by gene amplification (e.g., through formation of tandem arrays). After this, adaptive mutations lead to divergence and specialization of some of the duplicate copies. The EAC model, on the other hand, describes co-option rather as a subfunctionalization mechanism by which duplication allows a multifunctional gene to independently optimize conflicting subfunctions in different daughter genes.

Another aspect in which various models differ is the role of positive selection. Some models emphasize the importance of neutral drift, while in other models adaptive mutations play an important role. For example, in the "Duplication-Degeneration-Complementation" (DDC) model of subfunctionalization[103], degenerative mutations (accumulated by neutral drift) lead to complementary loss-of-function mutations in the duplicates, so that both copies become essential to perform all of the functions that were combined in the single preduplication gene. Whereas this type of subfunctionalization only involves genetic drift[103,159], other subfunctionalization models, such as the EAC model, attribute an important role to positive selection for the further functional optimization of the postduplication paralogs[98,231].

There is a sharp contrast between the large number of detailed theoretical models of evolution after gene duplication, on the one hand, and the lack of clear experimental evidence for the various predictions

made by these theories, on the other[98]. The key problem is the lack of knowledge about the functional properties of the ancestral, preduplication gene. Since these ancient genes and the proteins they encode no longer exist, many details in the chain of events that led from the ancestral gene to the present-day duplicates remain obscure. In most studies, the activities of the preduplication ancestor are inferred from unduplicated present-day outgroup genes that are assumed to have retained similar functional properties, but this is only an approximation. The central hurdle to surpass to obtain accurate experimental data on the evolution of gene duplicates involves rewinding the evolutionary record to obtain the sequence and activity of the ancestral proteins. Recent developments in sequencing and bio-informatics now enable us to reconstruct ancestral genes and proteins and characterize them in detail[202,203,236–242]. However, most ancestral reconstruction studies to date did not focus on the mechanisms that govern evolution after gene duplication.

In this study, we used the yeast *MALS* gene family as a model system to gain insight in the molecular mechanisms and evolutionary forces shaping the fate of duplicated genes. The *MALS* genes encode α-glucosidases that allow yeast to metabolize complex carbohydrates like maltose, isomaltose, and other α-glucosides[211,243]. Several key features make this family ideal to study duplicate gene evolution. First, it is a large gene family encompassing multiple gene duplication events, some ancient and some more recent. Second, the present-day enzymes have diversified substrate specificities that can easily be measured[243]. Third, the availability of *MALS* gene sequences from many fungal genomes enabled us to make high-confidence predictions of ancestral gene sequences, resurrect key ancestral proteins, and study the selective forces acting throughout the evolution of the different gene duplicates. Fourth, the crystal structure of one of the present-day enzymes, Ima1, has been determined[244]. Molecular modeling of the enzymes' binding pocket, combined with activity measurements on reconstructed and present-day enzymes, allowed us to investigate how mutations altered enzyme specificity and gave rise to the present-day alleles that allow growth on a broad variety of substrates. Combining these analyses, we were able to study the evolution and divergence of a multigene family to an unprecedented level of detail and show that the evolutionary history of the *MALS* family exhibits aspects of all three classical models of duplicate gene evolution proposed by Ohno (gene dosage, neo-, and subfunctionalization).

2.2 Material and methods

2.2.1 Phylogenetic tree construction

In total, the nucleotide and protein sequences of 169 extant maltases were collected for yeast species ranging from *Saccharomyces cerevisiae* to *Pichia* and *Candida* species. For *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, and *Kluyveromyces lactis*, sequences were downloaded from Génolevures (`www.genolevures.org`). Sequences for many of the *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* genes were obtained from the sequence assemblies provided by the Wellcome Trust Sanger Institute (`www.sanger.ac.uk/research/project\hskip\z@\relaxs/genomeinformatics/sgrp.html`). All of the remaining extant maltase sequences were downloaded from NCBI (`www.ncbi.nlm.nih.gov`). Sequences with greater than 92% pairwise protein sequence similarity to other sequences in the dataset were removed to reduce the phylogenetic complexity. All seven *Saccharomyces cerevisiae* S288c alleles

were kept, however, yielding a final dataset of 50 sequences. Sequences were aligned using MAFFT[245], and the resulting sequence alignment is depicted in supplementary figure D.1.

We used ProtTest 2.4[246] to score different models of protein evolution for constructing an AA-based phylogenetic tree. All possible models with all improvements implemented in the program were taken into account. An initial tree was obtained by Neighbor-Joining (BioNJ), and the branch lengths and topology were subsequently optimized for each evolutionary model independently. The LG+I+G model came out as best with a substantial lead over other protein models using $-$lnL, AIC, and AICc selection criteria (AICc=43,061.26 and AICw=1.00, while the second best model was WAG(+I+G) with AICc=43,158.00 and AICw=0.00). Consequently, an AA-based phylogeny for the 50 sequences was determined using MrBayes 3.1.2[247] with a LG invariant+gamma rates model (four rate categories). Since the LG model is not implemented by default in MrBayes, we used a GTR model and fixed the substitution rate and state frequency parameters to those specified by the LG model. The MCMC was run for 10^6 generations, sampling every 100 generations, with two parallel runs of four chains each. A burn-in of 2,500 samples was used, and the remaining 7,501 samples were used to construct a 50% majority-rule consensus phylogeny (see supplementary figure D.3). The AWTY program[248] was used to check proper MCMC convergence under the given burn-in conditions. MrBayes AA tree constructions were also performed under other evolutionary models (WAG, JTT). Additional tests were performed to exclude long branch attraction (LBA) artifacts (see supplementary information D.3.1). We also inferred a maximum likelihood (ML) tree using PhyML under the LG+I+G model with four rate categories[249]. The initial tree was again obtained by BioNJ; tree topology, branch lengths, and rate parameters were optimized in a bootstrap analysis with 1,000 replicates.

We also used MrBayes to construct a codon-based phylogeny, using a GTR codon model of evolution. The original dataset of 50 sequences contained 18 sequences for species that employ the alternative yeast nuclear genetic code (all of them outgroup species). These sequences were removed from the dataset, resulting in a reduced dataset of 32 sequences. The codon alignment was obtained by translating the AA alignment obtained earlier. MCMC analysis and consensus phylogeny construction were performed as described above for the AA trees. We contrasted models that did and did not allow for ω rate variation (i.e., the “Equal” versus “M3” codon model in MrBayes). AWTY analysis indicated that the latter was not able to converge properly, so we used the results of the Equal model.

2.2.2 Ancestral sequence reconstruction

The PAML package[250] was used to infer the posterior AA probability per site in the ancestors of interest under several commonly used models of protein evolution (LG, WAG, JTT), using the corresponding Bayesian consensus phylogenies. Both marginal and joint probability reconstructions were performed. The marginal reconstructions are presented in supplementary table D.1. Protein sequences resulting from marginal reconstructions under the JTT model were used to synthetize ancestral enzymes, and are depicted in supplementary figure D.2.

2.2.3 Positive selection tests

We performed tests for positive selection on the codon-based phylogeny obtained as described above. Various branch methods and branch-site methods included in the PAML[250] and HyPhy[251] packages were used.

Branch tests

We first explored the change in selective forces over time using the branch models implemented in the PAML package. The fit of the free-ratio model, which assigns an independent ω value for each branch, was found to be significantly better than that of null model assigning only one ω value to the whole tree (LRT stat=438.43; *df*=60; $p<0.0001$). This test confirms the presence of variability in selection pressure across branches of the codon tree, but its ω estimates are not reliable because the free-ratio model suffers from overparameterization.

We therefore applied the GA (Genetic Algorithm) Branch method, available as an extension to the HyPhy package[251,252], as described in[253]. This method uses a genetic algorithm to search through the space of possible models and divides the branches of the phylogenetic tree in subsets of branches that share the same ω estimate, reducing parametric complexity. We used the 012034 GTR nucleotide model, selected by a HyPhy model selection routine from all 203 available GTR models. We repeated the GA Branch procedure on five replicates and pooled results for postprocessing, after ensuring that all replicates reached similar solutions. The postprocessing resulted in a final branch partitioning model with four ω rate categories. Since the GA Branch method itself is focused on finding the best branch-clustering scheme rather than finding the best ω estimates, the estimated ω values obtained in the GA Branch analysis were further optimized using a HyPhy model optimization routine that allows for non-synonymous rate heterogeneity. The net effect was an increase of the estimated ω values for all four rate categories (see figure 2.4).

Branch-site tests

We used the modified branch-site model A implemented in PAML, which allows ω to vary both among sites in the sequence alignment and across branches on the tree, to screen for positive selection on sites along specific branches[207]. We used the anc*IMA1–4*, anc*MAL*, and anc*IMA5b* branches separately as the foreground branch, while the rest of the phylogeny was considered as the background, and assessed deviation from the null model (no positive selection) using a Likelihood Ratio Test following a χ^2_1 distribution[254]. A Bonferroni correction was employed to control for multiple testing[255], and a posteriori BEB (Bayes Empirical Bayes) inference technique was used to identify the sites that are most likely under positive selection[256].

We also used an alternative branch-site method that was recently implemented in the HyPhy package[257]. This method similarly identifies branches that are subject to episodic diversifying selection but differs from the branch-site tests implemented in PAML in that no background and foreground branches need to be specified *a priori*. Instead, the method fits a sequence of increasingly more complex models to the data, including a model that permits unrestricted combinations of selective regimes across sites

and branches. Subsequently, all branches with some proportion of sites with $\omega>1$ were tested for positive selection using a series of LRTs.

2.2.4 Co-evolving residue detection

Co-evolving residues in the *MALS* gene family were detected using the framework described by Brown et al.[258]. The NCBI Blast server was used to collect *Saccharomyces cerevisiae* S288c *MAL12* maltase homologs, with an E-value $<$10e-70, resulting in a set of 1,211 sequences. Proteins were removed that were shorter than 400 AAs, longer than 800 AAs, and more than 95% similar to another protein in the dataset. This resulted in a dataset of 640 maltase homologs with sequence similarity $>$40% compared to *Saccharomyces cerevisiae* S288c *MAL12*. These sequences were aligned with MAFFT and only the most reproducible residue–residue couplings (present in at least 90% of the splits) were retained.

2.2.5 Statistical analyses

A two-way ANOVA using log-transformed k_{cat}/K_m (to obtain values that are normally distributed) as the variable, and the different enzymes and sugars as factors, was performed using the aovSufficient function from the HH package in R. k_{cat} is the catalytic constant and represents the maximum rate of product formation, while K_m is the Michaelis dissociation constant that reflects how well the enzyme binds with its partner, so that k_{cat}/K_m (the specificity constant) is a measure for the efficiency of an enzyme. This analysis was followed by pairwise comparisons using the Games-Howell post-hoc test (since samples had unequal variances, as demonstrated by Levene's test). Results can be found in supplementary table D.3.

2.2.6 Microbial strains, growth conditions, and molecular techniques

Ancestral maltase genes were synthesized and cloned into vectors for overexpression in *E. coli* host cells by GENEART (`www.geneart.com`). Sequences can be found in supplementary table D.1. The inferred protein sequences were reverse translated in order to optimize their codon usage for *E. coli*. These gene sequences were synthesized including an N-terminal 6xHis tag (ATGGGCAGCAGCCATCATCATCATCAT-CACAGCAGCGGCCTGGTGCCGCGCGGCAGCCAT) and 5′UTR (TCTAGAAATAATTTTGTTTAACTT-TAAGAAGGAGATATACC), cloned into in-house vectors at GENEART, and then sequenced. Subsequently, the inserts were subcloned into pET-28(a) vectors (Merck) via XbaI/XhoI sites. All of the overexpression plasmids were transformed into *E. coli* strain BL21*. All *E. coli* strains were grown under selection in standard LB media+kanamycin (Sigma Aldrich). Details on protein expression and purification can be found in supplementary information D.3.3.

2.2.7 Enzyme assays and data analysis

The activities of the purified ancestral and present-day enzymes were determined by measuring glucose release from α-glucosides (maltose, sucrose, turanose, maltotriose, maltulose, isomaltose, palatinose, and methyl-α-glucoside) using a standard glucose oxidase/peroxidase coupled reaction. All sugars were

purchased in their highest available purity. More information on the purchased sugars as well as a detailed protocol can be found in supplementary information D.3.3.

For each protein and substrate, the reaction velocity (amount of glucose produced per time unit) was determined. Subsequently, reaction velocities normalized by enzyme concentration as a function of substrate concentration were plotted and fitted using a non-linear least squares fitting routine (Levenberg-Marquardt algorithm) both to Michaelis-Menten-style kinetics and Hill-style kinetics:

$$\frac{\nu}{[E]} = \frac{k_{cat}\,[S]^n}{(K_m)^n + [S]^n} \tag{2.1}$$

The data fits were compared using an *F* statistic (i.e., Michaelis-Menten is a specific case of Hill kinetics with *n*=1), and the Michaelis-Menten model was rejected with α=5%. From these fits, errors (standard deviations) were computed by jack-knifing over the individual substrate concentrations (12 data points in total). For numerical optimization, code was written in Python using NumPy. Model parameters of interest, along with their associated errors, were extracted (i.e., k_{cat} and K_m; see supplementary table D.2). Processing (`http://processing.org`) was used to draw figures 2.2 and 2.5F by writing code. Enzyme efficiencies were plotted (as vertical lines) at different points on the tree, and values between were interpolated.

2.2.8 Fitness measurements

Relative Malthusian fitness was determined by competing unlabelled WT (KV1042), *mal12* (KV1151), and *mal32* (KV1153) strains against a reference strain (KV3261), expressing GFP from the *TDH3p*. Details can be found in supplementary information D.3.3.

2.2.9 Molecular modeling

All molecular modeling was performed using the MOE 2010.10 package (The Molecular Operating Environment, The Chemical Computing Group, Montréal, Canada). The recently released crystal structure of the Ima1 protein (pdb entry: 3A4A), with glucose in the binding pocket, was used as a template to construct the different *MALS* homology models, with implementation of the Amber99 force field. Since the AAs contacting this glucose molecule are conserved within the different *MALS* subgroups, this glucose was used to model the different sugar substrates within the active sites, using the MOE 2010.10 ligX implementation.

2.3 Results

2.3.1 The present-day maltase enzymes arose from a functionally promiscuous ancestor

Some yeast species have evolved the capacity to metabolize a broad spectrum of natural disaccharides found in plants and fruits (see figure 2.1). The origin of this evolutionary innovation seems to lie in the duplication and functional diversification of genes encoding permeases and hydrolases[243]. The common

Saccharomyces cerevisiae laboratory strain S288c, for example, contains seven different *MALS* genes (*MAL12*, *MAL32*, and *IMA1–5*), which originated from the same ancestral gene but allow growth on different substrates[211,243].

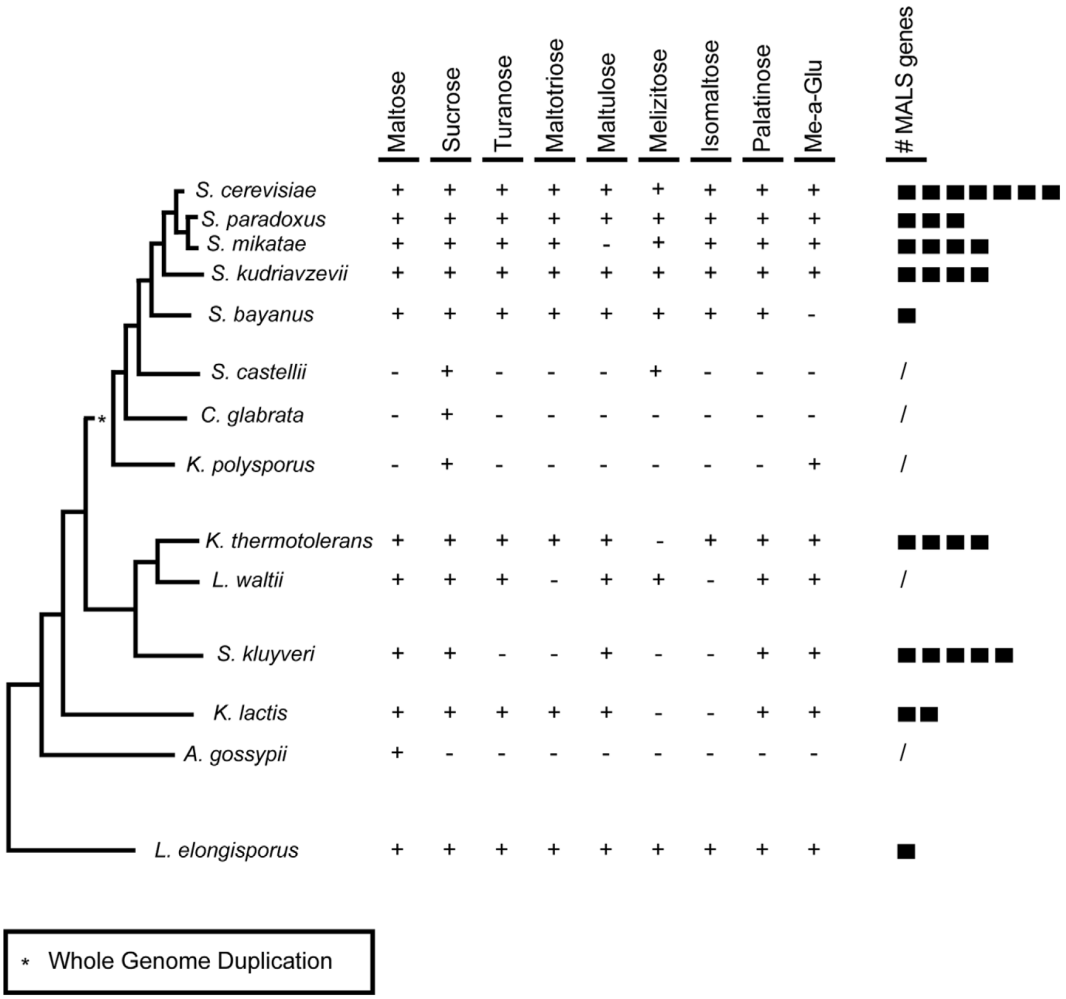


Figure 2.1: Yeast species can grow on a broad spectrum of α-glucosides. Serial dilutions of each species were spotted on medium (Yeast Nitrogen Base wihtout amino acids) with 2% of each sugar (Me-α-Glu = methyl-α-glucoside). Growth was scored after 3 days incubation at 22°C. +, growth; -, no growth; # *MALS* genes, the number of maltase genes found in each of these strains. Genotypes are listed in supplementary table D.5. Tree adapted from Kurtzman and Robnett[259].

To understand how duplications led to functionally different MalS enzymes, we reconstructed, synthesized, and measured the activity of key ancestral MalS proteins. We used the amino acid (AA) sequences of 50 maltases from completely sequenced yeast species, ranging from *Saccharomyces cerevisiae* to *Pichia* and *Candida* species, for phylogenetic analysis and ancestral sequence reconstruction (see Material and methods). A consensus amino-acid-based phylogenetic tree was constructed using MrBayes[247] under the LG+I+G model with four rate categories (see supplementary figure D.3). Trees constructed using MrBayes under other models of sequence evolution (WAG, JTT) generated largely identical results (unpublished data). To further check the robustness of the AA tree inferred by MrBayes, we inferred a maximum likelihood (ML) tree under the LG+I+G model using PhyML[249] (see supplementary figure D.4). With the exception of a few recent splits in the topology, the MrBayes and PhyML trees agree, increasing our confidence in the constructed tree. Codon-based tree reconstruction using MrBayes yielded similar results (see further). Additional tests were performed to control for potential long branch

attraction (LBA) artifacts, specifically to check the placement of the *K. lactis* branch as an outgroup to the *Saccharomyces* and *Lachancea* clades (see supplementary information D.3.1 and supplementary figures D.5, D.6, D.7, and D.8).

Next, we reconstructed the AA sequence of the ancestral maltases under several commonly used models of protein evolution (LG, WAG, JTT; see Material and methods). All models support roughly the same ancestral protein sequences, increasing our confidence in the reconstructed ancestral sequences. In particular, all models identified the same residues for variable sites within 10 Å of the active center (based on the crystal structure of the Ima1 protein), which are likely relevant sites with respect to enzymatic activity. The residues for a few other sites located further away from the active pocket vary between different models, but differences generally involve biochemically similar AAs (see supplementary table D.1).

Synthesis of the ancestral enzymes was based on the reconstructed ancestral sequences obtained with the JTT model. For ambiguous residues (i.e., sites for which the probability of the second-most likely AA is $>$0.2) within 7.5 Å of the binding pocket, we constructed proteins containing each possible AA, while for ambiguous residues outside 7.5 Å we considered only the most likely AA. There is one ambiguous residue close to the active center in the ancestral proteins ancMalS and ancMal-Ima, namely residue 279 (based on *Saccharomyces cerevisiae* S288c Ima1 numbering). We therefore synthesized two alternative versions of these proteins, one having G and one having A at position 279. Whereas these alternative proteins show different activities for some substrates, the relative activities are similar and our conclusions are robust. In the main figures, we show the variant with the highest confidence. Enzymatic data for all variants can be found in supplementary table D.2.

The activity of all resurrected ancestral enzymes was determined for different substrates (see figure 2.2; Materials and methods). The results indicate that the very first ancestral enzyme, denoted as ancMalS, was functionally promiscuous, being primarily active on maltose-like substrates but also having trace activity on isomaltose-like sugars. The activity data presented in figure 2.2 show how this promiscuous ancestral protein with relatively poor activity for several substrates evolved to the seven present-day enzymes that show high activity for a subset of substrates, and little or no activity for others. This confirms the existence of two functional classes of MalS enzymes that originated from ancient duplication events. First, Mal12 and Mal32 show activity against maltose-like disaccharides often encountered in plant exudates, fruits, and cereals, like maltose, maltotriose, maltulose, sucrose, and turanose (a signaling molecule in plants). The five MalS enzymes of the second class (Ima1–5), which in fact result from two independent ancient duplication events giving rise to the Ima1–4 and Ima5 clades, show activity against isomaltose-like sugars including palatinose (found in honey[260]) and isomaltose. Differences in hydrolytic activity between members of the same (sub)class are more subtle or even absent, which is not surprising since some of these recent paralogs are nearly identical (Mal12 and Mal32, for example, are 99.7% identical on the AA level).

The more recent ancestral enzymes also show a similar split in activity, with some enzymes (ancMal) showing activity towards maltose-like substrates, and others (ancIma1–4) towards isomaltose-like substrates. Moreover, activity on isomaltose-like sugars (isomaltose, palatinose, and methyl-α-glucoside) changes in a coordinate fashion when comparing different enzymes, and the maltose-like sugars also group together. Careful statistical analysis reveals that the maltose-like group consists of

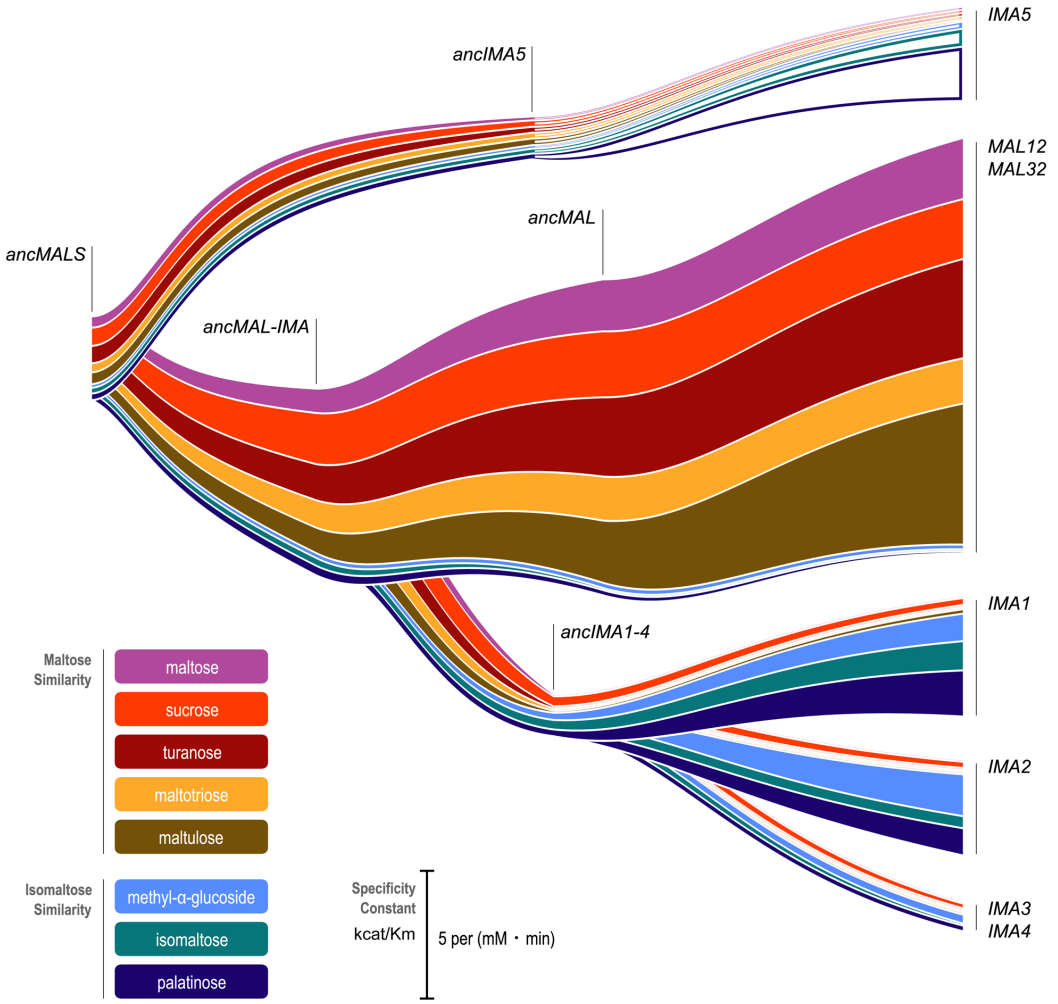


Figure 2.2: Duplication events and changes in specificity and activity in evolution of the *S. cerevisiae* MalS enzymes. The hydrolytic activity of all seven present-day alleles of Mal and Ima enzymes as well as key ancestral (anc) versions of these enzymes was measured for different α-glucosides. The width of the colored bands corresponds to k_{cat}/K_m of the enzyme for a specific substrate. Specific values can be found in supplementary table D.2. Note that in the case of present-day Ima5, we were not able to obtain active purified protein. Here, the width of the colored (open) bands represents relative enzyme activity in crude extracts derived from a yeast strain overexpressing *IMA5* compared to an *ima5* deletion mutant. While these values are a proxy for the relative activity of Ima5 towards each substrate, they can therefore not be directly compared to the other parts of the figure. For ancMalS and ancMal-Ima, activity is shown for the variant with the highest confidence (279G for ancMalS and 279A for ancMal-Ima). Activity for all variants can be found in supplementary table D.2.

two subgroups (maltose, maltotriose, maltulose, and turanose, on one hand, and sucrose, on the other) that behave slightly different, showing that the enzymes show quantitative differences in the variation of specificity towards these substrates (two-way ANOVA analysis followed by Games-Howell test on log-transformed k_{cat}/K_m values; *P*-values can be found in supplementary table D.3).

Interestingly, the most ancient ancestral enzymes do not show a clear split in activity towards either maltose-like or isomaltose-like sugars after duplication, and the transition of ancMalS to ancMal-Ima even shows an increase in activity for all substrates. This suggests that (slight) optimization for all substrate classes simultaneously was still possible starting from ancMalS. A clear divergence of both subfunctions occurred later, after duplication of ancMal-Ima, resulting in ancMal and ancIma1–4. AncMal shows a significant increase in activity on maltose-like sugars accompanied by a significant drop in activity on isomaltose-like sugars compared to ancMal-Ima, and the reverse is true for ancIma1–4 (see

supplementary table D.3 for exact *P*-values for each enzyme–enzyme comparison on the different sugars tested). Together, this illustrates how, after duplication, the different copies diverged and specialized in one of the functions present in the preduplication enzyme.

In two separate instances, a major shift in specificity is observed, from maltose-like sugars to isomaltose-like sugars (transition from ancIma5 to Ima5, and from ancMal-Ima to ancIma1–4). The shift in activity from ancMal-Ima to ancIma1–4 is particularly pronounced. The ancMal-Ima enzyme hydrolyzes maltose, sucrose, turanose, maltotriose, and maltulose but has hardly any measurable activity for isomaltose and palatinose, whereas ancIma1–4 can only hydrolyze isomaltose and palatinose (and also sucrose). For the evolution of the maltase-like activity from the ancestral MalS enzyme to the present-day enzyme Mal12, we see a 2-fold increase in k_{cat} and a 3-fold decrease in K_m for maltose, indicating an increase in both catalytic power and substrate affinity for this sugar. For the evolution of isomaltase-like activity in the route leading to Mal12, k_{cat} decreases more than 3-fold for methyl-α-glucoside. k_{cat} for isomaltose and palatinose and the affinity for isomaltose and palatinose are so low that they could not be measured (see supplementary table D.2 for the exact values of k_{cat} and K_m for each enzyme and each sugar; results of two-way ANOVA analysis followed by Games-Howell test comparing log-transformed k_{cat}/K_m values for different enzymes on each of the sugars can be found in supplementary table D.3).

2.3.2 Present-day enzymes from other yeast species show similar patterns of functional diversification

To further explore the evolution of *MALS* genes and consolidate the measured activities of the ancestral enzymes, we expressed and purified additional present-day α-glucosidase alleles from other yeast species and measured their activities (see figure 2.3). We focused primarily on enzymes that are directly related to one of the ancestral proteins but did not undergo any further duplication events, and therefore have a higher probability of having retained a similar activity as their (sub)class ancestor. Indeed, the only present-day MalS enzyme of the yeast *L. elongisporus* has a broad but relatively weak activity comparable to the very first ancestral MalS enzyme, providing extra support for the accuracy of our ancestral reconstructions. Also in *K. lactis*, which contains two Mal alleles, one of the paralogs retains the broad specificity of ancMalS. The other paralog (GI:5441460) has a deletion of five AAs close to the active pocket that likely explains the general lack of activity of this enzyme (see supplementary figure D.9). In contrast, yeasts that show multiple duplication events, like *K. thermotolerans* and *S. cerevisiae*, exhibit specialization, with some enzymes showing only activity for maltose-like substrates and others for isomaltose-like substrates. Moreover, the activities (maltase- or isomaltase-like) of homologs in *S. cerevisiae* and *K. thermotolerans* derived from the same intermediate ancestor are often similar, except in the *IMA5* clade. Here, the *K. thermotolerans* and *S. cerevisiae* homologs have very different substrate specificities, indicating species-specific evolutionary trajectories and/or reciprocal paralog loss in the different species (see figures 2.3 and 2.4).

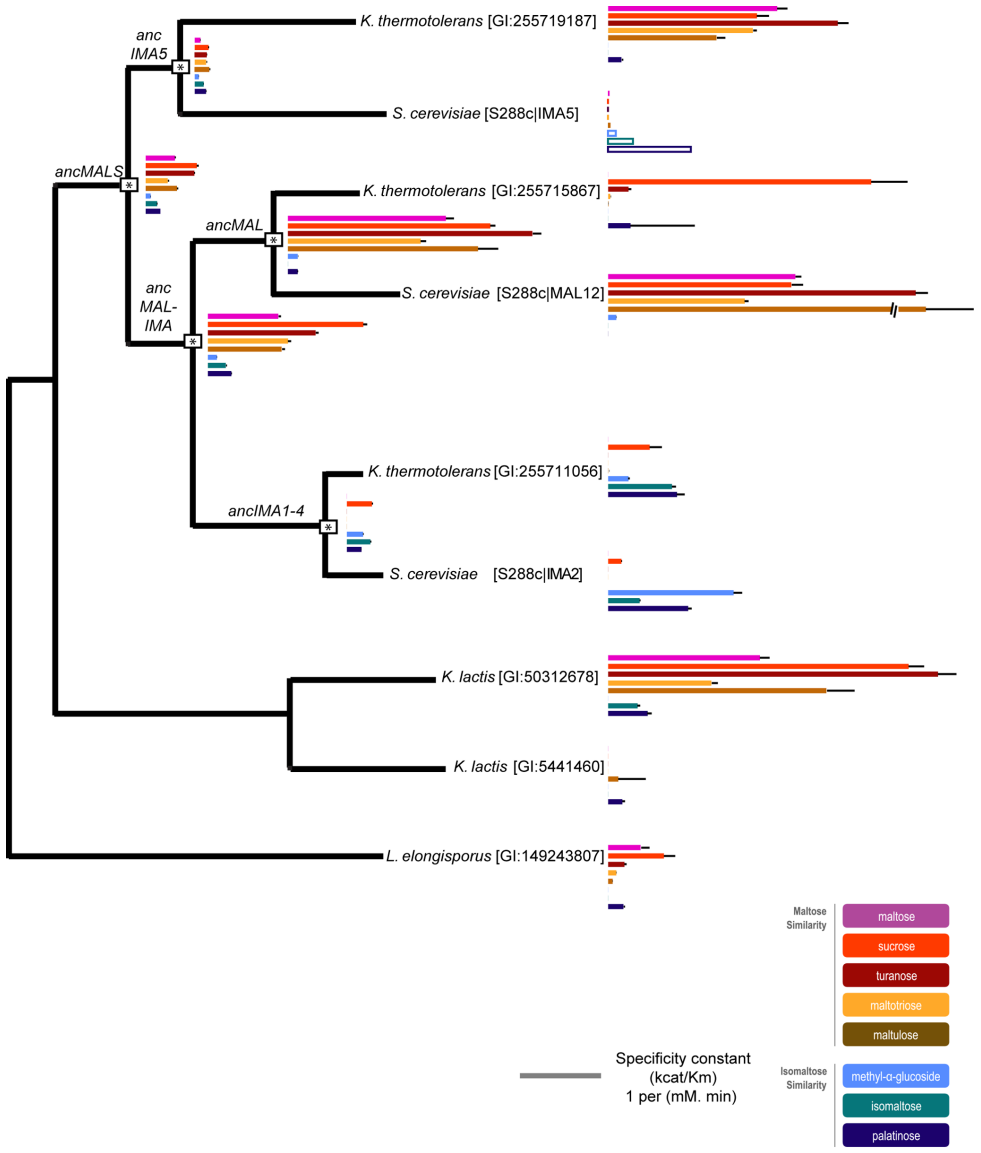


Figure 2.3: Activities of present-day MalS enzymes in distant fungi correspond well with activities of reconstructed ancestral enzymes. Basic phylogeny of the *MALS* gene family with different clades, showing the ancestral bifurcation points (indicated by *). Length of the colored bands corresponds to the measured k_{cat}/K_m of the enzyme for a specific substrate. Bands for Ima5 represent relative enzyme activity in crude extracts derived from a yeast strain overexpressing *IMA5* compared to an *ima5* deletion mutant. For ancMalS and ancMal-Ima, activity is shown for the variant with the highest confidence (279G for ancMalS and 279A for ancMal-Ima). Error bars represent standard deviations. Activity for all variants and the corresponding standard deviations can be found in supplementary table D.2.

2.3.3 Molecular modeling and resurrection of ancestral proteins identify residue 279 in the enzymes' binding pocket as a key determinant of substrate specificity

Next, we investigated which mutations underlie the observed functional changes. We used the recently resolved crystal structure of Ima1 (pdb entry 3A4A)[244] as a template to study the molecular structure of the enzymes' substrate binding pocket (see Materials and methods). All enzymes share a highly conserved molecular fold, suggesting that changes in activity or substrate preference are likely caused by mutations in or around the substrate binding pocket. We identified nine variable AA residues within 10 Å of the center of the binding pocket in the various paralogs (see figure 2.4, right panel). Site-directed mutagenesis and crystallographic studies by Yamamoto et al. confirmed the importance of several of

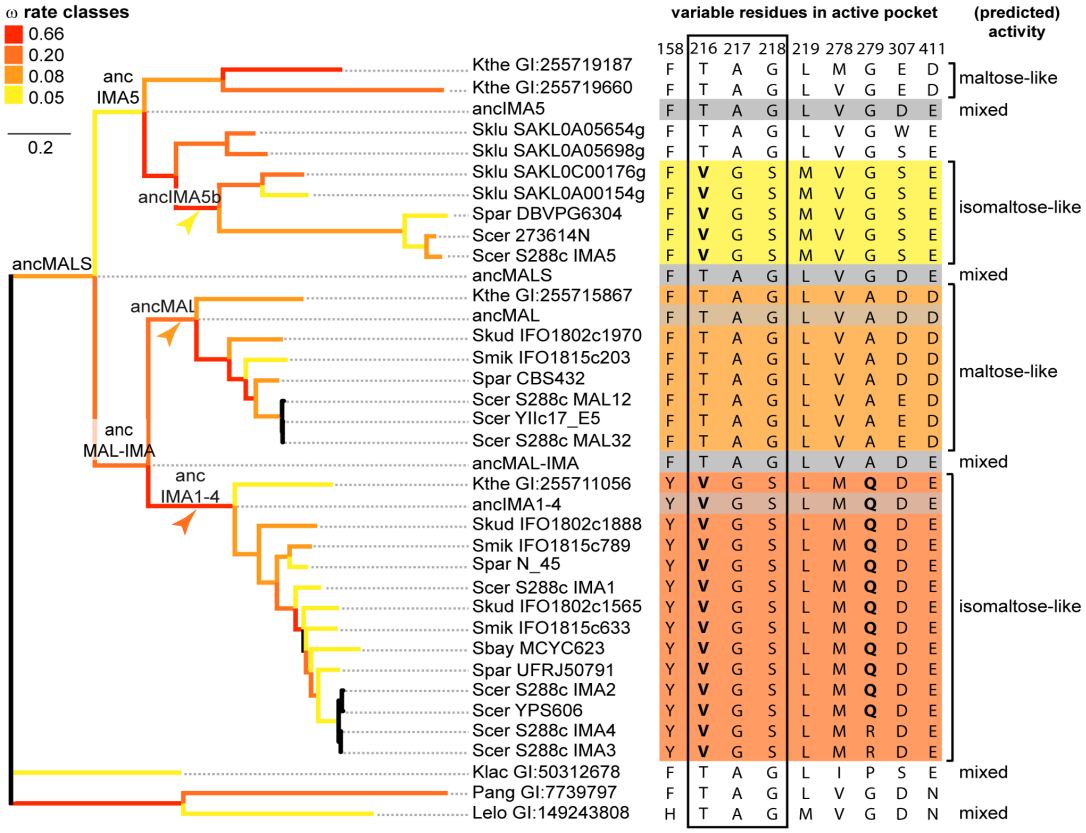


Figure 2.4: Positive selection on residues near the binding pocket resulted in distinct subgroups with different substrate preference. An unrooted codon-based phylogenetic tree of the *MALS* gene family is shown on the left. Branches are colored according to the ω (K_N/K_S) rate classes inferred from GA Branch analysis[252]. Branches for which branch-site tests for positive selection were performed are indicated by colored arrowheads. Since ω rate classes cannot be inferred reliably for very small branches, branches <0.01 are not colored. The right part of the figure shows the nine variable AA residues located near the substrate binding pocket of the respective enzymes (numbering based on Ima1 sequence). Sequences of ancestral enzymes are shaded in grey. Subgroups of enzymes that show similar substrate specificity are colored accordingly. Residues indicated in bold were found to be under positive selection by the branch-site tests. Perfectly co-varying residues are boxed. Substrate preference of extant and ancestral enzymes was deduced from enzyme assays on *S. cerevisiae*, *K. lactis*, *K. thermotolerans*, *L. elongisporus*, and reconstructed ancestral enzymes (see figure 2.3 and supplementary table D.4).

these residues for substrate specificity in the present-day Ima1 protein[261,262]. In particular, the latter characterized the influence of residues 216-217-218 (Ima1 numbering), which covary perfectly with each other and with the observed substrate specificity shifts across the phylogeny presented in figure 2.4. Sequence co-evolution analysis on 640 *MAL12* homologs identified another cluster of three co-evolving residues among these nine residues (positions 218, 278, and 279 in Ima1), which we investigate here in detail.

Together with residues 216 and 217, residues 218, 278, and 279 seem to contribute to the activity shift observed in the evolution of Ima1–4 (see figures 2.4, 2.5, 2.6, and D.10). Molecular modeling of the mutations at 218-278-279 on the branch leading to ancIma1–4 (see figure 2.4) suggests that the change from alanine to glutamine at residue 279 shifts the binding preference of the pocket from maltose-like to isomaltose-like sugars (see figure 2.5B–E). The two co-evolving residues at positions 218 and 278 are spatially close to AA 279 and cause subtle structural adaptations that help to better position the Q residue.

To investigate if changes at all three positions are necessary for the observed shift in substrate specificity from anc*MAL-IMA* to anc*IMA1–4* and to investigate the possible evolutionary paths leading

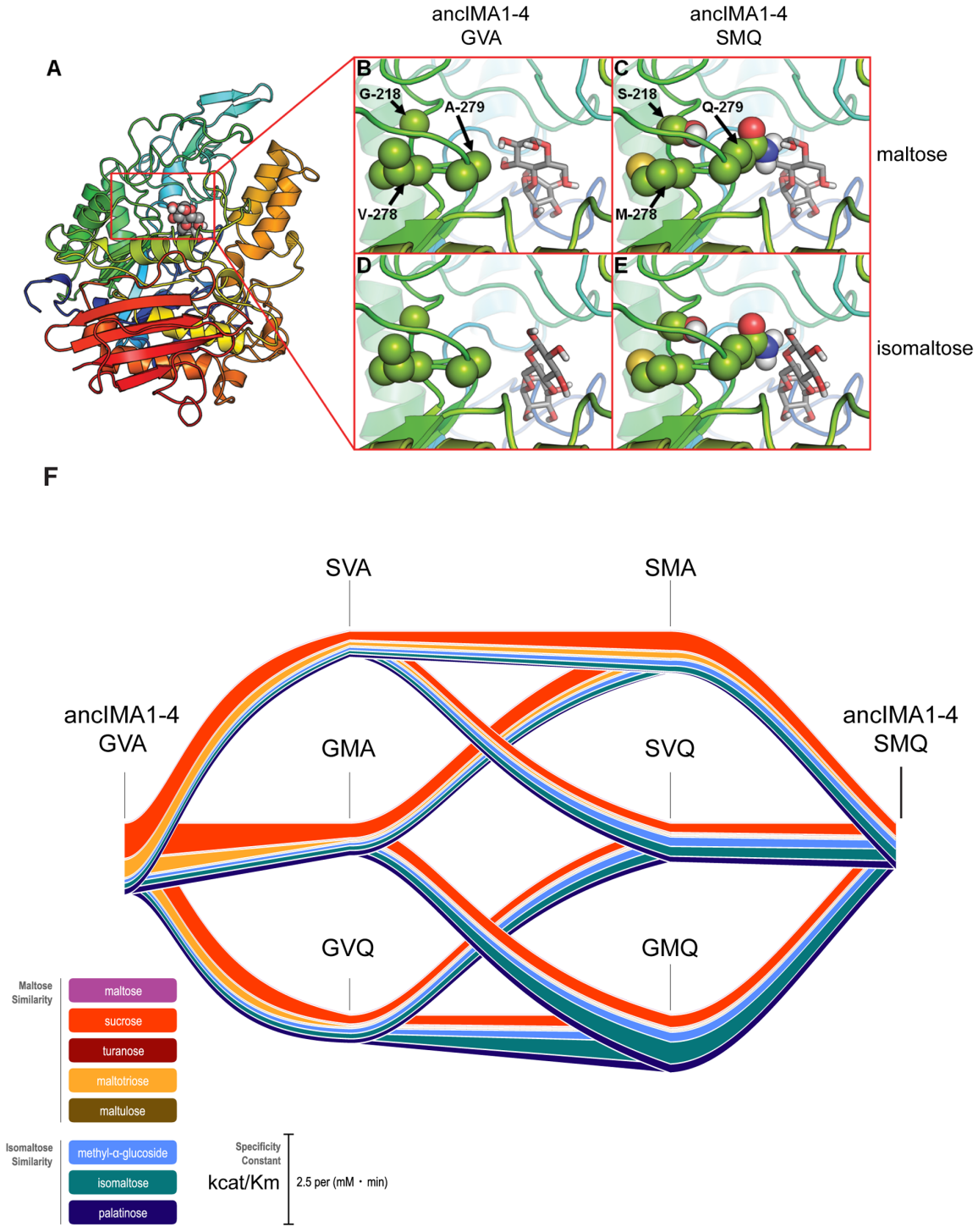


Figure 2.5: Three co-evolving residues determine the shift in activity observed in the evolution of Ima1–4. (A) Global structure of the MalS proteins with maltose, represented as spheres, bound in the active site. Panels (B–E) show details of the active site, with substrates as sticks (maltose in panels B and C; isomaltose in panels D and E). The variable AAs are shown as spheres. Structural analysis of the binding site suggests that the A279Q mutation affects substrate specificity the most. The side chain of Q279 sterically hinders binding of maltose but stabilizes isomaltose binding through polar interactions. The G218S and V278M changes cause subtle adaptations of the fold, causing Q279 to protrude further into the binding pocket, which allows optimal interaction with isomaltose. (F) Activity (k_{cat}/K_m) of all possible intermediary forms in the evolution of three
k_{cat} and K_m can be found in supplementary table D.2.

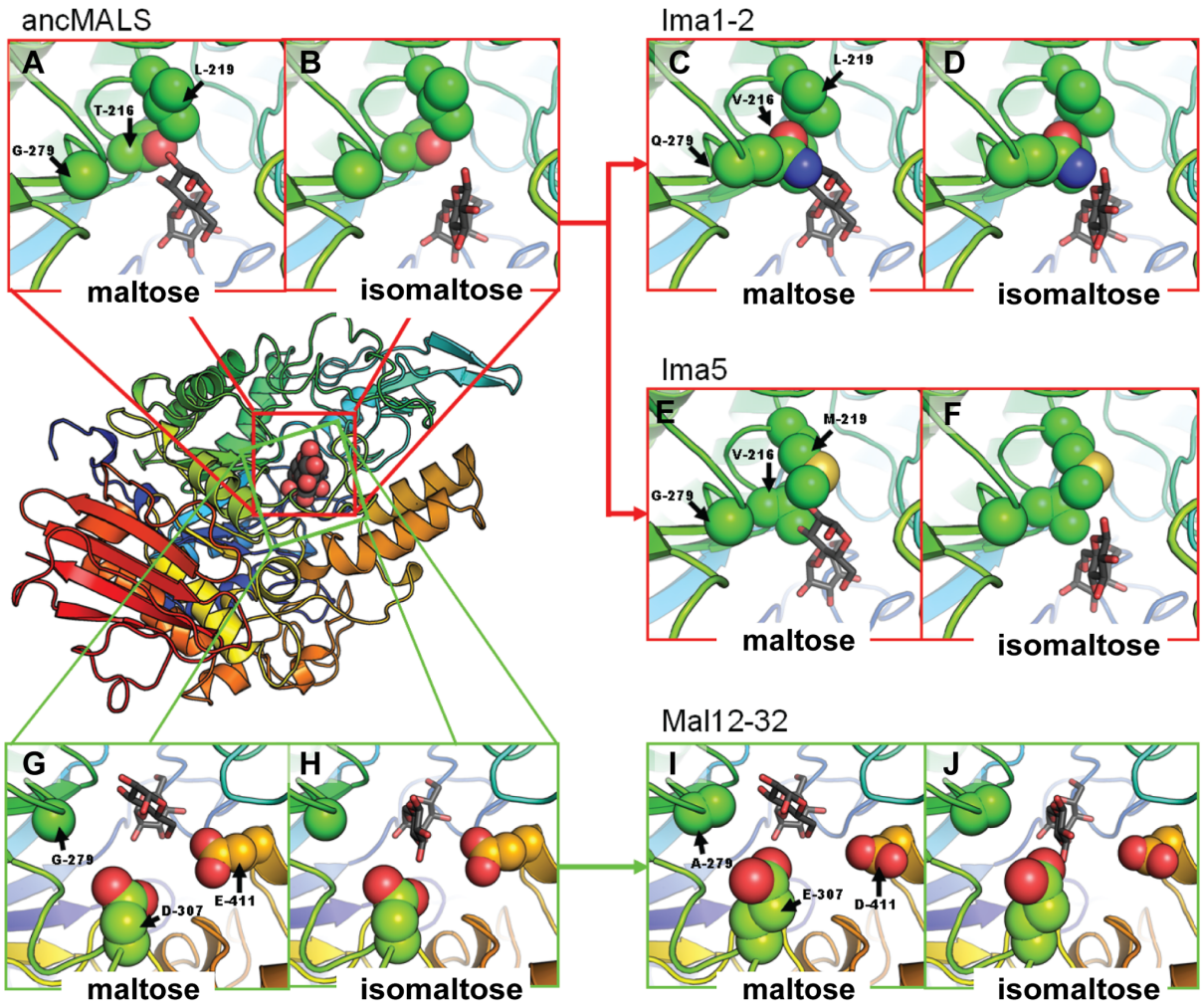


Figure 2.6: Evolution of the promiscuous ancMalS enzyme into isomaltose- and maltose-hydrolyzing enzymes. AncMalS is a promiscuous enzyme that hydrolyzes both maltose- and isomaltose-like substrates, whereas the present-day enzymes Ima1-2 and Ima5 preferentially hydrolyze isomaltose-like sugars and Mal12/32 preferentially hydrolyzes maltose-like sugars. First, the presence of a Thr or Val residue at position 216 affects the binding affinity of the enzyme through changes in the hydrophobic/hydrophilic interactions with the different substrate classes (panels A to D; see also supplementary figure D.10). The case of Ima1/2 and Ima5 (panels C to F) illustrates that an additional shift in substrate specificity can be obtained via different evolutionary routes. In the case of Ima1 and Ima2, the change of G279 to Q279 interferes with binding of maltose-like substrates, but the side chain of Gln can undergo polar interactions with isomaltose (panels C and D). The G218S and V278M changes cause additional subtle adaptations of the protein fold, causing Q279 to protrude further into the binding pocket, allowing optimal interaction with isomaltose (see also figure 2.2). The evolution of isomaltase activity in Ima5 also occurred via the introduction of steric hindrance in the binding pocket, although in this case the change involved was L219M (panels E and F). In ancMalS, residues D307 and E411 allow binding of both maltose- and isomaltose-like substrates (panels G and H). In the maltose-specific enzymes Mal12 and Mal32, however, these residues have evolved to E307 and D411 (panels I and J). These changes not only increase the affinity for maltose-like substrates but also make this site incompatible with isomaltose-like substrates. Subpanels are graphical representations of the binding pocket, with key amino acids depicted as spheres. Maltose and isomaltose are represented as sticks.

to these three interdependent mutations, we synthesized all possible intermediate anc*IMA1–4* enzyme variants with mutations at positions 218, 278, and 279. We subsequently expressed, purified, and measured activity of these enzyme variants. Figure 2.5 depicts the results of these enzyme assays and shows that these residues indeed affect substrate specificity, with the largest shift depending on the A to Q change at position 279, as expected from structural analysis. For one mutational path (GVA to GVQ to SVQ to SMQ), we observe a gradual increase in activity towards isomaltose and palatinose, demonstrating that there is a mutational path that leads to a consistent increase in isomaltase activity without traversing fitness valleys. Moreover, in keep with the stabilizing role of the mutations at positions

218 and 278, the A to Q change at position 279 along this path takes place before the two other mutations at positions 218 and 278.

Besides allowing the development of isomaltase activity in the Ima proteins, duplication also permitted further increase of the major ancestral function (hydrolysis of maltose-like sugars) in Mal12 and Mal32. Structural analysis reveals that this increase in maltase activity, from ancMalS to Mal12/32, is due to mutations D307E and E411D (see figure 2.6). These mutations increase the fit for maltose-like substrates but also completely block the binding of isomaltose-like substrates. Similar to what is seen for the evolution of ancMal-Ima to ancIma1–4, changes that increase the binding stability of one type of substrate cause steric hindrance that prevents binding of the other class of substrates. These signs of incompatibilities between substrates indicate that it is difficult to fully optimize one enzyme for both maltose-like and isomaltose-like substrates, with the highly suboptimal ancMalS being a notable exception. After partial optimization of ancMalS, duplication of anc*MAL-IMA* likely enabled further optimization of the conflicting activities in separate copies.

2.3.4 Different evolutionary routes can lead to similar changes in substrate specificity

Interestingly, the transition from ancMalS to Ima5 shows a similar shift in substrate specificity as the transition of ancMal-Ima to ancIma1–4. However, the residue at position 279, a key factor in the evolution of ancMal-Ima to ancIma1–4, remains unaltered in the evolution of ancMalS to Ima5. Instead, L219, a residue located proximal to position 279, has changed into M219 in the Ima5 enzyme (see figure 2.6). How can such seemingly very different mutations yield a similar change in substrate specificity?

Structural analysis shows that the L-to-M mutation at position 219 in Ima5 causes a very similar structural change as the G279Q change in ancIma1–4 (see figure 2.6), indicating that different evolutionary routes may produce a similar shift in activity. In both cases, the evolution of isomaltase-like activity involved introducing a residue that can stabilize isomaltose-like substrates but causes steric hindrance for maltose-like sugars in the binding pocket. Based on the phylogeny of binding pocket configurations and on our enzyme activity tests, this functional shift in the *IMA5* clade most likely occurred after a duplication in the common ancestor of *S. kluyveri* and *S. cerevisiae* (see figures 2.3 and 2.4).

2.3.5 Key residues in binding pocket of MalS enzymes show signs of positive selection

Next, we investigated the role of selective pressure during the different evolutionary transitions. We used MrBayes to construct a codon-based phylogeny under a GTR codon model of evolution, including 32 *MALS* genes that share the same nuclear genetic code. The resulting codon-based phylogeny was the same as the AA-based phylogeny generated using the LG+I+G protein model for all 50 sequences, apart from two exceptions in the anc*IMA1–4* clade. First, *S. mikitae IFO1815 c789* and *S. paradoxus N45* branch off separately from *S. kudriavzevii IFO1802 c1888* instead of together. Second, *S. kudriavzevii IFO1802 c1565* now branches off separately instead of multifurcating with *S. mikitae IFO1815 c633* and the branch leading to the *S. cerevisiae IMA2–4* genes. Relative branch lengths between genes

were similar to the branch lengths calculated under protein models of evolution. The topology of the codon-based tree is presented in figure 2.4.

GA Branch analysis[252] identified a branch class with an elevated ω (K_N/K_S) rate (ω = 0.66) but did not detect branch classes with $\omega>1$ that would be considered strong proof for positive selection (see figure 2.4; Materials and methods). These results, combined with our activity test results and the observed sequence configurations around the active center, suggest, however, that positive selection might have been operating on specific sites in three specific postduplication branches associated with enzyme activity shifts, namely the anc*IMA1–4*, anc*IMA5b*, and anc*MAL* branches, indicated with arrows on figure 2.4. We used the modified branch-site model A implemented in PAML[207] to assess positive selection along these branches (see Materials and methods). Results are presented in supplementary table D.4. For both the anc*IMA1–4* and anc*IMA5b* branches, *P*-values and parameter estimates suggest that a proportion of sites has strongly elevated ω values, consistent with the GA Branch results. On the branch from anc*MAL-IMA* to anc*IMA1–4*, four sites show signs of positive selection, with a posterior Bayes Empirical Bayes (BEB) probability $>$0.95, of which two, 216 and 279, are within 10 Å of the active center and known to be important for substrate specificity. On the anc*IMA5b* branch, four sites show signs of positive selection (BEB$>$0.95), including again site 216. For anc*MAL*, the null model (no positive selection) was not rejected at the 95% significance level. Both the corresponding parameter estimates and results of the GA Branch analysis, however, suggest relaxation of purifying constraints on this branch.

To get more support for the PAML branch-site test results, we performed an additional analysis using an alternative branch-site method that was recently implemented in the HyPhy package[257]. This method identified in total seven branches that possibly experienced positive selection: anc*IMA1–4* ($p<0.0001$), anc*IMA5b* (p=0.0232), anc*MALS* (p=0.0228), *S. kluyveri SAKL0A05698g* ($p<0.0001$), *K. thermotolerans GI: 255719187* ($p<0.0001$), the branch leading from anc*IMA5* to the anc*IMA5b* branch (p=0.0168), and finally the branch leading up to *S. cerevisiae IMA2*, *IMA3*, *IMA4*, and *YPS606* within the anc*IMA1–4* clade (p=0.0353). In other words, the anc*MALS*, anc*IMA1–4*, and anc*IMA5b* branches are suggested to have evolved under positive selection, together with four other branches. The branch-site method implemented in HyPhy currently does not allow the identification of specific sites that may have evolved under positive selection on these branches.

Together, our analyses indicate that some residues near the active pocket, in particular the key residues 216 and 279 that determine substrate specificity (see above), may have experienced positive selection in the postduplication lineages leading to isomaltose-specific enzymes. It should be noted, however, that the specificity and sensitivity of the currently available methods for detecting positive selection, in particular branch-site methods, is heavily debated[207–209,263,264]. Possible pitfalls include fallacies in the assumption that synonymous substitutions are neutral, a reported increase in the number of false positives due to sampling errors when the number of (non)synonymous substitutions and sequences is low, and potential inadequacies in the null and alternative models that are being compared, leading to difficulties with completely ruling out other explanations for perceived positive selection. For these reasons, the positive selection test results reported here should be approached as indications rather than definitive proof.

2.3.6 Recent duplicates *MAL12* and *MAL32* are maintained because of gene dosage effects

The previous results show how duplication of a promiscuous ancestral enzyme with limited activity towards two substrate categories allowed the evolution of separate enzyme clades that each show increased activities for a specific subset of substrates. The functional diversification of the different clades ensures their retention. However, why are recent, near-identical duplicates such as *MAL12* and *MAL32* conserved?

To investigate if selective pressure might protect the *MAL12*/*MAL32* duplicates, we determined the fitness effect of inactivating each of them. The results in supplementary figure D.11 show that strains lacking just one of the *MAL12* and *MAL32* paralogs show a considerable fitness defect compared to a wild-type strain when grown on maltose. These results suggest that gene dosage may play a primary role in preserving these recent paralogs[224]. Dosage effects increasing maltase and/or isomaltase activity may also have played a role after the earliest *MALS* duplications, before the duplicates were optimized for different activities.

2.3.7 Rapid expansion and functional divergence of the *MALS* subtelomeric gene family

Previous work has indicated that the *MALS* gene family is mainly present in the subtelomeric regions[243]. These are the repeat-rich and gene-poor regions proximal to the telomeres that are characterized by epigenetic silencing and increased rates of recombination and mutation[265]. An extensive study of these regions in different yeasts demonstrated that they are characterized by a high birth rate of new genes via small-scale duplications, most likely through increased recombination rates, which typically results in gene families that are larger compared to non-subtelomeric regions. Consequently, for *MALS* gene family members, there exist extensive differences in both the location and number of loci between different species and even strains within the same species[243]. Figure 2.1 demonstrates that several species exist that shared the *Saccharomyces* WGD but nevertheless do not possess any *MALS* genes, including *K. polysporus*, *S. castelii*, and *C. glabrata*. This indicates that their common ancestor had only few *MALS* genes that were completely lost in some lineages, but strongly expanded in others. Such changes can perhaps be linked back to life history traits, as *Candida* species for instance colonize mammals and presumably encounter enough simple preferred sugars in the blood and digestive tract[266]. The synteny of all seven present-day *S. cerevisiae* S288c loci (*IMA1-5* and *MAL12/32*) was therefore investigated using the Yeast Gene Order Browser[267] available at `http://ygob.ucd.ie`. Figure 2.7 illustrates the location of both *IMA1* and *MAL12* compared to the pre-duplication species *K. thermotolerans* (*Lachancea* clade) and *K. lactis* (*Kluyveromyces* clade), and the reconstructed pre-WGD *Saccharomyces* ancestor[268]. No apparent synteny was found with either the *Lachancea* or *Kluyveromyces* clade, nor with the reconstructed *Saccharomyces* pre-WGD ancestor. Searches using the other present-day *S. cerevisiae* S288c *MALS* loci lead to similar results (data not shown), suggesting that the syntenic signal of the pre-WGD *MALS* ancestor has been lost through the structural volatility of the subtelomeric regions. This confirms rapid

expansion and functional divergence of the *MALS* gene family most likely due to a selective advantage in some yeast species, whereas they were completely lost in others[243].

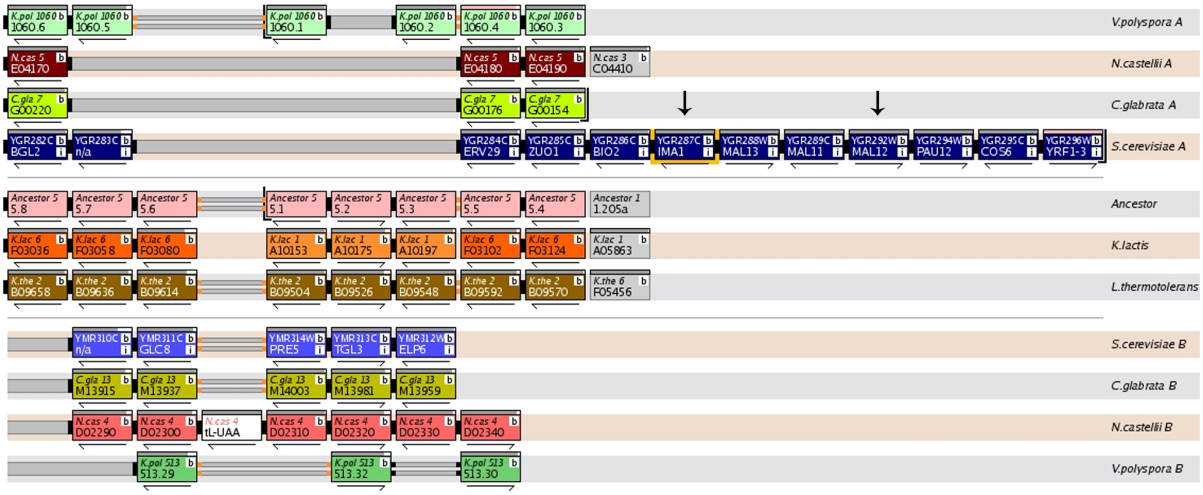


Figure 2.7: Synteny of *IMA1* and *MAL12* with other yeast species. Relationships are indicated based on searching the Yeast Gene Order Browser[267] available at `http://ygob.ucd.ie` with *IMA1* as a query gene whilst comparing the post-WGD *Saccharomyces* species *K. polysporus* (indicated here as *V. polyspora*), *S. castelii* (indicated here as *N. castelii*), *C. glabrata*, and *S. cerevisiae* for both post-WGD subgenomes (indicated here as A and B) with the reconstructed pre-WGD *Saccharomyces* ancestor[268] (indicated here as 'Ancestor'), and *K. thermotolerans* (indicated here as *L. thermotolerans*) and *K. lactis* from the *Lachancea* and *Kluyveromyces* clades, respectively. *IMA1* and *MAL12* are located on the *S. cerevisiae* A subgenome and are indicated with black arrows. Boxes in the same color represent loci from the same chromosome/contig per species track, whereas columns represent homologous genes in different species. No apparent pre-WGD corresponding loci can be found for either *IMA1* or *MAL12*, indicating that these loci originated through rapid expansion from other ancestral *MALS* loci in the subtelomeric regions[243].

2.4 Discussion

One of the major issues in the field of molecular evolution is the plethora of theoretical models and variants of models concerning the evolution of gene duplicates, with few of the claims supported by solid experimental evidence. On many occasions, inherent properties of the evolutionary process make it extremely hard to find or generate experimental evidence for a given model. However, recent developments in genome sequencing, evolutionary genomics, and DNA synthesis open up exciting possibilities. Using these new opportunities, we were able to resurrect ancient *MALS* genes and the corresponding enzymes to provide a detailed picture of the evolutionary forces and molecular changes that underlie the evolution of this fungal gene family. The *MALS* gene family is an ideal model for the study of duplicate gene evolution, since it underwent several duplication events and encodes proteins for which we could accurately measure different activities. The availability of multiple fungal genome sequences provided sufficient data to robustly reconstruct ancestral alleles and study the selective forces that propelled divergent evolution of the paralogs. Additionally, the existence of a high-quality crystal structure of one of the present-day enzymes made it possible to predict the functional effects of mutations and to study the mechanistic basis of suspected adaptive conflicts between the maltase-like and isomaltase-like subfunctions.

Our results paint a complex and dynamic picture of duplicate gene evolution that combines aspects of dosage selection and sub- and neofunctionalization (see figure 2.8). The preduplication ancMalS enzyme was multifunctional and already contained the different activities found in the postduplication

enzymes (the basic idea of subfunctionalization), albeit at a lower level. However, the isomaltase-like activity was very weak in the preduplication ancestor and only fully developed through mutations after duplication (increase of k_{cat}/K_m with one order of magnitude for isomaltose-like substrates from ancMalS to Ima1), which resembles neofunctionalization. The ancestral maltase-like activity also improved substantially but to a lesser extent (factor 6.9 on average from ancMalS to Mal12), which therefore perhaps fits better with the subfunctionalization model. Moreover, our activity tests on *mal12* and *mal32* mutants indicate that gene dosage may also have played a role in preserving *MALS* paralogs, especially right after duplication. This may not only have been the case for the recent *MAL12–32* and *IMA3–4* duplications but also for more ancient duplications involving multifunctional ancestors. In summary, whereas the classical models of dosage, sub-, and neofunctionalization are helpful to conceptualize the implications of gene duplication, our data indicate that the distinction between sub- and neofunctionalization is blurry at best and that aspects of all three mechanisms may intertwine in the evolution of a multigene family.

Although it is difficult to classify our results decisively under one of the many models of evolution after gene duplication, most of our findings agree with the predictions of the "Escape from Adaptive Conflict" (EAC) model[110,111,223,234], a co-option-type model in which duplication enables an organism to circumvent adaptive constraints on a multifunctional gene by optimizing the subfunctions separately in different paralogs. The EAC model makes three key predictions: (i) the ancestral protein was multifunctional, (ii) the different subfunctions could not be optimized simultaneously in the ancestral protein (or at least not in an evolutionarily easily accessible way), and (iii) after duplication, adaptive changes led to optimization of the different subfunctions in separate paralogs[111,230,269]. In general, our findings fit with these predictions: (i) we find that several of the ancestral preduplication maltase enzymes (ancMalS, ancMal-Ima, and ancIma5) were multifunctional; (ii) we provide evidence, through molecular modeling and activity tests of present-day enzymes, ancestors, and potential intermediates, that the maltase and isomaltase functions are difficult to optimize within one protein (but see also below); and (iii) we find that duplication resolved this adaptive conflict, and we find indications that positive selection might have driven key changes that optimized the minor isomaltase-like activity of the preduplication enzyme in one paralog, while the major maltase-like activity was further optimized in the other paralog.

Figure 2.2 and the statistical analysis in supplementary table D.3 indicate that the activity of the different enzymes changes significantly at certain points along the evolutionary path. Interestingly, the overall image that emerges suggests that the enzymes developed activity towards either maltose-like or isomaltose-like sugars, but not both. This pattern is most clear in the evolution of ancMal-Ima to ancMal and ancIma1–4. The postduplication improvement of the different activities present in the ancestral allele, with each of the new copies displaying increased activity for one type of substrate and concomitantly decreased activity towards the other substrate class, could be indicative of trade-offs in the evolution of the *MALS* gene family. However, the word "trade-off" implies that the two incompatible functions are both under selection, which is difficult to prove for the ancient enzymes. Moreover, our results indicate that for the ancient ancMalS enzyme, it is possible to simultaneously increase the activity towards both maltose-like and isomaltose-like substrates. Together, our analyses show that it is possible to optimize (to a certain extent) one function of a multifunctional enzyme without significantly reducing the other (minor) activity. However, analysis of the complete evolutionary path and molecular modeling of the active pockets of the enzymes shows that full optimization of both functions in a single enzyme is difficult to achieve,

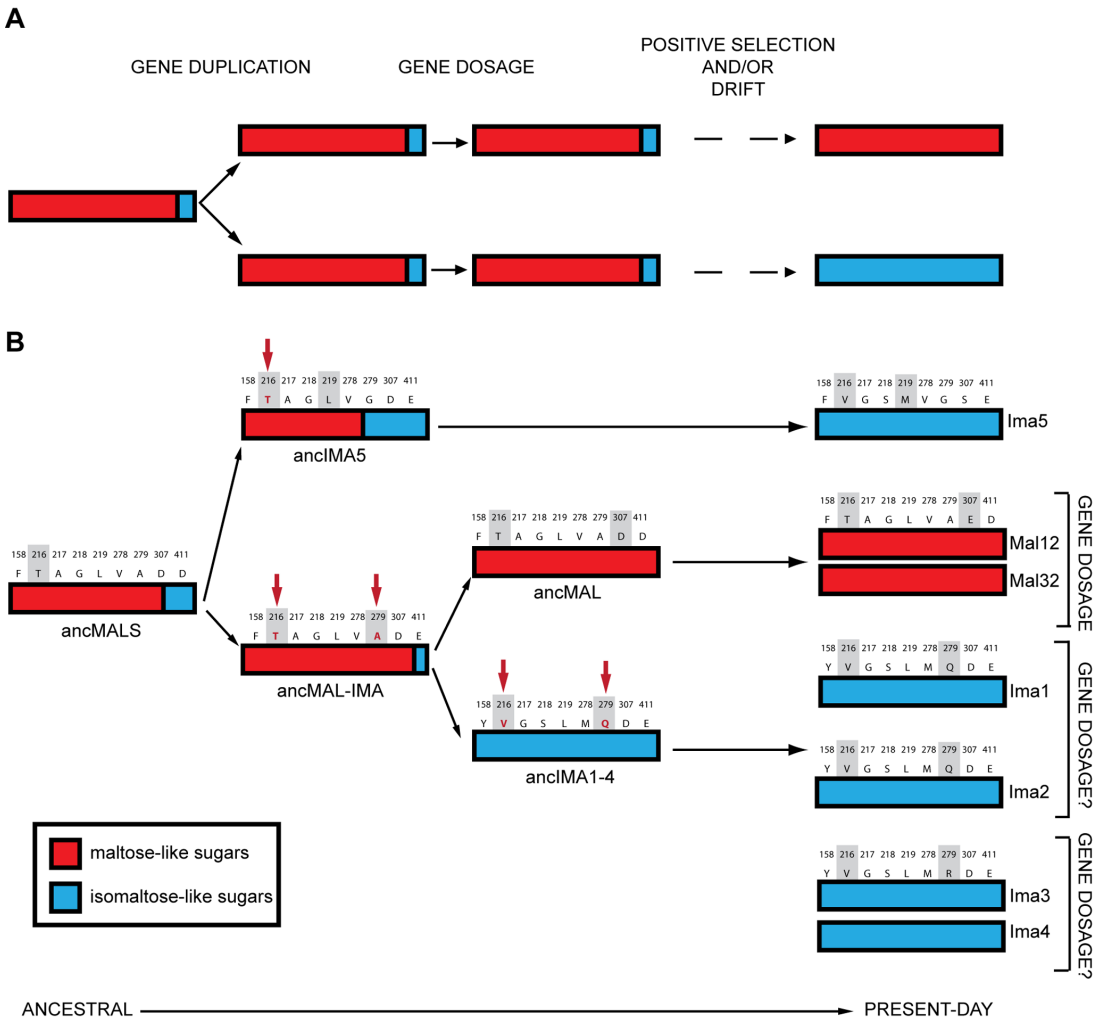


Figure 2.8: Multiple evolutionary mechanisms contributed to the evolution of the *MALS* gene family in *S. cerevisiae*. (A) Overview of evolutionary mechanisms in the evolution of an ancestral gene with two conflicting activities (major function, red; minor function, blue). Duplication can help resolve this 'adaptive conflict' by allowing optimization of these activities in two separate copies. Increased requirement for either of these activities, for example by changes in the environment, can first be met by duplication of the ancestral gene. Selection for increased gene dosage can help to preserve both copies until adaptive mutations optimize the different functions in separate copies. (B) Evolution of the promiscuous ancestral MalS enzyme into the seven present-day MalS alleles shows how different evolutionary forces contribute to the evolution of gene duplicates. Activity towards isomaltose-like sugars first existed only as a trace activity in the ancestral, preduplication enzyme. The nature of the binding pocket prevented simultaneous optimization of the major and minor function in the ancestral enzyme. Duplication allowed the (full) optimization of the two conflicting activities of the ancestral enzyme in separate copies. Several key residues in the enzymes' binding pocket responsible for these shifts in substrate specificity (shaded in grey) show signs of positive selection (indicated both in red and with red arrows; see also figure 2.4). Preservation of more recent, highly similar duplicate enzymes like Mal12 and Mal32 may be mediated through gene dosage effects (see also supplementary figure D.11). Sequences above each enzyme represent the nine variable residues in the binding pocket (numbering based on Ima1 sequence). AA changes that led to improvement of one of the hydrolyzing activities are shaded in grey.

due to steric hindrance for one substrate class when fully optimizing the active pocket for binding of the other substrate type. This problem can be most easily overcome by duplication of the enzyme, allowing optimization of the different subfunctions in different paralog copies, as can be seen in the transition of ancMal-Ima to ancMal and ancIma1–4.

While most aspects of our data fit with the EAC model, some results are more difficult to reconcile with the EAC theory. Specifically, one of the pillars of the EAC model is that positive selection drives the specialization of both paralogs after duplication. While our data demonstrate that duplication of anc*MAL-IMA* has led to optimization of both subfunctions in different duplicate lineages (maltase-like activity in anc*MAL* and isomaltase-like activity in anc*IMA1–4*), our selection tests only reveal indications

of positive selection in the anc*IMA1–4* lineage but not in the anc*MAL* lineage. Moreover, as discussed above, positive selection is difficult to prove[208,270], and we cannot exclude the possibility of both false positive and false negative artifacts.

Recently, some other likely examples of the EAC mechanism have been described[110,111,271–273]. These studies also presented plausible arguments for ancestral multifunctionality, adaptive conflict, and/or adaptive optimization of subfunctions in different paralogs, but as in the present case, none could provide strong experimental evidence for all three predictions made by the EAC model[269,274]. Instead of classifying the evolutionary trajectory of particular gene duplicates into one of the many models for gene duplication, it may prove more useful to distill a more general picture of duplicate evolution across a gene family that includes aspects of dosage selection, and sub- and neofunctionalization, like the one depicted in figure 2.8.

Our study is the first to investigate multiple duplication events in the same gene family in detail. Interestingly, we found that evolution has taken two different molecular routes to optimize isomaltase-like activity (the evolution of anc*MAL-IMA* to anc*IMA1–4* and anc*IMA5* to *IMA5*). In both cases, only a few key mutations in the active pocket are needed to cause shifts in substrate specificity. Some of these key mutations exhibit epistatic interactions. For example, the shift in substrate specificity occurring on the path from anc*MAL-IMA* to anc*IMA1–4* depends in part on mutations at three co-evolving positions (218, 278, and 279), but only one mutational path (279-218-278) shows a continuous increase in isomaltase-like activity. Interestingly, there is also a different path in the opposite direction (218-279-278) that shows a continuous increase in the ancestral maltase-like activity. This implies that the complex co-evolution at these three positions may be reversible. Interestingly, a recent study of the evolutionary history of plant secondary metabolism enzymes also identified AA changes that appear to be reversible[272], in contrast to the situation for, for example, glucocorticoid receptor evolution, where evidence was found for an “epistatic ratchet” that prevents reversal to the ancestral function[275].

It is tempting to speculate that complex mechanisms like those driving the evolution of the *MALS* gene family may be a fairly common theme. Many proteins display some degree of multifunctionality or promiscuity[276–278], just like the ancestral ancMal enzyme. Moreover, directed *in vitro* protein evolution experiments have shown that novel protein functions often develop from pre-existing minor functions[279,280]. Although the different functions within an enzyme often exhibit weak trade-offs, allowing optimization of the minor activity without affecting the original function of the enzyme[276,280,281], this may not always be the case. If there are stronger trade-offs between different subfunctions, duplication may enable the optimization of the conflicting functions in different paralogs.

While it is difficult to obtain accurate dating of the various duplication events, the duplication events studied here appear to postdate the divergence of *Saccharomyces* and *Kluyveromyces* clades, estimated to have occurred 150 mya[133], but predate the divergence of *Saccharomyces* and *Lachancea* and the yeast whole genome duplication, about 100 mya. *MALS* diversification may thus have happened around the appearance and spread of angiosperms (Early Cretaceous, between 140 and 100 mya[282]) and fleshy fruits (around 100 mya). Tentative dating results can be found in supplementary table D.6, but these should be approached with caution (see supplementary information D.3.2). The major shift in the earth’s vegetation caused by the rise of the angiosperms almost certainly opened up new niches, and it is tempting to speculate that duplication and diversification of the *MALS* genes may have allowed fungi

to colonize new niches containing sugars hydrolyzed by the novel Mal (Ima) alleles. In other words, the availability of novel carbon sources in angiosperms and fleshy fruits could have provided a selective pressure that promoted the retention of *MALS* duplicates and the ensuing resolution of adaptive conflicts among paralogs.

2.5 Acknowledgements

The authors thank Bodo Stern, Kevin Foster, Filip Rolland, Stijn Spaepen, Toon Nicolay, Bram Stynen, and all CMPG members for their help and suggestions. Statistical analyses were performed by Janick Mathys (BioInformatics Training and Services (BITS)-VIB).

2.6 Author contributions

I performed all the computational evolutionary analysis described in this chapter, with the exception of the strucutral modelling. I also contributed towards the analysis and description of all results within their proper evolutionary context for the resulting research article. Both were done under supervision and with significant contributions of Steven Maere.

Chapter 3

Inference of genome duplications

Kevin Vanneste, Yves Van de Peer, Steven Maere. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* **30**(1):177-190.

Abstract

Whole-genome duplications (WGDs), thought to facilitate evolutionary innovations and adaptations, have been uncovered in many phylogenetic lineages. WGDs are frequently inferred from duplicate age distributions, where they manifest themselves as peaks against a small-scale duplication background. However, the interpretation of duplicate age distributions is complicated by the use of K_S, the number of synonymous substitutions per synonymous site, as a proxy for the age of paralogs. Two particular concerns are the stochastic nature of synonymous substitutions leading to increasing uncertainty in K_S with increasing age since duplication and K_S saturation caused by the inability of evolutionary models to fully correct for the occurrence of multiple substitutions at the same site. K_S stochasticity is expected to erode the signal of older WGDs, whereas K_S saturation may lead to artificial peaks in the distribution. Here, we investigate the consequences of these effects on K_S-based age distributions and WGD inference by simulating the evolution of duplicated sequences according to predefined real age distributions and re-estimating the corresponding K_S distributions. We show that, although K_S estimates can be used for WGD inference far beyond the commonly accepted K_S threshold of 1, K_S saturation effects can cause artificial peaks at higher ages. Moreover, K_S stochasticity and saturation may lead to confounded peaks encompassing multiple WGD events and/or saturation artifacts. We argue that K_S effects need to be properly accounted for when inferring WGDs from age distributions and that the failure to do so could lead to false inferences.

For the author contributions, see page 74.

3.1 Introduction

The importance of gene duplication for evolutionary innovation has been widely recognized[93,94]. Small-scale gene duplications (SSDs) have been shown to be ubiquitous, and many eukaryotic genomes also contain traces of large-scale and even whole-genome duplications (WGDs)[128]. In particular, many plant species appear to have experienced one or more genome duplications in their evolutionary history[52,114,141,283]. Recent findings suggest that all extant seed plants are in fact paleopolyploids[136]. Examples of WGD events in other kingdoms include two rounds of WGD in the vertebrate ancestor and a third one in the teleost fish lineage[129–131], three WGDs in the ciliate *Paramecium tetraurelia*[132], and one WGD in the ancestor of the hemiascomycete *Saccharomyces cerevisiae* after its divergence from the *Kluyveromyces* clade[133,134]. In many species, duplicated transcriptional regulators and signal transducers have been retained in excess after WGDs, presumably because their loss is counteracted by dosage-balance effects[128,160,178,284,285]. Several authors suggest that this regulatory spandrel might have facilitated the evolutionary innovations and/or diversifications observed in many post-WGD lineages[52,128,160,161,193,283]. However, the occurrence and timing of WGDs and the precise nature of their link with evolutionary innovations and increased biological complexity remain important topics of discussion[52,150,215,286,287].

Lynch and Conery[227] were among the first to investigate the overall degree of duplicate loss and retention within eukaryotic genomes. They demonstrated that age distributions of duplicates retained from small-scale duplications are typically L-shaped, with many recent duplicates and fewer older duplicates, due to the fact that most newly created gene duplicates are eventually lost. Some age distributions exhibit additional peaks superimposed on the L-shaped background, representing sudden bursts of new gene duplicates created by larger-scale duplication events in the evolutionary past of the species, such as aneuploidy events or WGDs (see figure 3.1).

Although such WGD peaks can be very prominent, this is not always the case and they can sometimes hardly be distinguished from the small-scale duplication background[176]. Schlueter et al.[177] fitted mixtures of one to five normal components, representing WGD events, to empirical age distributions and compared different WGD scenarios by means of likelihood ratio tests. Cui et al.[135] first fitted a null model, that is, a constant rate duplicate birth–death model without WGDs, and applied mixture modeling techniques to detect WGDs if the null hypothesis was rejected. In addition to the aforementioned techniques, Barker et al.[182] used the program SiZer[187] to identify significant peak features in age distributions and boost confidence in the WGDs inferred by mixture modeling. Maere et al.[178] introduced a different approach to infer WGDs, simulating empirical age distributions with a quantitative duplicate population dynamics model that takes into account both SSD and WGD modes of gene duplication.

The use of age distribution-based methods for WGD inference offers several advantages. These methods generally have a relatively low computational cost, they have been shown successful if only a limited part of the paranome is available, for example, based on expressed sequence tag (EST) collections[135], and they do not require positional information on the paralogs. The latter is an important advantage over another type of methods frequently used to detect WGDs, namely synteny-based methods that search for syntenic gene blocks in and between different genomes to unravel their WGD history[167].

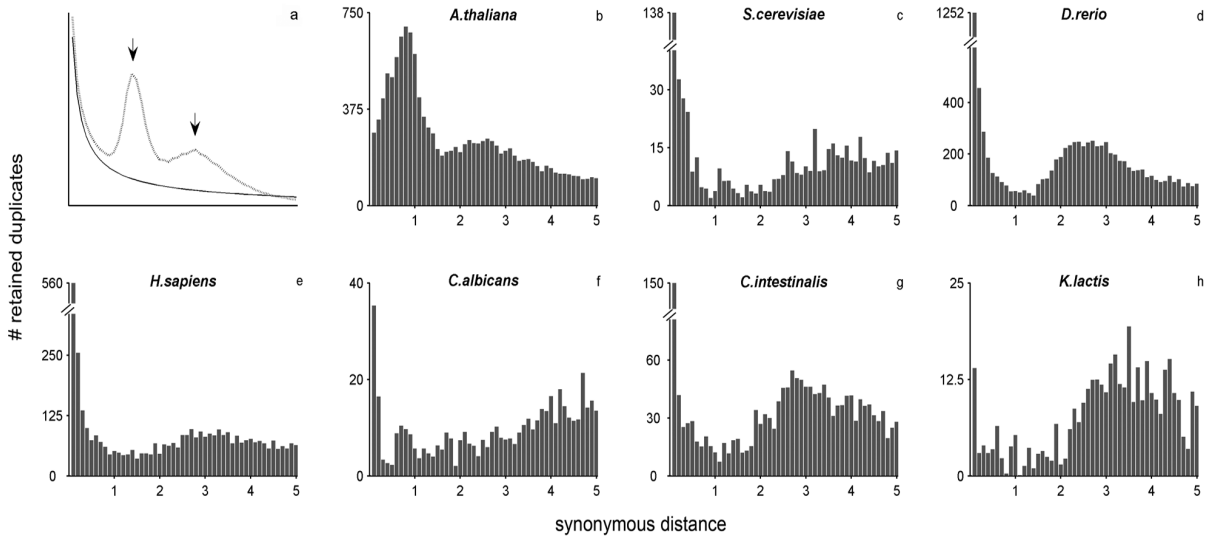


Figure 3.1: Examples of empirical K_S-based age distributions. (a) Illustration of two possible age distribution shapes. The solid line represents genomes impacted only by small-scale duplications (SSDs). The initial peak represents newly duplicated genes that are continuously being generated by SSD events (e.g., tandem duplications). The decreasing slope following this initial peak outlines the steady decrease of retained duplicates over time, reflecting the fact that most duplicates are eventually lost. The dotted line represents genomes impacted by one or more whole-genome duplications (WGDs). The SSD mode is distinctively present but superimposed are WGD components (indicated by black arrows). (b–h) Empirical age distributions for several species of interest.

Age distributions have therefore become a popular tool to investigate the (non-)occurrence of WGDs in species ranging from vertebrates[288,289] to arthropods[290], and especially plants[52,135,136,138,176–186].

There are, however, also intrinsic difficulties associated with the interpretation of duplicate age distributions, related to the use of proxies for the age of duplicated gene pairs. The use of such proxies is necessitated by the difficulties associated with absolute dating of duplication events. The most commonly used measure of age since duplication is the number of synonymous substitutions per synonymous site (K_S) between paralogs. Because synonymous substitutions do not change protein products and are therefore putatively neutral[175], they are thought to accumulate at an approximately constant rate. However, there are certain issues to take into account when using K_S as an age proxy. A first concern is the stochastic nature of synonymous substitutions, whereby the synonymous substitution levels of simultaneously duplicated paralogous pairs show increasing variation with time since duplication[5]. As a consequence, gene duplication peaks generated by older WGD events will be progressively flattened and dispersed in K_S-based age distributions, and they will gradually blend into the L-shaped SSD background, an effect that is exacerbated by ongoing duplicate loss[135,176,177]. The second concern is K_S saturation effects. With increasing age since duplication, paralogous pairs start to accumulate multiple substitutions per site, and the evolutionary models employed for K_S estimation are unable to fully correct for this, leading to K_S estimates that are systematically lower than the real synonymous substitution levels and eventually saturate[5]. Because of this saturation effect, older gene duplicates are wrongfully lumped together at lower K_S values, and an artificial saturation peak may be generated in the age distribution, which could be mistaken for a WGD peak[177,213]. The combination of these two factors could potentially lead to scenarios wherein a true older WGD peak is dispersed in the same range of the age distribution where saturated K_S estimates accumulate. None of the solutions devised so far for discerning WGD events account properly for stochastic and saturation effects on K_S. Most authors have avoided these issues by only considering

age distributions until a K_S cutoff of 1 or 2[135,176,177,182]. Usually, only K_S estimates lower than 1 are considered reliable, and beyond this threshold, saturation effects are expected to become important[5]. Discarding the tail of the age distribution after a relatively low cutoff value does, however, limit WGD inference to more recent events.

Here, we use a two-step approach to investigate how K_S stochasticity and saturation affect the shape of K_S-based age distributions for various species. First, we simulate the synonymous evolution of coding sequences (CDS) for different time spans, taking into account species-specific genome characteristics, and we re-estimate the corresponding synonymous distances under the same evolutionary model to quantify the aforementioned effects. Second, we incorporate these effects in a duplicate population dynamics model and simulate the K_S-based age distributions corresponding to predefined real age distributions with and without WGDs, to examine how K_S stochasticity and saturation interfere with the inference of WGDs.

3.2 Material and methods

3.2.1 Data collection and preparation

The complete genome sequences of *Arabidopsis thaliana*, *Candida albicans*, and *Kluyveromyces lactis* were obtained from the PLAZA platform (`bioinformatics.psb.ugent.be/plaza`)[171], the *Candida* Genome Browser (`www.candidagenome.org`)[291], and Génolevures (`www.genolevures.org`)[292], respectively. Genome sequences for other species (*S. cerevisiae*, *Homo sapiens*, *Ciona intestinalis*, and *Danio rerio*) were collected through Ensembl (`www.ensembl.org`)[293]. Only protein coding genes were kept for further analysis. All genes flagged as either suspected or known pseudogenes by the different platforms were removed. If alternative transcripts were available, only the one with the longest CDS was kept. This resulted in data sets of in total 27,363, 6,668, 6,006, 20,488, 22,826, 5,076, and 9,330 sequences for *A. thaliana*, *S. cerevisiae*, *C. albicans*, *H. sapiens*, *D. rerio*, *K. lactis*, and *C. intestinalis*, respectively.

3.2.2 Construction of empirical K_S age distributions

For each species, an all-against-all protein sequence similarity search was performed using BLASTP with an E-value cutoff of e$-$10. Species gene families were subsequently built through Markov Clustering[294] using the mclblastline pipeline (v10-201) (`micans.org/mcl`). For each gene family, a protein alignment was constructed using MUSCLE (v3.8.31)[295]. This alignment was used as a guide for aligning the DNA sequences of gene family pairs. Only gene pairs with a minimum gap-stripped alignment length of 100 amino acids were considered for further analyses. K_S estimates were obtained through maximum likelihood estimation (MLE) using the CODEML program[204] of the PAML package (v4.4c)[250]. Codon frequencies were calculated based on the average nucleotide frequencies at the three codon positions (F3x4), and a constant K_N/K_S (reflecting selection pressure) was assumed for every pairwise comparison (codon model 0), because a single pair of sequences generally does not provide sufficient information to detect variability in selection pressure. For each pairwise comparison, K_S estimation was repeated five times to avoid suboptimal estimates because of MLE entrapment in local maxima. Only K_S estimates

lower than 5 were considered in the construction of empirical age distributions. Gene families were subdivided into subfamilies for which K_S estimates between genes 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), an average linkage clustering approach was used as described in Maere et al.[178]. Briefly, for each gene family, a tentative phylogenetic tree was constructed by average linkage hierarchical clustering, using K_S as a distance measure. For each split in the resulting tree, corresponding to a duplication event, all m K_S estimates between the two child clades were added to the K_S distribution with a weight $1/m$, so that the weights of all K_S estimates for a single duplication event sum up to one.

3.2.3 Simulating synonymous evolution

Synonymous evolution model

Two major biases influencing synonymous evolution are documented to vary between different species. First, transition bias, that is, an excess of transitional over transversional substitutions, is a mutational bias that can be observed at synonymous sites[296,297]. Second, many species show a weak to strong preference for particular codons in a set of synonymous codons, an effect referred to as codon usage bias[298]. For the evolutionary simulations, we employed a simplified version of the codon model proposed by Goldman and Yang[204], as described by Yang and Nielsen[299], for the following reasons. First, as a codon model, it can account for both transition bias and codon usage bias[204,300]. Second, codon models are thought to outperform nucleotide and amino acid models in evolutionary analyses of protein coding genes[301]. Third, it is a mechanistic model allowing incorporation of features of the underlying process of evolution[302]. Fourth, estimation of K_S values between the original and synonymously evolved sequences under the same evolutionary model is straightforward, by virtue of its implementation in the CODEML program[204] of the PAML package[250].

Briefly, the substitution rate from codon i to codon j is given by the substitution rate matrix $Q=\{q_{ij}\}$, with $q_{ij}=\pi_j$ if i and j differ by a synonymous transversion, $q_{ij}=\kappa\pi_j$ if i and j differ by a synonymous transition, and $q_{ij}=0$ otherwise, because we only simulate synonymous evolution. π_j is the equilibrium frequency of codon j (reflecting codon bias), and κ is the mutational transition/transversion rate ratio (reflecting transition bias). For each species, the values of the 61 π_j parameters were calculated from all available protein coding genes, under the assumption that the observed codon frequencies do not differ drastically from the equilibrium frequencies[303]. To extract a genome-wide value for parameter κ, we averaged the κ values obtained from all possible pairwise comparisons among gene family members. Because previous work has indicated that likelihood-based methods outperform distance-based methods for calculating κ[304], we used the PAML package to extract κ for each pairwise comparison. The resulting κ values, however, still exhibited considerable heterogeneity. This has been observed before and has been attributed to the large estimation errors associated with κ estimation of short sequences, rather than true variance of κ between genes of the same genome[305]. We indeed observed a striking relationship between the variability of κ estimates and (stripped) sequence alignment length for all seven species, as illustrated in

supplementary figure E.1. Instead of taking the arithmetic mean to calculate a genome-wide κ value, we therefore calculated a weighted average of the κ estimates using the alignment lengths as weights:

$$\kappa = \frac{\sum n_i \kappa_i}{\sum n_i} \tag{3.1}$$

κ represents the genome-wide estimate for the transition bias, while n_i and κ_i represent the individual alignment lengths and estimates, respectively. For each species, the corrected value for κ is indicated on supplementary figure E.1.

By extracting the above information from the genome data sets, one derives the substitution rate matrix Q=$\{q_{ij}\}$. The diagonal elements of Q are determined by the requirement that the row sums are zero[299]:

$$q_{ii} = -\sum_{i \neq j} q_{ij} \tag{3.2}$$

Furthermore, the elements of Q are multiplied by a scaling factor to normalize the expected number of nucleotide substitutions per codon and per time unit to one, thereby ensuring that evolutionary simulation times t can be determined in terms of the desired expected number of substitutions (see further)[299]:

$$-\sum_i \pi_i q_{ii} = \sum_i \pi_i \sum_{i \neq j} q_{ij} = 1 \tag{3.3}$$

$q_{ij}\Delta t$ gives the probability that any given codon i will change to a different codon j in an infinitesimally small time interval Δt. The probability that a given codon i will change to a different codon j in a time interval $t>0$ is given by its transition probability $p_{ij}(t)$. The transition probability matrix $P(t)$=$\{p_{ij}(t)\}$ can be derived from Q by solving $P(t)$=e^{Qt}. We avoided numerically solving the matrix exponential by simulating the waiting times of a Markov chain, as described by Yang[306] for nucleotides and briefly summarized hereafter for codons. For a single codon position, let q_i=$-q_{ii}$=$\Sigma_{i \neq j} q_{ij}$ be the total exchange rate of the current codon i and t the total simulation time. A random waiting time s is drawn from an exponential distribution with mean $1/q_i$. If $s>t$, no change occurs in the time span t. If $s<t$, codon i is exchanged for another (synonymous) codon j with probability q_{ij}/q_i. Both the waiting times and transition probabilities are thus fully specified by the instantaneous rates given by Q. The remaining time t then becomes $t-s$, and a new random waiting time is drawn from an exponential distribution with mean $1/q_j$ (j being the new codon) until $s>t$. For a stretch of codons, the total rate of exchange q is equal to the sum of the rates across the individual codon positions in the sequence, and s is drawn from an exponential distribution with the mean equal to $1/q$. If $s<t$, the codon site to be mutated is randomly chosen with a probability proportional to its exchange rate q_i, and the codon i is exchanged for a codon j with probability q_{ij}/q_i, as before.

Running the simulations

The evolutionary simulation time t needed to produce a given expected number of (non-)synonymous substitutions per (non-)synonymous site (K_S and K_N, respectively) is given by[299]:

$$t = K_S \frac{3S}{(S+N)} + K_N \frac{3N}{(S+N)} \tag{3.4}$$

S and N are the number of synonymous and non-synonymous sites, respectively. Because we only simulate synonymous evolution, K_N equals zero, and the second part of the equation can be ignored. Furthermore, (S+N)/3 equals the total number of codons in the sequence, denoted L_c, so equation 3.4 can be rewritten as:

$$t = K_S \frac{S}{L_c} \tag{3.5}$$

For each species, we use the genome-wide average number of synonymous sites per codon S/L_c as the conversion factor to calculate the simulation time t needed to obtain a given K_S on average. We let the Markov chain run in time step equivalents corresponding to an expected K_S increase of 0.1 until a total simulation time $\sim K_S$ of 25, as we observed that the K_S estimates for all species had approximately reached complete saturation by then. More precisely, a real protein coding gene was taken as the 'ancestor gene' at time t=0. This gene was then synonymously evolved in time step equivalents corresponding to an expected K_S increase of 0.1. At each time step, the K_S between the ancestral and evolved gene was re-estimated with CODEML under the same evolutionary model as used for the simulations[299]. This was done for all available protein coding genes for each species, resulting in 27,363, 6,668, 6,006, 20,488, 22,826, 5,076, and 9,330 synonymously evolved genes at each time step for *A. thaliana*, *S. cerevisiae*, *C. albicans*, *H. sapiens*, *D. rerio*, *K. lactis*, and *C. intestinalis*, respectively. CODEML settings were the same as outlined earlier for the construction of empirical age distributions. Geometric means and standard deviations of the resulting K_S estimates for each simulation time were calculated on the log-transformed distributions because K_S estimates are expected to be lognormally distributed[307].

3.2.4 Incorporation of K_S characteristics in simulated age distributions

Duplicate population dynamics model

We use the duplicate population dynamics model described in Maere et al.[178] to simulate age distributions of duplicated genes. Briefly, the simulation starts from a number of founder genes G_0 and simulates the birth and death of gene duplicates in SSD and WGD duplication modes in time steps corresponding to an expected K_S interval of 0.1. The principal equations of the model are as follows:

$$D_0(1,t) = \nu \left(\sum_{x'=1}^{\infty} D_{tot}(x', t-1) + G_0 \right) \tag{3.6}$$

$$D_1(1,t) = \left[\sum_{x'=1}^{\infty} D_{tot}(x', t-1) + G_0 \right] \delta(t, t_1) \tag{3.7}$$

$$D_i(x,t) = D_i(x-1,t-1)\left[\frac{x}{x-1}\right]^{-\alpha_i} \quad x > 1 \qquad i = 0,1 \tag{3.8}$$

$$D_{\text{tot}}(x,t) = \sum_i D_i(x,t) \tag{3.9}$$

$D_i(x,t)$ stands for the number of retained duplicates in the ith duplication mode (i=0 for SSD and i=1 for WGD) having an age x (measured in 0.1 K_S equivalents) at time step t in the simulation. $D_{\text{tot}}(x,t)$ is the total number of duplicates of age x at time step t. Equation 3.6 describes the birth of duplicates in the continuous SSD mode at a birth rate of ν new duplicates per time step. Equation 3.7 models a discrete WGD at time point t_1 in the simulation. Equation 3.8 describes the loss of duplicates from one time step to the next, which follows a power law decay with constant α_0 for the SSD mode and α_1 for the WGD mode. Equation 3.9 couples equations 3.6, 3.7, and 3.8. A more detailed description of the model can be found in Maere et al.[178].

Age versus K_S distributions

The model described earlier produces 'real age' distributions without K_S stochasticity and saturation effects, featuring discrete WGD peaks. To convert these age distributions into K_S-based age distributions, we incorporated the K_S estimation biases gathered from our synonymous evolution simulations using the following smoothing procedure:

$$D'(x,t_n) = \sum_{\lambda=1}^{n} D_{\text{tot}}(\lambda,t_n).f_\lambda(x) \tag{3.10}$$

$D'(x,t_n)$ represents the K_S-based age distribution after smoothing. $D_{\text{tot}}(\lambda,t_n)$ is the modeled 'real age' distribution after n time steps, with λ the age bin. $f_\lambda(x)$ represents the species-specific frequency distribution of K_S estimates for genes that were synonymously evolved for a time interval corresponding to λ, as described before (see figure 3.2 and supplementary figures E.2–E.8). To investigate sample size effects, we used a second approach where for each age λ, $D'(x,t_n)$ K_S estimates were randomly sampled (with replacement) from $f_\lambda(x)$ to generate the K_S-based age distribution D'.

3.3 Results

3.3.1 Characterization of K_S stochasticity and saturation effects through synonymous evolution simulations

We simulated the synonymous evolution of sequences to characterize how the combined effects of K_S saturation and the stochastic nature of the synonymous substitution process influence K_S dating for different species. We used real protein CDS to generate data sets of synonymously evolved genes, artificially evolving them for certain amounts of time corresponding to predefined expected K_S values (hereafter referred to as synonymous ages). Afterward, the K_S distances between the real and synonymously evolved sequences were estimated under the same evolutionary model as used for the simulations,

using CODEML[204]. The results are summarized in figure 3.2, and detailed results are presented in supplementary figures E.2–E.8. The geometric mean and mode of the estimated K_S distributions can be used to assess K_S saturation effects, whereas the standard deviation of K_S estimates reflects the impact of K_S stochasticity and estimation errors.

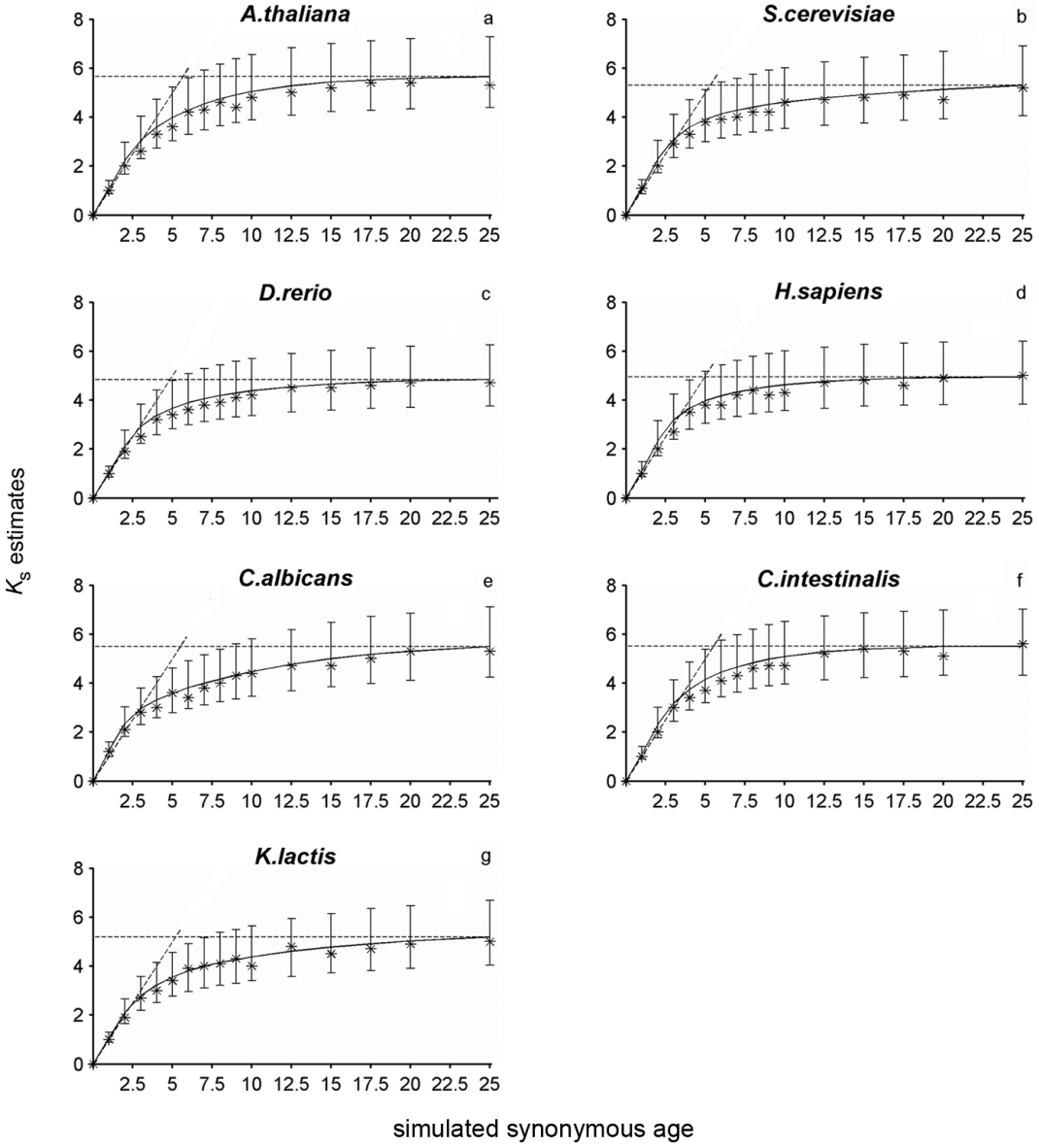


Figure 3.2: Summarized results of our artificial synonymous evolution approach for several species. Solid black lines connect the geometric means of K_S estimates for the simulated synonymous ages. Asterisks and flags represent the mode and standard deviations of K_S estimates, respectively. The horizontal dotted line indicates the position of the geometric mean for an evolutionary time span corresponding to an expected K_S of 25, whereas the second dotted line indicates the x=y linear. Full results are presented in supplementary table E.1.

Figure 3.2a depicts the trends for *A. thaliana*. For a synonymous age of 1, the mode of K_S estimates is equal to the expected K_S, with the geometric mean offset to 1.1, and a lower and upper standard deviation of 0.24 and 0.30, respectively. Most K_S estimates are thus found in the neighborhood of the expected K_S value, with only minor variation. At a synonymous age of 2, the mode of K_S estimates is still equal to the expected K_S, with the geometric mean offset to 2.2 and a lower and upper standard deviation of 0.55 and 0.74, respectively. At a synonymous age of 3, the mode of K_S estimates has shifted to 2.6, with a geometric mean of 3.1, and the lower and upper standard deviations increase to 0.75 and 0.99,

respectively. At this point, K_S saturation becomes noticeable. Saturation and K_S variability continue to increase for higher synonymous ages. At synonymous ages of 5, 10, 15, and 20, the mode (geometric mean) shifts to 3.6 (4.0), 4.8 (5.1), 5.2 (5.4), and 5.4 (5.6), whereas the lower and upper standard deviations increase to 0.94 and 1.24; 1.16 and 1.50; 1.21 and 1.56; and 1.26 and 1.62, respectively (see also supplementary table E.1). At higher synonymous ages, the K_S distribution characteristics stabilize as saturation becomes nearly complete.

Similar patterns are evident for the other six species presented in figure 3.2 (*S. cerevisiae*, *D. rerio*, *H. sapiens*, *C. albicans*, *C. intestinalis*, and *K. lactis*). The extent of K_S saturation and K_S variability seems to be within bounds until a synonymous age of 2, after which both start to manifest themselves increasingly. Although K_S estimates higher than 1 are generally considered unreliable[5], our results suggest that K_S saturation and stochastic effects remain fairly acceptable until at least a synonymous age of 2. There are, however, considerable differences between species in the onset and degree of K_S saturation. The K_S curves for *D. rerio*, *C. albicans*, and *K. lactis* flatten out more quickly than for other species, indicating that there is a quicker onset of K_S saturation. The *A. thaliana* and *C. intestinalis* curves saturate more slowly, whereas *H. sapiens* and *S. cerevisiae* exhibit intermediate saturation characteristics. At synonymous ages of 5/10, the geometric means for *D. rerio*, *C. albicans*, and *K. lactis* are located around 3.7/4.4, compared with values around 4.1/5.1 for *A. thaliana* and *C. intestinalis* (see also supplementary table E.1). Additionally, the K_S curves for some species, in particular *D. rerio* and *H. sapiens*, plateau at a considerably lower level than for other species. Interestingly, a quicker onset of saturation is not necessarily linked to a lower plateau level, as becomes evident when comparing, for example, the *C. albicans* and *H. sapiens* curves on figure 3.2.

3.3.2 The impact of saturation effects on age distributions

SSD age distributions are characterized by a saturation peak

We adapted the population dynamics model introduced by Maere et al.[178] to investigate how K_S stochasticity and saturation, as characterized by our synonymous evolution simulations, will affect the shape of K_S-based age distributions. The population dynamics model takes into account SSD and WGD events and simulates a 'real age' distribution at first, ignoring effects related to the use of age proxies such as K_S. K_S stochasticity and saturation effects were included by redistributing the duplicate counts in each age bin according to the distribution of K_S estimates obtained for that age in the synonymous evolution simulations. We first modeled age distributions considering only a SSD mode of evolution. The number of required parameters is minimal in this case (equations 3.6-3.9): a number of founder genes (G_0), the birth rate of new duplicates per time step (ν), and a power law decay constant for duplicate loss (α_0). G_0 was arbitrarily set to 10,000 genes. ν and α_0 were put to 0.03 and 0.80, respectively, based on parameter estimates obtained for *A. thaliana* by Maere et al.[178]. In total, we constructed four SSD age distributions for each species, running the simulation for increasing time spans corresponding to maximum duplicate ages (in K_S equivalents) of 5, 10, 15, and 20. Results for all seven species are presented in figure 3.3.

A striking observation is that for each species, and for each simulated time span, the simulated K_S distributions clearly deviate from the typical L-shape of real age distributions as advocated by Lynch and Conery[159,308]. In all cases, a secondary peak appears in the tail of the distribution. This peak results

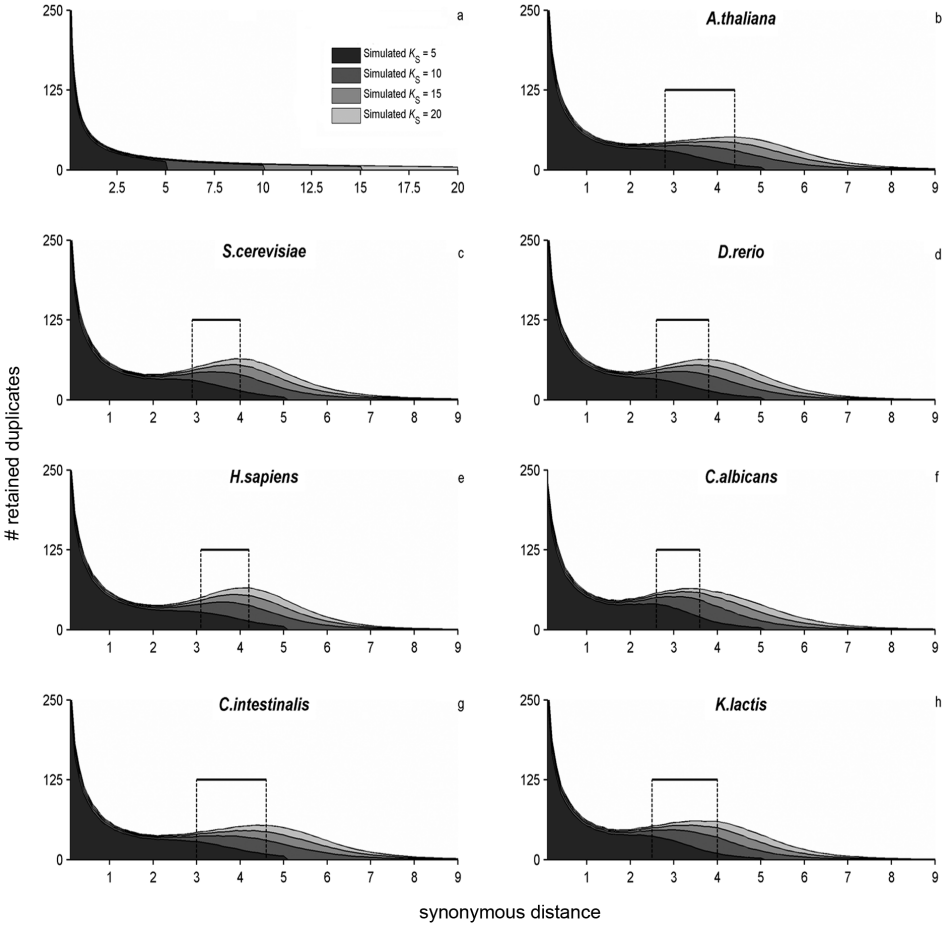


Figure 3.3: SSD age distributions are characterized by a saturation peak. (a) SSD 'real age' distributions generated by our population dynamics model over increasing evolutionary time spans, without correcting for the effects of K_S saturation and stochasticity. (b–h) SSD 'K_S-based age' distributions for the species indicated on top of the panels, generated from the real age distributions displayed in panel (a) by incorporating species-specific K_S saturation and stochasticity effects, as characterized by our synonymous evolution simulations. For all species, incorporation of K_S effects results in a SSD saturation peak. Solid black lines on top of the distributions indicate the range of the saturation peak mode across evolutionary time spans (see supplementary figure E.9).

from the fact that old duplicates are deposited at earlier synonymous distances because of K_S saturation effects, and it is therefore referred to as the saturation peak. Saturation peaks are generally spread out over a broad K_S range, reflecting the fact that for the older duplicates in the saturation regime, stochastic K_S variation and general uncertainty in K_S estimates become increasingly important. The occurrence of a saturation peak is independent of the exact model parameters used (see further).

For all seven species, age distributions considering longer time spans exhibit progressive displacement of the saturation peak to higher ages and higher elevation above the L-shaped background. For *A. thaliana*, for instance, the mode of the saturation peak shifts from ~2.8 to 4.4 for simulated evolutionary time spans going from 5 to 20. This is because an age distribution built over a longer evolutionary time span will contain more retained duplicates in the age range where K_S saturation is an issue, and the average saturation effects will progressively shift to the higher end of the saturation curves presented in figure 3.2.

Species-specific differences in the location of the saturation peaks can be reconciled with the results of the synonymous evolution simulations described in the previous section. It was noted above that *D. rerio*, *C. albicans*, and *K. lactis* saturate more quickly than, for example, *A. thaliana* or *C. intestinalis*.

Accordingly, the mode of the saturation peak is consistently located at a smaller K_S in these species than in other species (see figure 3.3). *H. sapiens* and *S. cerevisiae* again exhibit intermediate characteristics. The differences between species become more pronounced for age distributions considering longer evolutionary time spans, because more retained duplicates fall in the saturation regime.

Interestingly, for none of the species, the mode of the saturation peak reaches the saturation limit shown on figure 3.2, even for a simulation time span of 20 (see supplementary figure E.9). This reflects the fact that older duplicates close to the saturation limit are always outnumbered by younger duplicates in an earlier saturation stage, because of the dynamics of duplicate loss. Additionally, variation of the model parameters impacting duplicate birth (ν) and loss (α_0) over sensible ranges (ν from 0.01 to 0.05 and α_0 from 0.65 to 1.10) have little impact on the location of peak modes (see supplementary figures E.10–E.16). Therefore, the saturation peak in the empirical age distribution of a particular species (see figure 3.1) will likely be located in the corresponding peak mode interval depicted in figure 3.3 (see further). Where exactly in this interval empirical saturation peaks will manifest themselves is mainly dependent on how many ancient duplicates can still be identified. Indeed, unlike in our idealized model, older duplicate pairs may have diverged, for example, through (non-)synonymous substitutions, insertions, and deletions, to an extent that they can no longer be recognized as such. Assuming an average synonymous substitution rate in the order of 10 per synonymous site per billion years (from 2.5/ss/By for mammals to 15/ss/By for invertebrates[308]), duplicates with a synonymous age of 20 may be well over a billion years old.

The number of genes in the age distribution impacts its shape

Empirical age distributions often have a relatively rugged appearance because of the finite numbers of duplicates involved, especially in higher age bins. This is particularly the case for unsequenced organisms, for which age distributions are constructed from incomplete EST collections. To investigate the effects of limited sample size on the identifiability of saturation peaks, we used an alternative approach to include K_S stochasticity and saturation effects in the modeled age distributions, based on direct sampling of the K_S values for the duplicates in each age bin from the corresponding K_S estimate distribution obtained in our synonymous evolution simulations. The results of performing this sampling procedure on simulated *A. thaliana* age distributions with different numbers of founder genes (G_0) are presented in figure 3.4 (values for ν and α_0 were kept at 0.03 and 0.80 as before). Results for other species are presented in supplementary figure E.17.

A first observation is that the general characteristics of the shape of the age distribution do not change. A saturation peak is still present in the tail of the distribution. Age distributions considering longer evolutionary time spans still display a shift in the location of the saturation mode and a higher elevation of the saturation peak above the L-shaped background. Supplementary figure E.17 demonstrates that species-specific differences in the shape of the age distribution, due to differences in their synonymous evolution characteristics, also persist.

However, the number of founder genes has a strong effect on the smoothness of the distribution. For a low number of founder genes, G_0 = 1,000, and consequently a low number of duplicates in the age distribution (200/288 for evolutionary time spans of 5/20 on figure 3.4a), the saturation peak becomes barely discernible, especially for small evolutionary time spans, and locating the mode of the

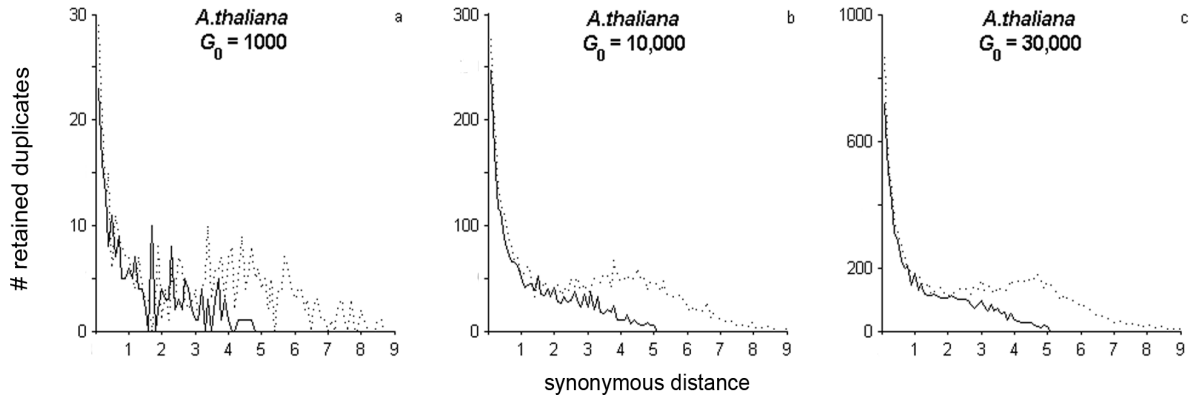


Figure 3.4: The number of genes in the age distribution impacts its shape. Results displayed for *A. thaliana* and different numbers of founder genes G_0. Age distributions simulated over evolutionary time spans of 5 and 20 are indicated by solid and dotted lines, respectively.

saturation peak becomes difficult. As the number of founder genes and consequently duplicate pairs grows (6,092/11,165 for evolutionary time spans of 5/20 on figure 3.4c), their K_S distribution will converge to the smooth distribution depicted in figure 3.3b.

K_S stochasticity and saturation also affect WGD peaks

So far, we only considered SSD age distributions, but many empirical age distributions contain superimposed peaks generated by WGD events (see figure 3.1). We investigated to which degree such WGD peaks are affected by K_S-related effects. We therefore employed our duplicate population dynamics model to simulate age distributions that contain a single WGD event on top of the SSD background. Relative to the SSD-only model, the WGD model contains an extra parameter, namely the power law decay constant α_1 for WGD duplicates, which was set to 0.90 for all scenarios. Values for the model parameters G_0, ν, and α_0 were kept at 10,000, 0.03, and 0.80, respectively. The results are qualitatively insensitive to the exact parameter values used. The results for *A. thaliana*, with simulated WGD events at synonymous ages of 1, 2.5, and 4, are presented in figure 3.5. Results for other species can be found in supplementary figure E.18.

As expected, WGD events of low synonymous age suffer minimally from K_S stochasticity and saturation effects, giving rise to a sharp K_S peak with the mode located at the expected synonymous distance. For higher WGD ages, the WGD peak becomes more dispersed, and the mode is offset to a lower synonymous distance because of saturation effects. For a WGD with a synonymous age of 2.5, close to the lower limit for the mode of the saturation peak in *A. thaliana* (see figure 3.3b), the WGD peak is still visibly discernible because of its location and amplitude. For higher WGD ages, however, it becomes increasingly more difficult to distinguish the WGD peak from the saturation peak. If only the complete distribution is considered in figure 3.5c, it appears that a single strong peak exists at a K_S of 3.1–3.2, which could easily have been generated through saturation effects alone, as can be seen by comparing figure 3.5c with the SSD-only distributions on figure 3.3b. The same trends are apparent for other species (see supplementary figure E.18).

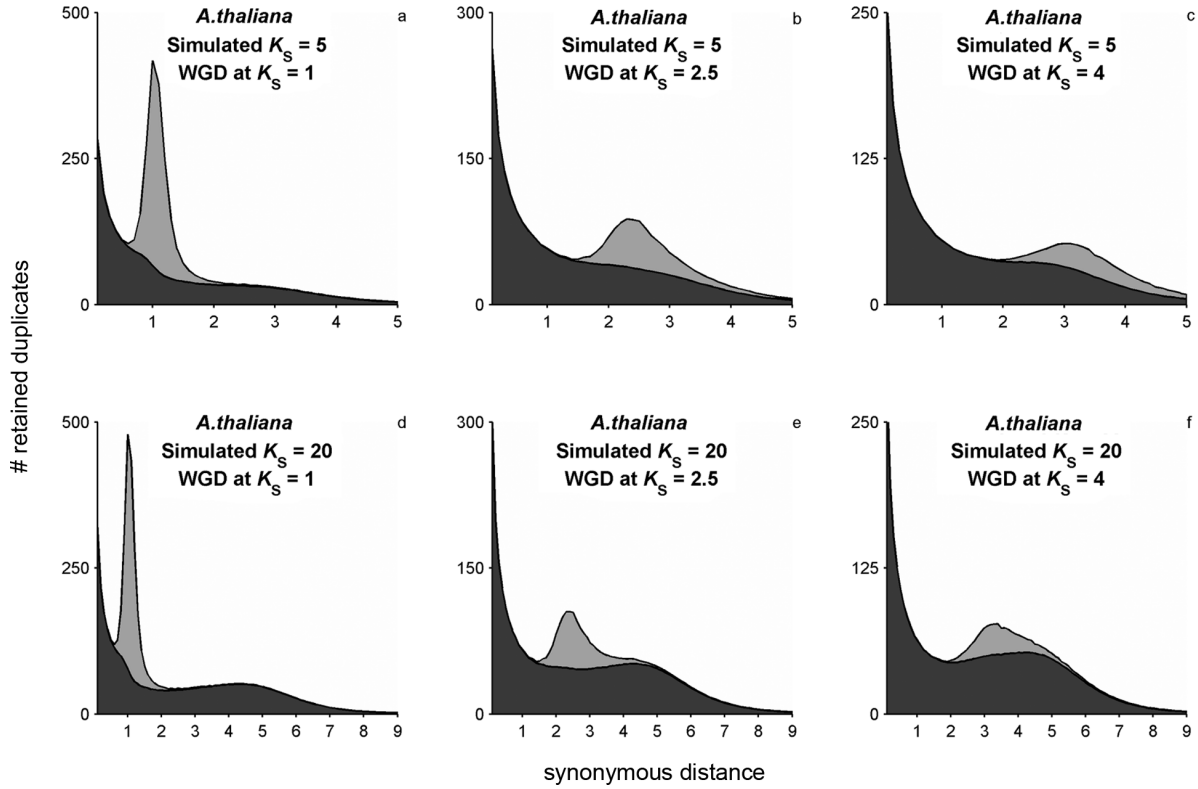


Figure 3.5: K_S stochasticity and saturation effects also affect WGD events. Results displayed for *A. thaliana*. The simulated evolutionary time spans and real WGD ages (in K_S time equivalents) are indicated on the panels. The light gray and dark gray represent the contribution of WGD and SSD duplicates, respectively.

3.4 Discussion

3.4.1 Synonymous evolution simulations characterize the effects of using K_S as a proxy for age since duplication

Saturating relationships between time since divergence and measured rate of change have been noted for a long time, for example, for mitochondrial DNA in animals[309], mammalian insulin genes[310], and enterobacterial genes[311]. These saturation patterns result from the inefficiency of the methods used to quantify (non-)synonymous changes when confronted with sequences that underwent multiple substitutions per site on average, rather than from true saturation of the synonymous substitution dynamics[303,312]. Although real sequences diverge through many other processes such as non-synonymous mutations, insertions, and deletions, evolutionary simulations focusing exclusively on synonymous evolution prove very useful to study K_S saturation dynamics. Our genome-wide simulation results are in qualitative agreement with previous smaller scale empirical examples and confirm that the observed saturation characteristics result from the fact that K_S estimation algorithms are unable to fully correct for the occurrence of multiple substitutions per site[303]. Although K_S estimates higher than 1 have generally been considered untrustworthy in literature[5], our simulations indicate that K_S estimates remain linearly related to the true synonymous distance until a synonymous age of at least 2. Complete K_S saturation for most species is only reached at a synonymous age of 20 or higher.

Although in general, the K_S of duplicate pairs becomes increasingly uncertain with age, and K_S estimates >2 can therefore not be relied upon as a proxy for the age of individual duplicates, K_S estimates still provide useful information at higher ages for large-scale duplication events such as WGDs that produce ensembles of same-aged duplicates. Such ensembles are expected to follow the distributional trends apparent in figure 3.2 and supplementary figures E.2–E.8. In support of this claim, the K_S stochasticity effects observed in our genome-wide simulations for *A. thaliana* at a synonymous age of 0.7–0.8 are in quantitative agreement with an empirical example of 242 simultaneously duplicated gene pairs remaining from the most recent WGD in the *A. thaliana* lineage[241] (see supplementary table E.2). Given a sufficient number of retained WGD duplicates, the mode of the ensemble K_S distribution is relatively stable to stochastic K_S variations for individual duplicates, and the true synonymous age of the WGD may be reconstructed by retracing the peak mode along a species-specific saturation curve as in figure 3.2.

The fact that different species exhibit different saturation curves can be explained by the differences in their substitution rate matrix Q used for the synonymous evolution simulations. Two major species-specific determinants of Q are the transition/transversion rate ratio κ, reflecting transition bias, and the equilibrium frequency π_j, reflecting codon bias (see Materials and methods). Because all other parameters in the synonymous evolution simulations were the same for all species, this confirms that species-specific transition and codon bias have a substantial impact on K_S estimation and saturation characteristics.

3.4.2 K_S stochasticity and saturation affect the shape of age distributions

Inference of WGD events from age distributions is based on the idea that peak-like deviations from an L-shaped distribution curve represent the signal of large-scale duplication events in the evolutionary history of the species of interest[176]. To avoid issues associated with K_S estimation, age distributions are often only evaluated until a K_S of 1 or 2[135,176,177,182]. This limits their use for WGD inference, however, to more recent events. It was previously unknown whether, where, and to what degree K_S saturation effects would manifest themselves in age distributions. We subjected simulated 'real age' distributions, generated by a duplicate population dynamics model, to a redistribution procedure that incorporates the K_S stochasticity and saturation effects learned from the synonymous evolution experiments discussed earlier. We demonstrated that K_S-related effects indeed result in a saturation peak in the tail of age distributions, irrespective of the species and the exact model parameters used. Both the amplitude and the mode of the saturation peak increase when the duplicate dynamics model runs over longer evolutionary time spans, because more and older duplicates are displaced to this saturation peak. The location and amplitude of the saturation peak are also influenced by species-specific differences in the saturation characteristics caused by differences in transition and codon bias.

The applicability of our simulation results on empirical age distributions hinges on the accuracy of the evolutionary model used in the simulations. However, the synonymous evolution strategy we employed corresponds to a special case of sequence evolution ($\omega=K_N/K_S=0$, absolute purifying selection) that is implausible, especially for recently duplicated genes, which are likely to undergo a period of relaxed selection. Moreover, non-synonymous evolutionary processes could have considerable impact

on the characteristics of synonymous sequence evolution trajectories, as well as on the K_S estimation performance of tools such as CODEML. In the supplementary information (see E.3), we consider a more complex scenario in which non-synonymous mutations are allowed, corresponding to the full form of the codon model as specified by Yang and Nielsen[299], and we demonstrate that allowing for non-synonymous mutations in the evolutionary simulations does not qualitatively change the results presented here, in particular regarding the occurrence of saturation peaks in K_S-based age distributions. Although no evolutionary model can capture all intricacies of real evolutionary processes[270,313], our simplified synonymous version of the full codon model outlined by Yang and Nielsen[299] seems to provide a reasonable approximation in the present context.

3.4.3 Impact on the use of mixture modeling techniques to detect WGDs

Mixture modeling techniques have proven successful in detecting even small deviations from a background distribution[182], which has led to their widespread use as tools for WGD inference. Given the power of these techniques, they should have little trouble detecting a saturation peak, which could be interpreted erroneously as evidence for the occurrence of an older WGD event. Based on the locations of saturation peaks observed in our simulations, mixture modeling techniques for inferring WGDs from age distributions are only reliable for synonymous distances lower than 2–2.5. There have, however, been attempts recently to use mixture modeling techniques over a wider K_S range, in an effort to elucidate older WGD events[136,138,183]. For example, Jiao et al.[136] evaluated the K_S distribution of the basal angiosperm *Amborella* until a synonymous distance of 3. Using mixture modeling techniques, they found evidence for subtle dispersed peaks around a synonymous distance of 1.5–2.0 and 2.5–3.0. These peaks were suggested to correspond to angiosperm and seed plant-wide ancient WGD events, respectively, which they also detected through an extensive phylogenomic approach. In light of our results, it remains difficult to discern whether the second peak in the *Amborella* distribution truly corresponds to the seed plant-wide WGD event, or whether it could be attributed to saturation effects, or both.

The fact that age distributions become less smooth as the number of incorporated duplicates decreases may also have implications for WGD inference. The ruggedness of small-sample distributions was observed in our simulations (figure 3.4), but it is also evident in some of the empirical age distributions presented in figure 3.1. Age distributions that include fewer duplicates (e.g., *K. lactis* and *C. albicans*) generally display a more rugged surface curve than age distributions that include a higher number of duplicates (e.g., *A. thaliana* and *H. sapiens*). Our simulations indicate that when the number of duplicates upon which the age distribution is based decreases sufficiently, the surface curve becomes rugged to such an extent that secondary small peaks appear over the whole distribution range. Mixture modeling techniques are prone to fit some of the bigger peak artifacts, even when using model selection criteria to determine the optimal number of fitted mixture components, such as the Akaike Information Criterion or Bayes Information Criterion[314]. The fitting of peak artifacts could be especially problematic when analyzing age distributions built from partial EST data sets. Cui et al.[135] investigated EST-based age distributions for several basal angiosperm lineages using mixture modeling techniques and found among other things evidence for two WGDs in the *Nuphar* lineage, with modes around a K_S of 0.5 and 1.25[52,135]. Both peaks are identified in a K_S range where the occurrence of saturation peaks should not be an

issue, but our simulations on small samples suggest that the second peak may include too few gene duplicates to confidently discern whether it originated from a true WGD event or through sample size effects, an issue that will soon be solved with more *Nuphar* sequence information becoming available[315]. In summary, our results suggest that the use of mixture modeling techniques for WGD inference should be limited to synonymous distances smaller than 2–2.5 and to age distributions containing a sufficient numbers of duplicates.

3.4.4 Empirical age distributions revisited

Our simulation results indicate that saturation peaks are to be expected in the tail of K_S-based age distributions. In our analyses, the location of the saturation peak is influenced by species-specific sequence biases, by the evolutionary time span and the number of founder genes considered, and to a lesser extent by the duplicate birth and death rates, which depend on the life history traits of the species under study[191] (see figures 3.3 and 3.4, and supplementary figures E.9–E.17). However, the precise location and magnitude of saturation peaks in empirical age distributions remain to be assessed, as well as their interplay with bona fide WGD peaks, which our simulations indicate can be considerable (see figure 3.5 and supplementary figure E.18).

In the empirical *A. thaliana* age distribution, a sharp peak is present at a synonymous distance of 0.8, and a more dispersed peak is found at a synonymous distance of 2.0–3.5. The first peak is located in a K_S range where stochasticity and saturation effects are minimal. This peak can therefore unambiguously be identified as a large-scale duplication peak, in this case corresponding to the documented α WGD event in the *A. thaliana* lineage[173,178]. The mode of the second peak is located at a K_S of 2.5, outside but close to the lower end of the range in which saturation peaks were observed in our simulations (see figure 3.3b), suggesting that it is not (primarily) caused by saturation effects. Indeed, previous modeling attempts[178] indicate that this peak covers two older polyploidization events (the β tetraploidization and γ hexaploidization events) that have been documented in the *A. thaliana* lineage[173,316]. The right flank of the older peak may also contain remnants of the recently uncovered angiosperm- and seed plant-wide WGDs[136], in addition to saturated K_S estimates from SSD duplicates. Clearly, the *Arabidopsis* age distribution, with two peaks covering at least three documented WGDs and a concealed saturation peak, demonstrates that dissection of age distributions without suitable mechanistic models is not evident.

A similar situation is encountered for the chordates. The age distributions of *D. rerio* and *H. sapiens* display a single peak with modes around a K_S of 2.7 and 3.3, respectively. This is in both cases at the lower end of the saturation peak mode range observed in our simulations (see figure 3.3d and e) but with an amplitude that appears too high to be caused by saturation alone. Indeed, the peak in the human distribution likely covers two WGDs that happened in close succession around the origin of the vertebrates[129–131]. The peak in the zebrafish distribution should additionally contain the remnants of a third round of genome duplication in the teleost lineage[317,318], which is in itself remarkable because this fish-specific duplication is separated from the two vertebrate WGDs by approximately 300 million years[128]. That the zebrafish peak conceals an extra WGD is also suggested by the higher peak amplitude in the *D. rerio* distribution compared with the *H. sapiens* distribution and the pronounced kink in the *D. rerio* curve around a K_S of 1.5. In contrast, the urochordate *C. intestinalis* is a documented preduplication

species[131], but its duplicate age distribution nevertheless contains a conspicuous peak around K_S of 2.5–3.0, which can only be ascribed to saturation effects. Indeed, although the peak manifests itself in the same range as the WGD-concealing peaks in the vertebrate distributions, close to the lower saturation threshold observed in our simulations, it exhibits a distinctively smaller amplitude. The age distribution of *C. intestinalis* therefore confirms that saturation peaks can be observed in the tail of empirical age distributions.

This conclusion is reinforced by investigation of the empirical age distributions of the preduplication yeast species *K. lactis* and *C. albicans*[134,319]. Given the absence of WGDs, the empirical age distributions of these small yeast paranomes only contain a limited number of gene duplicates generated by SSD events. Although displaying a rough surface curve typical for age distributions incorporating limited numbers of duplicates, both the *K. lactis* and the *C. albicans* distributions contain a sizeable peak in their tail with modes around a K_S of 3.5 and $>$4.0, respectively. In contrast to the previous examples, the *K. lactis* peak is situated well into the plausible range of saturation peaks for this species (figure 3.3h). The *C. albicans* peak even appears to overshoot this range (figure 3.3f), although establishment of the true peak location is difficult given the low sample size of duplicates (see also supplementary figure E.17). Both peaks can be considered unambiguous examples of saturation peaks. Intriguingly, the age distribution for the post-WGD species *S. cerevisiae*[133,134] contains a similar peak with mode around a K_S of 3.5 to 4.0. The fact that the amplitude of this peak is comparable to the amplitude of the saturation peaks in *K. lactis* and *C. albicans* suggests that it is also a saturation peak and that it does not cover the documented WGD in the *S. cerevisiae* lineage. Indeed, the age distribution for the WGD duplicate pairs found by Kellis et al.[134] peaks at a much lower K_S value, around 0.5, with a considerable number of paralogous pairs exhibiting a synonymous divergence close to zero (see supplementary figure E.19). This is consistent with the much higher initial peak in the empirical *S. cerevisiae* distribution compared with the *K. lactis* and *C. albicans* distributions (see figure 3.1). The early location of the WGD peak is puzzling, however, given that the yeast WGD is thought to be approximately 100 million years old[133,320], whereas a K_S of 0.5 translates to only 31 million years when assuming a silent substitution rate of 8.1/ss/By[159,308]. The apparent decelerated evolution of a sizeable proportion of yeast WGD duplicates has been observed before[134] and has been variously ascribed to long-term gene conversion[321,322] and strong codon usage bias[323], both in connection with selective pressure on retained duplicates for increased dosage.

3.5 Conclusion

Our simulation results indicate that K_S stochasticity and saturation have a large impact on duplicate age distributions and that saturation peaks are to be expected in the distribution tails. This is confirmed by investigating the empirical age distributions of non-WGD yeast species such as *K. lactis* and *C. albicans*, and non-WGD urochordate species such as *C. intestinalis*. However, documented post-WGD species also exhibit sizeable peaks in the saturation range, and in many cases, these peaks conceal one or multiple WGD events in addition to saturated K_S estimates. Elucidating the contribution of SSDs and WGDs to peaks in the saturation range of empirical age distributions will therefore require more elaborate methods than are currently in place. Mixture modeling approaches give good results for recent

genome duplications (K_S <2), but our results indicate that they are less suitable for discriminating older WGD events and for analyzing age distributions incorporating small numbers of duplicates. Without advanced modeling approaches, it remains difficult to learn more about the events that shaped empirical age distributions. Our results suggest that quantitative modeling approaches, incorporating the relative contribution of SSD and WGD duplication modes as well as K_S saturation and stochastic effects, will allow more reliable inference of the ancient WGDs that characterize many different lineages. In this respect, we are currently extending the duplicate population dynamics model introduced by Maere et al.[178] to incorporate species-specific synonymous evolution characteristics.

3.6 Acknowledgments

The authors thank two anonymous reviewers for constructive comments on the manuscript. This work was supported by Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”) and the Interuniversity Attraction Poles Programme (IUAP P6/25), initiated by the Belgian State, Science Policy Office (BioMaGNet). K. Vanneste and S. Maere are fellows of the Fund for Scientific Research Flanders (FWO).

3.7 Author contributions

I performed all the analyses described in this chapter, and wrote the resulting research article. Both were done under supervision and with significant contributions of Yves Van de Peer and Steven Maere.

Chapter 4

Dating of genome duplications

Kevin Vanneste, Guy Baele, Steven Maere, Yves Van de Peer. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Research*. *Accepted, pending revisions*.

Abstract

Ancient whole genome duplications (WGDs), also referred to as paleopolyploidizations, have been reported in most evolutionary lineages. Their attributed role remains a major topic of discussion, ranging from an evolutionary dead end to a road towards evolutionary success, with evidence supporting both fates. Previously, based on dating WGDs in a limited number of plant species, we found a clustering of angiosperm paleopolyploidizations around the Cretaceous-Paleogene (K-Pg) extinction event about 66 million years ago. Here, we revisit this finding, which has proven controversial, by combining genome sequence information for many more plant lineages and using more sophisticated analyses. We included 38 full genome sequences and three transcriptome assemblies in a Bayesian evolutionary analysis framework that incorporates uncorrelated relaxed clock methods and fossil uncertainty. In accordance with earlier findings, we demonstrate a strongly non-random pattern of genome duplications over time with many WGDs clustering around the K-Pg boundary. We interpret these results in the context of recent studies on invasive polyploid plant species, and suggest that polyploid establishment is promoted during times of environmental stress. We argue that considering the evolutionary potential of polyploids in light of the environmental and ecological conditions present around the time of polyploidization could mitigate the stark contrast in the proposed evolutionary fates of polyploids.

For the author contributions, see page 96.

4.1 Introduction

The omnipresence of whole genome duplications (WGDs) in evolution is striking. Both the angiosperm and vertebrate ancestors underwent at least two separate WGDs so that all their descendants are in fact ancient polyploids (paleopolyploids)[131,136]. In the vertebrate lineage, a third WGD occurred in the ancestor of the successful teleost fish[130]. In the angiosperm lineage, subsequent and sometimes repeated WGDs have been reported in all major clades[52,128]. WGDs have also been documented in other kingdoms, such as for instance three WGDs in the ciliate *Paramecium tetraurelia*[132], and one WGD in the ancestor of the hemiascomycete *Saccharomyces cerevisiae*[133]. A systematic overview of WGD in invertebrates, amphibians, and reptiles is lacking, but several examples have been described, contradicting the classical notion that paleopolyploidies are absent in these lineages[146,147].

Although the prevalence of WGDs has been firmly established[150], their attributed importance remains very controversial. Two long-standing opposite views regard polyploidy either as an evolutionary dead end[152,153], or as a road towards evolutionary success[154]. Much research has been dedicated to this topic, especially in the plant lineage because of the high frequency of WGD occurrence in plants, and studies have typically found ample support for both scenarios. Recently formed polyploids frequently display increased meiotic and mitotic abnormalities through improper pairing of both subgenomes during cell division, resulting in genomic instability that has detrimental effects on plant fertility and fitness[156]. The study of mutant *Arabidopsis thaliana tam-1* plants that cannot enter meiosis II and therefore increase in ploidy in subsequent generations, suggests that this genomic instability is polyploidy-associated, as *tam-1* plants with higher ploidy levels exhibit more detrimental effects coupled with a strong drive to revert to lower ploidy levels via genomic reductions[157]. Recently formed polyploid plants also need to cope with the minority cytotype disadvantage, a frequency-dependent reproductive disadvantage caused by ineffective matings of unreduced $2n$ gametes that cross with reduced n gametes from the diploid progenitor majority cytotype, which results in the formation of less fit and fertile triploid hybrids[122]. Consequently, even recently formed polyploids that are stable may be incapable of propagation because they simply cannot overcome the bottleneck of finding enough suitable mating partners to establish a viable population. Genomic and phenotypic instability, and the minority cytotype disadvantage, most likely contribute to the observation that polyploid plant species display lower speciation rates and higher extinction rates compared to diploids, and consequently an overall lower net diversification rate[158].

In contrast, the fact that all extant angiosperms[136] and vertebrates[131] are paleopolyploids indicates that polyploidization is not always a dead end. Moreover, an estimated 15% and 31% of speciations in flowering plants and ferns, respectively, were accompanied by a ploidy increase[142]. Most recent insights explaining the evolutionary success of polyploids have focused on their duplicated genome, which simultaneously provides thousands of novel genes for evolution to tinker with. Even though the large majority of these duplicated genes are lost through pseudogenization[159], the remaining fraction can lead to novel and/or expanded functionality through Ohno's classical models of neofunctionalization (the duplicated copy acquires a new function), subfunctionalization (the division and/or elaboration of pre-duplication functionality over the two daughter copies), and gene conservation due to dosage effects (the increased production of a beneficial gene product), and combinations thereof[93,98,212]. Interestingly, a fraction of WGD duplicates, including many regulatory and developmental genes, is most likely guarded

against loss through dosage-balance constraints on the stoichiometry of duplicated pathways and/or macromolecular complexes[160,161,178]. Resolution of dosage-balance constraints over time can thus provide polyploid species with an important toolbox that can be rewired to execute novel functionality[162], and may allow them to cope with new ecological opportunities and/or challenges[163,324]. The ecological conditions that allow the initial establishment and long-term success of polyploids have been a major question in early polyploidy research for a long time, but progress in this regard has shifted somewhat to the background due to the explosion in research on their genomic composition[125]. Recently formed polyploids are traditionally considered to be good colonizers that have a broad ecological tolerance, which gives them an adaptive advantage as invasive species[149,164]. The latter can be attributed to their phenotypic instability, which can also be viewed as increased phenotypic variability and plasticity[144]. Such generalizations should however be treated with caution because of the paucity of large-scale systematic data on the subject and the many exceptions that can be found[125].

In view of the contrasting WGD fates outlined above, it is perhaps not surprising that the precise nature of the link between WGD and evolutionary success remains heavily debated[52,150,287]. Previously, we performed absolute dating analyses on nine plant WGDs and proposed a link with the Cretaceous-Paleogene (K-Pg) extinction boundary[193], which took place 66 million years ago (mya) according to the most recent estimates[80], suggesting that polyploidization somehow contributed to enhanced plant survival at that time[151]. This study was however limited in terms of taxonomic sampling, due to the small number of plant genome sequences available at that time, and it relied on penalized likelihood inference methods that present inherent methodological challenges[215], such as for instance the assumption of an autocorrelated relaxed clock model that is most likely violated when taxon sampling is limited[195]. In the years since, the number of publicly available plant genomes has increased drastically, and the field of molecular dating has also progressed with the development of more powerful Bayesian methods of sequence divergence estimation that can incorporate advanced uncorrelated relaxed clock models and fossil age uncertainty[219].

Here, we revisit the previously proposed clustering of plant paleopolyploidizations around the K-Pg boundary using the latest genome sequence datasets and phylogenetic dating methods available. We analyzed data from in total 41 plant species, including 38 full genome sequences and three transcriptome assemblies, to date 31 WGDs in various species that correspond to 20 independent plant WGDs. We employed the BEAST software package, a state-of-the-art but computationally intensive Bayesian dating framework[220]. We tested whether these 20 plant WGDs follow a model where polyploid abundance simply increases randomly over time[325], or alternatively cluster statistically significantly in time in association with the K-Pg boundary[193], by comparing our WGD age estimates with a null model that assumes random WGD occurrence. We find a strongly non-random pattern with many WGDs clustering around the K-Pg boundary and we interpret our results in the light of new findings on recently formed plant polyploids that can help to explain this pattern. In particular, we argue that the environmental and ecological conditions during the time of polyploidization are of crucial importance.

4.2 Material and methods

4.2.1 Data collection

In total, sequence information from 41 species was collected, including 38 full genome sequences and three transcriptome assemblies. A concise overview of employed species and their data sources is provided in supplementary table F.1. For annotated full genome sequences, protein-coding genes were used as provided by their respective annotations (all genes flagged as either suspected or known pseudogenes were removed). If alternative transcripts were available, only the one with the longest CDS was kept. For transcriptome assemblies, unigene sets were employed as provided by their respective database. We used FrameDP (v1.0.3)[326] to extract the correct coding frame and putative coding sequence from the unigene sets, employing Swiss-Prot[327] as a reference database for the underlying HMM model and discarding genes shorter than 300 nucleotides.

4.2.2 Selection of homeologs

K_S age distributions for all species were constructed as described in Vanneste et al.[328]. For all species for which positional information was available, anchor pairs (i.e., duplicated gene pairs created by large-scale duplications that are positioned on duplicated segments) were extracted as follows. An all-against-all protein sequence similarity search was performed using BLASTP with an E-value cutoff of e-10. Paralogous gene pairs were retained if the two sequences were alignable over a length of more than 150 amino acids with an identity score of at least 30%[329]. Duplicated segments stemming from the most recent WGD were obtained by running i-ADHoRe (v3.0)[168,169]. i-ADHoRe parameters were set as follows: table_type=family, alignment_method=gg2, cluster_type=collinear, gap_size=35, cluster_gap=40, q_value=0.75, prob_cutoff=0.01, anchor_points=3, multiple_hypothesis_correction=FDR, max_gaps_in_alignment=40, and level_2_only=true. Peaks in the K_S age distribution supported by anchors were considered as valid WGD signatures. To ensure all reported anchors were created by the WGD in question, only anchors on duplicated segments with median K_S values (calculated based on all anchors) between the WGD peak boundaries were accepted as homeologs. Paranome K_S distributions with anchors mapped on them are presented in figure 4.1 for a few exemplary species, and in supplementary figure F.1 for all other species. WGD peak K_S boundaries are presented in table 4.1 for all species. For the Brassicaceae, we also tried to collect anchors for the older *beta* duplication[173] by rerunning i-ADHoRe with level_2_only=false, but this approach only resulted in enough quality orthogroups (see next section) for *A. thaliana* because of its high-quality genome information. *M. acuminata* is a special case because its peak in the K_S age distribution most likely represents two WGDs in very short succession[330] so that anchors reported by i-ADHoRe most likely stem from two WGDs. We therefore treated the *M. acuminata* WGD peak as a single event[330].

For species where no or few anchors could be collected through lack of positional information due to a fragmented assembly or in case of transcriptome data, we employed an alternative strategy to collect homeologs by selecting duplicate pairs from the WGD peak in the K_S age distribution. Although some of these duplicate pairs may not have been created by WGD, but rather by small-scale duplications in the same time frame, it can be safely assumed that the majority derives from the WGD[178]. Because multiple

paralogous pairs can descend from the same gene duplication due to subsequent duplications[193], we built amino acid-based phylogenies for all paralogous gene families in each species using PhyML (v3.0)[331] with default parameters, which were rooted using a mid-point rooting approach[332]. For duplication nodes with median K_S values (calculated based on all their terminals) between the WGD peak boundaries (see table 4.1), a random pair of descendent genes was taken as the representative homeologous pair. This strategy was applied for all species where fewer than 1,000 orthogroups (see next section) could be collected based on anchors, to increase the total number of homeologs used for obtaining a WGD age estimate.

4.2.3 Orthogroup construction

For each collected homeologous pair, an orthogroup was constructed consisting out of the homeologous pair and their orthologs in other plant species, since orthology relationships provide the most accurate representation of the followed evolutionary history[193,333,334]. We used Inparanoid (v4.1)[335] with default parameter settings to detect orthologs. Simply adding all identified orthologs from the other plant species to the homeologous pair was however not feasible because this would result in a plethora of possible tree topologies, for which applying the proper fossil calibrations and model specifications based on the BEAST XML syntax (see below) would be problematic. Additionally, this could also lead to systematic biases between different homeologous pairs from the same species caused by a different 'tree context'. Keeping the orthogroup topology fixed by requiring one ortholog to be present for every species listed in supplementary table F.1 proved however also problematic because this resulted in a drastic drop of the total number of recovered orthogroups, since most homeologs had to be discarded because orthologs could not be found in every other plant species. This is probably due to both species-specific ortholog loss and problems with orthology detection performance, since the latter decreases together with genome annotation quality, especially over large evolutionary distances[336], and many plant genomes have only been sequenced at relatively low coverage[337].

We therefore employed a strategy where different species were put together in species groups, each consisting of two to four members. For each species group, the best ortholog (based on the average score reported by Inparanoid to both paralogs of the homeologous pair) was selected as the representative ortholog for that species group, and added to the orthogroup. As a consequence, the orthogroup topology could be held constant, whereas for most homeologs at least one ortholog could be collected per species group so that the total number of recovered orthogroups for dating remained high and few homeologs had to be discarded. An extended description and justification for our used species grouping topology is provided in the supplementary information (see F.3.1). Table 4.1 summarizes the total number of collected orthogroups, separated into anchors and peak-based duplicates per species, where applicable. Lastly, the homeologous pair was always fixed to cluster together in all orthogroups by not allowing any speciation after duplication scenarios. The latter would entail identifying the correct orthology relationships in sets of outparalogs, which is notoriously difficult[338,339].

4.2.4 Orthogroup dating

All sequences in each orthogroup were aligned using MUSCLE (v3.8.31)[295]. Orthogroup alignments were cleaned up as described previously[288], and only orthogroups with a cleaned alignment of more than 100 amino acids were retained for further analysis. We used BEAST (v1.7.4)[220] to date the node joining the homeologous pair that represents the WGD of interest in each orthogroup. We set the underlying evolutionary model to be Le-Gascuel (LG), which is the most recent and large-scale amino-acid replacement matrix available[340], with gamma-distributed rate heterogeneity across sites using four rate categories[341]. To this end, we have implemented the LG model into the BEAST source code, as this model was not yet publicly available. We employed an uncorrelated relaxed clock model that assumes an underlying lognormal distribution (UCLD) on the evolutionary rates[219], which is more likely to yield accurate estimates than the uncorrelated relaxed clock model that assumes an exponential distribution (UCED) on the evolutionary rates[342]. A Yule pure birth process[343] was specified for the underlying tree model because contemporaneous sequences are considered in all orthogroups. We employed the following priors: a uniform prior between 0 and 100 for the Yule birth rate; an exponential prior with mean 0.5 on the rate heterogeneity parameter; an exponential prior with mean 1/3 on the standard deviation of the UCLD clock model; and a diffuse gamma prior with shape 0.001 and scale 1,000 on the mean of the UCLD clock model. Priors on the fossil calibrations are detailed extensively in the supplementary information (see F.3.2). A starting tree with branch lengths satisfying all the fossil prior constraints was manually constructed and is also presented in the supplementary information (see F.3.2). Operators on the tree model were disabled to keep the topology fixed so that only the branch lengths were optimized.

The MCMC analysis for each orthogroup was run for 10 million generations, whilst sampling every 1,000 generations, resulting in a total size of 10,000 samples per orthogroup. The quality of the approximation of the posterior distribution improves as the number of generations, i.e., the amount of computational time devoted to the MCMC, increases[344,345]. These methods are therefore computationally very intensive[346,347], especially since we had to process a total of 22,252 individual evolutionary histories across all collected orthogroups. There exist faster implementations incorporating relaxed clock methods in a Bayesian context, but we still preferred the use of BEAST because it scores very high on benchmarks[348], and also has a very rich XML language syntax. We employed a strategy where the separate orthogroups were run distributed over multiple CPU cores for independent evaluation[349]. We also made use of the BEAGLE library, which speeds up the MCMC by taking over part of the core likelihood calculations[347]. Since visual inspection of each individual trace file for each orthogroup was impossible, we employed LogAnalyser (part of the BEAST package) for automated evaluation of the orthogroups. A burn-in of 1,000 samples was used and orthogroups were only accepted if the minimum effective sample size (ESS) for all statistics was at least 200. Table 4.1 summarizes the total number of accepted orthogroups, separated into anchors and peak-based duplicates per species, where applicable.

4.2.5 Obtaining species-specific WGD age estimates

The age estimates for the node joining the homeologous pair in all accepted orthogroups were collected, and grouped into one or two absolute age distributions per species containing either age estimates based on anchors and/or peak-based duplicates, where applicable (see table 4.1). A consensus WGD age

estimate was obtained for each absolute age distribution by taking the mode of its kernel density estimate (KDE). The latter is much more flexible in comparison with traditional parametric distributions because it does not limit the shape of the estimated distribution to parameter-described forms, and therefore allows a much better exploration of the true underlying distribution and its trends[350]. We employed Matlab (vR2011a, The MathWorks Inc., Natick Massachusetts, United States) and the KDE toolbox (available at `http://www.mathworks.com/matlabcentral/fileexchange/17204-kernel-density-estimation` - retrieved 21th March 2013), which allows automatic bandwidth selection[350]. We used bootstrapping to obtain 90% confidence intervals (CIs) for all WGD age estimates[351]. For a dataset of age estimates $\{x_i; i=1...n\}$, n values are resampled with replacement to collect the bootstrap dataset $\{x_i^*; i=1...n\}$ and KDE is performed on x_i^* to obtain the bootstrap density estimate $\hat{p}^*$. This is repeated 1,000 times to collect a set of bootstrap density estimates $\{\hat{p}_j^*; j=1...1000\}$. The distribution of $\hat{p}_j^*$ around the original density estimate $\hat{p}$ mimics the distribution of $\hat{p}$ around the true density p, so that the modes for the 51th and 949th bootstrap density estimate (ranked in order of increasing value for their mode) give the lower and higher 90% CI boundary, respectively. Absolute age distributions are presented in figure 4.2 for a few exemplary species, and in supplementary figure F.2 for all other species. Exact values for species-specific WGD age estimates and their corresponding 90% CIs, separated into anchors and peak-based duplicates where applicable, are listed in table 4.1.

4.2.6 Clustering of WGDs in time

Assessing whether there exists a statistically significant grouping of WGDs in time was based on the median distance between WGD age estimates as described in Fawcett et al.[193]. Briefly summarized, smaller median distances indicate a tighter clustering. The observed median distance between WGDs was compared with a null model that is based on random WGD occurrence by assuming a background distribution where the probability of WGD occurrence at a certain point in time is proportional to the total number of species present at that time (see supplementary figure F.3). One million random samples were pulled from this null model to assess the probability that the observed median distance is significantly lower than the distribution of median distances based on random WGD occurrence. We considered a timespan between 0 and 100 mya, as both the identification and timing of older paleopolyploidizations is still uncertain. All WGD age estimates listed in table 4.1 were taken into account. Shared WGDs were only counted once by taking the average of WGD age estimates in all their descendant species (see figure 4.3), always using anchor-based WGD age estimates and only peak-based WGD age estimates if the former were not available. The observed median distance was significantly lower than expected under the null model (p=0.03, see supplementary figure F.3), indicating clustering of plant paleopolyploidizations in time. Moreover, this test is conservative because WGD age estimates in some woody species are most likely too young (see Results and discussion).

This evaluation of clustering does however not identify the exact location of the clustering. Because any *a priori* criterion to associate WGDs with the K-Pg boundary would be based on arbitrary cut-offs and is hence undesirable, we fitted a mixture of Gaussians (i.e., normal distributions) to the WGD age estimates (shared WGDs were only counted once as before) using the gmdistribution.fit function in Matlab. According to the Akaike Information Criterion (AIC)[352], a mixture with two components had the

best fit to the raw data (AIC=174.90 compared to AIC=180.33 and 177.96 for a mixture with one and three components, respectively). This mixture contained one very pronounced component at a location of 60.05 mya, corresponding to a clustering of WGDs close to the K-Pg boundary, while the second lesser component was located at 22.91 mya and most likely represents the background distribution (see supplementary figure F.4). Exclusion of the *M. acuminata* WGD in these analyses, because the latter most likely represents two WGDs in very close succession[330], did not significantly change these results (see supplementary figures F.3 and F.4).

4.3 Results and discussion

4.3.1 Massive absolute dating of homeologs created through WGDs reveals the timing of plant paleopolyploidizations

We focused on dating the most recent WGD in each plant species, because these can be most easily identified based on collinearity information (see Material and methods). One exception is *A. thaliana*, for which we were able to find a crude WGD age estimate for the older *beta* duplication, in addition to the more recent *alpha* duplication[173], because of the high-quality genome sequence information available for this model species. Another special case is *Musa acuminata*, which most likely experienced two separate WGDs in very close succession that are problematic to differentiate between and that were therefore treated as a single event[330]. We employed two approaches to collect homeologs (genes created by WGD) for absolute dating. First, we used positional information to select anchor pairs, i.e., homeologs located on duplicated segments generated through WGD, with ages corresponding to the WGD signature peak in the K_S age distribution[328]. Second, for species without positional information, or if fewer than 1,000 orthogroups (see below) could be constructed based on anchors, we supplemented the anchor pairs with 'peak-based' duplicates, which are non-anchor pairs that also map to the WGD signature peak in the K_S age distribution and therefore are assumed to consist mainly of homeologs[178]. The selection of homeologs for different plant species that experienced a WGD in the last ~100 million years is illustrated in figure 4.1 for a few exemplary species, and in supplementary figure F.1 for all other species. Next, all collected homeologs were combined with orthologs from other plant genomes to construct orthogroups (see Material and methods). The node joining the homeologous pair in each orthogroup phylogeny, representing the WGD of interest, was then dated using the uncorrelated lognormal (UCLD) relaxed clock model implemented in the BEAST package[219,220] based on several primary fossil calibrations (see below). The resulting absolute age estimates for all homeologs collected from the same species were afterwards grouped into one absolute age distribution, separated into anchors and peak-based duplicates where applicable. A consensus WGD age estimate was obtained for every species by taking the location of its peak in the absolute age distribution, as identified through kernel density estimation (KDE), while 90% confidence intervals (CIs) were obtained through a bootstrapping procedure (see Material and methods). Absolute age distributions for the species illustrated in figure 4.1 are presented in figure 4.2, and in supplementary figure F.2 for all other species. All WGD age estimates, their 90% CIs, and the number of dated orthogroups they were based on, are listed in table 4.1 per species, for both anchors

and peak-based duplicates. A general overview of all dated WGDs mapped on the green plant phylogeny is also presented in figure 4.3.

Figure 4.2 and supplementary figure F.2 demonstrate that WGD age estimates obtained from absolute age distributions based on anchors and peak-based duplicates are in good agreement within the same species. However, the left flanks of peak-based absolute age distributions are denser compared to their right flanks, i.e., their distribution has a higher total probability of containing younger age estimates. This is most likely because a fraction of peak-based duplicates, namely those that do not derive from the WGD but from small-scale duplications in the timeframe covered by the WGD signature peak, follow an asymmetrical power-law distribution[178]. As a result, the non-WGD pairs under the signature peak are slightly biased towards lower K_S values and younger ages. In contrast, anchor-based absolute age distributions exhibit a much more symmetrical shape. Nevertheless, KDE appears particularly well suited to correct for the different underlying shapes of anchor and peak-based absolute age distributions, and can accurately detect their peaks, which typically agree very well for both types of distributions within the same species. Their different shapes however prevent grouping both kinds of information into one absolute age distribution, despite the fact that anchors and peak-based duplicates theoretically describe the same species-specific WGD, since this would bias their resulting 90% CIs. Because anchor-based absolute age distributions are more symmetrical around their peak used for the WGD age estimate, and because they are based on actual duplicated segments, we consider them of higher quality, although peak-based duplicate WGD age estimates are clearly a good alternative for species where no or few anchors can be identified through lack of positional information.

In a few instances, we dated the same WGD in different descendant species. Figure 4.2 demonstrates for instance the anchor-based absolute age distributions and resulting WGD age estimates for four species that diverged after the Faboideae-specific WGD[353]: *Medicago truncatula* (66.01 mya), *Cicer arietinum* (63.66 mya), *Lotus japonicus* (63.26 mya), and *Cajanus cajan* (56.96 mya). Note that although *Glycine max* also shares this WGD, it underwent an additional more recent polyploidization, which we dated instead. The above four independent estimates converge on a WGD age of ~63 to 66 mya, and also indicate that the *C. cajan* estimate most likely constitutes an underestimate, which might be due to either gene conversion or a strong genome-wide decelerated evolutionary rate that could not be completely corrected for (see below). Since all anchors from these four species describe the same event, an alternative strategy could have been to group them into one absolute age distribution to obtain a single WGD age estimate, which could however lead to misleading results. Since there are 361 dated anchors for *C. cajan* compared to 308 for all three other species combined (see table 4.1), pooling them would introduce a systematic bias by pulling the whole absolute age distribution towards a younger WGD age estimate, and would also prevent us from inferring that the *C. cajan* WGD age most likely represents an underestimate. The same applies to peak-based duplicates that describe a shared WGD in other species. We expect that as new plant genomes become available, continued efforts in dating shared WGDs will help to pinpoint their exact age more precisely.

It should be noted that because allopolyploids result from the merger of two different species, in contrast to autopolyploids, their WGD age estimate could be slightly overestimated, since the latter reflects the time at which both contributing parental genomes started to diverge rather than the polyploidization itself[354]. Distinguishing between auto- and allo-paleopolyploidizations is however notoriously difficult.

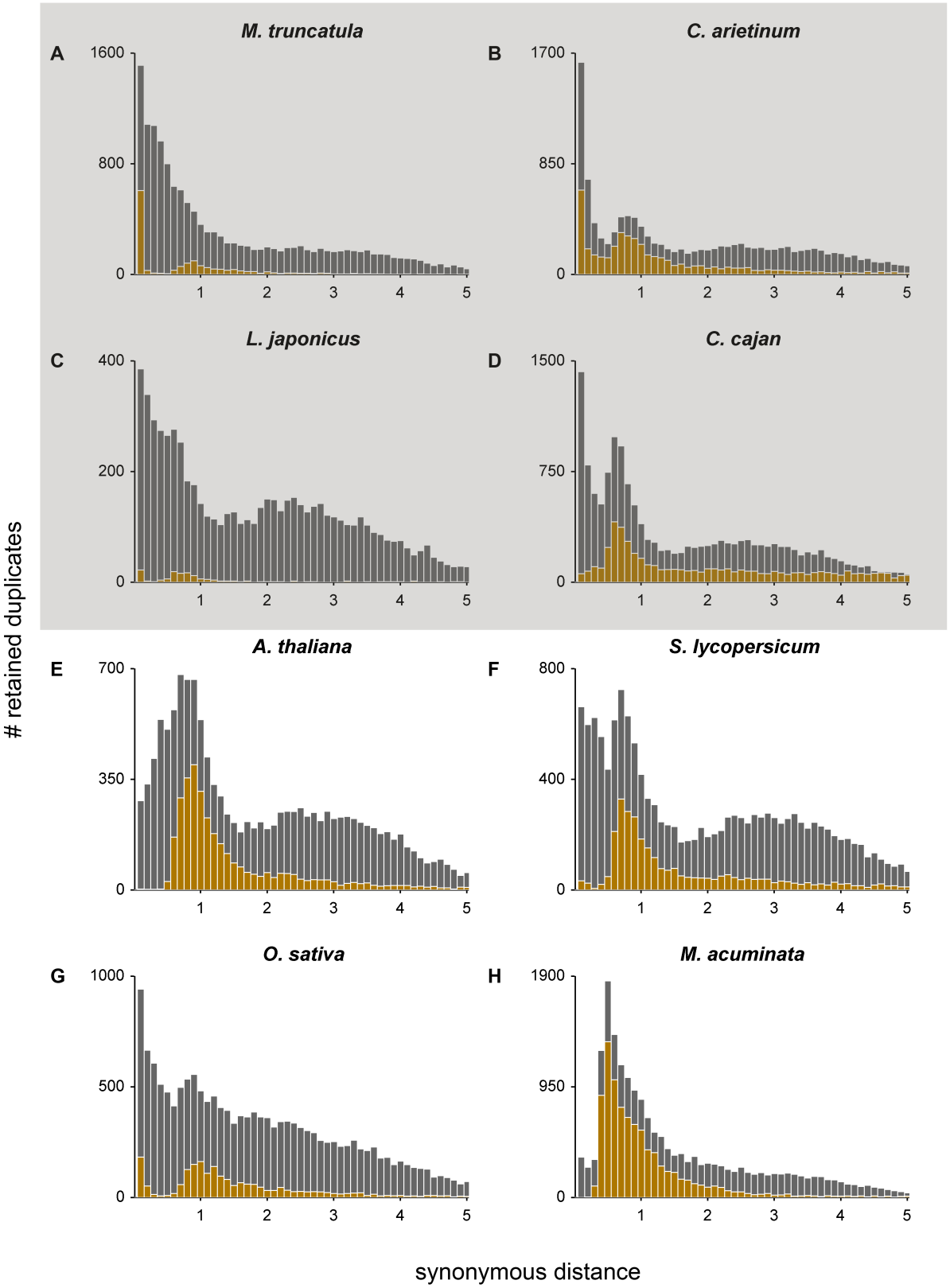


Figure 4.1: K_S age distributions for several species of interst. K_S age distributions for (A) *M. truncatula*, (B) *C. arietinum*, (C) *L. japonicus*, (D) *C. cajan*, (E) *A. thaliana*, (F) *S. lycopersicum*, (G) *O. sativa*, and (H) *M. acuminata*. The grey and beige bars represent the distribution of the paranome and duplicated anchors identified with i-ADHoRe, respectively. Anchors and peak-based duplicates used as homeologs for absolute dating were extracted between the WGD peak boundaries (see table 4.1). The grey box surrounding (A-D) indicates that these four species represent the same Faboideae-specific WGD.

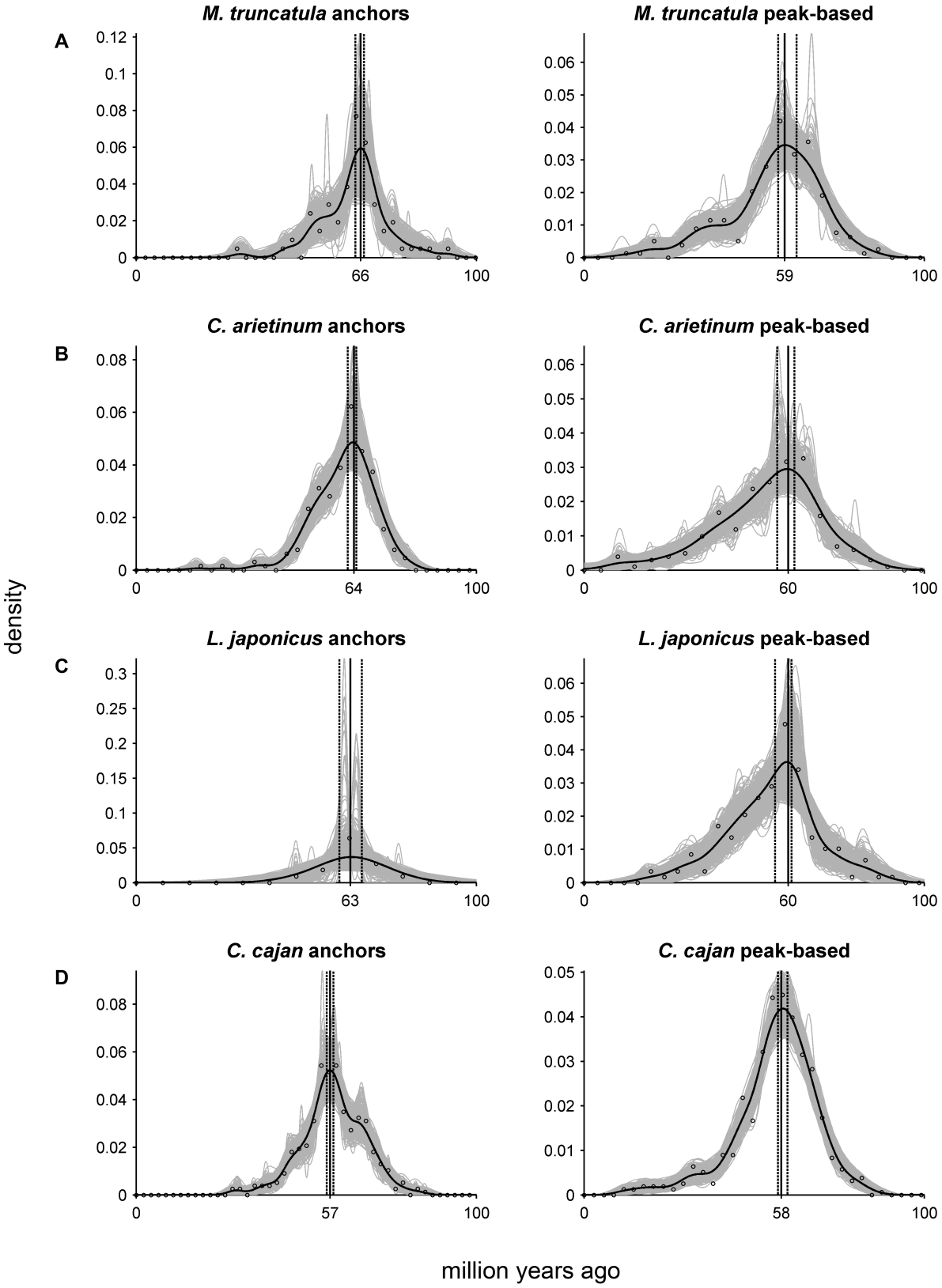


Figure 4.2: Absolute age distributions for several species of interest. Absolute age distributions of the dated anchors (left panel) and peak-based duplicates (right panel) for (A) *M. truncatula*, (B) *C. arietinum*, (C) *L. japonicus*, and (D) *C. cajan*. The non-vertical black solid line represents the kernel density estimate of the dated homeologs, while the vertical black solid line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical black dashed lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries. The distributions for (A-D) represent the same Faboideae-specific WGD.

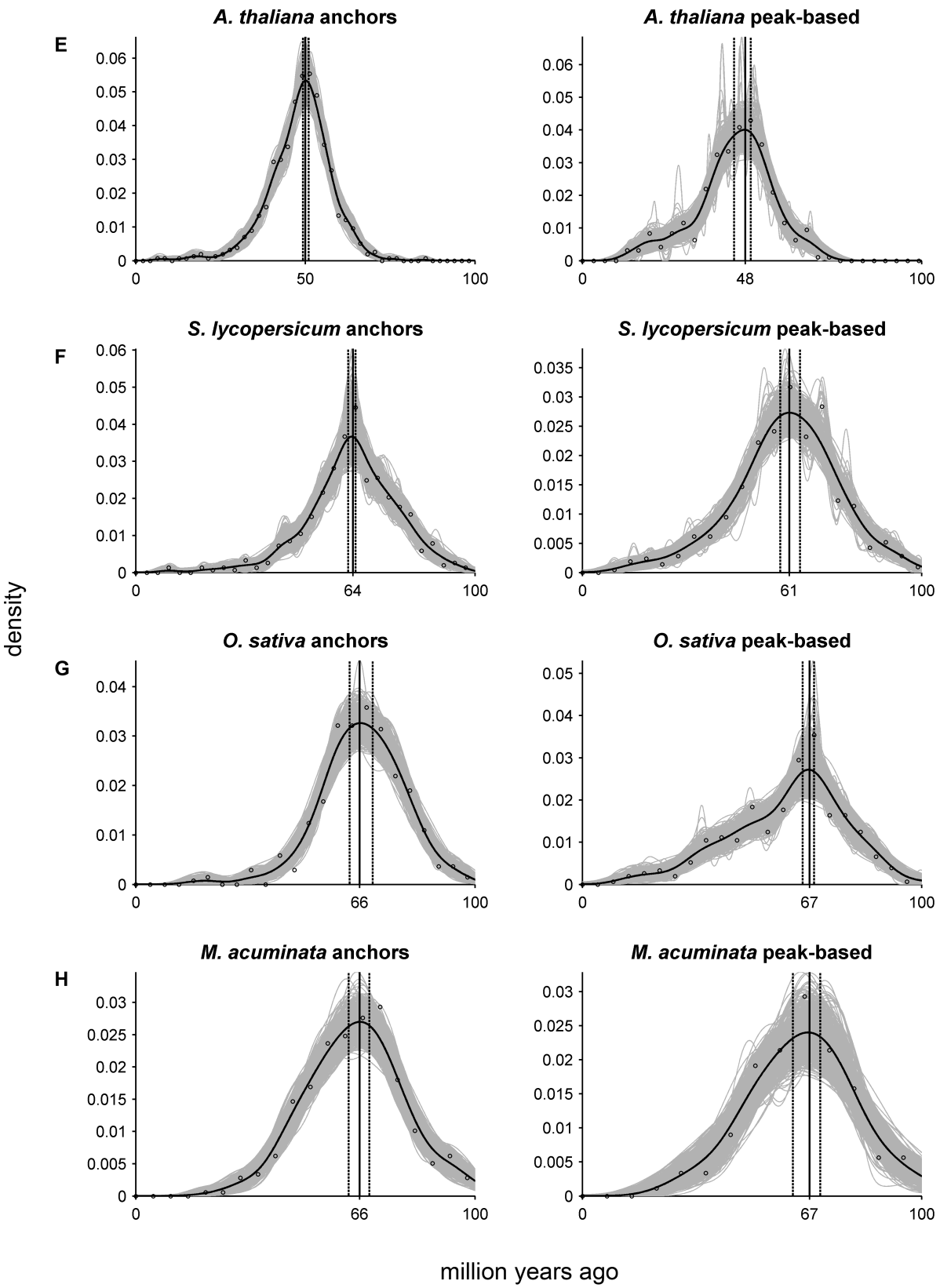


Figure 4.2: Absolute age distributions for several species of interest - Continued. Absolute age distributions of the dated anchors (left panel) and peak-based duplicates (right panel) for (E) *A. thaliana alpha* duplication, (F) *S. lycopersicum*, (G) *O. sativa*, and (H) *M. acuminata*. The non-vertical black solid line represents the kernel density estimate of the dated homeologs, while the vertical black solid line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical black dashed lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries.

Table 4.1: Overview of WGD age estimates presented in this study. Overview of WGD peak K_S boundaries used for selecting homeologs in each species, number of dated and accepted orthogroups based on anchor pairs (APs) and peak-based duplicates (PBs), and their resulting WGD age estimates with respective 90% confidence intervals (CIs).

Species	K_S range	# Dated (accepted) APs	APs WGD Age (90% CI)	# Dated (accepted) PBs	PBs WGD Age (90% CI)
Pyrus bretschneideri	0.1-0.3	1,000 (982)	19.85 (18.83-20.77)	0 (0)	n/a
Glycine max	0.05-0.15	1,000 (989)	13.59 (11.87-13.99)	0 (0)	n/a
Cajanus cajan	0.4-1.0	361 (355)	56.96 (56.04-58.02)	542 (534)	58.42 (57.03-59.85)
Medicago truncatula	0.6-1.2	79 (77)	66.01 (64.43-67.00)	201 (191)	59.08 (57.11-62.49)
Cicer arietinum	0.5-1.1	210 (203)	63.66 (62.23-64.76)	208 (204)	59.71 (56.81-61.83)
Lotus japonicus	0.4-1.0	19 (14)	63.26 (59.74-66.37)	155 (149)	59.60 (56.19-61.03)
Manihot esculenta	0.2-0.6	1,000 (977)	40.44 (38.72-42.12)	0 (0)	n/a
Linum usitatissimum	0.1-0.3	1,000 (988)	10.66 (9.93-11.87)	0 (0)	n/a
Populus trichocarpa	0.15-0.4	1,000 (986)	34.73 (32.60-36.34)	0 (0)	n/a
Brassica rapa	0.3-0.5	1,000 (978)	26.78 (24.76-28.57)	0 (0)	n/a
Thellungiella parvula	0.5-1.1	779 (758)	48.72 (47.55-52.27)	264 (258)	50.37 (47.73-51.58)
Arabidopsis thaliana α*	0.5-1.1	754 (736)	50.07 (49.27-50.99)	293 (289)	47.80 (44.76-49.67)
Arabidopsis thaliana β*	1.5-3.0	9 (9)	61.21 (54.58-69.38)	198 (110)	62.97 (56.04-70.01)
Arabidopsis lyrata	0.5-1.1	706 (687)	48.75 (47.55-49.85)	290 (282)	49.96 (44.43-52.05)
Gossypium raimondii	0.3-0.75	1,000 (978)	58.02 (56.48-59.12)	0 (0)	n/a
Solanum lycopersicum	0.4-1.0	479 (471)	63.66 (62.64-64.84)	463 (449)	61.03 (58.35-64.18)
Solanum tuberosum	0.4-1.0	478 (466)	59.56 (57.47-63.19)	487 (480)	63.77 (61.87-64.84)
Lactuca sativa	0.6-1.2	0 (0)	n/a	451 (445)	58.32 (55.64-60.04)
Aquilegia formosa x *pubescens*	0.4-1.2	0 (0)	n/a	55 (50)	51.10 (44.84-60.40)
Brachypodium distachyon	0.6-1.2	319 (302)	69.56 (67.58-71.21)	300 (276)	71.58 (69.19-74.51)
Hordeum vulgare	0.6-1.0	0 (0)	n/a	323 (306)	72.45 (69.46-74.47)
Phyllostachys heterocycla	0.1-0.3	503 (487)	19.71 (18.75-20.95)	497 (472)	18.46 (17.14-20.92)
Oryza sativa	0.6-1.0	334 (322)	66.23 (63.08-69.89)	350 (335)	66.67 (64.98-68.32)
Zea mays	0.1-0.3	948 (918)	20.40 (19.71-20.99)	52 (48)	15.68 (13.92-18.75)
Sorghum bicolor	0.6-1.3	170 (162)	69.67 (65.93-73.11)	379 (362)	69.05 (66.26-70.77)
Setaria italica	0.6-1.2	309 (298)	67.66 (65.38-70.48)	425 (401)	67.66 (63.52-70.88)
*Musa acuminata***	0.3-0.7	367 (345)	66.08 (62.78-68.86)	126 (122)	66.52 (62.05-70.11)
Phoenix dactylifera	0.2-0.4	32 (28)	53.70 (48.53-57.77)	809 (749)	49.85 (47.99-51.68)
Nuphar advena	0.2-0.6	0 (0)	n/a	119 (116)	72.78 (67.88-76.78)
Physcomitrella patens	0.5-0.8	319 (263)	60.55 (54.95-73.44)	681 (577)	68.97 (58.13-76.92)

*α and β refer to the *A. thaliana alpha* and *beta* duplication, respectively[173].
**This event most likely represents 2 separate WGDs in close succession[330].

Another caveat in estimating WGD ages is the influence of gene conversion, which may preserve WGD duplicates in an undiverged sequence state over extended time periods[134,322], and would result in erroneously young WGD age estimates[355]. Effects of such processes are very difficult to quantify for the large time scales considered in our dataset, and their precise influence remains unknown.

4.3.2 A substantial sequence compendium and state-of-the-art Bayesian evolutionary analysis framework increase confidence in our dating results

Our current study employs a substantially larger sequence compendium compared to our previous work[193], because only six full plant genomes (*A. thaliana*, *Populus trichocarpa*, *M. truncatula*, *Vitis vinifera*, *Oryza sativa*, and *Physcomitrella patens*) were available at that time, supplemented with a few transcriptome assemblies. We now incorporate sequence data from in total 38 full genome sequences and three transcriptome assemblies (see supplementary table F.1). We originally included all transcriptome assemblies from the previous study, including *Eschscholzia californica* and *Acorus americanus*[193], but were unable to obtain unambiguous WGD age estimates for the latter with the methods used in this study (see supplementary information F.3.7). In total, we could date 31 WGDs in various species that correspond to 20 independent WGDs in the plant lineage, compared to nine independent plant WGDs previously. Additionally, the typical orthogroup phylogeny size increased to a total of 14 to 15 sequences,

compared to seven previously[193]. The orthogroup size does not scale linearly with the total number of full plant genomes, because several species were grouped into species groups for which only one representative ortholog was included, in order to increase the total number of recovered orthogroups for dating (see Material and methods). The doubling of sequence information per orthogroup, in combination with a much broader coverage of the green plant phylogeny, are expected to improve the quality of the sequence signal that guides the molecular sequence divergence estimation[67,198,216,356].

Our previous work employed the penalized likelihood inference method[197], as implemented in the r8s package[196], to date individual orthogroups[193], while the current study is based on a state-of-the-art Bayesian approach as implemented in the BEAST package, which incorporates several important methodological advances[219,220]. In particular, Markov chain Monte Carlo (MCMC) methods used in Bayesian sequence divergence estimation allow for much more parameter-rich and complex models of sequence evolution, and can also incorporate prior evidence and/or beliefs[357]. This allows for instance for orthogroup branch lengths to be estimated together with other parameters during the MCMC, instead of having to estimate them *a priori* with other methods/software to avoid propagation of branch length errors[358]. Of special importance is however the more explicit modeling of both the underlying clock model and fossil calibration uncertainty[359].

Considering the underlying clock model, it is now generally accepted that molecular evolution does not follow a strict clock[190], and this is in particular the case for the evolutionary histories of the orthologs in the random orthogroups used here, which are expected to display a much larger degree of rate variation compared to the conserved house-keeping genes that are used in traditional molecular dating studies[334]. Since rates of evolution are linked to certain life history traits such as generation time[191], relaxed clock methods are preferable[194]. Our previous work employed an autocorrelated relaxed clock model[193], which assumes that adjacent branches share similar substitution rates because the latter are correlated with mutation rates that are affected by heritable life history traits. These assumptions are however violated in case of sparse taxon sampling and when other forces such as selection are involved[65,195]. Moreover, even the very closely related *A. lyrata* and *A. thaliana* genomes exhibit a large degree of rate variation that can be attributed to other factors such as gene length, GC content, codon bias, and others[192]. Similarly, large rate variation has been reported for homeologs stemming from the *alpha* WGD in *A. thaliana*[360] and the WGD in *S. cerevisiae*[361]. Violation of the assumption of autocorrelation may however lead to inconsistent estimates when using the penalized likelihood inference method[216]. Here, we use the UCLD relaxed clock model implemented in the BEAST package, which assumes an uncorrelated lognormal distribution of evolutionary rates[219,220]. The latter is a more realistic assumption in light of the above[65,195], although a general consensus is still absent as at least one study found that autocorrelated clocks outperform uncorrelated clocks[362], while another study found that both resulted in similar posterior age estimates[67]. Bayesian model testing methods that allow comparison of their performance exist[342,363], but applying them proved infeasible in terms of the required computational resources on the scale needed here[364].

Considering fossil calibration uncertainty, a substantial body of literature demonstrates that proper modeling of such uncertainty is of paramount importance because it allows to separate the contribution of the evolutionary rate and total time to the overall observed divergence, which can heavily influence the posterior time estimates[66,67,198,216,359,365–367]. Our previous work necessitated the use of mostly secondary point calibrations that were based on other molecular dating studies, because only limited opportunities

for inserting primary calibrations based on direct fossil evidence were available[193]. Secondary calibrations carry however the risk of propagating dating errors over different studies[222], while point calibrations result in illusionary precision of the final age estimates[217]. Our current study employs only primary fossil calibrations, modeled as flexible lognormal calibration priors that mimic the associated error in fossil calibration in an intuitive way[67,222]. Orthogroup dating was always based on at least two calibrations. More calibrations allow for more rate corrections, and therefore help to guide molecular sequence divergence estimation[368]. At least one rate-correcting calibration was always present between the homeologous pair and root in all orthogroups, with the sole exception for dating the WGDs in *Nuphar advena* and *P. patens*, since their basal position necessitated a direct branch between the root and duplicate pair. Furthermore, the WGD age estimates presented in table 4.1 are robust against differences in the employed calibrations (see supplementary information F.3.3).

4.3.3 Some drastic rate shifts are not fully corrected for

Concerns have been raised that uncorrelated relaxed clocks still might not be able to correct completely for drastic rate shifts[65]. To investigate the possibility of remaining rate shift artifacts in our WGD age estimates, we performed pairwise Relative Rate Tests (RRTs) between the different plant orders, employing their respective full plant genomes that experienced a WGD where available, and found a mostly consistent pattern with in particular the orders Malvales, Malpighiales, and Rosales displaying a strong shift towards slower evolutionary rates (see supplementary information F.3.4). This has been observed before as these three orders contain only woody species in our dataset, while in particular woody status, large size, and long generation time have been associated with a strong decrease in evolutionary rate[191,369–371]. Since the first angiosperms most likely were woody species themselves[55], this apparent deceleration might however rather be viewed as an artefact due to the inclusion of multiple herbaceous species with a strongly accelerated evolutionary rate in such analyses, i.e., woody species did not strictly undergo any deceleration but herbaceous species rather underwent an acceleration. The latter does however not prevent that the lower rate of evolution of woody species most likely will lead to underestimation of their true age in analyses based pre-dominantly on herbaceous species, such as is the case here.

There is evidence that at least two WGDs for woody species in our dataset most likely represent an underestimate. First, the *P. trichocarpa* (poplar tree) WGD constitutes a shared event of the genera *Populus* and *Salix*, which both are members of the family Salicaceae within the order Malpighiales[372]. The oldest known *Populus* fossils are leaves from the Middle Eocene Evacuation Creek at Green River Formation (Utah, USA)[373,374], and are estimated to be at least 47.4 million years old[375]. Our estimate of 34.7 mya for the *P. trichocarpa* WGD (see table 4.1) thus underestimates this boundary with at least 12.7 million years. The latter is moreover conservative because there exists an additional timespan between the shared WGD and divergence of *Populus* and *Salix* itself[376]. Second, the *Malus domestica* (apple tree) and *Pyrus bretschneideri* (pear tree) WGDs similarly constitute a shared event of the genera *Malus* and *Pyrus*, which both are members of the family Rosaceae within the order Rosales[377]. Fossil *Malus* and *Pyrus* leaves from the Eocene Orchards at Republic (Washington, USA) are however estimated to be at least 48.7 million years old[378]. This age should be interpreted with due caution because fossil rosaceous leaves of closely related species are difficult to differentiate between[379], but it is supported by at least

one molecular dating analysis focusing on these genera that estimated the divergence between *Malus* and *Pyrus* to be between ~45 to 59 million years old[380]. Our two independent estimates for this shared WGD, 18.32 mya and 19.85 mya in *M. domestica* and *P. bretschneideri*, respectively, thus underestimate this boundary with at least ~28 million years. The latter is again conservative because of the timespan between the shared WGD and actual divergence of both genera[377].

The above two examples demonstrate, perhaps not surprisingly, that strong rate shifts are still difficult to fully correct for by the uncorrelated relaxed clock model when taxon sampling is limited, but it remains difficult to quantify the effects thereof. We investigated this by specifically re-dating the *P. bretschneideri* WGD based on more complete taxon sampling and additional fossil calibrations that could be implemented for this particular species, and obtained a new WGD age estimate of 30.1 mya (see supplementary information F.3.5). This constitutes an increase of more than 10 million years with respect to the original estimate, but still falls short 18.6 million years of the previously described fossil minimum bound of 48.7 million years. This result suggests that breaking up long branches in orthogroup phylogenies through better taxon sampling, in combination with better rate-correcting fossil calibrations, will allow to correct for drastic rate shifts when more full plant genome sequences become available in the future. Note that the original WGD age estimate of *P. bretschneideri* is used in table 4.1 and figure 4.3 to allow consistent comparison with the other WGD age estimates.

4.3.4 Polyploid establishment was most likely enhanced at and/or after the K-Pg boundary

Plant paleopolyploidizations cluster statistically significantly in association with the K-Pg extinction

It has been proposed that a simple ratcheting process can explain the prevalence of polyploids. In essence, because polyploidization is an irreversible process, polyploid abundance is expected to increase over time[325]. This ratcheting theory provides a null hypothesis to study paleopolyploid occurrence[325]. In particular, it predicts that successful paleopolyploidizations are distributed randomly over time. We find however, in line with previous results[193], that WGD age estimates exhibit a statistically significant clustering in time compared to a null model that assumes random WGD occurrence (p<0.05, see Material and methods; supplementary figure F.3). We fitted a mixture of Gaussians to the WGD age estimates to estimate around which age they cluster, and identified a very pronounced component at 60.05 mya (see Material and methods; supplementary figure F.4). Note that these analyses are based on the 20 independent plant WGDs by taking the average of anchor-based species-specific WGD age estimates, or peak-based if the former were not available, that describe the same shared event (see Material and methods).

This places many plant paleopolyploidizations at but especially also after the K-Pg extinction, which is the most recent of the five major mass extinctions of the Phanerozoic eon, during which an estimated ~75% of all living species became extinct[79]. Several factors probably contributed to this large-scale extinction for an extended timespan, such as increased volcanism, greenhouse warming, and in particular the bolide impact near Chicxulub (Mexico) that marks the K-Pg boundary itself at 66.0 mya[80]. Recent

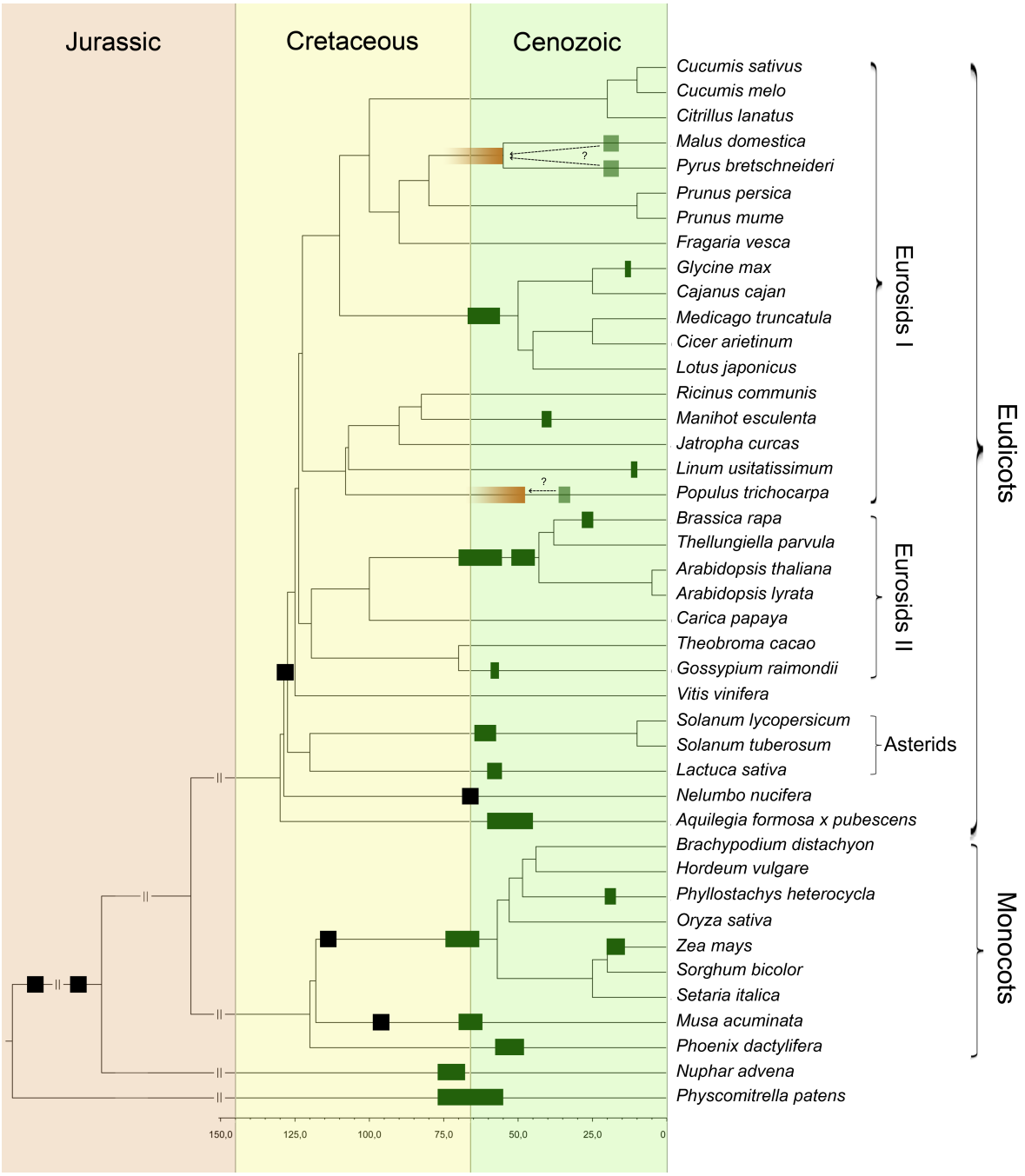


Figure 4.3: Phylogenetic tree of the green plant with all dated WGDs indicated. Phylogenetic tree of the green plants incorporating all species used in this study, with the exception of *N. nucifera* as a public annotation was not yet available upon completion. In total, sequence information from 38 full genome sequences and three transcriptome assemblies was employed (see supplementary table F.1). Bars indicate all known WGDs. Black bars indicate WGD age estimates from literature and are not to scale (see supplementary information F.3.6). Green bars indicate estimates for WGDs dated in this study, with right and left boundaries corresponding to the youngest and oldest 90% confidence interval boundary found in the complete set of species-specific WGD age estimates that descend from each independent WGD (see table 4.1). Some WGDs in woody species such as *G. raimondii* (Malvales), *P. trichocarpa* and *M. esculenta* (Malpighiales), and the WGD shared by both *M. domestica* and *P. bretschneideri* (Rosales), are most likely underestimated through strong rate deceleration that is not fully corrected for (see supplementary information F.3.4). The fading brown bars for the WGD in *P. trichocarpa*, and the WGD shared by *M. domestica* and *P. bretschneideri*, indicate corrected WGD age suggestions based on fossil evidence and/or other dating studies. The green bar for *M. acuminata* most likely represents two separate WGDs in close succession[330]. A possible WGD at the base of the monocots is not indicated because its exact phylogenetic placement remains unclear[140]. Branch lengths are truncated after 150 mya to improve clarity.

evidence indicates that this cataclysmic impact resulted in high levels of infrared radiation in the earth's higher atmosphere, which led to worldwide firestorms that set whole ecosystems ablaze and created global dust clouds that blocked sunlight for an extended period of time[81]. This was most likely especially problematic for stationary plant communities, as evidenced by the extinction of about one-third to three-fifths of plant species[84] and global deforestation[85]. The time interval for full plant community recovery was in the order of millions of years, and most early Paleogene localities are consequently characterized by an exceptionally low plant diversity[82]. The overabundance of plant paleopolyploidizations at, and/or not long after, the K-Pg boundary indicates that polyploid establishment was enhanced during this period of mass extinction and/or recovery with respect to the simple ratcheting background model, which calls for potential explanations.

Enhanced polyploid establishment through increased adaptive potential under challenging conditions

Several adaptive advantages of possessing a polyploid genomic heritage for evolutionary innovations and/or species diversifications are being untangled[324], but this long-term adaptive potential fails to explain why polyploids formed around the K-Pg boundary may have had a higher chance of establishment in the short term. Most explanations for the success of recently formed polyploids focus on their unstable genomic background that, despite most often leading to negative phenotypic effects through chromosomal abnormalities, also can infer the necessary plasticity to react quickly in a changing environmental context[200]. Typical short-term advantages include transgressive segregation and increased hybrid vigor, by which recently formed polyploids can display more extreme phenotypes than their diploid progenitors[128]. This propensity for a broader ecological tolerance and increased invasive success in vacant and perturbed habitats was previously suggested as a potential explanation for the clustering of plant paleopolyploidizations at the K-Pg boundary[193].

There are some recent indications in favor of these adaptive hypotheses. Newly formed polyploids frequently display profound morphological and physiological differences[144], and may have a higher capacity for phenotypic plasticity[381,382] compared to their diploid progenitors. For instance, despite very low genetic diversity of the founder population, increased phenotypic plasticity most likely allowed polyploid *Ceratocapnos claviculata* species to recolonize northern European habitats after the last glacial maximum[383]. Similarly, polyploid *Centaurea stoebe* species most likely displayed 'pre-adaptation' for some traits that predisposed them for colonization success upon introduction in North America ~120 years ago[384]. Polyploid *A. thaliana* plants have a broader salt tolerance, which may provide them with a fitness advantage that allows improved establishment in saline environments[385]. Polyploids may even have a higher chance of being invasive, and diploids of being endangered, on a worldwide scale[143]. Such observations support the hypothesis that recently formed polyploids possess a propensity for a higher adaptive potential under challenging conditions, whereas the cost of increased phenotypic variability and genomic plasticity is most likely too high under 'standard' conditions. This would explain why the signature of enhanced polyploid establishment upon drastic ecological upheaval, such as at the K-Pg boundary, is prominent enough to be picked up by our current, admittedly still limited, data and methods.

Enhanced polyploid establishment through mitigation of the minority cytotype disadvantage

A series of recent findings sketch an alternative explanation for enhanced polyploid establishment at the K-Pg boundary. The formation of unreduced 2*n* gametes is considered the main route towards polyploidization in plants[114,118]. Despite being traditionally viewed as too restrictive because of the low levels of unreduced gametes observed in natural plant populations, unreduced gamete production nevertheless appears adequate for cytotype coexistence in natural populations[124]. For instance, polyploid *Melampodium cinereum* populations originated recurrently since the last glacial maximum 12,000 years ago in the southwestern United States[386], illustrating that polyploids are indeed being formed continuously at an appreciable rate in stable environments. It is furthermore well established that environmental stress and/or fluctuations can even increase unreduced gamete formation in plants[114]. The underlying molecular processes are being unraveled[119], and it appears that many of their associated components are thermosensitive[387]. For instance, both heat stress in *Rosa* species and cold stress in *A. thaliana* led to increased unreduced gamete formation through alterations in spindle formation during meiosis II[388], and alterations in post-meiotic cell plate formation and cell wall establishment[389], respectively. Similar observations exist in interspecific *Brassica* hybrids subject to cold stress[390], while most hybrids already exhibit increased levels of unreduced gamete formation[114]. Recent evidence supports that environmental stress and/or fluctuations could also have increased unreduced gamete levels at previous large-scale extinctions, as demonstrated by the increased number of unreduced fossil pollen found in the now extinct conifer family Cheirolepidiaceae at the Triassic-Jurassic transition 201.3 mya[391]. Abnormal gymnosperm pollen[392] and lycophyte spores[393] have also been reported at the Permian-Triassic transition 252.3 mya[394]. The former and latter boundary correspond to the second and third most recent mass extinctions in the Phanerozoic, respectively[79].

These observations indicate that environmental stress and/or fluctuations can enhance plant polyploidization by promoting unreduced gamete formation. Alternatively, even in the absence of the latter, massive extinction of both diploid and polyploid cytotypes can decrease the overall plant population sizes markedly, which increases the role of stochastic drift in allowing to overcome the minority cytotype disadvantage by random chance events[148]. Both stress and extinction therefore have the potential to mitigate the polyploid minority cytotype disadvantage by increasing their chances of finding suitable mating partners. Enhanced polyploid establishment under such conditions therefore does not necessarily require any direct adaptive advantage that promotes polyploid survival, but may rather be based on higher polyploid formation. This more neutral scenario is supported by modeling approaches that do not assume any *a priori* adaptive advantages of newly formed polyploids, but nevertheless find increased replacement of diploids by polyploids under a changing environment[395]. Empirical observations also indicate that recently formed polyploids are much more abundant in stressful environments such as the Arctic[145], which might be due to both their adaptive potential and/or increased unreduced gamete formation[146]. Mitigating the minority cytotype disadvantage by increasing the polyploid minority cytotype frequency through increased unreduced gamete formation, and/or the influence of stochastic drift through overall background extinction of plant populations, does therefore constitute an alternative neutral explanation for the clustering of plant paleopolyploidizations at the K-Pg boundary that was not previously considered. Moreover, there exists a lag phase in the order of millions of years between the extremely stressful

environmental conditions and the massive extinction associated with the K-Pg boundary itself, and plant population recovery afterwards[82,84], which effectively opens up an extended timespan during which the polyploid minority cytotype disadvantage was most likely alleviated. This would also explain why, apart from underestimated WGD ages through drastic rate shifts in some woody species (see before), plant paleopolyploidizations appear to cluster somewhat after the K-Pg boundary in a period characterized by slow recovery of plant population structure and size.

4.4 Conclusion

In this study, we dated 20 independent plant paleopolyploidizations. In line with previous results[193], we find that plant paleopolyploidizations in the last ~100 million years are not distributed randomly over time but that many of them cluster in association with the K-Pg extinction boundary, which defies the hypothesis that successful polyploid establishment can be explained entirely by a simple ratcheting process. Given that our results are based on a substantial plant sequence information compendium with broad taxonomic coverage and a state-of-the-art Bayesian evolutionary analysis approach that incorporates uncorrelated relaxed clock models and fossil calibration uncertainty, this establishes the association of plant paleopolyploidizations with the K-Pg boundary as a legitimate hypothesis that warrants further investigation to either falsify or establish potential mechanistic explanations. In particular, we suggest that apart from traditional explanations for the success of recently formed polyploids that focus on their adaptive potential under sufficiently challenging conditions, more neutral mechanisms involving increased unreduced gamete formation and/or the influence of stochastic drift through background extinction merit further attention. We emphasize that our results do not support, nor do we claim, that WGD was either a prerequisite or guarantee for plant survival at the K-Pg boundary. Similarly, extinction and stress should not be viewed as absolute prerequisites or guarantees for successful polyploid establishment. We argue however that the establishment potential of polyploids should be viewed in light of the environmental and ecological challenges and opportunities at the time of polyploidization, with in particular stress and extinction being good candidate factors for promoting polyploid establishment. We believe that such a perspective will help to mitigate some of the conflicting hypotheses and observations on the proposed evolutionary fates of polyploids.

4.5 Acknowledgements

The authors thank three anonymous reviewers for their constructive comments on a previous version of the manuscript. This work was supported by Ghent University [Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”]. Kevin Vanneste and Steven Maere are fellows of the Fund for Scientific Research Flanders (FWO). Guy Baele receives funding from the European Union Seventh Framework Programme [FP7/2007-2013] under ERC Grant agreement no. 260864. Yves Van de Peer acknowledges support from the European Union Seventh Framework Programme [FP7/2007-2013] under ERC Advanced Grant Agreement no. 322739 - DOUBLE-UP. This work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University,

the Hercules Foundation, and the Flemish Government Department EWI. The authors would like to acknowledge Michiel Van Bel for assistance with extraction and manipulation of data in the PLAZA platform (`http://bioinformatics.psb.ugent.be/plaza`); Kenneth Hoste, Ewald Pauwels, and Luc Van Wiemeersch for assistance in setting up the high-performance computing dating analysis; Jens Hollunder for fruitful discussions regarding orthology detection; and Stephane Rombauts, Lieven Sterck, and Yao-Cheng Lin for fruitful discussions regarding genome annotation data input and quality.

4.6 Author contributions

I performed all the analyses described in this chapter, and wrote the resulting research article. Both were done under supervision and with significant contributions of Guy Baele, Steven Maere, and Yves Van de Peer.

Chapter 5

A burst of WGDs at the end of the Cretaceous and the consequences for plant evolution

Kevin Vanneste, Steven Maere, Yves Van de Peer. Tangled up in two: A burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences. In press*.

Abstract

Genome sequencing has demonstrated that besides frequent small-scale duplications, large-scale duplication events such as whole genome duplications (WGDs) are found on many branches of the evolutionary tree of life. Especially in the plant lineage there is evidence for recurrent WGDs, and the ancestor of all angiosperms was in fact most likely a polyploid species. The number of WGDs found in sequenced plant genomes allows us to investigate questions about the roles of WGDs that were hitherto impossible to address. An intriguing observation is that many plant WGDs seem associated with periods of increased environmental stress and/or fluctuations, a trend that is evident for both present-day polyploids and paleopolyploids formed around the Cretaceous-Paleogene (K-Pg) extinction 66 million years ago. Here, we revisit the WGDs in plants that mark the K-Pg boundary, and discuss some specific examples of biological innovations and/or diversifications that may be linked to these WGDs. We review evidence for the processes that could have contributed to increased polyploid establishment at the K-Pg boundary, and discuss the implications on subsequent plant evolution in the Cenozoic.

For the author contributions, see page 112.

5.1 Introduction

Flowering plants typically have large genome sizes and contain many genes, the majority of which evolved during the past 250 to 300 million years through gene duplication[159]. A particularly striking feature of plant genomes, also explaining their large sizes, is the large number of whole genome duplications (WGDs) that have been uncovered[52,128,135]. It is now commonly accepted that one WGD occurred in the ancestor of all seed plants, and an extra one in the ancestor of all flowering plants, so that every extant angiosperm is in fact a paleopolyploid containing the remnants of at least two WGDs[136]. Furthermore, a hexaploidy event predates the origin of all core eudicots, which make up approximately 75% of extant angiosperm diversity[137–139], while traces of a WGD at the base of the monocots also suggest a WGD shared by most, if not all, monocots[140]. In addition, several more recent independent WGDs have been unveiled in many different plant lineages. As a result, the genomes of some extant plant species carry the remains of up to six successive genome duplications[141]. Here, we focus on the more 'recent' paleopolyploidizations that occurred in the last 100 million years, a large fraction of which seemingly took place around the Cretaceous-Paleogene (K-Pg) extinction event, 66 million years ago (mya)[193]. We have an in-depth look at this wave of WGDs associated with the K-Pg boundary, many of which predate lineage diversifications that resulted in some of the largest and arguably most successful present-day plant families, often characterized by particular biological innovations. Finally, we review processes that can explain these observations, and discuss how these paleopolyploidizations could have influenced plant evolution in the Cenozoic.

5.2 A burst of genome duplications at the K-Pg boundary

In 2009, we described a tentative link between many of the known paleopolyploidization events in plants and the K-Pg boundary, and speculated that WGD was linked to plant survival around that time[193]. Although many found this an interesting hypothesis[215], most remained sceptical, in particular because of the limited amount of data available at that time and because dating ancient events that occurred tens of millions of years ago is often problematic. Only six complete genome sequences and a few transcriptome assemblies were available for analysis in 2009, limiting both the taxon sampling and possibility to implement proper primary fossil calibrations. Dating was done using a penalized likelihood inference method that incorporates an autocorrelated relaxed clock model, which assumes that branches that share a direct common ancestor also share similar evolutionary rates[197]. This assumption seems however unlikely in light of the sparse taxon sampling considered[195], and violation thereof may lead to inconsistent age estimates[216]. Calibrations were typically implemented as fixed secondary point calibrations, which may lead to illusionary precision of the time estimates[217].

Recent years have seen a huge increase in plant (whole genome) sequence data[218], in addition to the development of more powerful Bayesian methods for sequence divergence estimation[219–221], as well as more powerful high-performance computing systems that allow such intensive Bayesian algorithms to be run on a massive scale. We therefore recently revisited the hypothesized link between the K-Pg mass extinction and successful WGDs[396]. We used plant genome sequence information from a total of 41 species representing a broad coverage of the overall angiosperm phylogeny, incorporating 38 full

genome sequences and three transcriptome assemblies, greatly improving taxon sampling with respect to the previous study[193]. In total, 20 independent WGDs could be dated compared to nine previously by dating all their identifiable homeologs created by the WGD event. For WGDs for which genome sequence information was available for several descendant species (e.g., WGDs preceding the divergence of Solanaceae, Fabaceae, or Poaceae - see further), this WGD was dated independently for each species to assess their individual age estimates. Absolute age distributions were then constructed for each species WGD, for which a consensus WGD age estimate was obtained by taking the mode of its kernel density estimate, which is more flexible in comparison with traditional parametric distributions because it allows a better exploration of the true underlying shape of the distribution[350], while 90% confidence intervals were obtained through a bootstrapping procedure[351]. Dating itself was done with the BEAST package[220], using an uncorrelated relaxed clock model that assumes a lognormal distribution on evolutionary rates[219], and therefore should be better equipped to deal with rate shifts between different branches compared to autocorrelated relaxed clocks when taxon sampling is limited[65]. Proper calibration priors in Bayesian time estimation are of paramount importance as they can have a profound impact on the posterior age estimates[67,198,216,365,366]. Primary fossil calibrations were implemented as flexible lognormal calibration priors that represent the error associated with the age of the fossil in a more intuitive manner[67,222]. Fossils have a hard minimum bound corresponding to the earliest age to which the fossil can reliably be attributed to. The peak mass probability can be put at some distance after this earliest age to accommodate for the lag between first fossil occurrence and the actual divergence event the fossil is used to describe. Lastly, the lognormal distribution has an infinite extending but small probability tail that can be used as a soft maximum bound to account for the uncertainty associated with choosing proper maximum bounds for fossil calibrations. More detailed information can be found in Vanneste et al.[396].

An updated overview of paleopolyploidizations is summarized in figure 5.1[396]. Although dating of such ancient events surely remains a challenging exercise, and WGD dates are subject to change as more plant sequence data and powerful dating methods become available[214,215,397], many plant paleopolyploidizations were again found to cluster at the K-Pg boundary[396], supporting our previous observations[193].

5.3 Implications of genome duplications associated with the K-Pg boundary

The increased long-term survival of WGDs around the K-Pg boundary appears indicative of enhanced polyploid plant establishment at that time, either because WGDs provided a selective advantage for polyploids compared to their diploid progenitors, or alternatively, because the cataclysmic events that took place 66 mya were responsible for the production of an excess of polyploids (see further). However, whether cause or effect, many of these WGDs predate the radiation of some very large and successful plant families with particular biological innovations. Similar observations can be done in other parts of the tree of life, where WGDs are often found at branches leading to species-rich clades, such as $>$25.000 species of teleost fishes and $>$350.000 species of flowering plants[52,318]. On the other hand, one should be cautious not to over-interpret the importance of WGDs for species radiations. For instance, in vertebrates

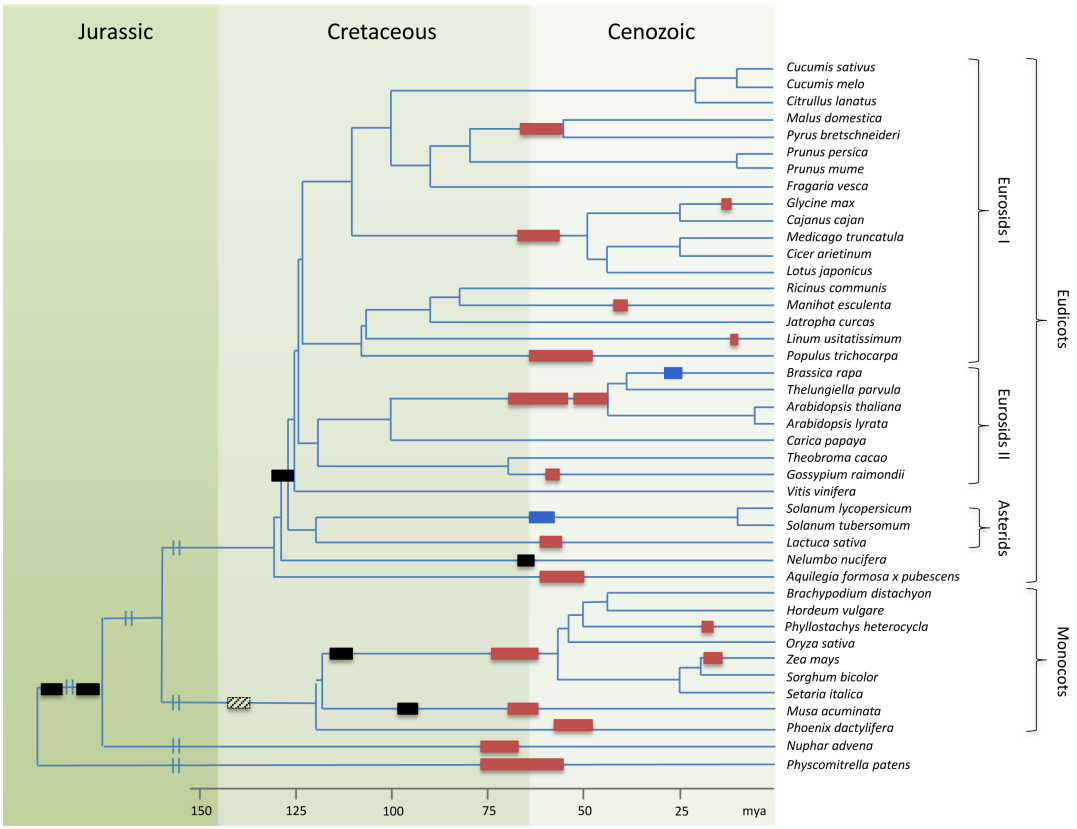


Figure 5.1: A wave of WGDs is associated with the K-Pg boundary ~66 million years ago. The figure illustrates the tree topology for the green plants with all known WGDs indicated by bars. Red and blue bars represent 90% confidence intervals on dated tetraploidies and hexaploidies, respectively. Black bars represent WGD age estimates from literature[396]. A possible WGD at the base of the monocots is indicated by a dashed bar because its exact phylogenetic placement remains unclear[140]. The WGD for *Populus trichocarpa* and the one shared by *Malus domestica* and *Pyrus bretschneideri* are corrected WGD age estimates based on fossil minimum boundaries and/or other dating studies[396]. Branch lengths are truncated after 150 million years ago to allow a better overview. Figure adapted from Vanneste et al.[396].

it was suggested that the often quoted correlation between the teleost fish WGD and increased post-WGD diversity and/or complexity does not hold when extinct basal lineages were considered because pre-WGD extinct teleost lineages demonstrate a strong diversification similar to post-WGD extant lineages[398]. However, since those pre-WGD lineages are mostly extinct while post-WGD lineages still thrive, this demonstrates that teleost fish evolution rather fits a more nuanced pattern of reduced extinction risk after WGD, resulting in a lag period between WGD and its effect on species diversity and/or complexity[199]. Additionally, it was recently demonstrated that an extended period of about 40 to 50 million years passed between the salmonid-specific WGD and strong lineage diversification, suggesting the latter was probably mostly driven by climatic factors[399]. Below, we will first examine a few examples of biological innovations (or better said, elaborations thereof[141]) that can reliably be traced back to WGDs located at the K-Pg boundary in plants, focussing on fleshy fruits in the Solanaceae and advanced nodulation characteristics in the papilionoids, before taking a deeper look at evidence whether or not these WGDs could have directly enhanced speciation.

5.3.1 Biological novelty

Fleshy fruits

The fleshy fruits observed in some plant lineages are an important biological innovation that serves to enhance seed distribution by attracting vertebrate frugivores for long-distance seed dispersal, and hence increases plant success[400]. Specialization of the fleshy fruit for particular (groups of) vertebrates may also enhance speciation[401]. Based on the recently published genome of tomato (*Solanum lycopersicum*), a genome triplication event in the Solanaceae shared with potato (*Solanum tuberosum*) was firmly established[402] and dated at the K-Pg boundary (see figure 5.1). Many new gene family members with important fruit-specific functions were created through this WGD. Figure 5.2a illustrates several genes in the fruit ripening control network that are paralogs with different physiological roles generated through the genome triplication. These include for instance the transcription factors and enzymes necessary for ethylene biosynthesis (*MADS1*/*RIN*, *CNR*, and *ACS2*/*ACS6*), red light photoreceptors influencing fruit quality (*PHYB1*/*PHYB2*), and also some effector genes mediating lycopene biosynthesis (*PSY1*/*PSY2*) that control fruit pigmentation. Endogenous ethylene receptors (*ETR3*/*ETR4*) created by the eudicot-wide genome duplication also participate in this network. Similarly, fruit texture is controlled in part by over 50 genes that encode proteins involved in modification of cell wall structure and composition, and show differential expression during fruit development and ripening. Figure 5.2b for instance illustrates the expansion, through genome triplication and subsequent tandem duplications, of a family of xyloglucan endotransglucosylase/hydrolases (*XTHs*) involved in determining fruit texture. Differential loss between tomato and potato of one of the triplicated members, *XTH10,* suggests that genetic specialization, and hence diversification between the different members of the Solanaceae, was facilitated by the triplication event[402]. It should however be noted that fleshy fruits exist in many different plant lineages, many of which are not marked by a specific polyploidy, emphasizing that the Solanaceae-shared WGD contributed several genes that were later incorporated into more elaborate fleshy fruit development, so that the latter represents an 'elaboration' rather than a true 'innovation'[128].

Rhizobial nodulation

A common feature of most papilionoid legumes is rhizobial nodulation, the formation of specialized organs called root nodules, which host nitrogen-fixing rhizobial symbionts. Nodulation is a biological innovation that allows to grow on nitrogen-deprived soils because plants receive fixed nitrogen from their symbionts, in return for a steady supply of carbon and energy sources[403]. Specialization for different rhizobial symbionts may also have aided papilionoid speciation[404]. Analysis of the genome sequence of *Medicago truncatula* confirmed that the papilionoid-shared WGD, also located at the K-Pg boundary (see figure 5.1), has played an important role in the evolution and elaboration of rhizobial nodulation[405]. Nodulation is initiated when the plant signalling system comes into contact with specific bacterial Nod factors, which in papilionoids evolved a distinctly nodulation-specific function[406]. Analysis of the *M. truncatula* genome learned that both the Nod factor receptor *NFP* and transcription factor *ERN1* have paralogs, *LYR1* and *ERN2* respectively, that originated through the papilionoid WGD. Figure 5.3 illustrates that both gene pairs show divergent expression patterns, reflecting functional specialization. *NFP* and *ERN1*

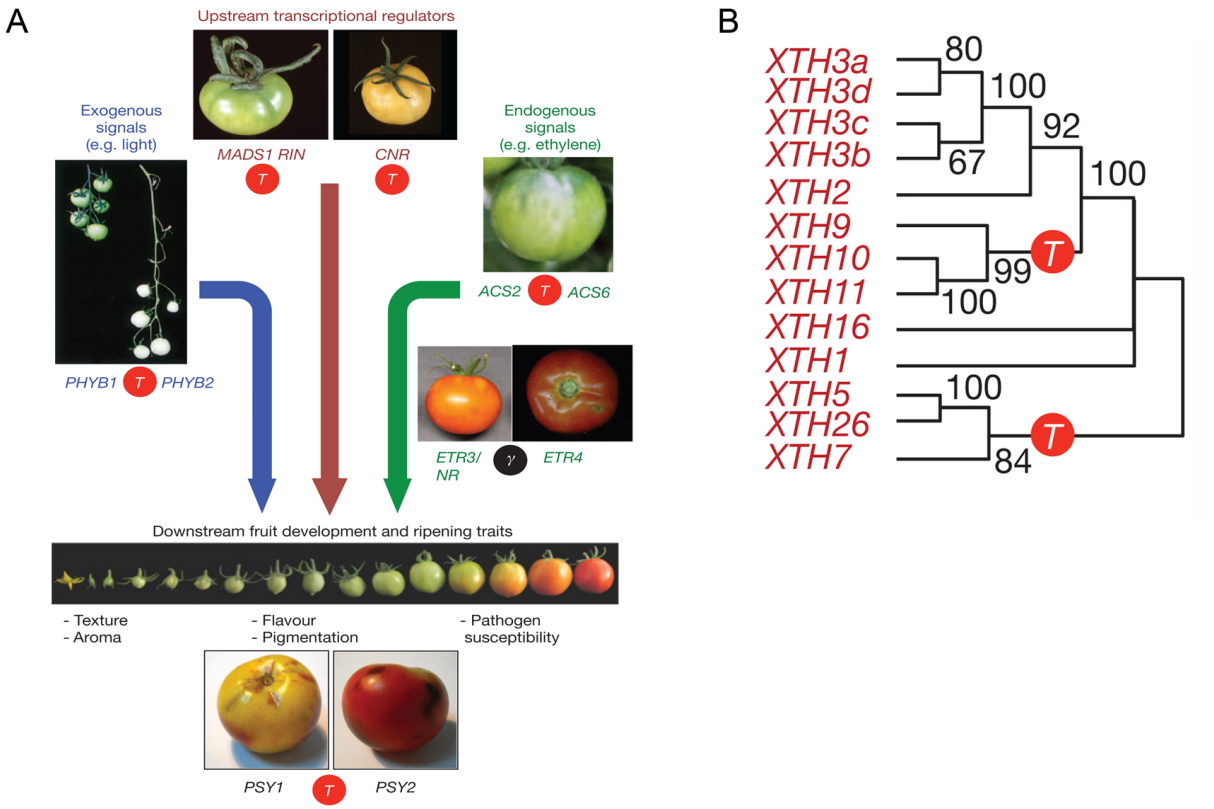


Figure 5.2: The Solanaceae-specific genome triplication contributed to the evolution of the tomato fruit. (A) Illustration of the fruit ripening control network. The upstream transcriptional regulators *MADS-RIN* and *CNR*, in combination with the enzyme ACC synthase (*ACS*), control the production of the ripening hormone ethylene. Ethylene receptors (*ETR*) drive expression changes in several output genes, including phytoene synthase (*PSY*), which is the rate-limiting step in carotenoid biosynthesis. Light influences fruit pigmentation through an ethylene-independent pathway mediated by phytochromes (*PHY*). Several key component paralogous gene pairs (*MADS1*/*RIN*, *PHYB1*/*PHYB2*, *ACS2*/*ACS6*, *PSY1*/*PSY2*) were generated by the genome triplication (*T*, red circle), while *ETR3*/*ETR4* was created by the core eudicot shared hexaploidy (γ, black circle). (B) Illustration of the expansion by both genome triplication (*T*, red circle) and tandem duplications of a family of xyloglucan endotransglucosylase/hydrolases (*XTHs*), which control fruit ripening through modification of cell wall structure and composition. Figure adapted from Sato et al.[402].

are expressed predominantly in the nodule and are known to be active in nodulation[407], whereas *LYR1* and *ERN2* are highly expressed during mycorrhizal colonization. This suggests that these nodulation-specific signalling components are derived from more ancient genes originally functional in mycorrhizal signalling that evolved new transcriptional functionality after the papilionoid WGD[405]. Additional support for this conclusion comes from the observation that the ortholog of *NFP* in a nodulating non-legume outgroup, *Parasponia andersonii*, functions both in nodulation and mycorrhizal signalling[408]. Interestingly, a nodulating legume outgroup that did not share the papilionoid WGD, *Chamaecrista fasciculata*, exhibits ancestral nodule characteristics in comparison with most nodulating papilionoids[409]. *Parasponia* diverged somewhere between 100 and 120 mya from the papilionoids[75], whereas *Chamaecrista* diverged ~60 mya from the papilionoids[409]. Independent from whether their last common ancestor could already perform nodulation or whether this trait evolved independently in both lineages, this would suggest that the ability for advanced nodulation characteristics was not able to evolve for about 40 to 60 mya, whereas it did so very rapidly after the papilionoid WGD[409]. This emphasizes that although the papilionoid WGD was not an absolute prerequisite for the evolution of nitrogen-fixing nodulation, it most likely facilitated the development of several elaborate papilionoid nodule forms.

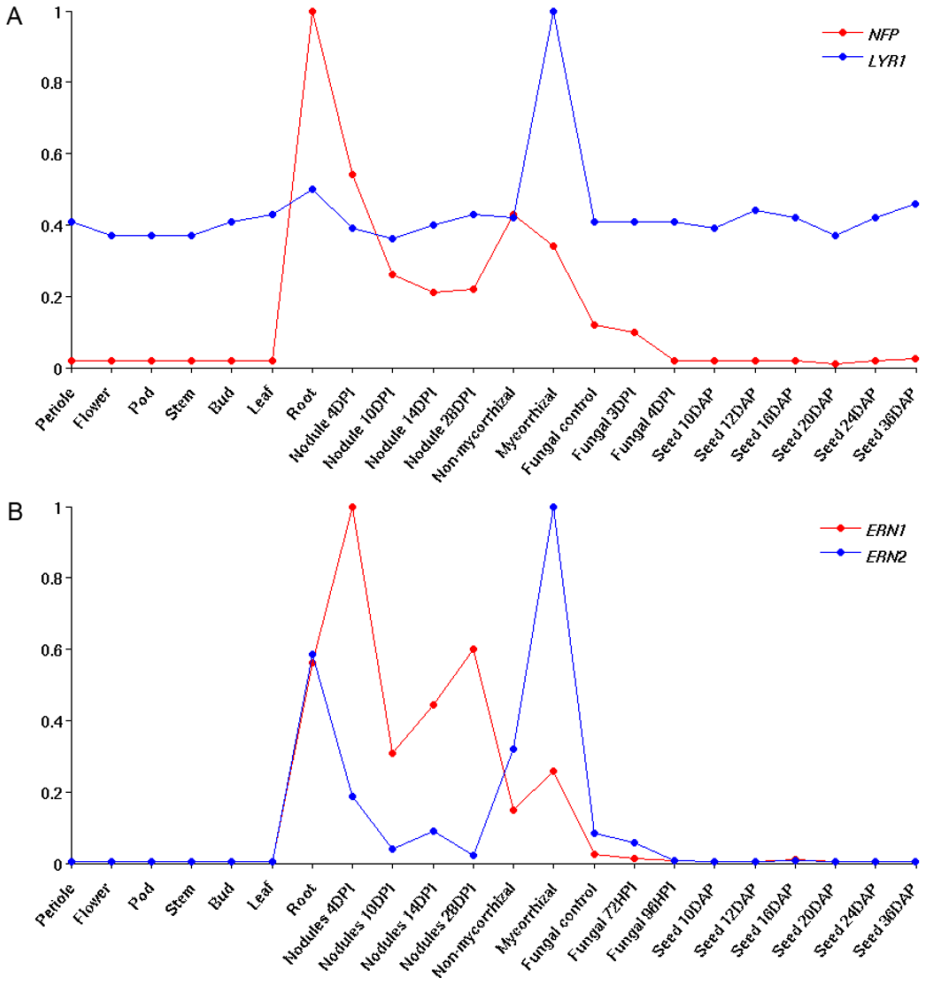


Figure 5.3: The papilionoid genome duplication contributed to the evolution of nodulation. Paralogs created by WGD, (A) *NFP* and *LYR1*, and (B) *ERN1* and *ERN2*, display contrasting expression patterns, suggesting functional specialization. *NFP* and *ERN1* are expressed predominantly in the nodule, whereas *LYR1* and *ERN2* are highly expressed during mycorrhizal colonization. The average transcript levels of three replicates are shown, scaled by dividing each data point by the maximum mean transcript level across all experiments. DPI = days past inoculation. DAP = days after pollination. Figure adapted from Young et al.[405].

To assess the contribution of the papilionoid WGD to *M. truncatula* nodulation in more detail, Young et al.[405] also investigated the expression of 618 homeologous gene pairs from six different organs based on RNA-seq data for one or both homeologs, to determine the number of genes showing organ-enhanced expression (defined as having expression in a single organ that is at least twice the level in any other). A large fraction of homeologs demonstrated organ-specific enhanced expression. Among homeologous gene pairs with nodule-enhanced expression, a single paralog was nodule-enhanced in 43 out of 51 gene pairs, with the other eight gene pairs showing nodule-enhanced expression for both gene pairs. Out of 142 transcription factors derived from the papilionoid WGD for which RNA-seq data was available, 11 showed such enhanced nodule expression. These results indicate that many homeologous genes, in particular signalling components and regulators, were retained after the papilionoid WGD and gained specialized roles in nodulation afterwards. However, some other nodule-related genes were found to derive from the core-eudicot specific hexaploidy. This confirms a more complex model wherein the capacity for primitive interaction with new symbionts evolved quite early, derived from the existing

mycorrhizal machinery, explaining the evolution of nodulation in multiple plant lineages[403,410], after which the papilionoid WGD allowed the creation of additional genes that were incorporated into the development of more advanced nodulation characteristics[405]. A recent integrated comparative genomic approach based on the sequenced genomes of four papilionoid species (*M. truncatula*, *Lotus japonicus*, *Glycine max*, and *Cajanus cajan*) supports this by demonstrating that many of the approximately 25% of WGD-derived duplicate pairs that have been retained, show high levels of expression divergence and function in different processes required for successful nodulation[411].

5.3.2 Speciation

The previous examples of biological innovations originating through the retention of WGD duplicates suggest that WGDs, through assisting biological innovations and diversifications, might also facilitate speciation. For instance, as stated previously, specialization for interactions between particular vertebrate frugivores for seed dispersal in fleshy fruits or with specific rhizobial symbionts in nodulation, might aid speciation. However, the question remains whether WGD itself can also actively promote speciation. Some of the WGDs associated with the K-Pg boundary (see figure 5.1) predate extremely successful plant lineages characterized by species radiations following the WGD event. These include the Brassicaceae (∼3,700 species), Poaceae (∼10,000 species), Asteraceae (∼23,600 species), Solanaceae (∼2,460 species), and Fabaceae (∼19,500 species). Many of these however have a species-poor sister group that shared the WGD event, which led to the development of the WGD-Radiation Lag Time model that emphasizes that the success of these plant families should be viewed in light of their specific evolutionary routes taken[324]. Even the limited set of species in figure 5.1 demonstrates that many present-day plant families, such as for instance the Cucurbitaceae, represented by *Cucumis melo*, *Cucumis sativus*, and *Citrullus lanatus*, did not undergo any WGD in the last ∼100 million years. Using the number of species as a simple, albeit admittedly crude, measure for success, this family of about 950 - 980 species can also be considered fairly successful[412]. Alternatively, some plant families with a paleopolyploid history, such as the Nymphaeaceae, have arguably not been very successful in terms of species radiation, counting only around 70 species[413]. Such observations emphasize the importance of ecological opportunity for realizing plant evolutionary potential, irrespective of polyploidization[15,141,193,324].

Nevertheless, the success of many plant families that have undergone a WGD suggests that their strong diversification may be ascribed, at least partly, to their polyploid ancestry. In an attempt to gauge the effect of WGD on speciation, Soltis et al.[52] tested whether such post-WGD clades displayed higher diversification rates, while accounting for the confounding effects of extinction. Although the results were considered preliminary, due to the lack of reliable genomic data for paleopolyploidy in combination with insufficient taxon sampling to place WGDs confidently on plant family phylogenies, a highly statistically significant relationship between diversification and the WGD was found for four of the five aforementioned successful plant families. The fifth plant family, the Asteraceae, was not considered and a statistical relationship hence remains untested. It should however be noted that the latter constitutes the single largest present-day angiosperm family[414].

The molecular mechanisms that might promote speciation after WGD are still not very well understood. One often quoted mechanism is reciprocal gene loss (RGL), the genetic isolation of separated

populations through loss of different gene copies that lead to incompatibilities when the populations encounter each other again[159,415]. Through WGD, a very large pool of loci becomes available simultaneously for divergent resolution between subpopulations, which could quickly result in reproductive isolation if essential genes are involved. Scannell et al.[416] demonstrated that the pattern of duplicate gene pair loss differs at 20% of all loci between three different yeast species that shared a WGD. Similarly, about 8% of ancestral *Tetraodon* and zebrafish loci were subjected to RGL after the teleost fish WGD[417]. For plants, the situation is less clear. Schnable et al.[418] separated the two subgenomes of modern grasses derived from the WGD shared by the Poaceae. In contrast to the aforementioned studies in yeast and teleost fishes, strong evidence of RGL between homeologs of the different subgenomes was lacking, suggesting post-WGD RGL was unlikely to be a driving force in the radiation of the grasses[418], although systematic studies about RGL in plants are still missing.

Genes however do not necessarily need to get lost or silenced, as other neutral scenarios after gene duplication might also promote speciation. Many genes perform multiple functions through differential expression at different developmental stages and/or tissues. Duplication of such genes often leads to subfunctionalization, the division of the subfunctions over the two daughter copies[103,159]. Alternatively, genes can have trace activity for a second function whose optimization is constrained by adaptive conflicts with the primary function, which can be resolved by optimizing the functions separately in different paralogs after duplication, see for instance Voordeckers et al.[419]. Reproductive isolation of such a population, for instance driven by geological phenomena that lead to geographical barriers, could lead to orthologs of the two isolated populations acquiring different subfunctions. Although F1 hybrids in contact zones from the two populations would develop correctly because each (sub)function is performed by one of the genes from each population, 1/8th of the F2 zygotes will lack one of the (sub)functions, which could be lethal if such functions are essential[420,421]. As for RGL, this effect would be exacerbated in the case of WGD, which generates a much larger number of duplicate loci that can be divergently subfunctionalized[128]. Lineage-specific subfunctionalization could therefore in theory accelerate speciation, but remains untested.

5.4 Both neutral and adaptive processes most likely contribute towards enhanced polyploid establishment under stressful conditions

Above, we discussed new evidence that seems to provide further support for the association between plant paleopolyploidizations and the K-Pg boundary, some of which can be linked to particularly successful biological innovations and increased diversification rates. The K-Pg boundary is especially known for its associated extinction event, which constitutes the last of the five major mass extinctions in the Phanerozoic eon[78]. This cataclysmic event most likely resulted from the combination of several factors such as increased volcanism, greenhouse warming, and in particular the bolide impact near Chicxulub (Mexico)[80], resulting in a challenging unstable environment impairing the survival of most living organisms[81]. The question remains, why, at a time when an estimated ~75% of all species went extinct[79], many of the plant

species we are all so familiar with likely underwent a WGD? Similar observations are done for present-day polyploids, which are often encountered in unstable and stressful environments[422]. For instance, there is an overabundance of recently formed polyploids in the Arctic[145]. Below, we will discuss two, not mutually exclusive, processes that could help explain this pattern and the implications thereof for plant evolution.

5.4.1 The adaptive scenario

The adaptive scenario explaining polyploid success has been explored extensively in the past decade[52,127,128,144,200,423], and will therefore only be covered concisely here. This scenario is mostly based on a characteristic often displayed by newly formed polyploids, namely transgressive segregation, i.e., the formation of more extreme phenotypes in the resulting hybrid populations compared to their diploid parents[200]. The latter becomes more pronounced as the two parental genomes contributing to the polyploid become more diverged, especially so in allopolyploids that result from the merger of two different species, which may display strong hybrid vigour (heterosis) by virtue of possessing novel allelic combinations not found in either parent[424]. The exact molecular mechanisms behind hybrid vigour are however still largely unknown[425], although is has been suggested recently that cells can maybe distinguish between parental alleles based on their relative protein and mRNA stability, which therefore conserves energy otherwise required for removal of such unstable products that can be used to promote growth and expression of new favourable traits[426].

Irrespective of the exact molecular mechanisms, genomic instability and gene expression changes soon after polyploid formation may result in increased phenotypic variability of the polyploids with respect to their diploid progenitors[141]. Genomic instability refers to the extensive structural changes of the chromosomal DNA that typically take place in the first few generations after polyploidization, such as fusions, fissions, duplications, inversions, translocations, and eliminations[427], often coupled to mitotic and meiotic abnormalities[157,428]. Gene expression typically changes markedly[429], in conjunction with widespread epigenetic repatterning[430], in the first few generations after polyploidization. These structural and expression changes have collectively been described as genomic shock, and in the case of allopolyploids seem to be attributable to both the hybridization process[431] and the genome doubling itself, with the latter possibly having a calming effect[432]. Although these extensive changes often result in decreased polyploid fitness and increased offspring sterility, in light of increased phenotypic variability, they can also confer plasticity to the polyploid genome to allow quick adaptation to new environments and changing conditions[127,144,200,433,434].

Other potential advantages of newly formed polyploids include the masking of deleterious recessive alleles leading to increased genetic redundancy[435], network redundancy on a larger scale[436], and possibly even an increased capacity for phenotypic plasticity itself[381,382]. Polyploids also often exhibit traits that promote their establishment through mitigating the minority cytotype disadvantage, which is a strong negative frequency-dependent selection on the polyploid through a large proportion of ineffective matings with the diploid progenitor majority cytotype[122]. Such traits include the loss of self-incompatibility, which enables selfing, and the gain of apomixis, which enables asexual reproduction. Polyploidization is also sometimes associated with a shift from annual to perennial habit, which opens up a longer time window for successful mating. Lastly, their fast morphological and/or physiological differentiation can

enhance the number of successful matings through sympatric niche separation from the diploid progenitor population[144,155,437].

5.4.2 The neutral scenario

A series of recent findings point to the possibility of a more neutral scenario to explain the apparent association between paleopolyploidizations and the K-Pg boundary[396]. It has been acknowledged for a long time that the formation of unreduced gametes is the main mode of polyploid formation in plants, but the low estimates of unreduced gamete production in natural populations typically seemed too restrictive for the establishment of polyploids[114,118]. Although the chance of two unreduced 2*n* gametes meeting is very low, tetraploid occurrence is most likely facilitated by a triploid bridge, the creation of an intermediate triploid stage through the combination of an unreduced 2*n* and reduced *n* gamete[438]. Such triploids often display large fertility and fitness defects, but also produce enhanced levels of unreduced 3*n* gametes that can form tetraploids through backcrosses with reduced *n* gametes from the diploid progenitor population, and hence alleviate the minority cytotype disadvantage[390,439]. Accordingly, a recent general gametic modelling approach for diploid-polyploid systems that predicts equilibrium ploidy frequencies based on empirical estimates of unreduced gamete formation, demonstrated that these low levels can be adequate to explain a drift towards higher ploidy[124].

Another well-documented observation is that levels of unreduced gamete formation can be increased by external stimuli such as stress and a fluctuating environment[114,146,149,387,440–442]. Especially temperature has a pronounced effect on unreduced gamete formation. Increasing temperatures to extreme levels in *Rosa* species resulted in more unreduced gametes being produced through alterations in spindle formation during meiosis II[388]. Similarly, inducing cold stress increased unreduced gamete formation in *A. thaliana* through alterations in post-meiotic cell plate formation and cell wall establishment[389]. Although hybridization itself typically also increases the levels of unreduced gamete formation in plants[147], temperature levels can potentially also enhance this hybrid trait, as witnessed in some *Brassica* interspecific hybrids after cold treatment[390]. Moreover, it became recently clear that the effect of the environment on unreduced gamete formation is most likely not limited to present-day plants. Increased levels of fossil unreduced pollen were observed in the now extinct conifer family Cheirolepidiaceae at the Triassic-Jurassic transition, which corresponds to the fourth of the five major extinction events[391]. Abnormal gymnosperm pollen[392] and lycophyte spores[393] have also been reported during the Permian-Triassic transition, corresponding to the third of the five major extinction events.

Increased unreduced gamete production during times of environmental stress and/or fluctuation could thus be an important factor in explaining the apparent clustering of paleopolyploidizations at the K-Pg boundary[396]. It could also explain why many present-day polyploids often are more abundant in stressful environments, such as the Arctic[145] or habitats created by anthropogenic disturbance[443]. For both the K-Pg boundary and present-day examples, the association between increased polyploid establishment and environmental stress and/or fluctuation would not require any explicit adaptive advantage, but could be explained by a neutral mechanism[146] such as increased unreduced gamete formation. This is in agreement with modelling approaches that predict increased replacement of diploids by polyploids under a changing environment, without assuming any *a priori* adaptive advantage of the polyploids[395]. The

effect of increased unreduced gamete production during environmental stress and/or fluctuation is even expected to be intensified through higher background extinction levels of the diploid populations[199], increasing the overall relative frequency of unreduced gametes to the total gamete pool, which would enhance the chance of successful unreduced gamete matings.

Accumulating evidence for a more prominent role of the neutral scenario does however not preclude a role for the adaptive one. Figure 5.4 summarizes an intertwined situation wherein environmental stress and/or fluctuation drive polyploid formation through increased unreduced gamete production, after which adaptive processes act to ensure polyploid establishment. Dependant upon specific circumstances, either the neutral or adaptive component could carry more weight. The apparent association of paleopolyploidizations with the K-Pg boundary[396], and present-day polyploids with stressful habitats[145,443], in combination with evidence that unreduced gamete formation is a major route towards polyploidization[124] that may be intensified through environmental stress and/or fluctuations as witnessed at several large-scale extinction events[391], hints at a strong role for the neutral component. There are however many observations that also argue in favour of the adaptive component[144]. Although one has to remain cautious with generalizations about the distribution and prevalence of recent polyploids, because many exceptions can be found[125], some trends are apparent. For instance, recent polyploids appear to have larger habitat distributions, suggesting they can tolerate more ecological conditions[164,385,444]. Most strikingly, they are less likely to be endangered and more likely to be invasive on a worldwide scale compared to diploids[143]. Such observations would be difficult to explain purely through neutral mechanisms.

The genetic component of unreduced gamete production merits some more attention. Traditional breeding studies established that diploid gamete production is a highly heritable trait that can be enhanced in as few as two to three cycles of recurrent selection in species such as alfalfa[121] and red clover[120]. In *Arabidopsis*, a surprisingly strong tolerance of gametes to both trisomy and several other complex karyotypes exists[445], while several genetic players that can influence unreduced gamete production through their effect on the orientation of the spindle apparatus in male meiosis have recently been identified[119], such as *AFH14*[446], *JAS*[389], and *AtPS1*[447]. Stress-induced altered functionality of these genetic components may explain the effect of the environment on unreduced gamete production[387]. These observations open up the possibility that polyploidization might even constitute an inducible evolutionary mechanism by which plants cope with ecological disasters, much akin to the stress-inducible mutator systems such as the SOS response in bacteria[448]. The latter is a transient response to stress and changing environments by means of a set of 'evolution genes' that decrease replication fidelity and increase mutation rates to generate genetic diversity upon which natural selection can act[449,450]. Such evolution genes are thought to undergo biological evolution themselves through indirect selection, and their presence in higher organisms has been hypothesized[451]. Since all extant angiosperms shared at least two rounds of WGD[136], with an extra shared WGD at the base of the core eudicots[137] and possibly also the monocots[140], recurring WGD events[52,128,135] could have maintained residual heritable genetic variation in diploid plants for the ability to produce unreduced gametes and form polyploids in times of ecological upheavel. Despite a genetic component, this does not need to be necessarily under the direct control of any adaptive program, as it could just as well primarily be an 'evolutionary spandrel' that received secondary functionality[20]. In any case, such a system could provide an alternative for

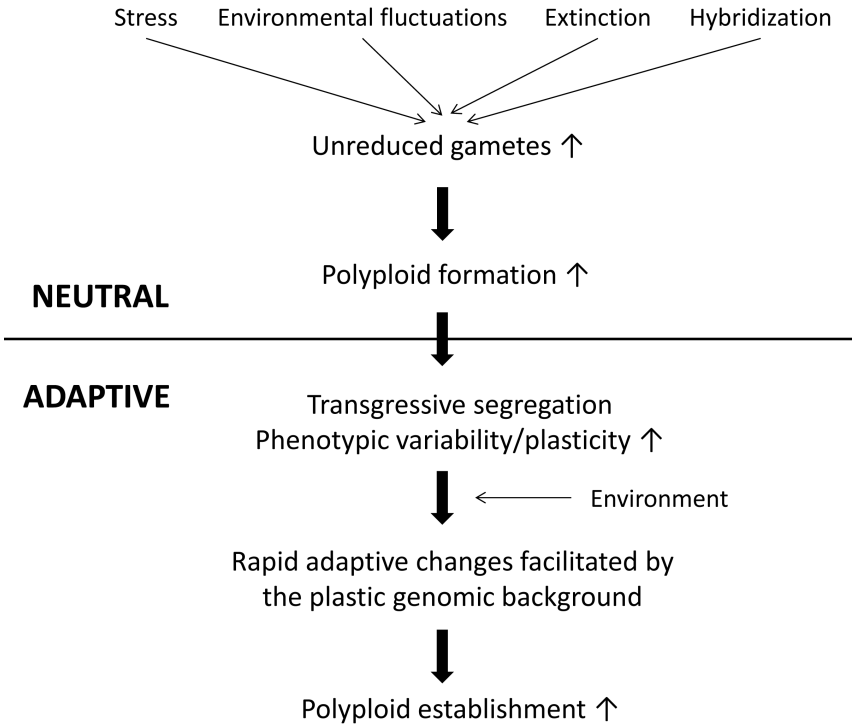


Figure 5.4: Both neutral and adaptive processes probably contribute to enhanced polyploid establishment under environmental stress and/or fluctuations. The latter likely increase the formation of unreduced gametes, while other processes such as hybridization and extinction of the background diploid population can also contribute to an overall increase of unreduced gametes to the total gamete pool. This is expected to lead to more polyploids being formed even in absence of any active adaptive advantage. Transgressive segregation and genomic instability of polyploids on the other hand may lead to heterotic phenotypes, increased phenotypic variability, and plasticity that, if beneficial under the changing environment, can be rapidly selected for, which is expected to lead to more polyploids being established even in the absence of increased polyploid formation. Note that irrespective of which scenario carries more weight, the environment plays an important role in polyploid establishment.

the mutator systems in bacteria, which would be less efficient in plants due to their smaller effective population sizes and longer life cycles, but remains currently however entirely hypothetical.

5.5 Enhanced polyploid establishment at the K-Pg boundary may have paved the way for angiosperm success in the Cenozoic

The neutral and adaptive processes described above offer a framework for the apparent clustering of WGDs at the K-Pg boundary, but fail to explain their long-term success in terms of speciation and biological novelty. For all examples we considered, it was apparent that the duplication of the whole genome provided an increase in raw genetic material on which evolution could work. In accordance with Ohno's classical models[93,94], the newly created gene copies could undergo neofunctionalization (the creation of a new function), subfunctionalization (the division of an ancestral function or functions over the daughter copies), or be kept for dosage amplification (the production of more of a beneficial gene product), or any combination thereof as explained by more complex population genetic models[98]. Although the fate of most duplicated genes is in fact loss through pseudogenization[227], WGDs provide a massive number of contemporarily created gene duplicates, of which only a small fraction seems to have contributed to some major biological innovations and/or elaborations.

It has become increasingly clear that rather than just the functional divergence of the coding regions and/or regulatory sequences of individual genes, especially the rewiring of the regulatory network containing these individual components following WGD is of major importance[162,452]. A body of literature exists demonstrating that particularly regulatory and developmental genes are retained in excess after WGDs. This is most likely due to dosage-balance constraints, i.e., selection against loss of individual components of completely duplicated macromolecular complexes and/or pathways because this would disrupt their overall stoichiometry[160,178,284,453,454]. Retention of balance-sensitive duplicates thus does not provide an immediate evolutionary advantage, but results from the fact that their loss would lead to an immediate disadvantage. In this respect, the retained regulators may be considered an evolutionary spandrel[20,160], which might later on have facilitated the evolutionary innovations and/or diversifications observed in many of these post-WGD lineages[52,128,161]. Selection to maintain dosage balance eventually relaxes over time allowing functional divergence in the context of the environment[453,455] so that part of the duplicated network can be rewired to execute novel functions[162]. However, the underlying mechanisms are currently unclear. Gene duplication has been shown to contribute to innovations even after prolonged periods between the original duplication event and the origin of novelty[253], suggesting that individual components of these duplicated networks can undergo neo- and subfunctionalization in accordance with Ohno's classical models[93,98] even long after the duplication event itself. Some of these processes could have caused network-rewiring events that could help explain the vast post-WGD success observed in some of the plant families that experienced a WGD at the K-Pg boundary.

There are many examples that support the role of network rewiring over time. The ability for anaerobic fermentation in yeast has been associated with global rewiring of its transcriptional network after genome duplication, involving changes in the promoter regions of several genes such as the loss of specific regulatory motifs[320,456]. Similarly, the abundance of teleost fish pigmentation synthesis pathways has been attributed to the teleost WGD through rewiring in combination with subfunctionalization of existing pathways[457]. In plants, the *gamma* hexaploidy at the base of the core eudicots resulted in expansion of MADS-box gene families, key regulators of reproductive development, which through rewiring of their interaction network in combination with neo- and subfunctionalization, acquired roles in several major plant developmental processes[139,458].

5.6 Conclusions

Advances in plant genomics, molecular sequence divergence estimation and high-performance computational solutions, allow us to address questions about the role of genome duplication that were previously impossible to investigate. It should be emphasized that the fate of most newly formed polyploids appears an evolutionary dead end through outcompetition by their diploid specialized progenitors[152,153,158], because of a whole range of associated negative effects such as minority cytotype exclusion[122], severe meiotic and mitotic abnormalities[156], and ploidy-associated genomic instability[157]. Nevertheless, it appears that there exists a strong link between environmental stress and/or fluctuation and genome duplication, as currently supported for both present-day polyploids and paleopolyploids at the K-Pg boundary. Could unreduced plant gamete production have increased polyploid formation at the K-Pg

boundary? Alternatively, can the apparent prevalence of polyploids at the K-Pg boundary be explained by their increased adaptability? Or do we observe the signature of another mechanism and/or pattern that currently remains elusive, perhaps because both dating of such ancient events and making generalizations about current polyploids remain particularly problematic? In any case, this polyploid heritage may afterwards have fuelled evolution of biological innovations and speciation in the context of newly encountered conditions during the Cenozoic through extensive network rewiring and functional diversification of regulatory and developmental genes that were originally guarded against loss through mechanistic dosage-balance constraints. Polyploids in some sense thus seem reminiscent of the 'hopeful monsters' advocated by Richard Goldschmidt[42] (M. Freeling, personal communication), at least at the genomic level, while their full potential at the phenotypic level can only be realized given time and the right conditions[163]. It thus appears that especially the role of the environment in both polyploid establishment and their evolutionary success constitutes an important aspect that merits further investigation.

5.7 Acknowledgements

This work was supported by Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"). The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 322739 - DOUBLE-UP. Part of this work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Hercules Foundation, and the Flemish Government Department EWI. The authors would also like to acknowledge the many fruitful discussions with scientists working on polyploidy at the Plant Genome Evolution (Amsterdam, September 2013) meeting. K. Vanneste and S. Maere are fellows of the Fund for Scientific Research Flanders (FWO).

5.8 Author contributions

I composed the evolutionary framework described in this chapter and wrote the resulting review article. The evolutionary framework built very heavily on previous work and ideas from Steven Maere and Yves Van de Peer and was developed through extensive discussion with the both of them. They also both contributed significantly towards writing the resulting article.

Chapter 6

Conclusion and future perspectives

“When we try to pick out anything by itself,
we find it hitched to everything else in the Universe.”

John Muir (Scottish-American naturalist),

“My First Summer in the Sierra”

For the author contributions, see page 122.

6.1 Gene duplicates don't care about our attempts for categorization

Dissection of the duplication history of the *MALS* genes in yeast allowed to assess both pre-duplication functionality, and the contribution of neutral and adaptive processes that drove post-duplication divergence. Both are important deterministic features amongst the theoretical models that explain the evolutionary fates of genes (see 1.4). In particular, the evolutionary history of the *MALS* gene family discussed in chapter 2 indicates that all three classical models of duplicate gene evolution as proposed by Ohno contributed towards evolutionary innovation and/or diversification[93]. These include gene conservation (e.g., *MAL12* and *MAL32*, which both need to be retained despite their similar function and sequence for optimal fitness), subfunctionalization (e.g., the distribution of isomaltose- and maltose-like functionality over both daughter paralogs), and neofunctionalization (e.g., the remarkable increase in isomaltose-like functionality of the anc*IMA1-4* clade compared to the largely maltose-like functionality of the pre-duplication ancestor).

Despite the technical hurdles to overcome in creating ancestral genes and detecting positive selection, we found that especially the EAC model was able to explain the overall divergence of the *MALS* gene family because it conformed largely to the main predictions thereof: the ancestral gene demonstrated promiscuous activity for a minor secondary function that could not be optimized within the same locus, but for which gene duplication most likely allowed to resolve this adaptive conflict by episodic positive selection on specific residues in both post-duplication paralogs.

Nevertheless, the *MALS* gene family illustrates that the three basic trajectories for gene duplication cannot be separated easily. Rather than strictly following a certain scenario, we found a particularly dynamic and complex interplay between the different outlined fates after duplication. Gene conservation seems especially important for initial duplicate retention, after which a combination of both positive selection and neutral genetic drift led to the long-term divergence of post-duplication paralogs that demonstrated aspects of both sub- and neofunctionalization. Despite the EAC model being a good candidate to concisely describe their overall evolutionary trajectory, the *MALS* gene family demonstrates that a strict classification into one of the many detailed theoretical models is particularly difficult. Rather, it may prove more useful to distil a more general picture of duplicate evolution across a gene family. Additionally, it may be worthwhile to put the production of new theoretical models on hold for a while, at least until experimental studies have had a chance to catch up with the plethora of models that currently exist.

6.2 Neither do genome duplications

Two long-standing viewpoints regard WGD either as a road towards evolutionary success[154], or as an evolutionary dead end[152,153], a dichotomy that permeates many of the discussions about the evolutionary significance of WGD to this day[287]. As mentioned in the introduction of this dissertation, this does however seem outdated. Recently formed polyploids experience a large array of chromosomal abnormalities that lead to irregularities during cell division, resulting in phenotypes that are often less fit and fertile[157]. Even particularly stable neopolyploids need to overcome the minority cytotype disadvantage, which may

prevent them from becoming successfully established[155]. There is a strong discrepancy between the low number of known successful paleopolyploidizations and the very large number of described neopolyploids. Despite the need for more systematic evaluation of both paleo- and neopolyploid abundance, this stark contrast entails that the vast majority of these neopolyploids will not stand the test of time[151]. These factors render the strict categorization of WGD as a road towards evolutionary success hardly justifiable.

On the other hand, several former tentatively described paleopolyploidizations have become well established. When the hidden duplication past of *Arabidopsis thaliana* was first described[316], the latter being chosen as a model species for plant biology partly because of its compact genome, it was difficult to imagine that such a small genome could harbour any WGD at all. It is now known that this small genome contains the remnants of at least five WGDs during its evolutionary past[136]. Several other successful paleopolyploidizations are now also well established, especially in the plant lineage, but also in other complex eukaryotic lineages[52,128], and render the strict categorization of WGD as an evolutionary dead end equally unjustifiable. The former lack of appreciation for paleopolyploid abundance was probably to a large extent due to the fact that advanced computational approaches are required to detect ancient genome duplications that underwent diploidization and extensive fractionation[167]. There has been a continuous effort in the development of more powerful tools, such as collinearity-based methods[169], tree-based methods[136], and paranome age distributions[178], rendering them more apt to detect increasingly older paleopolyploidizations. Especially paranome age distributions are a popular tool for WGD inference, but detection of very old paleopolyploidizations is plagued by the confounding effects of both K_S stochasticity and saturation. In chapter 3, we investigated their impact on age distributions in more detail. In particular, by performing artificial evolutionary simulations that evolve real protein-coding genes while accounting for species-specific genome characteristics, we were able to quantify K_S stochasticity and saturation in empirical sequence data. Incorporation of these effects in predefined age distributions demonstrated that their tails contain a diffuse SSD saturation peak. Separation of real WGD peaks from the SSD saturation peak seems therefore particularly troublesome, for which current standard-practice mixture modelling techniques cannot account properly. Rather, quantitative modelling approaches that separate the contribution of both the SSD and WGD mode of duplication in the tail of paranome age distributions will be required. In this regard, we are currently testing the population dynamics model introduced by Maere et al.[178], by using a simulated annealing approach that optimizes model parameters to empirical age distributions while accounting for species-specific K_S stochasticity and saturation.

Such an effort seems especially valuable in the context of the continuous genome sequencing by the broad scientific community, where many labs will soon be able to afford low-coverage sequencing of their 'pet genome'[337]. Low-coverage sequencing entails positional information required for collinearity-based methods remains problematic, while the computational resources required for applying tree-based methods on a large scale also render them prohibitive, making age distributions an ideal exploratory tool for WGD inference. The increase in genome sequence data will of course most likely not be paired with a similar increase in paleopolyploid discovery. For instance, although the total number of considered plant species tripled between the study of Fawcett et al. in 2009[193] and the one in this dissertation, the total number of (dated) paleopolyploidizations 'only' doubled. Additionally, many of the WGDs that are shared by large phylogenetic clades are slowly becoming known because many of them will soon have at least

one representative species sequenced. Nevertheless, there is ample room for discovery of more 'recent' lineage-specific paleopolyploidizations in several angiosperm plant families. For instance, a series of recently sequenced genomes that were published in 2013 such as sacred lotus[459] and bamboo[460], or that are under active development such as eucalyptus and orchid (unpublished data), all found evidence for previously undescribed lineage-specific paleopolyploidizations. Other plant lineages, such as ferns where polyploidization is known to be a frequent phenomenon[142], remain almost completely unexplored at the moment, suggesting that the vast expense of sequence data that is coming our way will reveal many more hidden successful paleopolyploidizations in the plant lineage.

6.3 Because not all answers can be found in their genome itself

Perhaps the dichotomy described above is still widely used for the simple reason a suitable alternative is mostly lacking. One hypothesis is that polyploid abundance simply increases over time through a ratcheting process, wherein a very small fraction of polyploids become established constantly over time[325]. The latter explanation is however unsatisfactory because it does not explain what factors are responsible for that very small fraction becoming established. In chapter 4, we described evidence that a substantial fraction of known paleopolyploidizations cluster statistically significantly in time in association with the K-Pg boundary. Although these results will need to be updated in light of newly discovered paleopolyploidizations and increasingly powerful methods for sequence divergence estimation in order to evaluate whether this pattern stands the test of time (see also further), at the moment, our results support the association between paleopolyploidization and the K-Pg boundary as first tentatively suggested by Fawcett et al.[193]. The K-Pg mass extinction was a culmination of different factors such as greenhouse warming, volcanic activity, and a bolide impact, for which all available evidence indicates that it constituted a very drastic event that affected all life on earth[80,81]. Strikingly, the association between stress and/or extinction and polyploid establishment also appears valid for neopolyploidizations. Neopolyploids are traditionally considered as colonizing invasive species because they possess a broad ecological tolerance[149,164]. They are for instance especially prevalent in stressful environments such as the Arctic[145]. Although proper precautions need to be taken when interpreting trends in neopolyploids because only a very small fraction has been properly assessed[125], more recent and larger-scale studies also support that neopolyploids are more often invasive species compared to diploids on a world-wide scale[143].

There appears thus a strong link between polyploid establishment and stress for both paleo- and neopolyploidizations. In chapter 5, we incorporated this into an evolutionary framework that has the potential to mitigate the stark contrast in the proposed evolutionary fates of polyploids by explicitly accounting for the effect of environmental stress on polyploid establishment by both neutral and active processes. It is known that environmental stress often increases the formation of unreduced gametes. This is well described for neopolyploids[114], and the underlying molecular components and processes responsible for this are being unravelled[119]. Similar evidence for paleopolyploidizations is necessarily more anecdotal, but increased unreduced and aberrant fossil pollen has been observed during previous mass extinction events[391,392], hinting at least at the possibility thereof. Even if increased unreduced gamete production under stress is a relatively novel trait in plants that was not at play during the K-Pg

mass extinction, the latter event also severely impacted diploid progenitor populations of whom many went extinct[82], increasing the relative contribution of unreduced gametes to the overall gamete pool. This purely neutral mechanism thus explains how stress and/or extinction ameliorate the severe minority cytotype disadvantage that recently formed polyploids have to cope with. On the other hand, the genomic instability and phenotypic variability that is frequently displayed by neopolyploids[144], can be an important adaptive advantage in stressful and perturbed environments that allows them to react more quickly to newly created vacant niches by exploiting their potential as invasive colonizing species[384]. Convincingly demonstrating that paleopolyploids formed around the K-Pg boundary had a higher adaptive potential will most likely remain impossible forever, but at least there is little doubt that completely new and vacant niches were being created on a massive scale, which would be more easy to cope with for colonizing species with a broad ecological tolerance (whether those were diploid or polyploid).

As illustrated in figure 6.1, stressful environments and extinction may thus alleviate the minority cytotype disadvantage, and increase the chance that the otherwise typically unstable polyploid phenotypes become advantageous, increasing polyploid establishment. Afterwards, mechanisms such as RGL or lineage-specific subfunctionalization might tentatively explain why post-WGD clades often experience enhanced speciation rates[128]. Importantly, a large set of developmental and regulatory genes seem guarded against loss after WGD through mechanistic dosage-balance constraints on the stoichiometry of completely duplicated pathways and/or macromolecular complexes[161], which might provide plants with a polyploid heritage a toolbox that allows them to react more adequately to newly encountered ecological opportunities and/or challenges through extensive sub- and neofunctionalization of individual components after resolution of dosage-balance constraints[163]. The resulting extensive network rewiring coupled with increased speciation rates could thus explain the increased species diversity and/or complexity observed in many post-WGD clades.

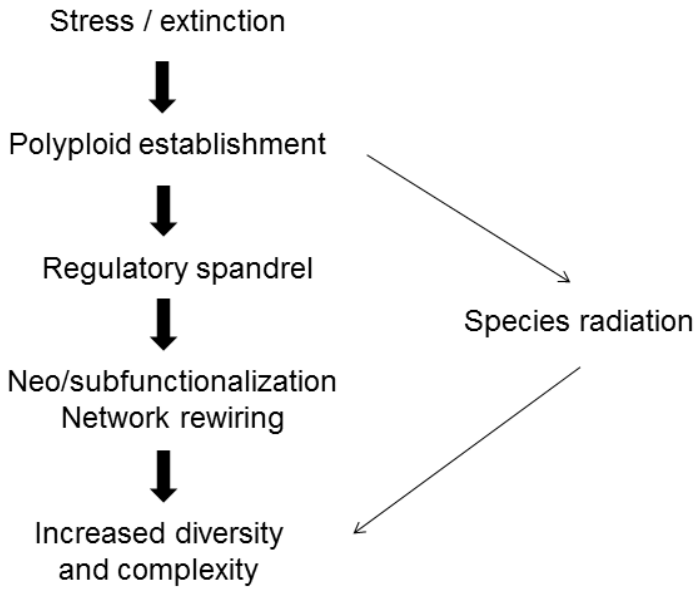


Figure 6.1: Updated view on polyploid succes. Stress and extinction increase polyploid establishment, after which increased species radiation and network rewiring through neo- and subfunctionalization of regulatory and developmental genes, which were retained through dosage-balance constrains and hence form an evolutionary spandrel, can lead to increased species diversity and/or complexity.

Polyploidy in some sense could thus be viewed as a saltational event through the creation of hopeful monsters. WGD in any case seems reminiscent of the systematic mutations that affect the whole genome, as described by Goldschmidt, but whether they truly adhere to the definition of being hopeful monsters that lead to the saltational origin of a completely different clade is open for discussion. Allopolyploids for instance often display large phenotypic differences compared with their diploid progenitors, suggesting that they can give rise to a relatively drastically different evolutionary lineage, once successfully established. Autopolyploids on the other hand often display similar phenotypes to their diploid progenitors, hinting that they give rise to relatively similar evolutionary lineages, if successfully established. In either case, their polyploid heritage may allow them increased diversity and/or complexity, but sometimes only long after the initial WGD[324]. Polyploids could therefore perhaps be considered as hopeful monsters at the genomic level, while their full potential at the phenotypic level can only be realized given time and/or the right conditions. The importance of the proper time and place for realizing evolutionary potential is not a particularly ground-breaking insight as it goes straight back to the modern synthesis[15], but has perhaps been too absent in the genomics era of polyploidy research.

6.4 So cherish the past

As stated before, our results will need to be updated in light of newly discovered paleopolyploidizations and increasingly powerful methods for sequence divergence estimation in order to evaluate whether the putative framework depicted in figure 6.1 will stand the test of time. Our current temporal framework for paleopolyploidizations (see figure 4.3) indicates that there are also quite some paleopolyploidizations that are not found in association with the K-Pg boundary. Because the total number of dated paleopolyploidizations remains fairly limited, only the clustering of their majority with the K-Pg boundary could be verified, most likely because this event was so drastic that it left a sufficiently large signature we were able to pick up with our current data and methods. Figure 6.1 however also makes the prediction that many of the WGDs not found in association with the K-Pg boundary, can be linked to other lesser periods of stress and extinction. Figure 6.2 depicts a background profile of extinction intensity in the Phanerozoic eon for marine genera, which through their ease of fossilization serve as a good proxy for the overall extinction intensity, including terrestrial genera[78]. Not surprisingly, extinction intensity is closely linked to several major and minor extinction events that are mostly driven by changing geographical and climatic factors[79]. Figure 6.1 tentatively suggests that periods of lesser and greater extinction should similarly have resulted in less and more pronounced periods of increased polyploid establishment, respectively, so that a broader sampling of dated paleopolyploidizations would exhibit a similar trend as the extinction profile depicted in figure 6.2.

There is some limited anecdotal evidence that could be interpreted as in support of this prediction. As mentioned earlier, increased levels of unreduced fossil pollen were found in the now extinct conifer family Cheirolepidiaceae at the large-scale Triassic-Jurassic extinction event 201.3 mya[391], while the angiosperm-shared WGD was dated putatively at ~192 mya[136], not very far from this boundary. Clearly, evidence for increased fossil pollen production from one conifer family in combination with a vague estimate for the angiosperm-shared WGD offers little in terms of definite proof, but at least opens up

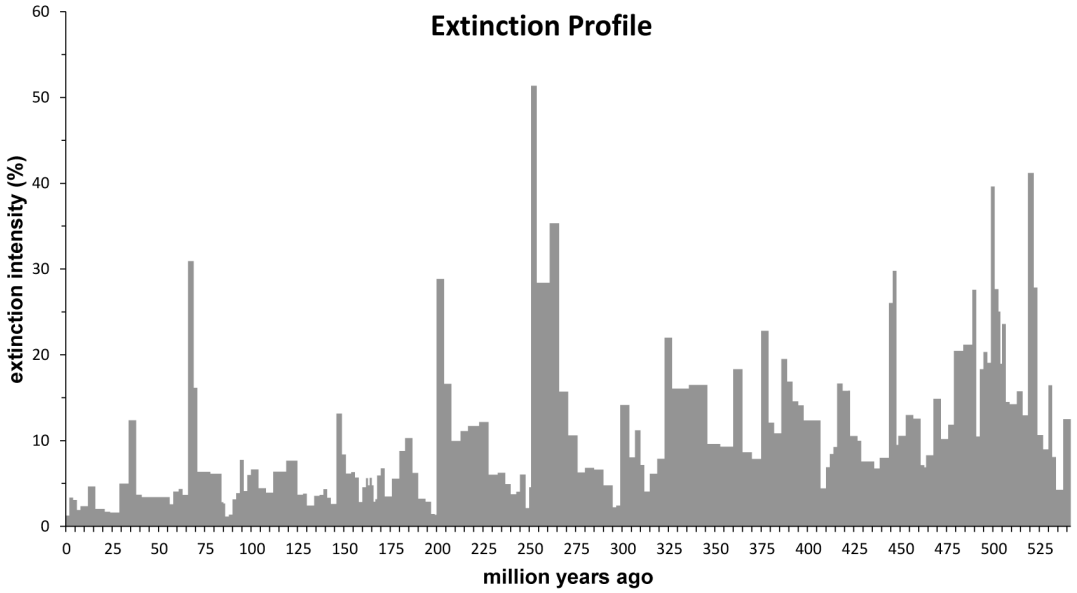


Figure 6.2: Profile of extinction intensity during the Phanerozoic. Extinction intensity of marine genera, serving as a proxy for all life on earth, throughout the Phanerozoic eon. Figure based on publicly available data by Rohde et al.[78].

the possibility that polyploid establishment might also have been increased during the fourth of the five major mass extinction events. Similarly, increased levels of aberrant gymnosperm pollen[392] and lycophyte spores[393] have been found at the Permian-Triassic boundary 252.2 mya, the third largest extinction event. The eudicot shared hexaploidy event was dated around ~117 mya[138], close to the early Aptian extinction event ~120 mya, which constitutes one of the many lesser extinction events throughout the Phanerozoic[139].

To go beyond such vague and questionable anecdotal evidence, it would therefore be very interesting to provide an updated temporal framework for paleopolyploid abundance in a few years, once a sufficiently large amount of new plant sequence data are available. However, rather than just more sequence data, some methodological and technical hurdles will also need to be surpassed. Despite the fact that uncorrelated lognormal relaxed clocks should be better equipped to deal with rate shifts in comparison with autocorrelated relaxed clocks, concern has been raised that they still might not be able to cope properly with drastic rate shifts[65], which was also evident in our study (see 4.3.3). Improper correction for such drastic rate shifts could be especially problematic, because not even exhaustive sequence data will be able to fill in all phylogenetic gaps. For instance, many of the discussions about the exact age of the angiosperm stem (see 1.3.2) arise because there are no intermediate extant representatives between the gymnosperms and first angiosperms, so that this single long branch supporting all angiosperms experienced multiple rate shifts that are very difficult to account for in absence of proper sequence information that can break up their overall contribution to the total branch[61]. Research in this area is however actively progressing with for instance the development of uncorrelated inverse Gaussian relaxed clocks, which theoretically should handle such drastic rate shifts better[461], although the latter still needs to be evaluated more thoroughly. The technical hurdles in scaling up such a dating analysis will most likely also be substantial. Our current study relied on state-of-the art high-performance computational infrastructure, but still required several months to run to full completion, indicating further technical developments will be equally important. The availability of computational resources was hence a bottleneck in our current work.

For instance, Baysian model testing methods that allow comparison of different types of relaxed clocks exist, but applying them proved infeasible in terms of the required computing time. Similarly, dating of individual orthogroups could have benefited by running the MCMC for a longer time, for instance 100 million generations instead of 10 million, to ensure better convergence of orthogroups that currently had to be discarded because the ESS for all their statistics was not equal or higher than 200. Multiple independent repeats for each orthogroup instead of one single run could have helped to ensure that replicates for the same orthogroup converge on the same solution to boost confidence in their age estimates. There is however also progress in this regard, with for instance libraries such as BEAGLE[347] that speed up the MCMC component of the sequence divergence estimation, while especially speed gains on HPC systems that contain GPUs (Graphics Processing Units) are promising. Such systems are now still rare and expensive but expected to become common standard over the next few years. Lastly, new methodologies for evaluating the association between paleopolyploidizations and extinction intensity will also be required. Figure 6.2 illustrates extinction is an on-going process in evolution where periods of more and less intensive extinction alternate frequently. Associating individual paleopolyploidizations with particular lesser extinction waves may be tempting, but the inherent uncertainty involved in dating any paleopolyploidization, especially very old ones, makes this a very undesirable approach. Any randomly picked WGD age estimate can most likely always be associated with some lesser extinction wave that is situated in close proximity. Rather, it may be more worthwhile to focus on a relatively young period, for instance the last 66 million years since the K-Pg event, and devise a sliding-window approach that systemically evaluates every age bin for overrepresentation of paleopolyploid abundance.

6.5 But look forward to the future

The vast number of current neopolyploids provides a similar opportunity for evaluation, as they are predicted to show an association with habitats that are characterized by extensive environmental stress and/or fluctuations, independent from the consideration whether they will eventually stand the test of time or not. Confirming this link would not require any advanced computational approaches and a vast array of sequence data as for paleopolyploidizations, but rather just good ‘old-fashioned’ large-scale and systematic neopolyploid sampling. On the one hand, the latter could be considered as providing more accurate validation, since it is not plagued by the inherent uncertainties associated with sequence divergence estimation of ancient events. On the other hand, the sheer number of both diploid and polyploid plant species makes this a very difficult exercise, as substantial labour-intensive efforts will be required to avoid the confounding effects of sampling biases. Furthermore, even if a particularly strong and unambiguous association could eventually be demonstrated, it is well known that correlation does not necessarily imply causation.

Neopolyploidizations however offer the advantage that they are contemporary events and therefore can be subject to manipulation. A direct experimental approach could hence be especially rewarding. For instance, it is well described that subjecting plants to stressors increases their unreduced gamete formation[387]. However, although increased unreduced gamete formation in combination with background extinction has been theoretically demonstrated through mathematical modelling to alleviate the minority

cytotype disadvantage[395], as far as we known, this has never been validated empirically. Similarly, despite the fact that there are many examples of neopolyploids displaying a higher adaptive potential under stressful conditions[144], as far as we know, there are no studies that systematically evaluate whether neopolyploids, once formed, on average have a higher chance of survival under stressful conditions. Testing this in plants is of course also particularly difficult, since it would require a controlled environment wherein all individuals constantly need to be phenotypically quantified. Even very rapidly generating model plant species such as *Arabidopsis thaliana*, require two to three months between generations and substantial infrastructure. Rather, a unicellular model system close to plants that shares many of their genomic characteristics and also their inherent capacity for polyploidization, such as for instance some green algal species[462], seems an interesting alternative through their fast generation times and ease of culture. A direct large-scale experimental approach that follows the formation and fate of both diploid and polyploid individuals of such a model system under a variety of induced stressors could consequently allow direct evaluation of whether stress leads to increased polyploid establishment.

Furthermore, even if it would be convincingly demonstrated that stress and/or extinction lead to increased polyploid establishment through a combination of mitigating the minority cytotype disadvantage and/or a higher adaptive potential, as remarked in chapter 5 (see 5.5), this does not explain their long-term success in terms of increased biological diversity and complexity. There are some indications that WGD may indeed increase speciation rates but conclusive evidence is still lacking[52]. Similarly, the over-retention of regulatory and developmental genes, most likely through dosage-balance constraints, is well substantiated[160,178,284,453,454], and an increasing set of examples are becoming known where network rewiring with components of this toolbox has led to expanded functionality[139,162,163,456,457]. All this evidence is however mostly based on the present-day genomes of paleopolyploids, for which the current snapshots that are sequenced corroborate the model presented in figure 6.1. Nevertheless, these genomes evolved for a period of several millions of years after their WGD event, during which their behaviour remains almost entirely 'black-box'. For instance, the over-retention of developmental and/or regulatory genes after WGD is a phenomenon that has been encountered in almost every sequenced paleopolyploid (plant) genome, and for which a classical genetic concept such as dosage-balance can adequately explain the observed pattern[160]. There is however no conclusive evidence available yet that convincingly proves that such genes are indeed being guarded continuously during the evolution of polyploids, and a few species are in fact known where this pattern of over-retention has not been corroborated[182]. Similarly, the exact mechanisms that drive network rewiring remain obscure. Consequently, this dissertation hopefully demonstrated that environmental factors have good potential to help move forward beyond some of the seemingly conflicting observations of the genomic era, but also that many, the large majority probably, of the evolutionary forces involved in polyploidy still await our discovery.

6.6 Author contributions

The content of this chapter was written by myself. It resulted from the many fruitful discussions with both my promoters and all partners I had the chance to work with during my PhD studies.

Appendices

Appendix A

Summary

The prevalence of both continuous small-scale duplications (SSDs) and whole genome duplications (WGDs) during evolution is well established. Their evolutionary significance is however most certainly not. Especially the fate of WGD remains vividly debated, and depending on research context has been labelled either as an evolutionary dead end or as a road towards evolutionary success. This dissertation presents research that contributes towards the notion that both SSD and WGD have played a major role in the evolution of increasing biological complexity and/or diversity. In particular, our research findings present a framework that focuses on the environmental context for initial successful polyploid establishment, which has the potential to mitigate some of the conflicting statements about the fate of polyploids found in literature.

Concerning SSD, there does exist a general consensus about its importance through the creation of new gene loci, which are largely freed from selectional constraints because the original gene can maintain the original functionality, while the copy is free for evolution to tinker with. Several theoretical models have been developed in the last decennia that explain how the new gene loci can obtain novel or specialized functionality through different underlying molecular mechanisms. There is however a sharp contrast between the detailed theoretical predictions and limited experimental evidence found for the outlined trajectories under these different models. We studied a family of fungal glucosidases that expanded through repeated SSDs by resurrecting their ancestral enzymes. Through a combination of structural analysis, activity measurements, and extensive computational molecular evolutionary analysis, we provided a detailed picture of the molecular mechanisms that drove the divergence of these duplicated enzymes. In particular, we found that the expansion of this gene family did not follow one strict model, but rather exhibited a dynamic and complex interplay between different mechanisms such as dosage amplification (i.e., the creation of more of a beneficial gene product), subfunctionalization (i.e., the division of ancestral functionality over the two daughter copies), and neofunctionalization (i.e., the creation of a new function). Our results thus demonstrate how the basic outlined trajectories for gene duplicates intertwine into a complex evolutionary path that leads to innovation.

Concerning WGD, increasing evidence indicates that WGDs occurred at least once during the evolution of most major lineages, but their precise phylogenetic position and timing often remain obscure. This has major ramifications for the interpretation of their evolutionary significance, since it determines whether their successful establishment was merely a random chance event, or alternatively, a deeper underlying evolutionary principle is at play.

To obtain accurate inference of WGDs, we developed a duplicate population dynamics model that uncovers the contribution of WGDs in empirical duplicate age distributions, where they manifest themselves as peaks against an exponential SSD background. The interpretation of duplicate age distributions is however complicated by the use of K_S, the number of synonymous substitutions per synonymous site, as a proxy for the age of paralogs. The stochastic nature of synonymous substitutions leads to increasing uncertainty in K_S with increasing age since duplication, while the inability of evolutionary models to fully correct for the occurrence of multiple substitutions at the same site leads to K_S saturation. The former erodes the signal of older WGDs, whereas the latter leads to artificial WGD peaks in the distribution. We investigated the consequences of these effects by performing evolutionary simulations of synonymous evolution based on a codon model that incorporate both codon usage and transition/transversion rate bias, and applied the observed K_S stochasticy and saturation effects thereafter onto predefined real age

distributions. We demonstrated that the tail of duplicate age distributions may indeed encompass multiple WGD events and/or K_S artefacts. Hence, our duplicate population dynamics model provides a much more powerful quantitative modelling framework compared to commonly used mixture modelling techniques that can only infer WGDs based on deviations from the background SSD distribution, especially for very old paleopolyplodiziations found in the tail of duplicate age distributions.

To obtain accurate dating of WGDs, we developed a Bayesian absolute dating framework. Taking full advantage of the boost in plant genome sequencing, we could incorporate data from in total 41 species, including 38 full genome sequences and three transcriptome assemblies. This resulted in an extensive coverage across the angiosperm phylogeny, allowing the implementation of several reliable primary fossil calibrations, modelled as flexible lognormal calibration priors that represent the error associated with the age of the fossil in a more intuitive manner. Dating itself was done using the BEAST package, which allows the implementation of an uncorrelated relaxed clock model that assumes a lognormal distribution on evolutionary rates, and therefore should be able to deal with rate shifts between the different branches when large taxonomic distances are considered. Our approach confirmed a previously proposed tentative clustering of paleopolyploidizations with the Cretaceous-Paleogene (K-Pg) boundary ~66 million years ago, supporting increased polyploid establishment around this time.

Our inference and dating results of plant paleopolyploidizations, in combination with recent data on newly formed invasive polyploid plant species, led us to propose that both neutral and adaptive processes probably contributed to the enhanced establishment of polyploids at the K-Pg boundary. Stress and environmental fluctuations likely increase the formation of unreduced gametes, as witnessed both for present-day and even ancient plants at other major extinction events, while other processes such as hybridization and extinction of the background diploid population can also contribute to an overall increase of unreduced gametes to the total gamete pool. This neutral process is expected to lead to more polyploids being formed even in absence of any active adaptive advantage. Transgressive segregation and genomic instability of polyploids on the other hand may lead to heterotic phenotypes, increased phenotypic variability, and plasticity that, if beneficial under the changing environment, can be rapidly selected for, which is expected to lead to more polyploids being established even in the absence of increased polyploid formation. Our framework thus emphasizes the environmental context as having a major influence on initial successful polyploid establishment, and explains why polyploids are sometimes successfully established, despite most often being an evolutionary dead end because of outcompetition by their diploid specialized progenitors through of a whole range of associated negative effects such as minority cytotype exclusion, severe meiotic and mitotic abnormalities, and ploidy-associated genomic instability.

Strikingly, some of the WGDs we dated at the K-Pg boundary are found in plant families that are characterized by particular biological innovations, and/or extensive post-WGD lineage diversifications. Furthermore, genome sequencing of such paleopolyploid species has indicated strong over-retention of genes with developmental and/or regulatory roles after WGD, which can be explained by mechanistic dosage-balance constraints that guard such genes against loss through limitations on the overall stoichiometry of the macromolecular complexes and/or pathways they are part of. Moreover, many of these retained duplicates were later co-opted into existing basic processes to allow novel and expanded functionality through extensive network rewiring. An intriguing hypthesis is therefore that after successful

polyploid establishment promoted by specific environmental contexts, the possession of a double complement of developmental and/or regulatory genes might have facilitated evolution of particular biological innovations and/or diversifications throughout the Cenozoic.

Bijlage B

Samenvatting

De aanwezigheid van zowel continue kleinschalige genduplicaties en volledige genoomduplicaties in evolutie is goed omschreven in de literatuur, maar over hun precieze evolutionaire significantie bestaat daarentegen nog veel discussie. In het bijzonder de uitkomst van polyploïdisatie blijft hevig gedebatteerd, en wordt afhankelijk van de precieze context beschouwd als ofwel een evolutionair dood einde, ofwel een weg naar evolutionair succes. Dit proefschrift presenteert onderzoek dat bijdraagt aan de notie dat zowel kleinschalige gen- als volledige genoomduplicaties een belangrijke rol in de evolutie van toenemende biologische complexiteit en/of diversiteit gespeeld hebben. Onze onderzoeksresultaten schetsen in het bijzonder een referentiekader dat focust op het belang van de omgeving in de initiële succesvolle vestiging van polyploïde species, dat het potentieel heeft om sommige van de conflicterende uitspraken over het lot van polyploïdisatie in de literatuur te verklaren.

Met betrekking tot kleinschalige genduplicaties bestaat er in feite een algemene consensus over hun evolutionair belang omdat ze nieuwe genen creëren die grotendeels vrij zijn van selectie, zodat het originele gen zijn oorspronkelijke functie kan behouden terwijl de kopij nieuwe functionaliteit kan verwerven. Er zijn dan ook verschillende theoretische modellen ontwikkeld gedurende de laatste decennia die beschrijven hoe de nieuwe kopij evolueert aan de hand van allerlei onderliggende moleculaire mechanismen. Er is echter een sterk contrast tussen de gedetailleerde theoretische voorspellingen en het gelimiteerde eigenlijke experimentele bewijs dat bestaat voor deze verschillende modellen. Wij onderzochten een genfamilie van gist glucosidases die voornamelijk geëxpandeerd zijn door herhaaldelijke kleinschalige genduplicaties. Door een combinatie van structurele analyse, activiteitsmetingen, en een uitgebreide moleculaire evolutionaire analyse, hebben we een gedetailleerd beeld kunnen schetsen van de moleculaire mechanismen die een rol speelden bij de expansie van deze genfamilie. In het bijzonder vonden we dat geen enkel welbepaald model gevolgd wordt, maar eerder een dynamische en complexe wisselwerking tussen verschillende mechanismen zoals doserings-amplificatie (de creatie van meer van een voordelig genproduct), sub-functionalisatie (de verdeling van de ancestrale functionaliteit over de twee dochter kopijen), en neo-functionalisatie (de creatie van een nieuwe functie). Onze resultaten demonstreren dus dat de verschillende (basis) modellen die het lot van gedupliceerde genen beschrijven in feite ineenstrengelen tot een complex evolutionair pad dat tot innovatie leidt.

Met betrekking tot genoomduplicaties is er toenemend bewijs dat ze minstens eenmalig in de evolutie van de belangrijkste fylogenetische lijnen plaatsgevonden hebben. Hun precieze fylogenetische locatie en tijdstip in evolutie blijven daarentegen vaak onduidelijk. Deze hebben echter belangrijke gevolgen voor de interpretatie van hun evolutionaire significantie, omdat ze bepalen of de succesvolle polyploïdisaties te wijten zijn aan random factoren, of alternatief een meer diepgaand overkoepelend evolutionair principe aanwezig is.

Om genoomduplicaties accuraat te kunnen identificeren, hebben we een populatiedynamiek model ontwikkeld dat de contributie van genoomduplicaties in empirische leeftijds-distributies beschrijft, waar ze zichzelf manifesteren als pieken tegen een exponentiële achtergrond van kleinschalige duplicaties. De interpretatie van dergelijke distributies wordt echter gecompliceerd door het gebruik van K_S, het aantal synonieme substituties per synonieme site, als een benadering voor de leeftijd van de paralogen in de distributie. De stochasticiteit waarmee synonieme substituties plaatsvinden leidt tot toenemende onzekerheid in de K_S schatting met toenemende leeftijd sinds de duplicatie, terwijl het onvermogen van evolutionaire modellen om volledig te corrigeren voor meerdere substituties op dezelfde site leidt tot

K_S saturatie. Het eerste erodeert het signaal van oudere genoomduplicaties, terwijl het tweede leidt tot artificiële pieken in de distributie. We onderzochten de gevolgen van deze effecten door evolutionaire simulaties uit te voeren gebaseerd op een codon model dat zowel bias in het gebruik van codons, als bias in het aantal transities ten opzichte van transversies, kan incorporeren. Door vervolgens de waargenomen effecten van K_S stochasticiteit en saturatie toe te passen op voor-gedefinieerde leeftijds-distributies konden we demonstreren dat hun staart inderdaad meerdere genoomduplicaties en/of K_S artefacten kan bevatten. Incorporatie van deze artefacten in ons populatiedynamiek model laat dus een kwantitatieve dissectie toe, in het bijzonder voor de paleo-polyploïdisaties die zich in de staart bevinden, en is bijgevolg veel krachtiger dan de standaard gebruikte technieken die een combinatie van meerdere normale distributies op de gehele distributie proberen toe te passen op basis van afwijkingen in het oppervlak van de curve in de leeftijds-distributie.

Om genoomduplicaties accuraat te kunnen dateren hebben we een Bayesiaans absolute daterings platform ontworpen. Met behulp van de vooruitgang in sequenering, konden we in totaal 41 planten incorporeren, waarvan 38 volledige genomen en drie transcriptomen. Dit resulteerde in een uitgebreide dekking over de gehele angiosperm fylogenie, wat op zijn beurt de implementatie van verschillende betrouwbare primaire fossiele kalibraties toeliet, die gemodelleerd werden als flexibele lognormale priors die toelaten de onzekerheid geassocieerd met de leeftijd van het fossiel intuïtief te beschrijven. De dateringen zelf gebeurden met behulp van de BEAST software. Deze laat de implementatie toe van een niet-gecorreleerd gerelaxeerd klok model, dat ervan uitgaat dat er een lognormale distributie op de evolutionaire snelheden tussen de verschillende takken zit, en daarom beter zou moeten kunnen omgaan met drastische verschillen in evolutionaire snelheden te wijten aan grote taxonomische afstanden. Deze aanpak bevestigde een voordien voorgestelde clustering van paleo-polyploïdisaties rond de Krijt-Paleogeen (K-Pg) overgang 66 miljoen jaar geleden, wat sterke indicaties geeft voor een toegenomen vestiging van polyploïde planten rond deze tijd.

Onze resultaten over de identificatie en datering van plant paleo-polyploïdisaties, in combinatie met recent beschikbare gegevens over nieuw gevormde invasieve polyploïde planten, lieten ons toe een referentiekader te schetsen bestaande uit zowel neutrale als adaptieve processen, die de toegenomen vestiging van polyploïde planten rond de K-Pg overgang kan verklaren. Stress en omgevingsfluctuaties leiden waarschijnlijk tot een toegenomen vorming van ongereduceerde gameten, een fenomeen dat wordt waargenomen bij zowel hedendaagse planten die leven in stressvolle gebieden als planten die leefden ten tijde van andere grootschalige massa extincties. Andere processen zoals hybridisatie en extinctie van de diploïde moeder populatie kunnen ook bijdragen aan een toename van de totale proportie van ongereduceerde gameten die beschikbaar zijn. Een dergelijk neutraal proces kan de vorming van polyploïde soorten bevorderen zonder dat enige actieve adaptieve voordelen vereist zijn. De transgressieve segregatie en genomische instabiliteit van recent gevormde polyploïde soorten kan aan de andere kant leiden tot heterotische fenotypes in combinatie met toegenomen fenotypische variabiliteit en plasticiteit. Als dergelijke fenotypes toevallig voordelig zijn onder de stressvolle omgeving, kan er snel voor geselecteerd worden, wat kan leiden tot een hogere succesvolle vestiging van polyploïde soorten zelfs in afwezigheid van de hierboven beschreven neutrale processen. Onze resultaten benadrukken dus de vooraanstaande rol van de omgeving in het initiële stadium van de succesvolle vestiging van polyploïde soorten, en verklaren waarom deze onder bepaalde omstandigheden toch een verhoogde

kans op succes hebben, ondanks het feit dat polyploïdisatie wel degelijk meestal een evolutionair dood einde is. Ze zijn immers onderhevig aan een hele reeks van geassocieerde negatieve effecten zoals minderheids cytotype exclusie, zware meiotische en mitotische abnormaliteiten, en ploïdie-geassocieerde genomische instabiliteit, waardoor ze meestal niet in staat zijn succesvol te competeren met hun diploïde sterk gespecialiseerde moeder populatie.

Het is daarnaast ook opvallend dat sommige van de genoomduplicaties die we rond de K-Pg overgang gedateerd hebben, teruggevonden worden in planten families die gekenmerkt zijn door hun eigen speciale biologische innovaties en/of extensieve diversificaties. Sequenering van dergelijke paleo-polyploïde genomen heeft aangetoond dat er een sterk behoud is van genen met ontwikkelings en/of regulatorische rollen na genoomduplicatie. Dit kan verklaard worden door mechanistische doserings-beperkingen op gehele gedupliceerde macromoleculaire complexen en/of pathways die voorkomen dat individuele genen verloren geraken. Bovendien blijkt dat veel van deze gedupliceerde genen later ingelijfd worden in bestaande processen wat toelaat nieuwe of geëxpandeerde functionaliteit te verkrijgen. Een intrigerende hypothese is dus dat na de initiële vestiging van polyploïde soorten, het bezit van een dubbel complement van ontwikkelings en/of regulatorische genen de evolutie van bepaalde biologische innovaties en/of diversificaties in het Cenozoïcum gefaciliteerd kan hebben.

Appendix C

Academic CV

Personal information

Name	Kevin Vanneste
Adress	Tuinwijklaan 67B, 9000 Gent, Belgium
E-mail	keneste@gmail.com
Phone	+32 478 47 76 10
Date of birth	23/08/1986
Place of birth	Leuven, Belgium

Education

2009-2014 **Doctor of Science, Bioinformatics**
Ghent University / Flanders Institute for Biotechnology - Ghent, Belgium
Fellow of the Flanders Fund for Scientific Research (FWO)

2004-2009 **Master of Science, Biology**
Ghent University - Ghent, Belgium
Specializations: Major Research, Minor Functional Biology
Received Pierre Verkerk award for best master's dissertation
Graduated summa cum laude

1998-2004 **Greek-mathematics**
Sint-Bernardus College - Oudenaarde, Belgium

Publications

*contributed equally

11. *Cai J, *Liu L, *Proost S, ***Vanneste K**, *Tsai W-C, *Liu K-W, Chen L-J, He Y, Xu Q, Bian C, Zheng Z, Sun F, Liu W, Hsiao Y-Y, Pan Z-J, Hsu C-C, Yang Y-P, Hsu Y-C, Chuang Y-C, Xu X, Wang J-Y, Wang J, Xiao X-J, Zhao X-M, Du R, Zhang G-Q, Wang M, Su Y-Y, Xie G-C, Liu C-H, Li L-Q, Van de Peer Y, Luo Y-B, Chen H-H, Huang L-Q, and Liu Z-Y (2014). The genome sequencing of an orchid, *Phalaenopsis equestris*, provides insights into CAM photosynthesis. *Under review*.

10. **Vanneste K**, Baele G, Maere S, and Van de Peer Y (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications at the Cretaceous-Paleogene boundary. *Genome Research*. *Accepted, pending revisions*.

9. Myburg A, Grattapaglia D, Tuskan G, Hellsten U, Hayes R, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein D, Dubchak I, Poliakov A, Mizrachi E, Kullan A, van Jaarsveld I, Hussey S, Pinard D, Silva-Junior O, Togawa R, Pappas M, Faria D, Sansaloni C, Petroli C, Yang X, Ranjan P, Tschaplinski T, Ye C-Y, Li T, Sterck L, **Vanneste K**, Murat F, Soler M, San Clemente H, Saidi N, Cassan-Wang H, Dunand C, Hefer C, Bornberg-Bauer E, Kersting A, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd A, Liston A, Spatafora J, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim C, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel S, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones R, Steane D, Vaillancourt R,

Potts B, Barry K, Pappas Jr G, Strauss S, Jaiswal P, Grima-Pettenati J, Salse J, Van de Peer Y, Rokhsar D, and Schmutz J (2014). Genome sequence of *Eucalyptus grandis*: A global tree crop for fiber and energy. *Nature*. *In press*.

8. **Vanneste K**, Maere S, and Van de Peer Y (2014). Tangled up in two: A burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*. *In press*.

7. *Gadeyne A, *Sanchez-Rodriguez C, Vanneste S, Di Rubbo S, Zauber H, **Vanneste K**, Van Leene J, De Winne N, Eeckhout D, Persiau G, Van De Slijke E, Vercruysse L, Adamowski M, Ehrlich M, Schweighofer A, Bednarek S, Ketelaar T, Maere S, Friml J, Gevaert K, Witters E, Russinova E, Persson S, De Jaeger G, and Van Damme D (2014). The TPLATE adaptor complex drives clathrin-mediated endocytosis in plants. *Cell* **156**(4):691-704.

6. Nystedt B, Street NR, Zuccolo A, Lin Y-C, Wetterbom A, Vezzi A, Scofield DG, Delhomme N, Alexeyenko A, Giacomello S, Vicedomini R, Sahlin K, Sherwood E, Elfstand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Olson A, Niittylä, T, Ritland K, Rilakovic N, Rosselló JA, Sena J, Svensson T, Talavera-López C, Theißen G, **Vanneste K**, Tuominen H, Zhang J, Wu Z, Zerbe P, Bhalerao RP, Bohlmann J, Arvestad L, Bousquet, J, Garcia Gil R, de Jong PJ, Hvidsten TR, MacKay J, Ritland K, Morgante M, Sundberg B, Van de Peer Y, Lee Thompson S, Nilsson O, Andersson B, Lundeberg J, and Jansson S (2013). The 20 Gbp Norway spruce genome sheds light on conifer genome evolution. *Nature* **497**:579-584.

5. **Vanneste K**, Van de Peer Y, and Maere S (2013). Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* **30**(1):177-190.

4. De Wilde K, De Buck S, **Vanneste K**, and Depicker A (2013). Recombinant antibody production in *Arabidopsis* seeds triggers an unfolded protein response. *Plant Physiology* **161**(2):1021-1033.

3. *Voordeckers K, *Brown A, **Vanneste K**, Van Der Zande E, Voet A, Maere S, and Verstrepen K (2012). Reconstruction of ancestral maltase enzymes reveals molecular mechanisms underlying sub- and neofunctionalization in gene duplicates. *PLoS Biology* **10**(12):e1001446.

2. *Vekemans D, *Proost S, **Vanneste K**, Coenen H, Viaene T, Ruelens P, Maere S, Van de Peer Y, and Geuten K (2012). Gamma paleohexaploidy in the stem-lineage of core eudicots: Significance for MADS-box gene and species diversification. *Molecular Biology and Evolution* **29**(12):3793-3806.

1. *Vandesteene L, *López-Galvis L, **Vanneste K**, Feil R, Maere S, Lammens W, Rolland F, Lunn J, Avonce N, Beeckman T, and Van Dijck P (2012). Expansive evolution of the *TREHALOSE-6-PHOSPHATE PHOSPHATASE* gene family in *Arababidopsis thaliana*. *Plant Physiology* **160**(2):884-896.

Selected meetings

Plant Genome Evolution 2013 (PGEV2013) - Amsterdam, The Netherlands
Oral talk: A burst of whole genome duplications in flowering plants at the end of the Cretaceous and the beginning of the Paleogene

Society for Molecular Biology and Evolution 2013 (SMBE2013) - Chicago, USA
Poster: A burst of WGDs in flowering plants at the Cretaceous-Paleogene boundary

Society for Molecular Biology and Evolution 2012 (SMBE2012) - Dublin, Ireland
Poster: Detection of WGDs by modeling duplication dynamics

Frontiers in plant biology: From discovery to applications 2012 - Ghent, Belgium
Poster: Detection of WGDs by modeling duplication dynamics

Plant Genome Evolution 2011 (PGEV2011) - Amsterdam, The Netherlands
Poster: Detection of WGDs by modeling duplication dynamics

Benelux Bioinformatics Conference 2009 (BBC2009) - Liège, Belgium

Selected training - Specialist courses

Advanced statistics and data mining - Madrid, Spain
Two week summer school by University of Madrid

Optimization techniques in systems biology and bionformatics - Amsterdam, The Netherlands
One week summer school by Netherlands Bioinformatics Centre

Data structures and algorithms - Ghent, Belgium
Regular course unit by Ghent University

Introduction to MySQL - Ghent, Belgium
VIB research training course

Selected training - Transferable skills courses

Tech transfer & science based entrepreneurship - Ghent, Belgium
VIB research training course

Communication and presentation techniques - Ghent, Belgium
VIB research training course

Effective scientific writing - Ghent, Belgium
VIB research training course

Communication skills - Ghent, Belgium
Ghent University doctoral schools course

Appendix D

Supplementary material - Functional innovation through gene duplication

D.1 Supplementary figures

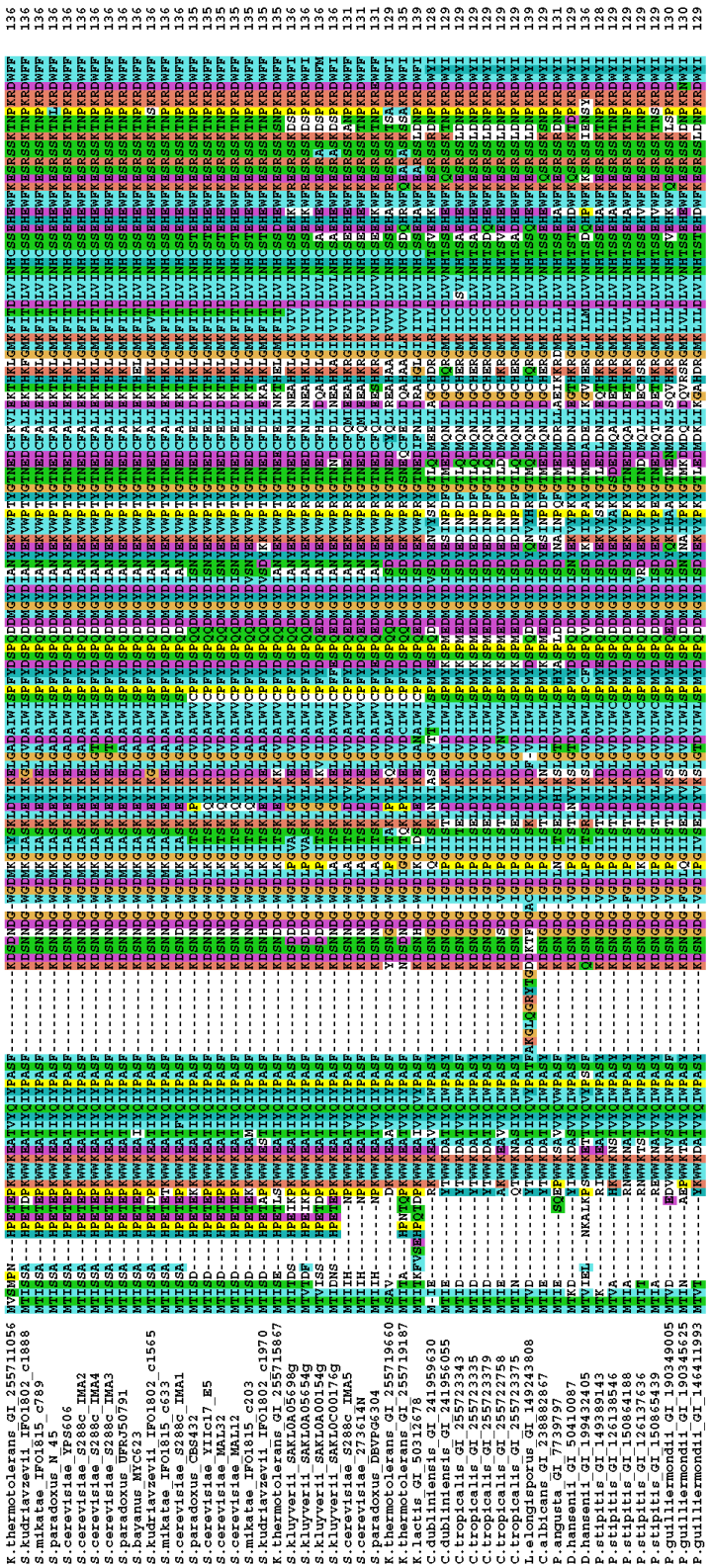


Figure D.1: Alignment of *MALS* genes. The alignment contains all genes that were used for phylogenetic tree construction (see 2.2.1), and ancestral sequence reconstruction (see 2.2.2).

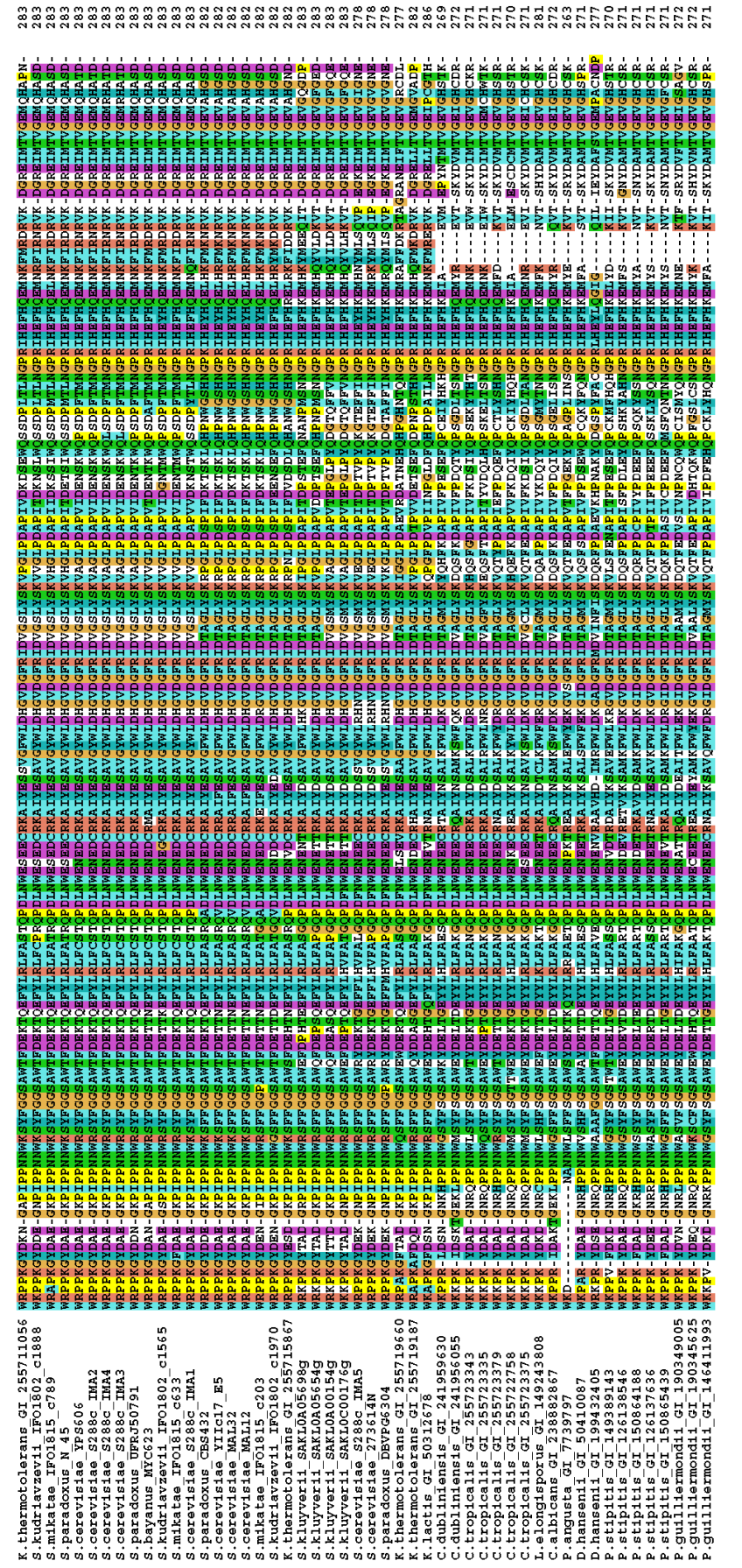


Figure D.1: Alignment of *MALS* genes - Continued. The alignment contains all genes that were used for phylogenetic tree construction (see 2.2.1), and ancestral sequence reconstruction (see 2.2.2).

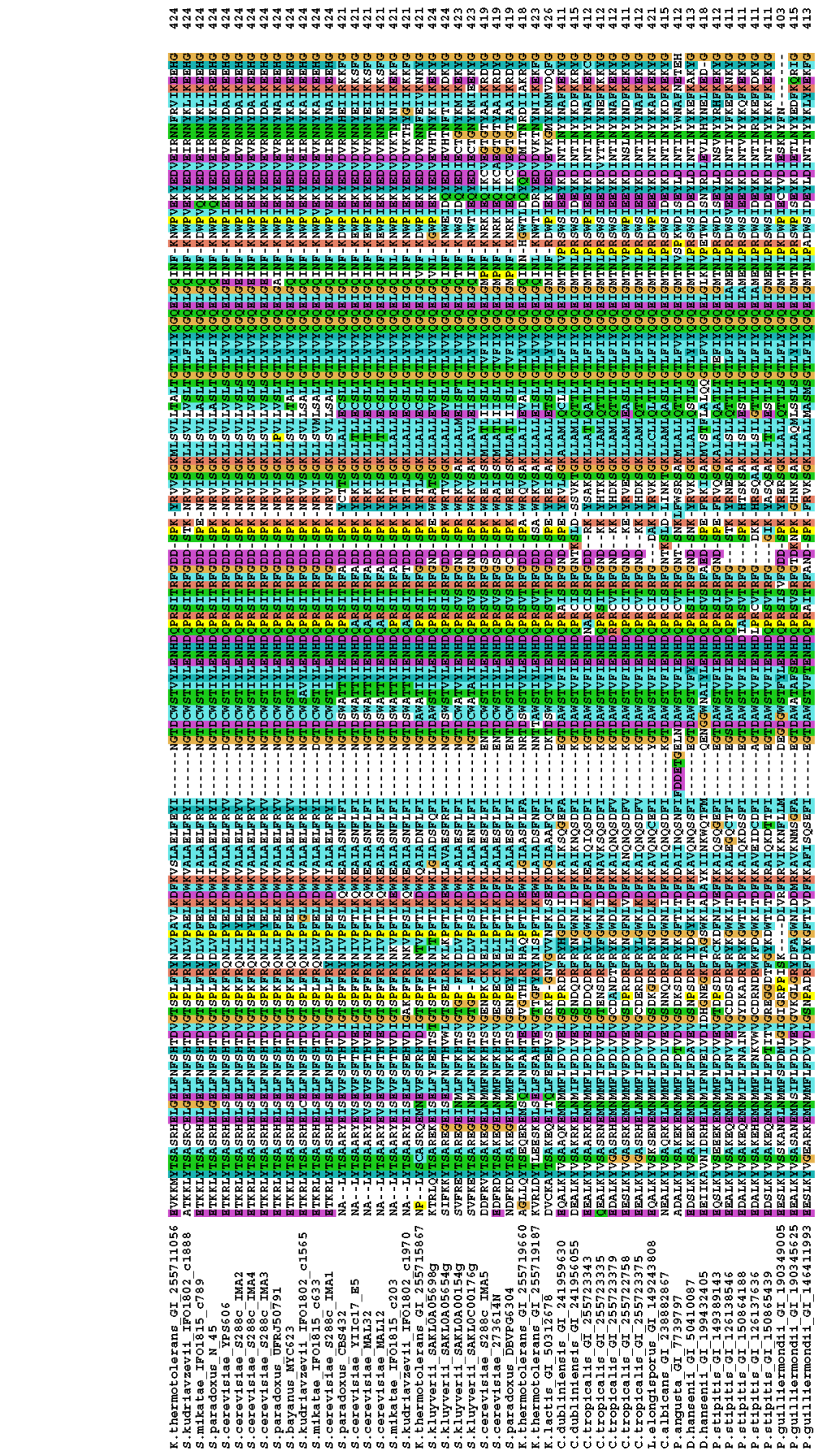


Figure D.1: Alignment of *MALS* genes - Continued. The alignment contains all genes that were used for phylogenetic tree construction (see 2.2.1), and ancestral sequence reconstruction (see 2.2.2).

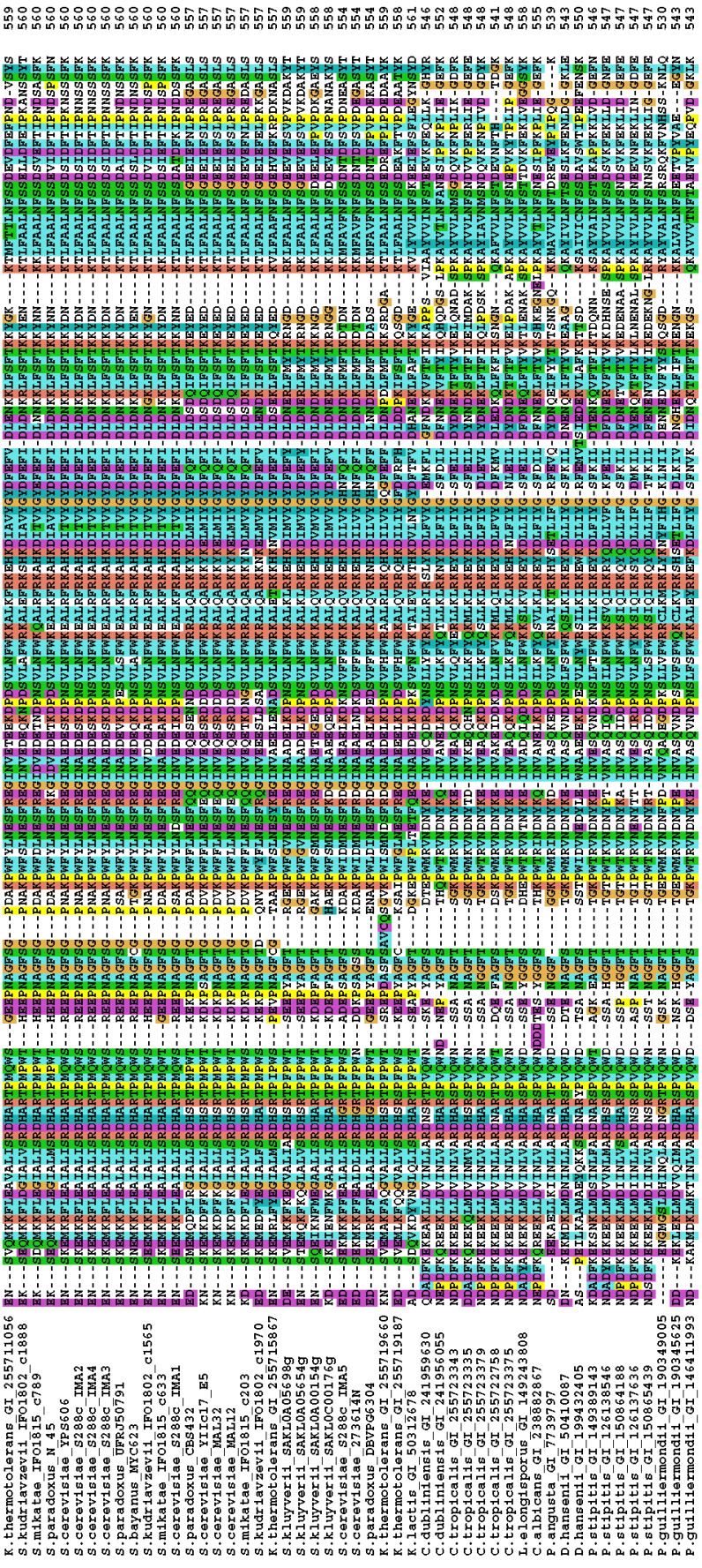


Figure D.1: Alignment of *MALS* genes - Continued. The alignment contains all genes that were used for phylogenetic tree construction (see 2.2.1), and ancestral sequence reconstruction (see 2.2.2).

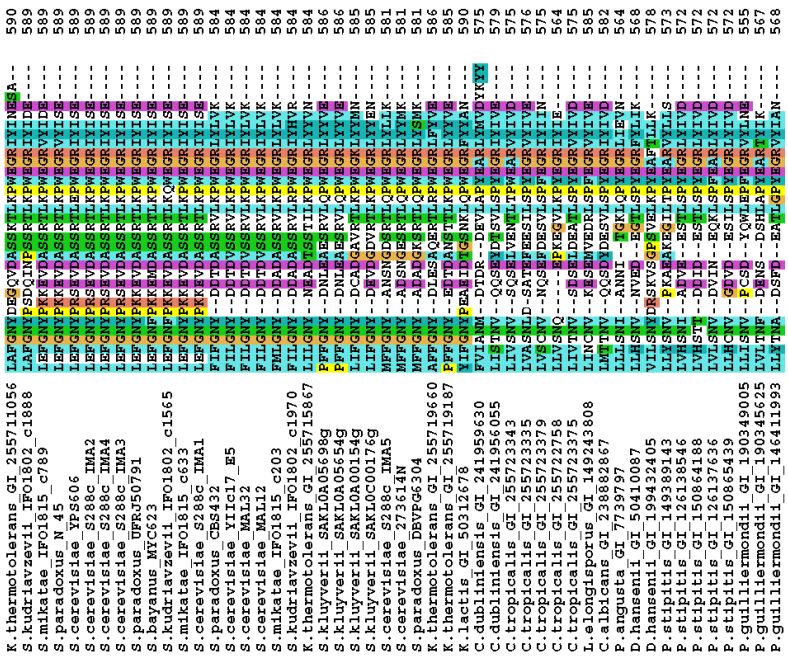


Figure D.1: Alignment of *MALS* genes - Continued. The alignment contains all genes that were used for phylogenetic tree construction (see 2.2.1), and ancestral sequence reconstruction (see 2.2.2).

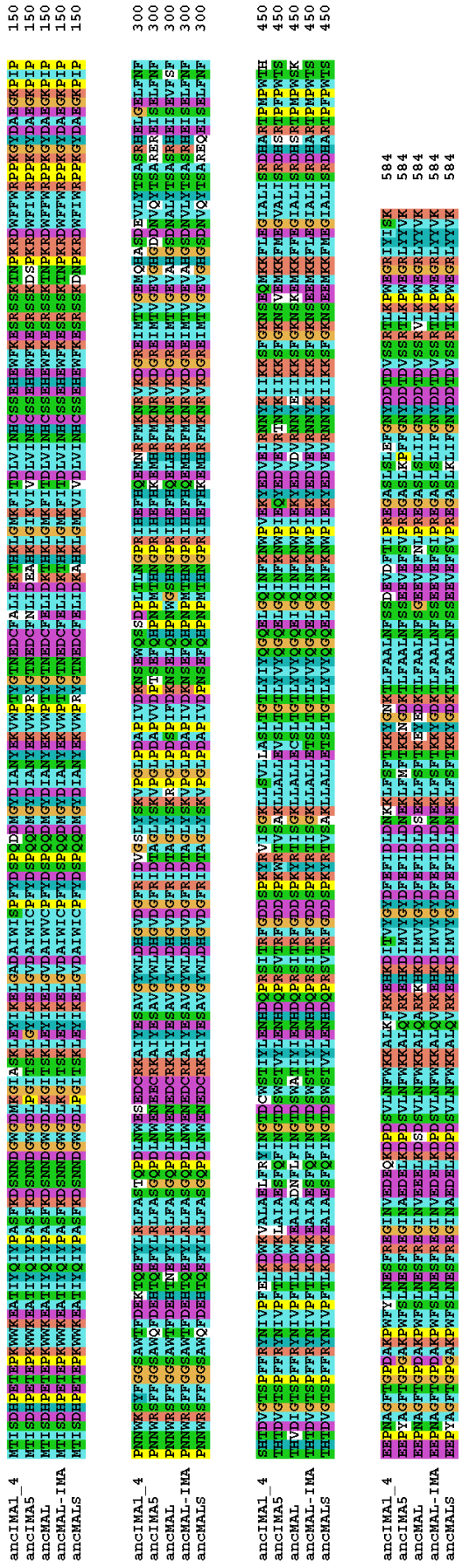


Figure D.2: Reconstructed ancestral sequences. Protein sequences resulting from marginal reconstructions under the JTT model that were used for synthesis of ancestral enzymes (see 2.2.2).

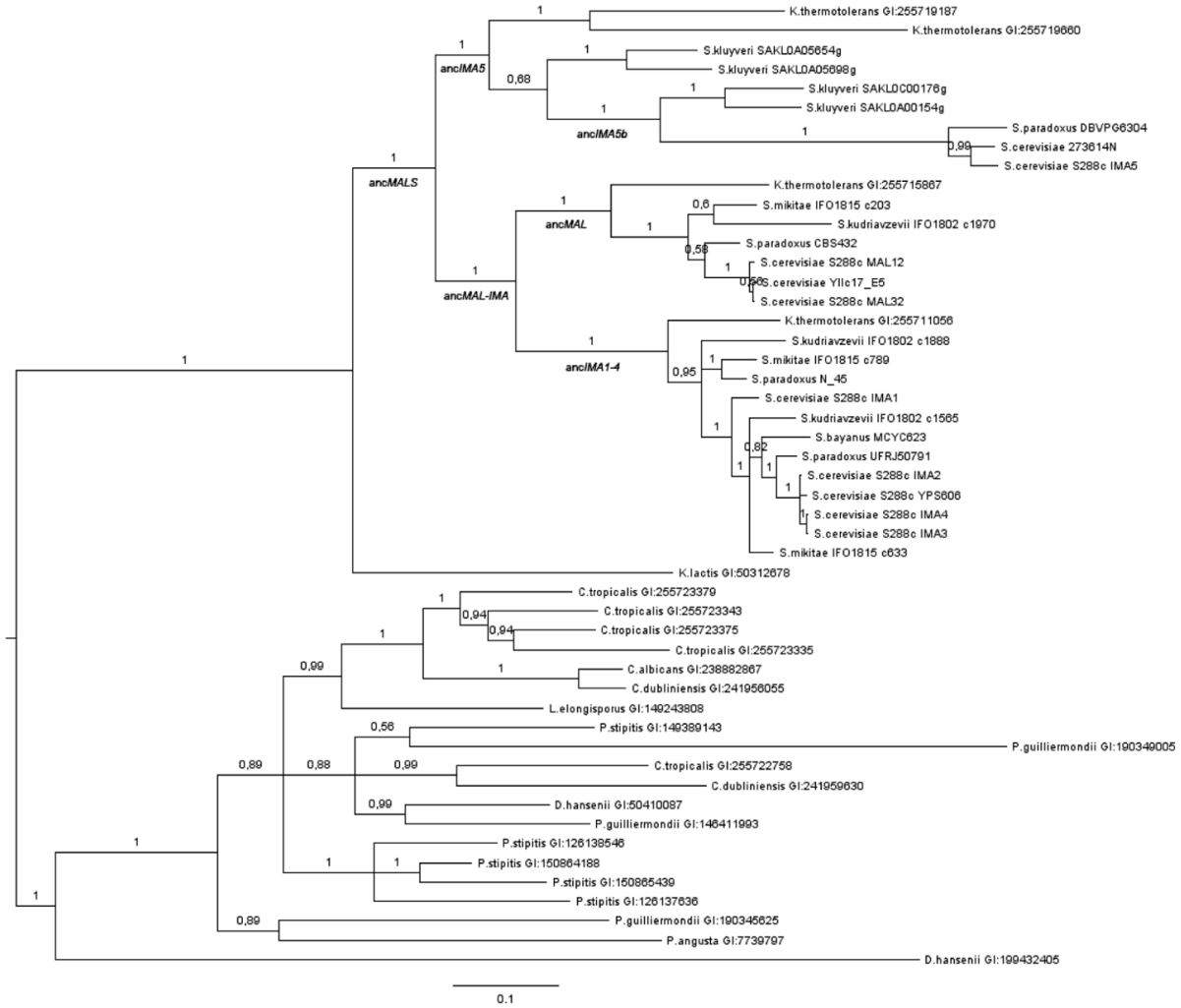


Figure D.3: Bayesian consensus topology of the 50 *MALS* genes. MrBayes consensus tree of the 50 *MALS* genes (AA-based, LG+I+G model with four rate categories). Posterior probabilities are indicated on the branches.

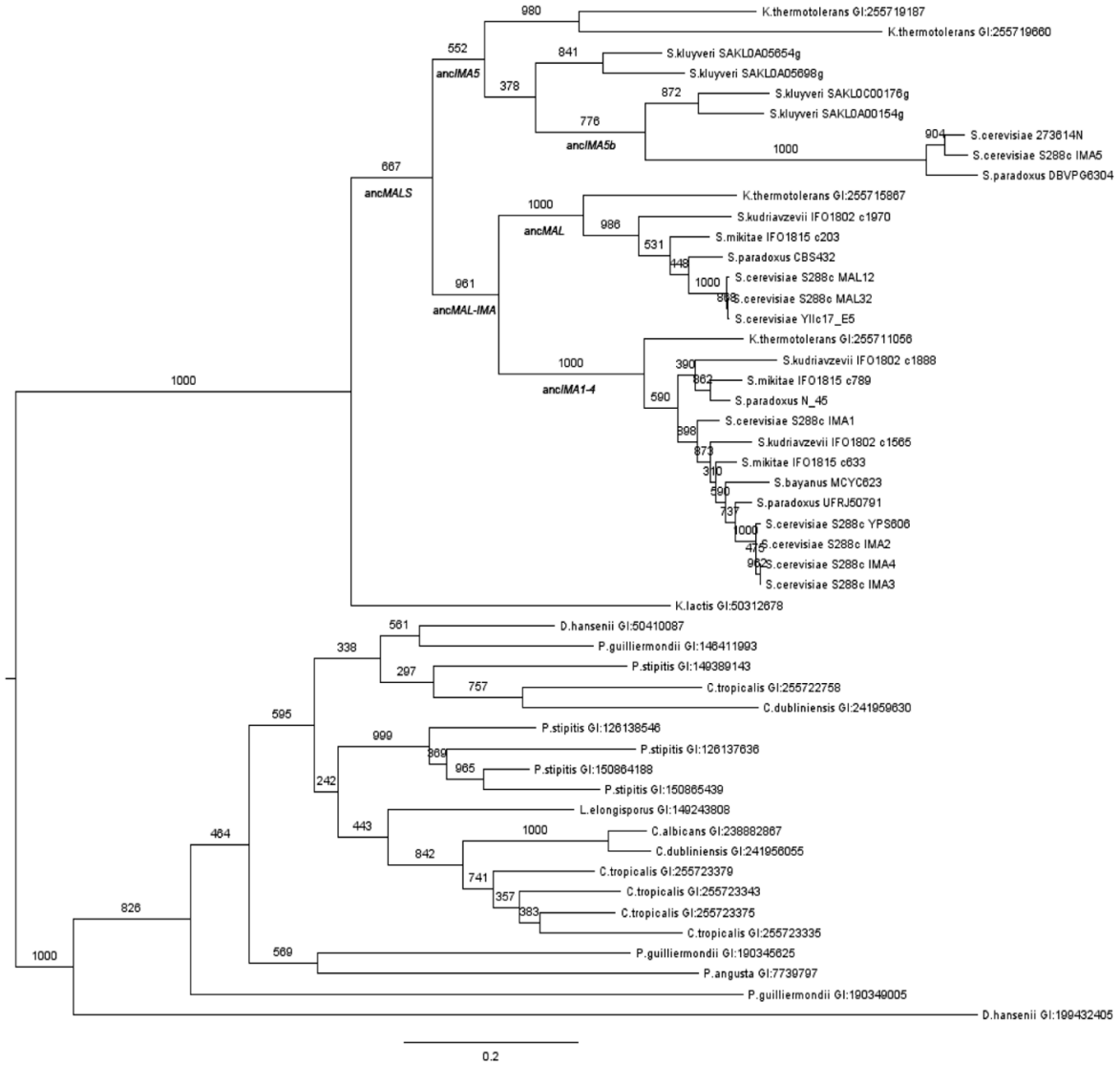


Figure D.4: Maximum likelihood topology of the 50 *MALS* genes. Maximum likelihood phylogeny of the 50 *MALS* genes calculated with PhyML (AA-based, LG+I+G model with four rate categories, 1,000 bootstraps). Bootstrap values are indicated on the branches.

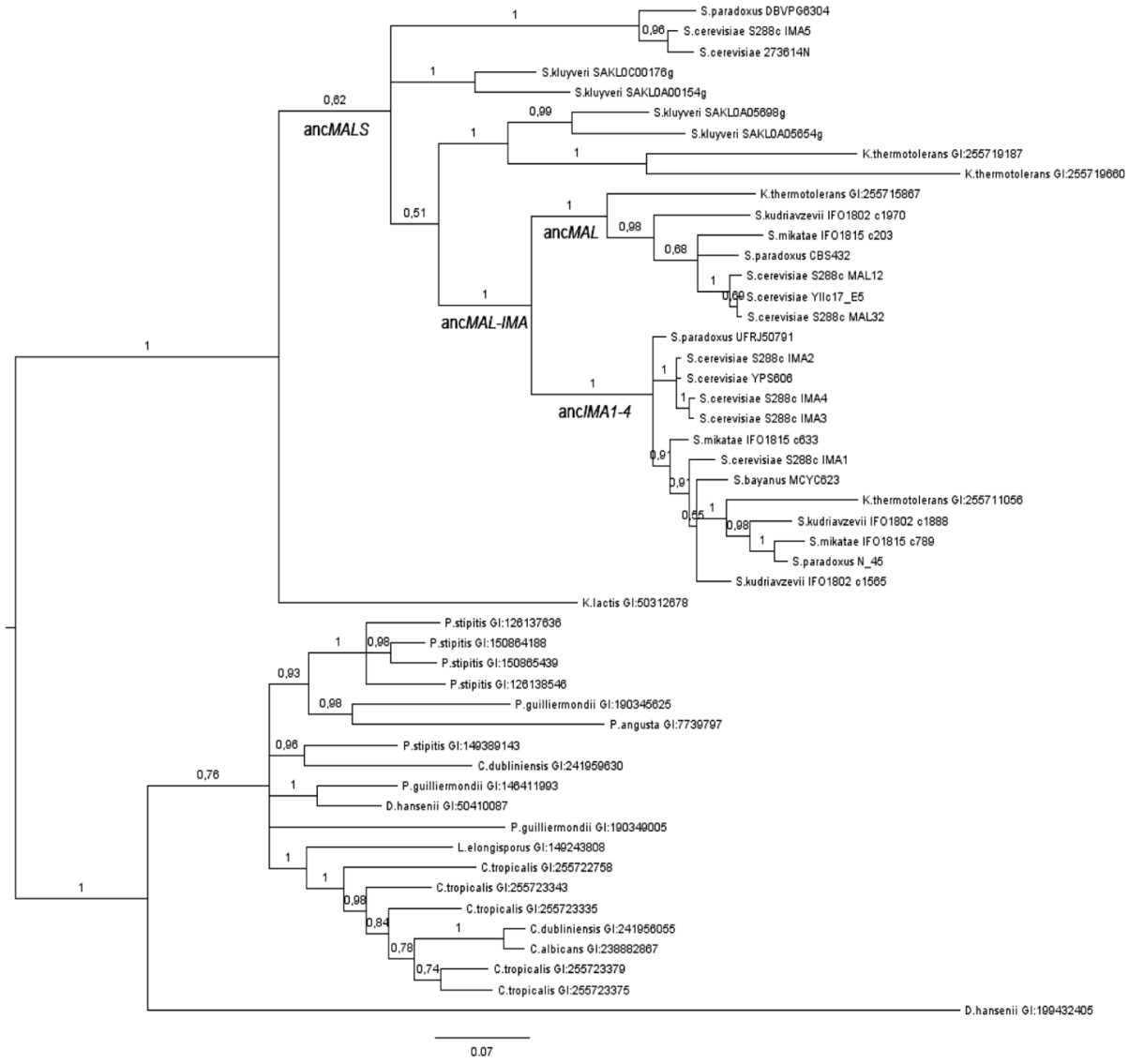


Figure D.5: Bayesian consensus topology of the 50 *MALS* genes with fast evolving sites removed. MrBayes consensus tree of the 50 *MALS* genes (AA-based, LG+I+G model with four rate categories). All AA sites with more than three variable AAs in the outgroup were stripped from the alignment. Posterior probabilities are indicated on the branches.

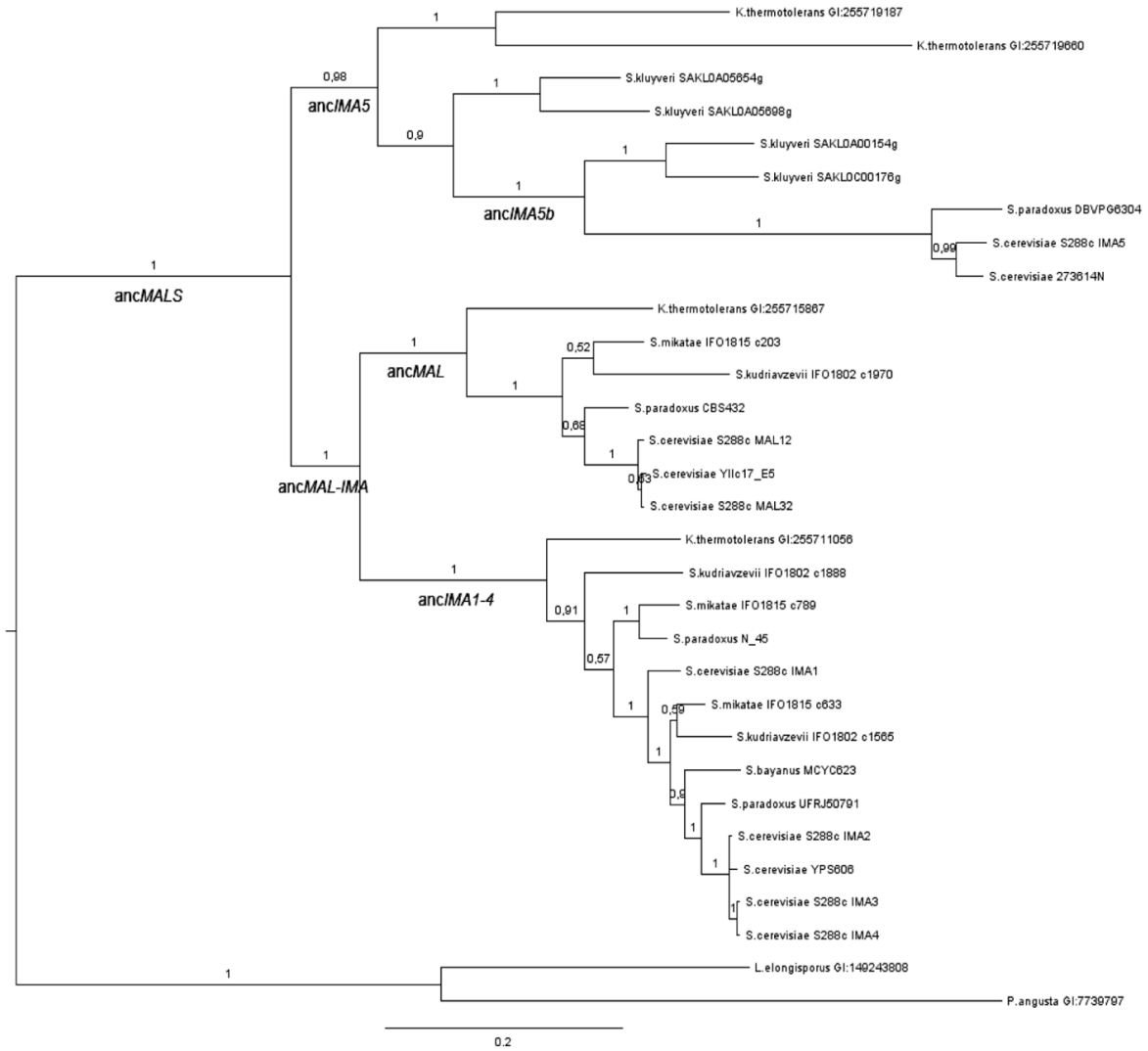


Figure D.6: Bayesian consensus topology of the *MALS* genes without *K. lactis*. MrBayes consensus tree of the *MALS* genes (AA-based, LG+I+G model with four rate categories). The *K. lactis* branch was not included in the tree reconstruction. Posterior probabilities are indicated on the branches.

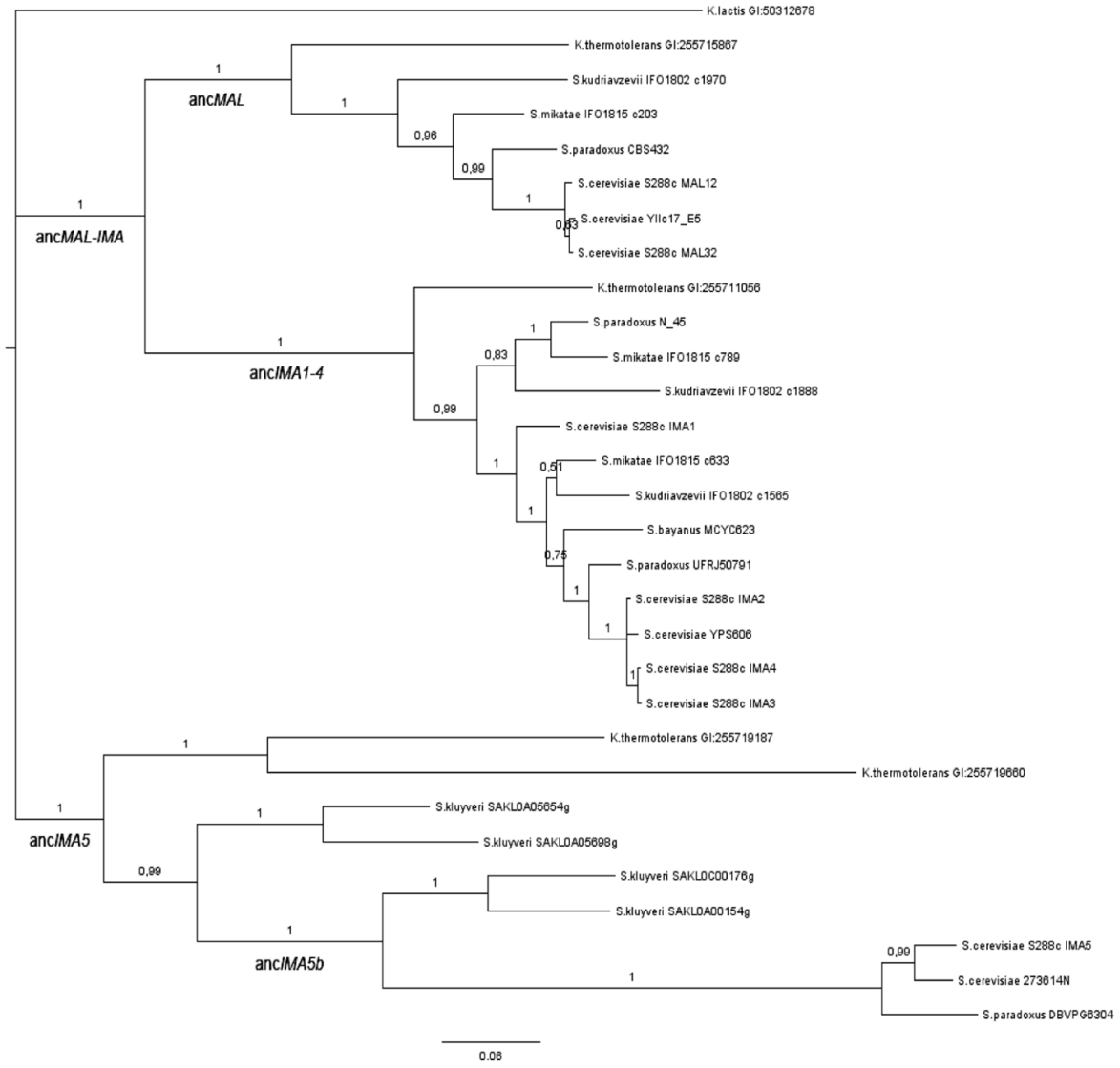


Figure D.7: Bayesian consensus topology of the *MALS* genes without the outgroup. MrBayes consensus tree of the *MALS* genes (AA-based, LG+I+G model with four rate categories). The outgroup branches were not included in the tree reconstruction. Posterior probabilities are indicated on the branches.

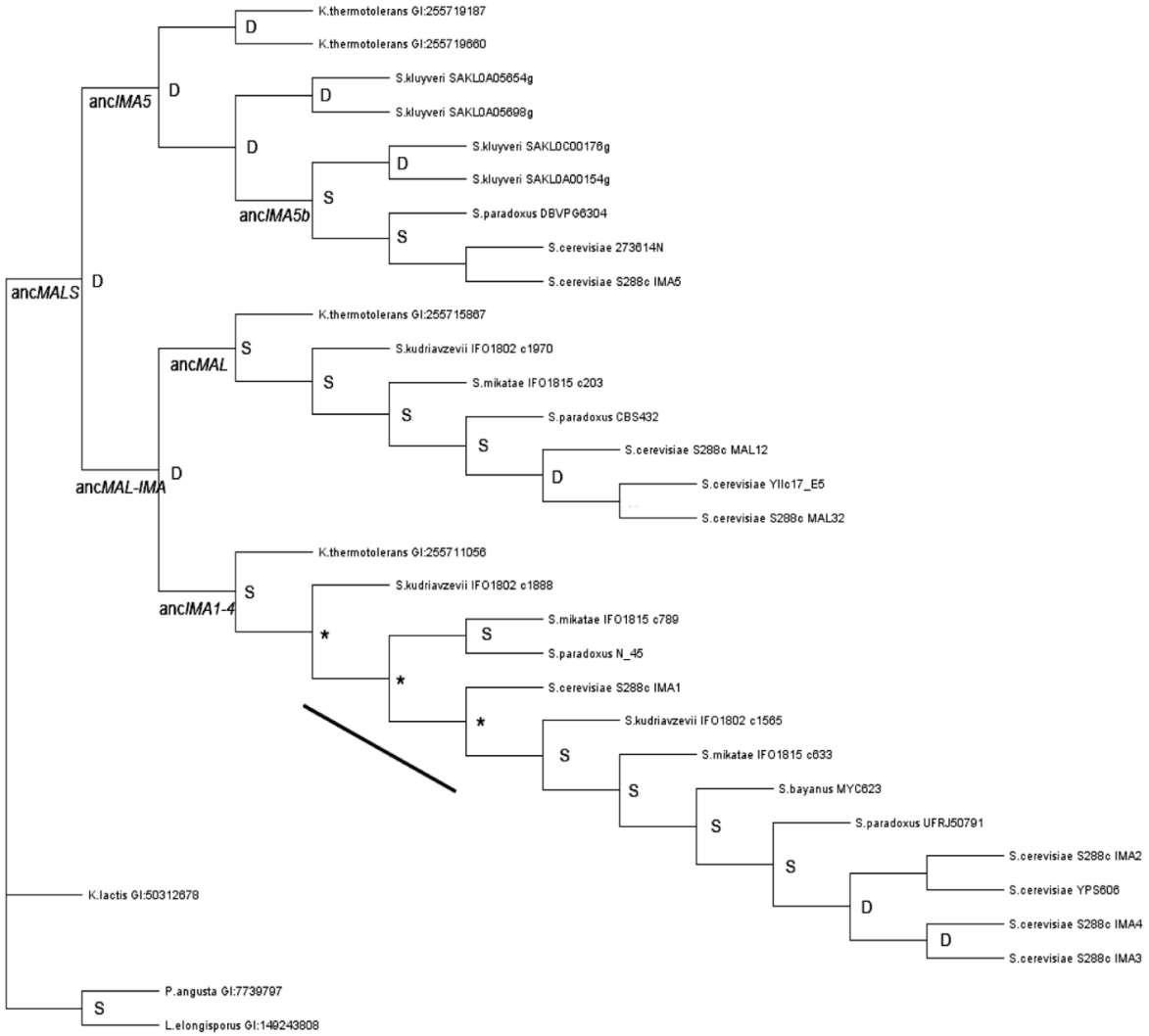


Figure D.8: Schematic tree showing inferred orthology-paralogy relationships between the different *MALS* genes. A schematic version of the codon-based phylogenetic tree inferred with MrBayes (see figure 2.4) is shown. Duplication events, D; speciation events, S. Asterisks denote nodes along a segment with ambiguous speciation/duplication history.

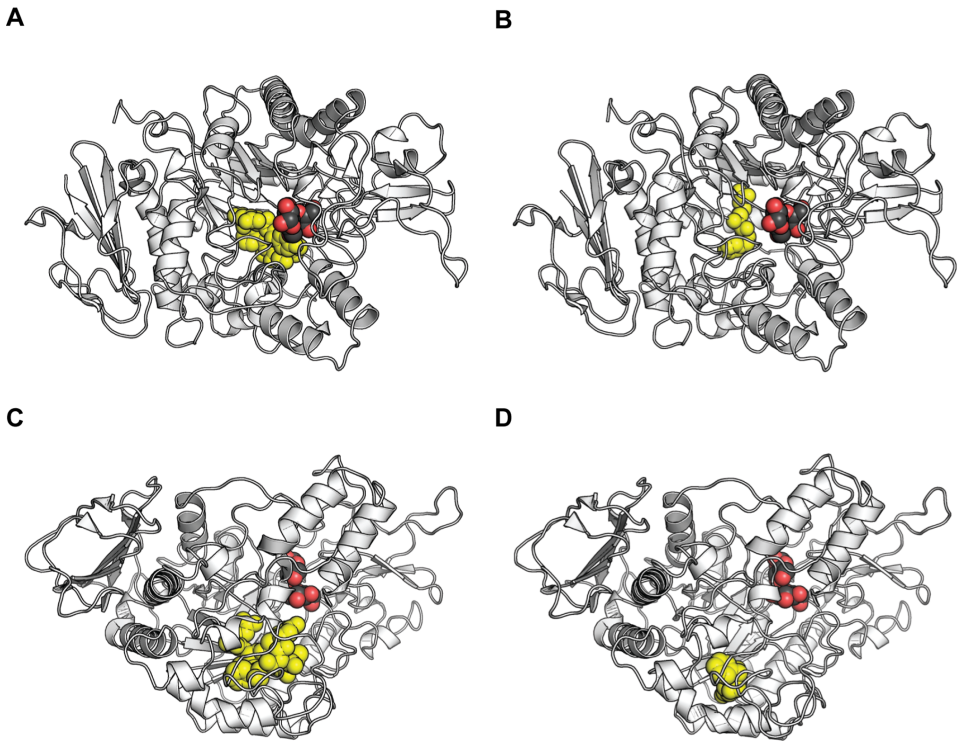


Figure D.9: Structural differences between *K. lactis GI:50312678* and *K. lactis GI:5441460* can explain lack of glucosidase activity in the latter enzyme. Cartoon representation of *K. lactis GI:50312678* (A and C) and *K. lactis GI:5441460* (B and D) in two different orientations (A and B result in C and D, respectively, after a 90° rotation) with maltose represented as black and red spheres. Comparing the sequence of *K. lactis GI:50312678* and *K. lactis GI:5441460* reveals the absence of five AAs in the latter protein. Mapping the position of these residues (the five AAs as well as two flanking residues are shown as yellow spheres in A and C; in B and D only the flanking residues are shown) shows that this region is located below the active site of the enzyme. Its deletion creates a larger cavity. This in turn could be compensated in the tertiary structure and explain the lack of activity detected for maltose- and isomaltose-like substrates for *K. lactis GI:5441460*.

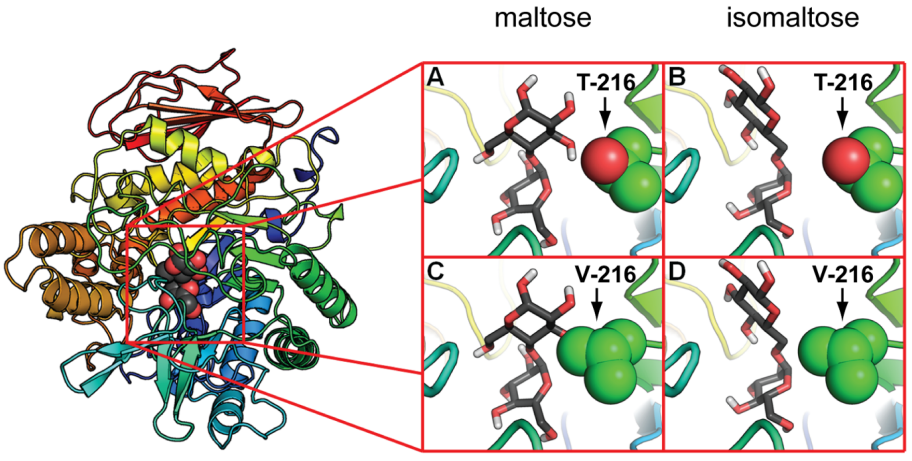


Figure D.10: Crucial role for the residue at position 216 in determining substrate affinity. Structural analysis of the active site reveals a crucial role for position 216 in determining substrate affinity, by affecting the hydrophobic/hydrophilic interactions with the different substrate classes. Subpanels are graphical representations of the binding pocket, with the residue at position 216 shown as spheres. Panels A and B depict an active site with threonine at position 216, whereas C and D depict an active site with valine at position 216. Maltose (A and C) and isomaltose (B and D) are represented as sticks. This structural analysis shows that threonine is able to form a hydrogen bond with a hydroxyl of the secondary glucose in maltose (A). The secondary glucose of isomaltose, however, is positioned in such a way that it causes unfavorable interactions (B). On the other hand, when residue 216 is a valine, it can form hydrophobic interactions with isomaltose (D).The hydrophobic side chain of valine is incompatible with the hydrophilic binding mode of maltose (C).

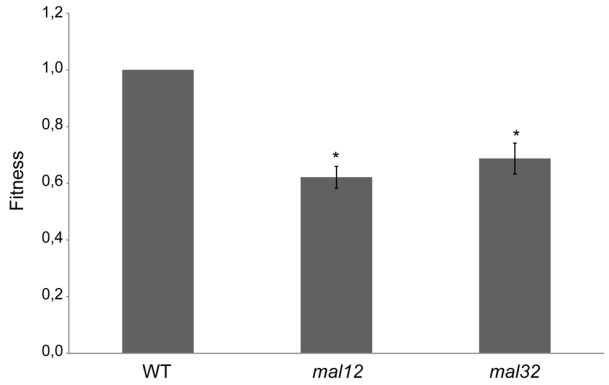


Figure D.11: Strains lacking one of the *MAL12*/*MAL32* paralogs have a fitness defect on maltose compared to the wild type. *mal12* (KV1151) and *mal32* (KV1153) strains show a significant fitness defect compared to the wild-type strain (KV1042) on maltose. A *mal12 mal32* double deletion strain does not grow on maltose. Asterisks indicate significant differences between mutant and wild-type strains (α=0.05). Error bars represent 95% confidence intervals.

D.2 Supplementary tables

Table D.1: Results of ancestral sequence reconstruction assuming different models of protein evolution. Results are available as a mutli-sheet Excel file online (doi:10.1371/journal.pbio.1001446.s014).

Table D.2: k_{cat} and K_m values for different enzymes on different sugars. Results are available as a multi-sheet Excel file online (doi:10.1371/journal.pbio.1001446.s015).

Table D.3: Results of two-way ANOVA analysis on log-transformed k_{cat}/K_m. Results are available as a multi-sheet Excel file online (doi:10.1371/journal.pbio.1001446.s016).

Table D.4: Results of PAML branch-site tests. Values show the result of PAML branch-site tests to identify residues that are under positive selection on three specific branches of the *MALS* phylogeny. Branch identifiers follow the nomenclature of figure 2.4. The location of positively selected sites is based on *IMA1* numbering.

branch	H_0	H_A	LRT	*P*-value	parameter estimates	pos. selected sites (BEB$>$0.95)
anc*IMA1-4*	-28326.54	-28320.88	11.32	$p<0.01$	$\hat{p}_0$=0.934, $\hat{p}_1$=0.028, $\hat{\omega}_0$=0.082, $\hat{\omega}_2$=5.466	216, 279, 333, 562
anc*MAL*	-28334.80	-28333.21	3.18	p=0.22	$\hat{p}_0$=0.953, $\hat{p}_1$=0.029, $\hat{\omega}_0$=0.083, $\hat{\omega}_2$=4.738	n/a
anc*IMA5b*	-28330.08	-28322.96	14.24	$p<0.001$	$\hat{p}_0$=0.950, $\hat{p}_1$=0.029, $\hat{\omega}_0$=0.083, $\hat{\omega}_2$=11.245	216, 299, 315, 414

Table D.5: Genotypes of yeast strains used in *MALS* gene study.

Strain name	Genotype
KV1042	S288c Mata MAL13::HYG-RM11_MAL63c9
KV1444	S288c Mata MAL13::HYG-RM11_MAL63c9 TEFp-IMA5
KV2498	S288c Mata MAL13::HYG-RM11_MAL63c9 IMA5::KanMX
KV1151	S288c Mata MAL13::HYG-RM11_MAL63c9 MAL12::KanMX
KV1153	S288c Mata MAL13::HYG-RM11_MAL63c9 MAL32::KanMX
KV1774	S288c Mata MAL13::HYG-RM11_MAL63c9 MAL12::KanMx MAL32::KanMX
KV3261	S288c Mata MAL13::HYG-RM11_MAL63c9 TDH3p::GFP-KanMX
KV3002	*Lodderomyces elongisporus* CBS2605
KV1983	*Ashbya gossypii* ATCC 10895
KV3000	*Kluyveromyces lactis* ATCC 8585
KV3190	*Saccharomyces kluyveri* CBS3082
KV3191	*Lachancea waltii* CBS6430
KV2817	*Kluyveromyces thermotolerans* CHCC5657
KV3192	*Kluyveromyces polysporus* CBS263
KV3193	*Saccharomyces castellii* CBS4309
KV1980	*Candida glabrata* CBS138
KV1556	*Saccharomyces bayanus* CBS7001
KV1981	*Saccharomyces kudriavzevii* IFO 1802
KV1982	*Saccharomyces mikatae* IFO 1815
KV1557	*Saccharomyces paradoxus* NCYC2600

Table D.6: Dating results for key splits in the *MALS* gene tree. Mean, median, and geometric mean refer to different average age estimates obtained from the sampled traces across the different MCMC chains, and 95% HDP upper and lower can be regarded as 95% confidence intervals (see BEAST documentation). The effective sample size (ESS) is a measure of convergence (higher is better).

	anc*IMA1-4*	anc*IMA5*	anc*MALS*	anc*MAL-IMA*	calibration2	calibration1	anc*MAL*
mean	55.9373	94.1671	118.6754	87.9487	170.155	149.5962	55.5027
stderr of mean	9.1298E-2	9.8827E-2	8.6328E-2	0.102	1.5688E-2	1.5756E-2	0.1055
median	55.3395	94.348	119.4	87.8065	170.1588	149.5978	54.8716
geometric mean	55.3002	93.5564	118.2222	87.2616	170.1289	149.5666	54.603
95% HPD lower	39.5439	73.1691	97.9651	66.8271	164.3415	143.7635	36.9593
95% HPD upper	72.4399	114.4806	137.1406	109.0841	176.0083	155.4541	75.6895
auto-correlation time	41719.8218	31274.6714	25781.4103	31386.5077	10000	10115.93	40035.6677
effective sample size	8629.9506	11512.1913	13965.101	11471.1711	36004	35591.3891	8992.981

D.3 Supplementary information

D.3.1 Additional tests to exclude long branch attraction (LBA) artifacts

Despite high support for our inferred topology by both Bayesian and Maximum Likelihood methods, the position of *K. lactis GI: 50312678* warranted further investigation. Our topology of the *MALS* gene family supports that this gene branched off before the *S. kluyveri* - *S. cerevisiae* split. However, according to another commonly accepted view of ascomycete evolution, the *Kluyveromyces* (*K. lactis*) and *Lachancea* (*S. kluyveri* and *K. thermotolerans*) clades branched off together from *Saccharomyces*[463]. The topology of the ascomycete species tree is currently insufficiently resolved to be considered final, and earlier studies have provided conflicting results regarding the branching order of the *Saccharomyces*, *Lachancea*, and *Kluyveromyces* clades[463–466]. Nevertheless, the position of the *K. lactis* branch in our topology could potentially have been impacted by long branch attraction (LBA) between the *K. lactis* branch and long outgroup branches[467]. Since *K. lactis* serves as most recent outgroup to the anc*MALS* clade, it has a big influence on ancestral sequence reconstruction and requires confidence in its placement. Bayesian and maximum likelihood methods as used in our tree reconstruction have been found to be less susceptible to LBA artifacts but nevertheless are not invulnerable to it. Improved taxon sampling around the *K. lactis* branch could mitigate possible LBA artifacts[467], but this proved impossible as all relevant *MALS* sequences known to date were already included in the tree reconstruction. We therefore ran 2 extra analyses that help to detect LBA artifacts. First, we removed all fast evolving sites in our protein alignment

by discarding all sites in the alignment that had more than 3 variable amino acids in the outgroups (defined here as all sequences not belonging to *Saccharomyces*/*Lachancea*/*Kluyveromyces* species). A different placement of the *K. lactis* branch would then be indicative of LBA artifacts caused by fast evolving sites in the alignment[467]. The phylogeny was determined using MrBayes 3.1.2 with a LG+I+G model with 4 rate categories. The resulting topology (see figure D.5) is consistent with the topology presented in figures D.3 and D.4. Although confidence in more recent splits is lower and results in more multifurcations (most likely due to the loss of information associated with removing data from the alignment), *K. lactis* still branches off before the *S. kluyveri* - *S. cerevisiae* split with high posterior probabililty. For the second analysis, we ran 2 separate phylogenies with *K. lactis* and outgroup sequences excluded, respectively. Excluding one of the 2 potential long branch attractors should result in a correct placement of the other branch and is therefore also indicative of LBA artifacts[467]. Both phylogenies were constructed using MrBayes 3.1.2 as described before. Figure D.6 presents the phylogeny with the *K. lactis* branch removed and outgroup representatives included. The topology of the ingroup corresponds with the topology of the ingroup in figures D.3 and D.4 for all major splits. Figure D.7 presents the phylogeny with the *K. lactis* branch included and the outgroups removed. The topology of the ingroup again corresponds almost completely with those presented in figures D.3 and D.4. The *K. lactis* branch multifurcates together with the anc*IMA5* and anc*MAL-IMA* clades but is not pulled inside one of the *Lachancea* clades. In conclusion, both excluding fast evolving sites and excluding potential long branch attractors did not change the major ingroup topology of the *MALS* genes (i.e., the anc*MALS* clade in figure 2.4), and provide support that *K. lactis GI: 50312678* indeed does not belong to one of the three *Lachancea* - *Saccharomyces* clades in the *MALS* gene phylogeny.

D.3.2 Dating of *MALS* duplications

We estimated the age of the major divergences in the *MALS* phylogeny (i.e., anc*MALS*, anc*IMA5*, anc*MAL*, anc*MAL-IMA*, and anc*IMA1-4*) using a Bayesian approach as implemented in the BEAST v1.6.1 program[468]. We employed the general GTR+I+G model of DNA substitution with four rate categories. For the clock model, we selected the lognormal relaxed-clock model, which allows rates to vary among branches without any *a priori* assumption of autocorrelation between adjacent branches. For the tree prior, we employed a Yule process of speciation, with the topology specified as in figure 2.4 of the main text (without branch lengths specified). The ingroup was considered monophyletic with respect to the outgroup consisting of *P. angusta* and *L. elongisporus*. The posterior distribution of the estimated divergence times was obtained by specifying 2 calibration points based on literature. The first calibration point is the divergence of the *Saccharomyces* from the *Kluyveromyces* clade, estimated at 150 mya[133]. The second calibration point is the divergence of *C. albicans* from *S. cerevisiae*, estimated at 170 mya[469]. Both of these calibration points are however molecular-based age estimates themselves, instead of fossil and/or geological-derived. They are therefore prone to biases induced by the possible inadequacy of the molecular data, the model of molecular evolution, and the methods used to derive these estimates. Use of such calibration points in divergence dating is therefore generally discouraged but nonetheless required in this case since no other viable calibration points were available[470]. Results should however be interpreted with due caution in the present context. For both calibration age estimates, we used a

normal prior with as mean the estimate and as standard deviation 3 million years. In total, 4 independent MCMC runs were run for 100 million generations, sampling every 10,000 generations to reach a total of 1,0000 samples per individual run. Log files from each run were analyzed with Tracer v1.5 (part of the BEAST package) using a burn-in of one million generations, and demonstrated strong equilibrium with effective sample sizes (ESS) of all parameters far exceeding 200. Convergence of run replicates was confirmed by visual inspection of traces within and between traces, and the results of the combined traces are presented in table D.6.

D.3.3 Microbial strains, growth conditions, and molecular techniques

Protein expression

Overnight cultures of *E. coli* were diluted 1:20 into 500 mL of LB + kanamycin. These cultures were grown at 37°C for 3 hours, after which cells were induced with 1mM IPTG (Sigma Aldrich) and then grown at 30°C for another 5 hours. Cells were harvested by spinning at 6000g for 10 minutes at 4°C. The cell pellets were then frozen at -80°C.

Protein purification

Frozen cell pellets were thawed and resuspended in 10 mL of Eq. Wash Buffer (50 mM phosphate + 300 mM NaCl + 5% glycerol at pH 7). Cell suspensions were incubated with gentle agitation at room temperature with 7.5 mg lysozyme (Sigma Aldrich) for 15 minutes. The cell suspensions were sonicated 4x1 minute with 1 minute breaks on ice in between. The raw cell lysate was fractioned into 2 mL test tubes and spun at 10,000g for 10 minutes at 4°C. The supernatant was added to 6 mL of pre-equilibrated (3 mL packed bead volume washed twice with 15 mL of Eq. Wash Buffer) TALON (Westburg) resin in a 5 mL polypropylene column (Qiagen). The column was incubated at room temperature with gentle agitation for 20 minutes in order to bind the 6xHis-tagged proteins. After binding the resin, the column was washed twice by incubating at room temperature with gentle agitation for 10 minutes with 15 mL of Eq. Wash Buffer. The bound protein was eluted with 2.5 mL of Elution Buffer (Eq. Wash Buffer + 200 mM imidazole (Sigma Aldrich)) by incubating for 10 minutes at room temperature. The protein concentration was quantified by using a Protein Quantification Kit-Rapid (Fluka) and qualified by running on a NuPage Novex Bis-Tris Mini Gel (Invitrogen).

Enzyme assays and data analysis

The following sugars were purchased from Sigma in their highest available purity (number in brackets corresponds to catalogue number): maltose (M5885), sucrose (84097), turanose (T2754), maltotriose (M8378), maltulose (50796), melezitose (M5375), methyl-α-glucoside (M9376), isomaltose (I7253), and palatinose (P2007). For maltose, sucrose, turanose, maltotriose, and maltulose, stock concentrations of 0.5, 1, 2, 3, 5, 10, 15, 20, 40, 60, 80, and 100 mM were prepared in Enzyme Assay Buffer (50 mM phosphate buffer + 300 mM NaCl + 5% glycerol at pH 6). For methyl-α-glucoside, isomaltose and palatinose, stock concentrations of 5, 10, 20, 30, 40, 50, 75, 100, 125, 150, 175, and 200 mM were prepared in Enzyme Assay Buffer. Reaction mixtures were prepared by adding 3 μL of purified

protein to 27 μL of stock sugar solution in a 96-well plate, such that the final concentration of protein was $\sim$100 μg/mL. The reaction plates were incubated at 30°C from 15-30 minutes (depending on activity of the enzyme tested), then inactivated at 98°C for 2 minutes. The final glucose concentration was measured by adding 90 μL of GOD-PAP reagent (Dialab), incubating at 30°C for 10 minutes, and measuring the absorbance at 505 nm. A negative control of *E. coli* strain BL21* (purified equivalently to the other proteins), incubated with the sugars, was included for each substrate concentration. The values obtained with this negative control were subtracted from the values obtained with the purified enzymes. The concentration of hydrolyzed substrate was determined by normalizing the measured glucose concentration by the number of glucose molecules per substrate, assuming that all glucose molecules liberated are assayable (e.g., for maltose divide measured concentration of glucose by 2).

Fitness measurements

Cultures were inoculated with equal numbers of labeled reference and unlabeled strains ($\sim 10^6$ cells of each) and allowed to grow for several generations. The experiment was carried out in SC maltose (2%) medium. The ratio of the two competitors was quantified at the initial and final time points by flow cytometry. Measurements were corrected for the small percentage of labeled, non-fluorescent cells that occurred even when the reference strain was cultured separately as well as for the cost of GFP expression in the labeled reference strain. This correction is made before feeding data in the “S formula”. For each fitness measurement, three independent replicates were performed. The selective advantage, s, of each strain was calculated as s=(ln(U_f/R_f)-ln(U_i/R_i))/T where U and R are the numbers of unlabeled and reference strain respectively, the subscripts refer to final and initial populations, and T is the number of generations that reference cells have proliferated during the competition. The fitness of the unlabeled WT strain was designated 1, and the fitness of *mal12* and *mal32* strains is 1+s.

Appendix E

Supplementary material - Inference of genome duplications

E.1 Supplementary figures

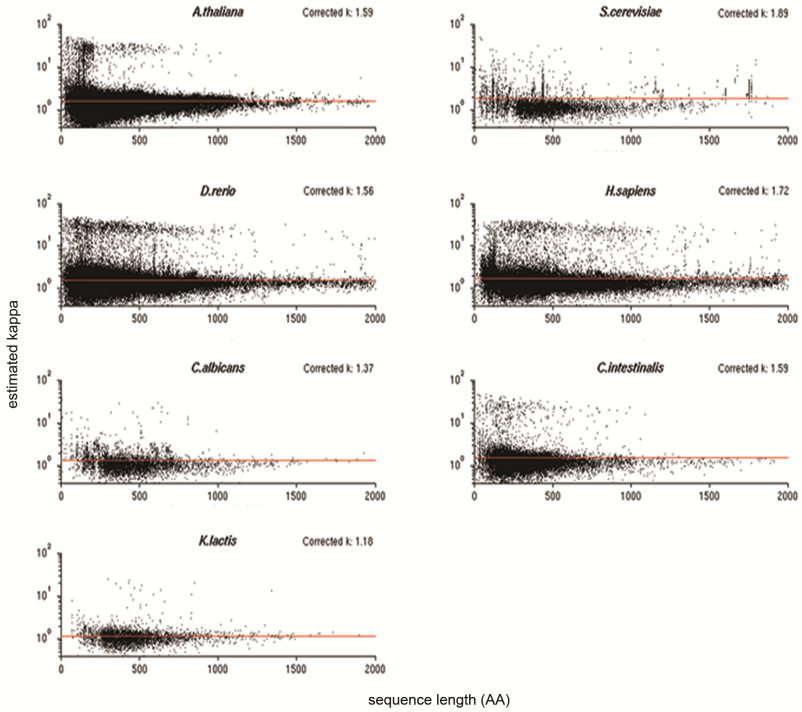


Figure E.1: Scatterplots demonstrating the relationship between the estimated κ and (stripped) sequence alignment length of all pairwise combinations for all seven species. The average corrected value for κ per species is printed in the upper right corner of each scatterplot, and is also indicated by a horizontal line.

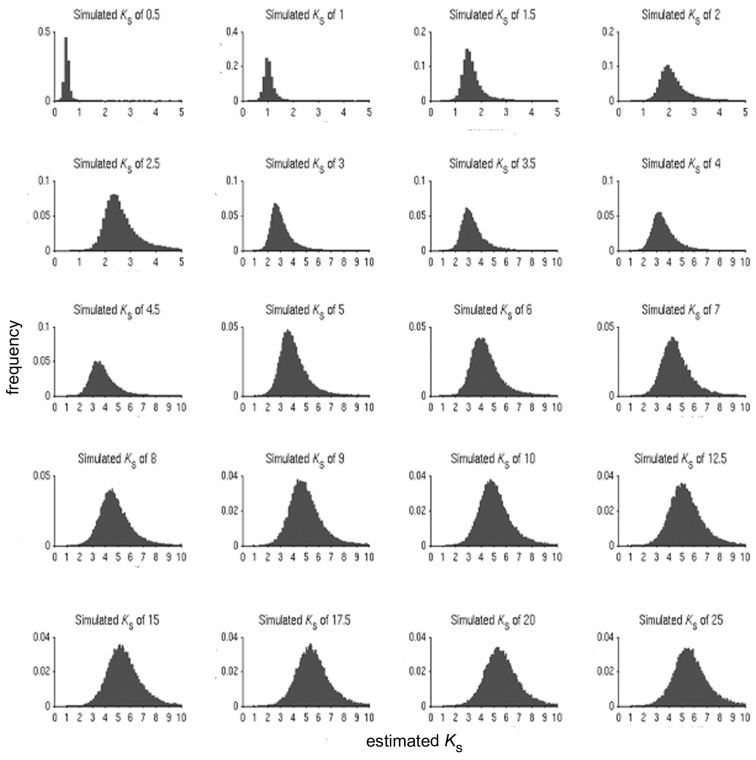


Figure E.2: Detailed results of synonymous evolution simulations for *A. thaliana*. 27,363 protein coding genes were synonymously evolved for evolutionary time intervals corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and synonymously evolved sequences were calculated. The panels display the resulting K_S estimate frequency distributions. The ordinate scale varies between panels.

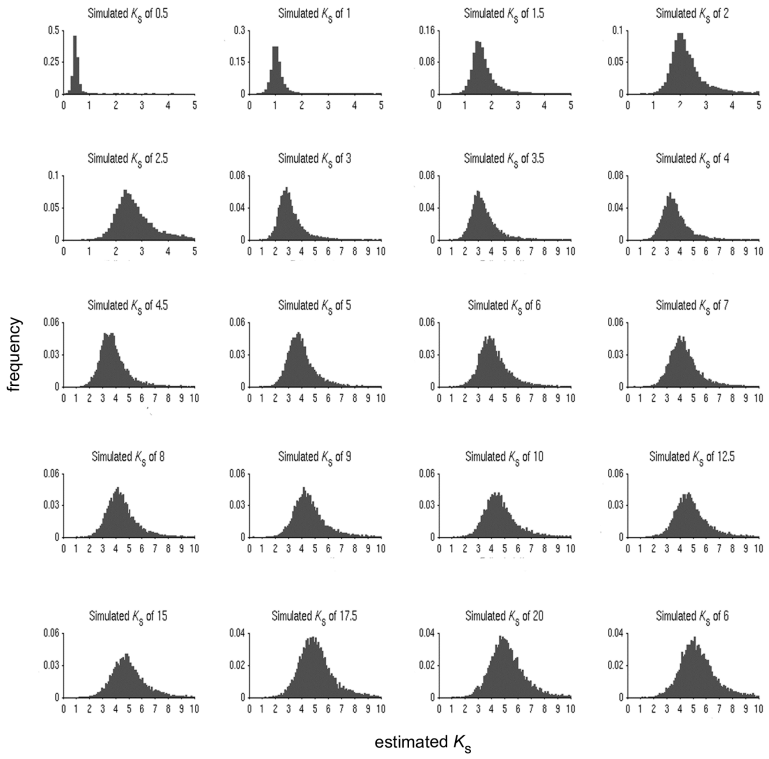


Figure E.3: Detailed results of synonymous evolution simulations for *S. cerevisiae*. 6,668 protein coding genes were synonymously evolved for evolutionary time intervals corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and synonymously evolved sequences were calculated. The panels display the resulting K_S estimate frequency distributions. The ordinate scale varies between panels.

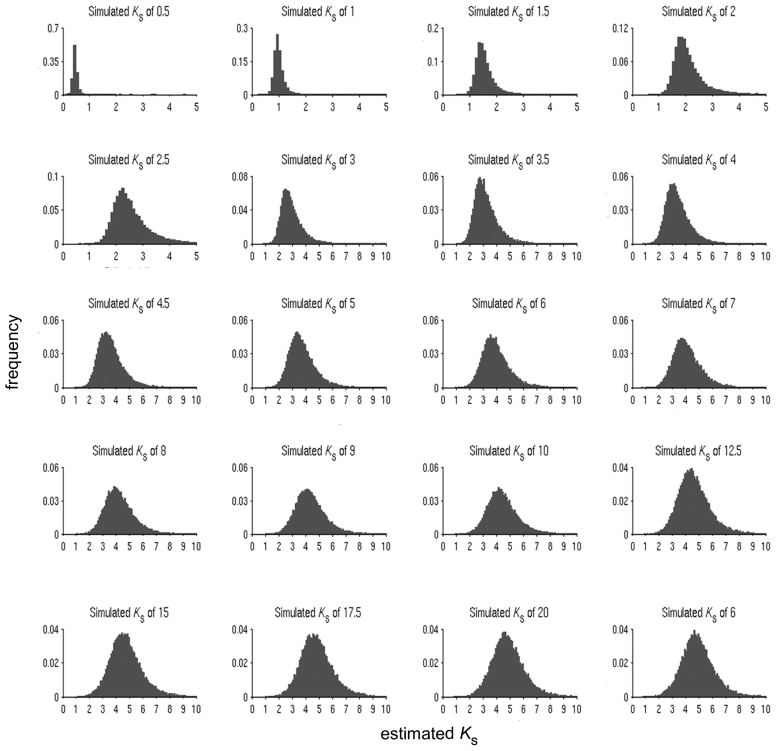


Figure E.4: Detailed results of synonymous evolution simulations for *D. rerio*. 22,826 protein coding genes were synonymously evolved for evolutionary time intervals corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and synonymously evolved sequences were calculated. The panels display the resulting K_S estimate frequency distributions. The ordinate scale varies between panels.

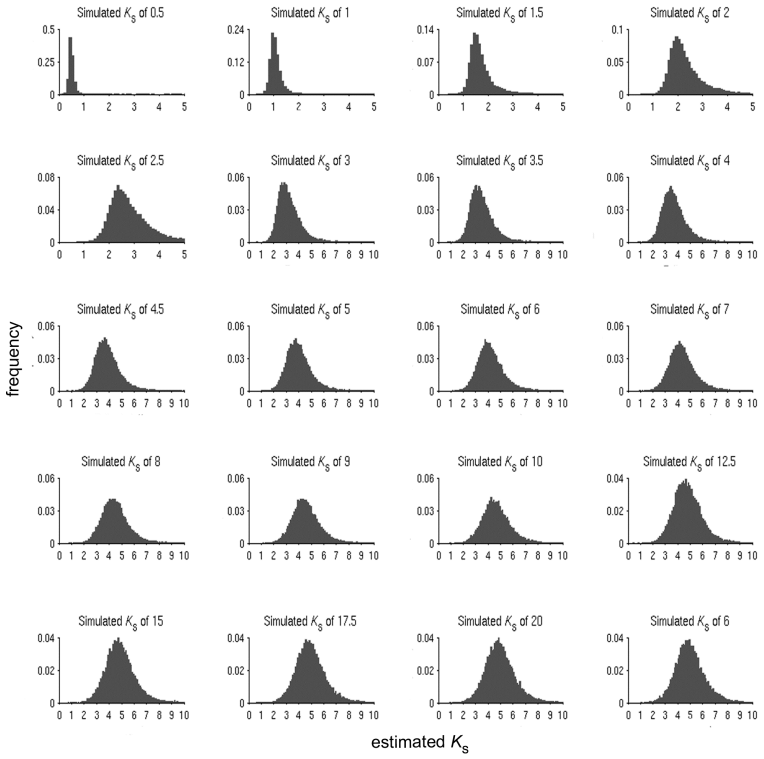


Figure E.5: Detailed results of synonymous evolution simulations for *H. sapiens*. 20,488 protein coding genes were synonymously evolved for evolutionary time intervals corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and synonymously evolved sequences were calculated. The panels display the resulting K_S estimate frequency distributions. The ordinate scale varies between panels.

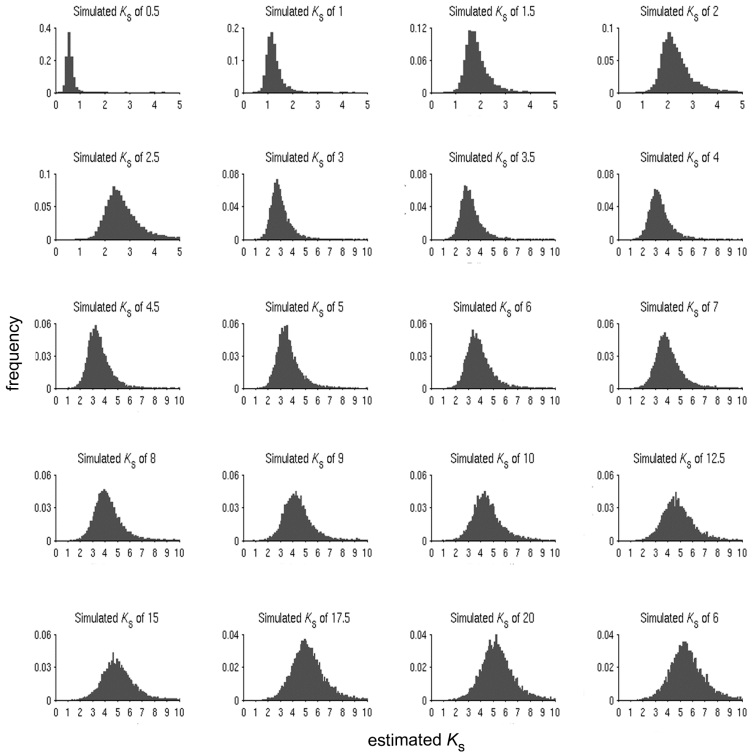


Figure E.6: Detailed results of synonymous evolution simulations for *C. albicans*. 6,006 protein coding genes were synonymously evolved for evolutionary time intervals corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and synonymously evolved sequences were calculated. The panels display the resulting K_S estimate frequency distributions. The ordinate scale varies between panels.

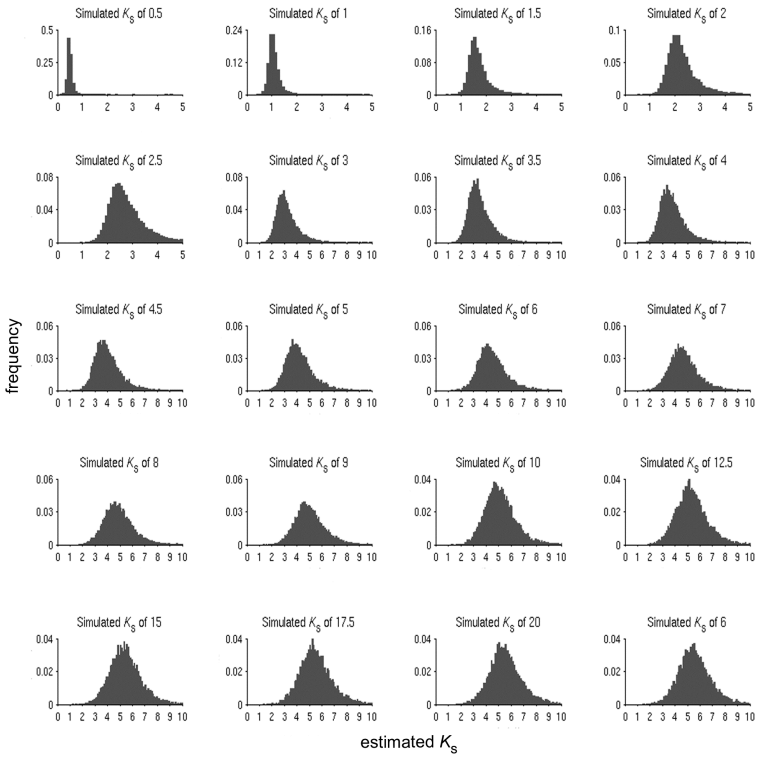


Figure E.7: Detailed results of synonymous evolution simulations for *C. intestinalis*. 9,330 protein coding genes were synonymously evolved for evolutionary time intervals corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and synonymously evolved sequences were calculated. The panels display the resulting K_S estimate frequency distributions. The ordinate scale varies between panels.

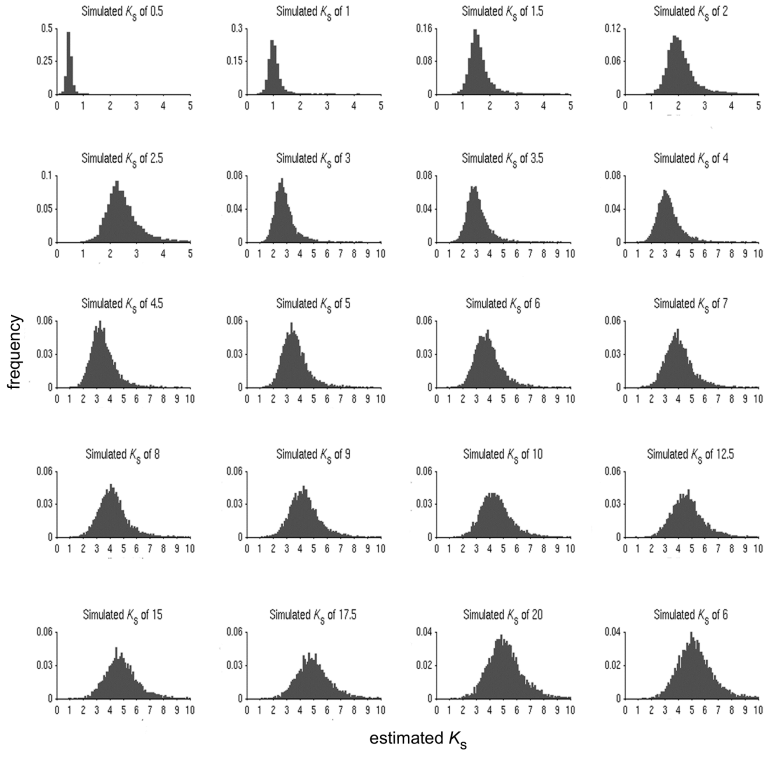


Figure E.8: Detailed results of synonymous evolution simulations for *K. lactis*. 5,076 protein coding genes were synonymously evolved for evolutionary time intervals corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and synonymously evolved sequences were calculated. The panels display the resulting K_S estimate frequency distributions. The ordinate scale varies between panels.

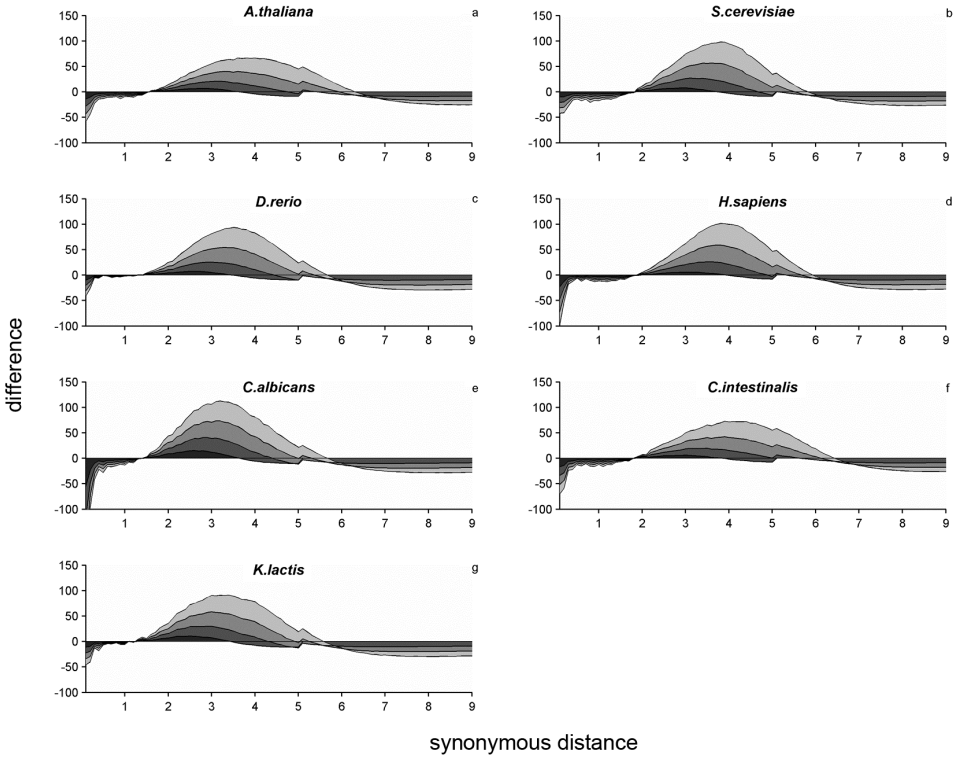


Figure E.9: Difference between simulated real and K_S-based age distributions for different species as indicated on top of the panels. The K_S-based age distributions of figure 3.3b-h were subtracted from the 'real age' distribution of figure 3.3a to obtain their difference, and to locate the precise location of the saturation peak for increasing evolutionary timespans as indicated on the color legend of figure 3.3a.

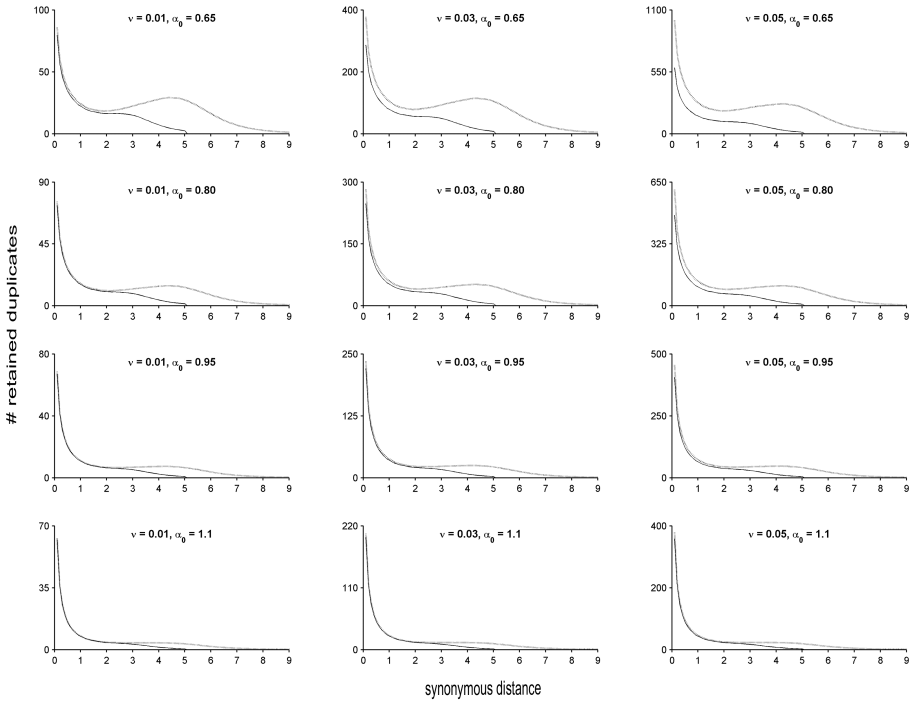


Figure E.10: The occurrence of SSD saturation peaks is not dependent on duplicate birth and death rates for *A. thaliana*. K_S-based age distributions were simulated using the duplicate population dynamics model and smoothing procedure described in the material and methods. The birth rate of new genes, ν, was varied over the range [0.01, 0.03, 0.05], and combined with variation of the power law decay constant of SSD duplicates, α_0, over the range [0.65, 0.80, 0.95, 1.10]. G_0 was kept at 10,000 for all simulations. Age distributions simulated over evolutionary timespans of 5 and 20 are indicated by solid and dotted lines, respectively.

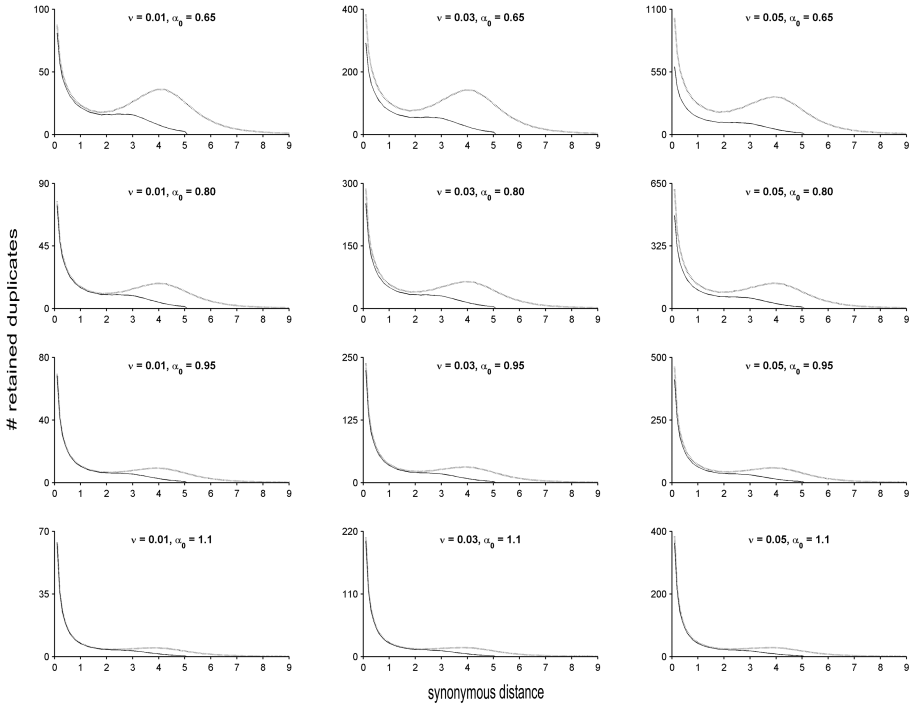


Figure E.11: The occurrence of SSD saturation peaks is not dependent on duplicate birth and death rates for *S. cerevisiae*. K_S-based age distributions were simulated using the duplicate population dynamics model and smoothing procedure described in the material and methods. The birth rate of new genes, ν, was varied over the range [0.01, 0.03, 0.05], and combined with variation of the power law decay constant of SSD duplicates, α_0, over the range [0.65, 0.80, 0.95, 1.10]. G_0 was kept at 10,000 for all simulations. Age distributions simulated over evolutionary timespans of 5 and 20 are indicated by solid and dotted lines, respectively.

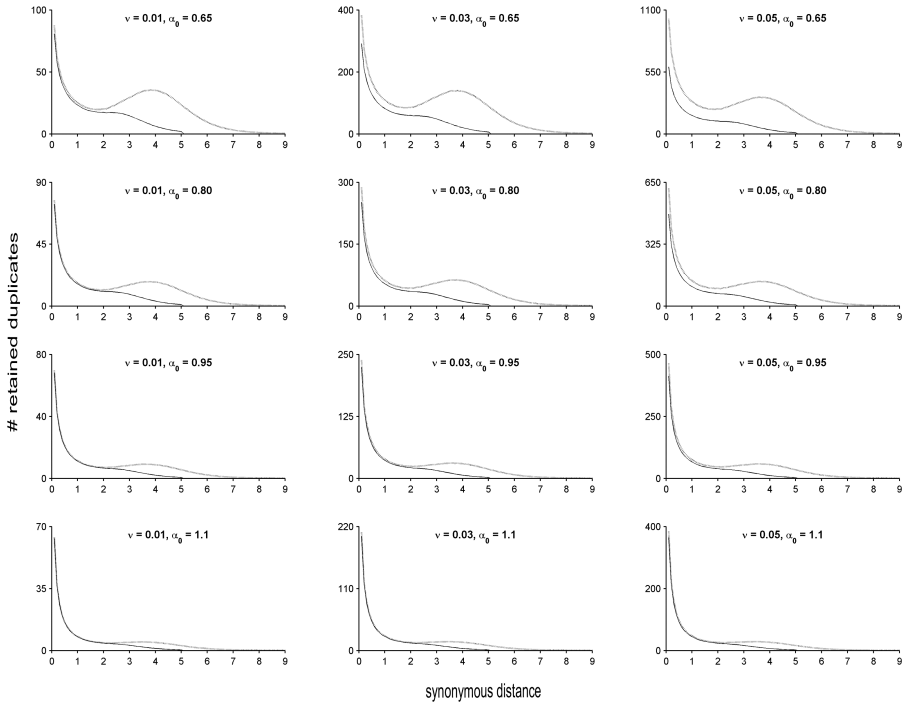


Figure E.12: The occurrence of SSD saturation peaks is not dependent on duplicate birth and death rates for *D. rerio*. K_S-based age distributions were simulated using the duplicate population dynamics model and smoothing procedure described in the material and methods. The birth rate of new genes, ν, was varied over the range [0.01, 0.03, 0.05], and combined with variation of the power law decay constant of SSD duplicates, α_0, over the range [0.65, 0.80, 0.95, 1.10]. G_0 was kept at 10,000 for all simulations. Age distributions simulated over evolutionary timespans of 5 and 20 are indicated by solid and dotted lines, respectively.

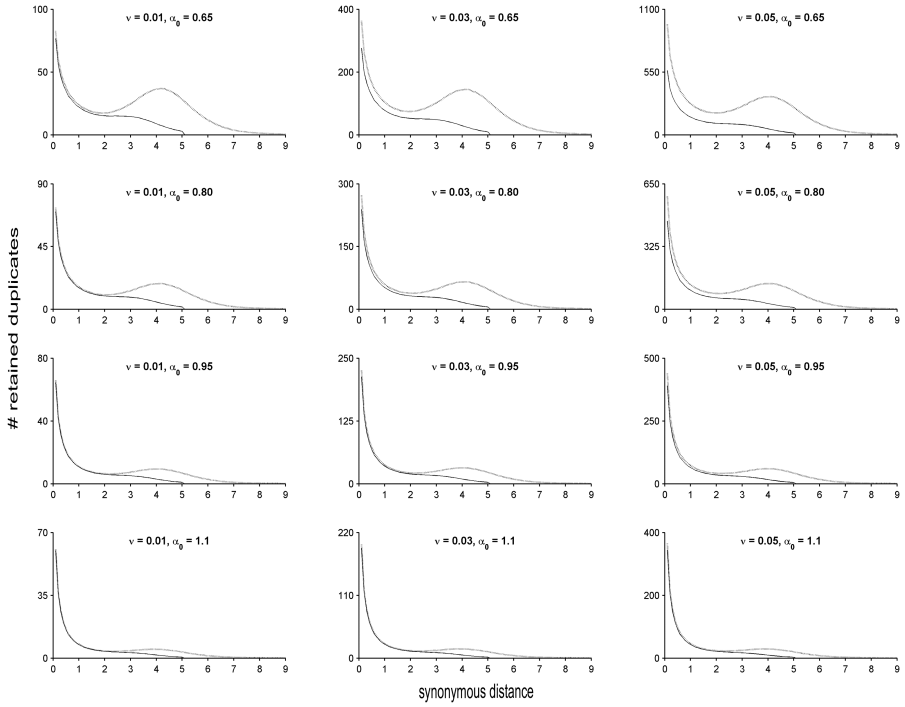


Figure E.13: The occurrence of SSD saturation peaks is not dependent on duplicate birth and death rates for *H. sapiens*. K_S-based age distributions were simulated using the duplicate population dynamics model and smoothing procedure described in the material and methods. The birth rate of new genes, ν, was varied over the range [0.01, 0.03, 0.05], and combined with variation of the power law decay constant of SSD duplicates, α_0, over the range [0.65, 0.80, 0.95, 1.10]. G_0 was kept at 10,000 for all simulations. Age distributions simulated over evolutionary timespans of 5 and 20 are indicated by solid and dotted lines, respectively.

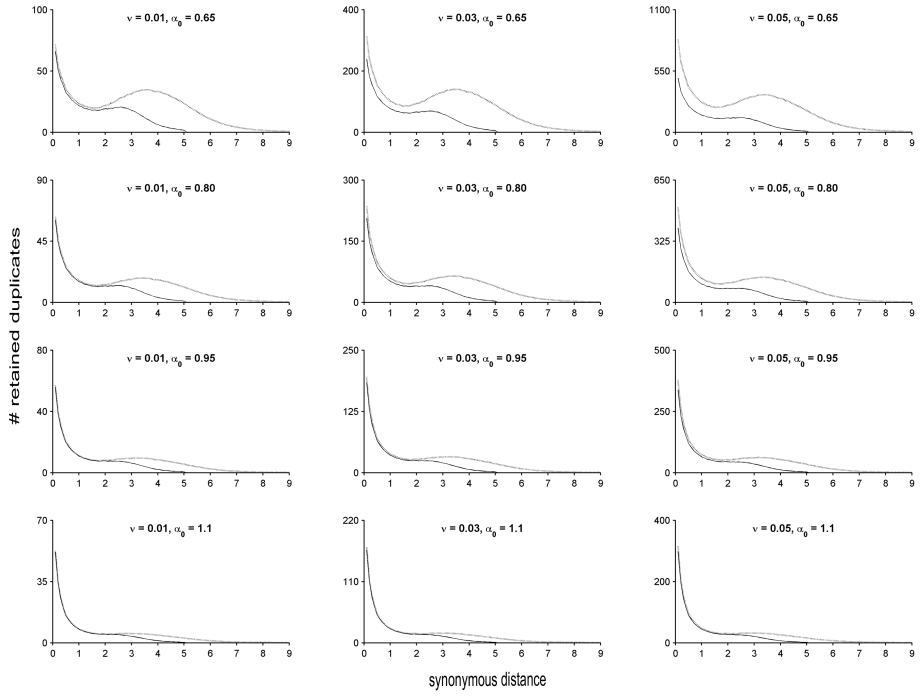


Figure E.14: The occurrence of SSD saturation peaks is not dependent on duplicate birth and death rates for *C. albicans*. K_S-based age distributions were simulated using the duplicate population dynamics model and smoothing procedure described in the material and methods. The birth rate of new genes, ν, was varied over the range [0.01, 0.03, 0.05], and combined with variation of the power law decay constant of SSD duplicates, α_0, over the range [0.65, 0.80, 0.95, 1.10]. G_0 was kept at 10,000 for all simulations. Age distributions simulated over evolutionary timespans of 5 and 20 are indicated by solid and dotted lines, respectively.

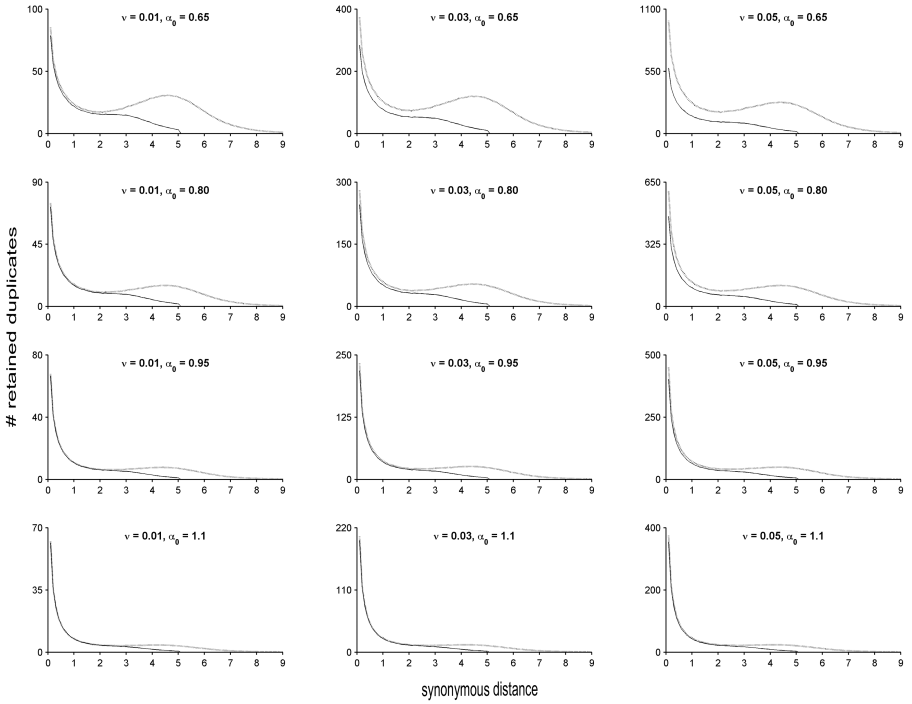


Figure E.15: The occurrence of SSD saturation peaks is not dependent on duplicate birth and death rates for *C. intestinalis*. K_S-based age distributions were simulated using the duplicate population dynamics model and smoothing procedure described in the material and methods. The birth rate of new genes, ν, was varied over the range [0.01, 0.03, 0.05], and combined with variation of the power law decay constant of SSD duplicates, α_0, over the range [0.65, 0.80, 0.95, 1.10]. G_0 was kept at 10,000 for all simulations. Age distributions simulated over evolutionary timespans of 5 and 20 are indicated by solid and dotted lines, respectively.

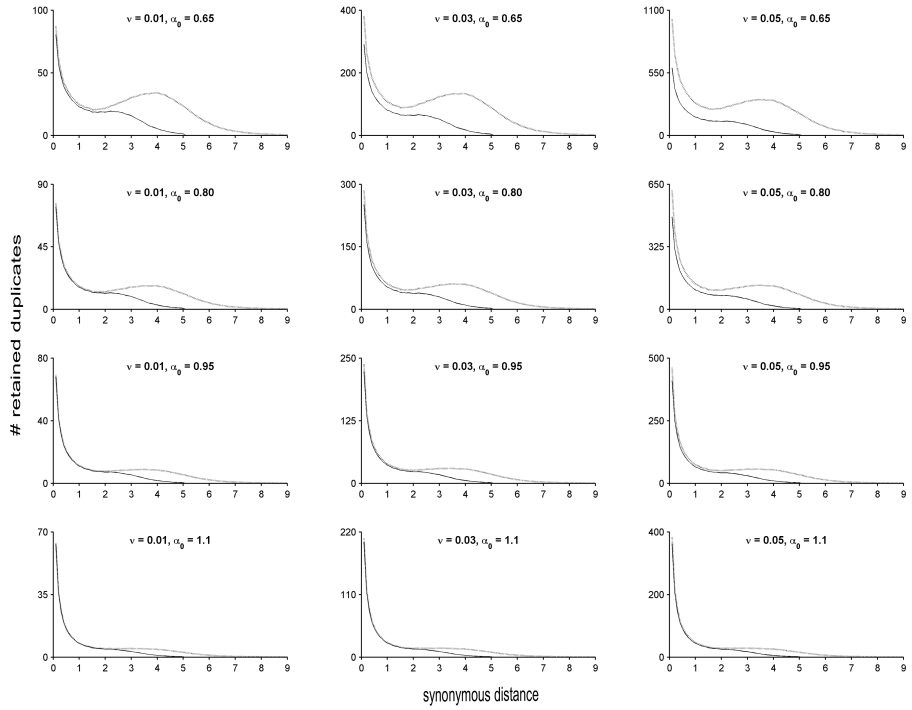


Figure E.16: The occurrence of SSD saturation peaks is not dependent on duplicate birth and death rates for *K. lactis*. K_S-based age distributions were simulated using the duplicate population dynamics model and smoothing procedure described in the material and methods. The birth rate of new genes, ν, was varied over the range [0.01, 0.03, 0.05], and combined with variation of the power law decay constant of SSD duplicates, α_0, over the range [0.65, 0.80, 0.95, 1.10]. G_0 was kept at 10,000 for all simulations. Age distributions simulated over evolutionary timespans of 5 and 20 are indicated by solid and dotted lines, respectively.

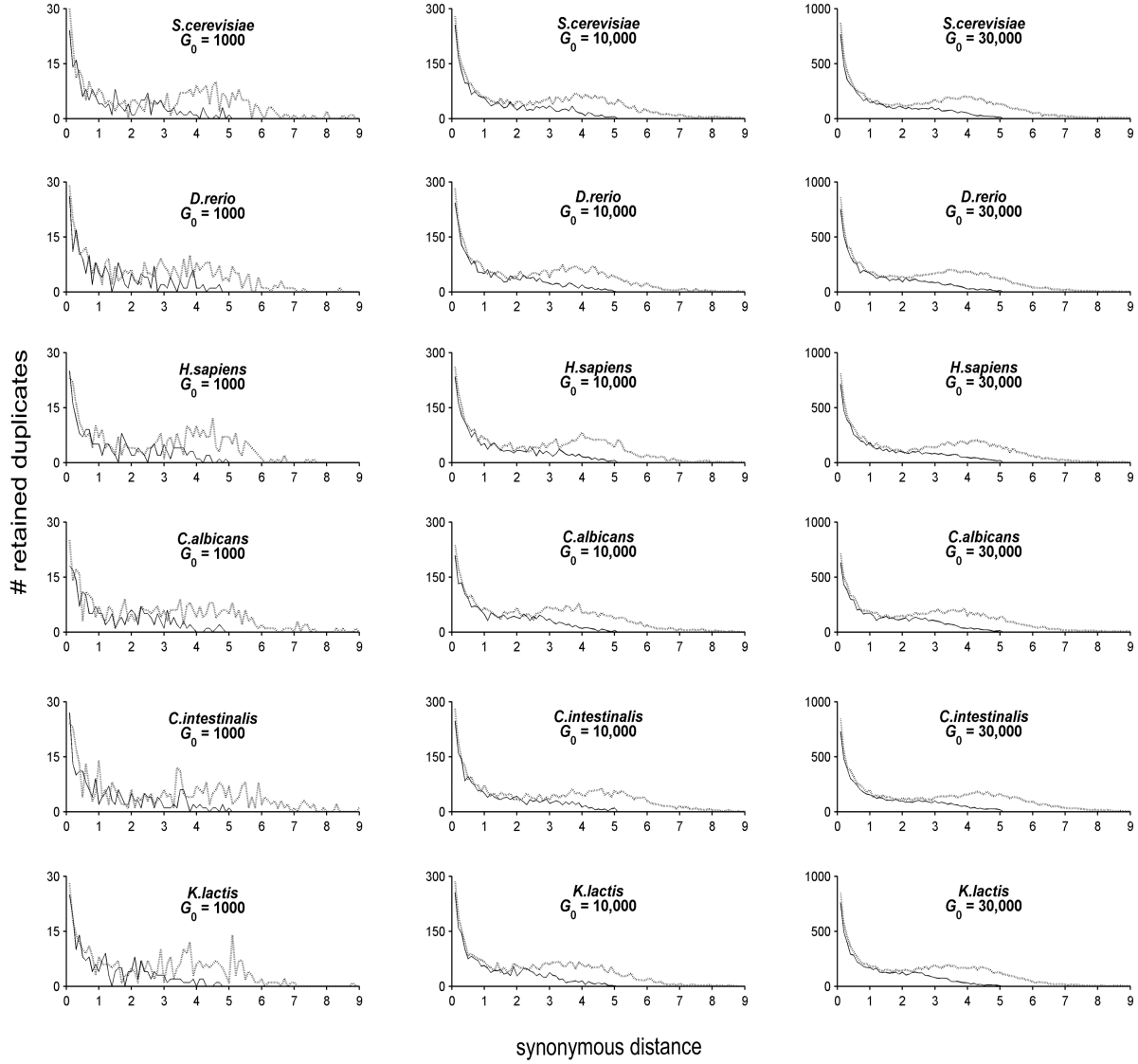


Figure E.17: The number of genes in the age distribution impacts its shape. Results displayed for *S. cerevisiae*, *D. rerio*, *H. sapiens*, *C. albicans*, *C. intestinalis*, and *K. lactis*. SSD 'real age' distributions incorporating increasing numbers of duplicates were simulated by increasing the number of founder genes G_0 from 1000 to 10,000 and 30,000. Values for α_0 and ν were kept at 0.80 and 0.03. Afterwards, a K_S estimate resampling procedure was performed to incorporate K_S saturation and stochasticity effects, as described in the material and methods. In each panel, age distributions simulated over evolutionary timespans of 5 and 20 are indicated by the solid and dotted lines, respectively.

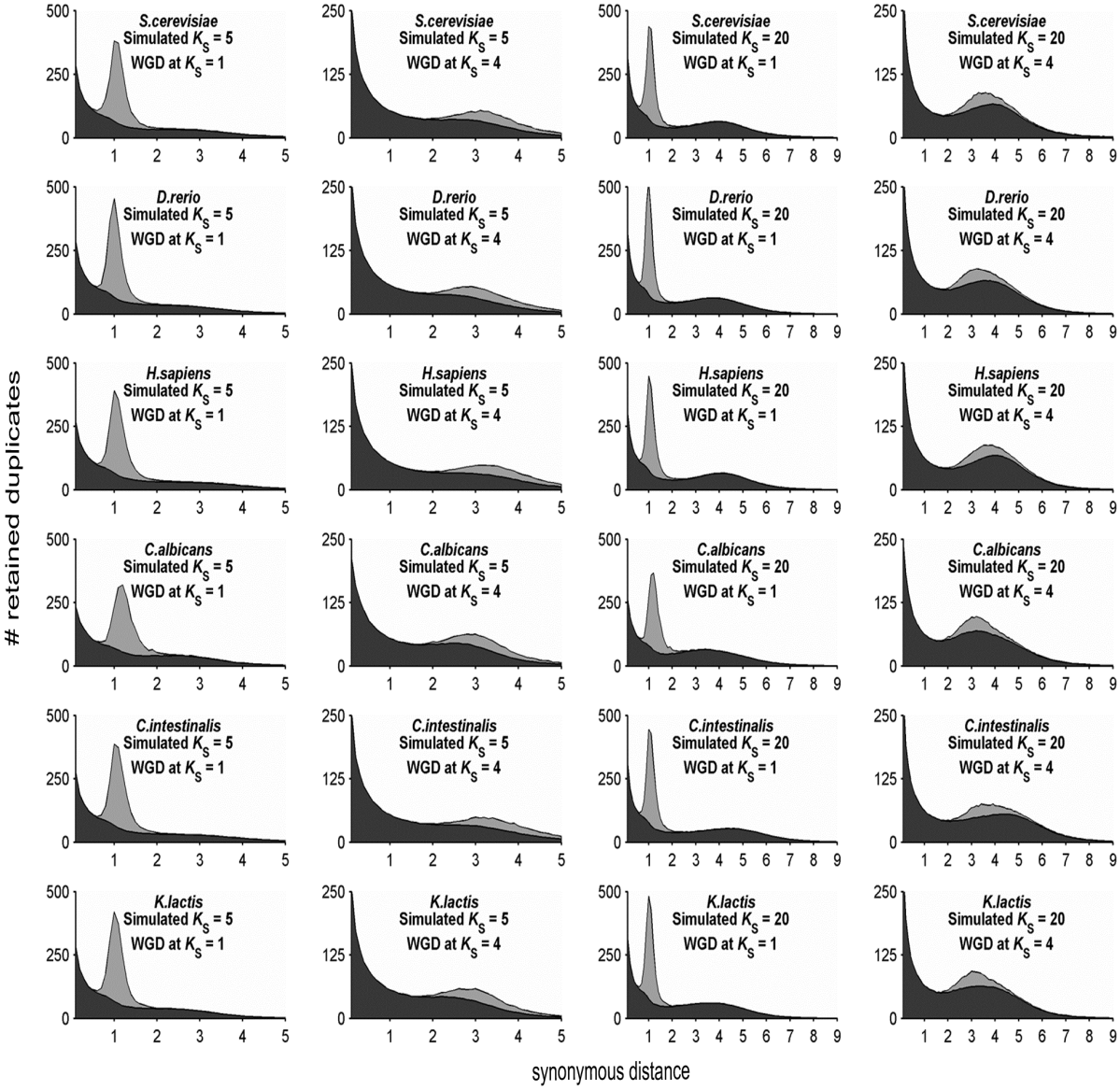


Figure E.18: K_S **stochasticity and saturation also affect WGD events.** Results displayed for *S. cerevisiae*, *D. rerio*, *H. sapiens*, *C. albicans*, *C. intestinalis*, and *K. lactis*. Simulated age distributions over evolutionary timespans of 5 and 20 were created using our duplicate population dynamics model, taking into account an SSD background duplication mode as well as WGD events at synonymous ages of 1 or 4. Other model parameters were set as follows: G_0=10,000, α_0=0.80, α_1=0.90, and ν=0.03 for all scenarios.

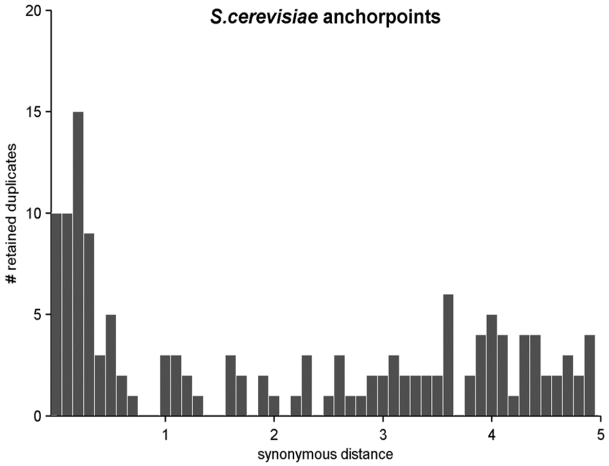


Figure E.19: *S. cerevisiae* anchorpoint K_S distribution. *S. cerevisiae* WGD duplicate pairs and their alignments were taken from Kellis et al.[134]. K_S estimates were obtained through maximum likelihood estimation (MLE) using the CODEML program[204] of the PAML package (v4.4c)[250], as described in the material and methods. Since each anchorpoint pair represents one duplication event, no correction for redundancy in K_S estimates was however carried out afterwards.

E.2 Supplementary tables

Table E.1: Detailed results of simulating synonymous evolution for all species

Species	Simulated K_S	Lower SD	Geometric mean	Upper SD	Modus
	1	0.87	1.1	1.41	1
	2	1.68	2.23	2.97	2
	3	2.3	3.05	4.04	2.6
	4	2.73	3.59	4.72	3.3
	5	3.04	3.98	5.22	3.6
	6	3.29	4.3	5.61	4.2
	7	3.49	4.55	5.93	4.3
A. thaliana	8	3.65	4.75	6.17	4.6
	9	3.78	4.91	6.39	4.4
	10	3.89	5.05	6.55	4.8
	12.5	4.08	5.28	6.84	5
	15	4.23	5.44	7.01	5.2
	17.5	4.28	5.52	7.11	5.4
	20	4.33	5.59	7.21	5.4
	25	4.39	5.65	7.28	5.3
	1	0.87	1.12	1.44	1.1
	2	1.72	2.29	3.06	2
	3	2.33	3.09	4.1	2.9
	4	2.73	3.59	4.72	3.3
	5	2.99	3.92	5.13	3.8
	6	3.15	4.14	5.44	3.9
	7	3.28	4.28	5.59	4
S. cerevisiae	8	3.38	4.41	5.75	4.2
	9	3.47	4.53	5.91	4.2
	10	3.54	4.61	6.01	4.6
	12.5	3.67	4.79	6.25	4.7
	15	3.75	4.92	6.44	4.8
	17.5	3.86	5.03	6.54	4.9
	20	3.93	5.13	6.69	4.7
	25	4.06	5.3	6.92	5.2
	1	0.86	1.05	1.29	1
	2	1.62	2.12	2.78	1.9

Continued on next page

Table E.1 – Continued from previous page

Species	Simulated K_S	Lower SD	Geometric mean	Upper SD	Modus
	3	2.22	2.92	3.83	2.5
	4	2.59	3.38	4.42	3.2
	5	2.82	3.68	4.79	3.4
	6	3	3.9	5.08	3.6
	7	3.12	4.06	5.28	3.8
D. rerio	8	3.22	4.19	5.44	3.9
	9	3.3	4.29	5.58	4.1
	10	3.37	4.38	5.69	4.2
	12.5	3.51	4.55	5.9	4.5
	15	3.59	4.66	6.03	4.5
	17.5	3.67	4.74	6.12	4.6
	20	3.7	4.79	6.19	4.7
	25	3.75	4.84	6.26	4.7
	1	0.86	1.13	1.49	1
	2	1.72	2.34	3.17	2
	3	2.4	3.2	4.25	2.7
	4	2.81	3.68	4.82	3.5
	5	3.05	3.98	5.18	3.8
	6	3.21	4.18	5.44	3.8
	7	3.34	4.33	5.62	4.2
H. sapiens	8	3.43	4.46	5.78	4.4
	9	3.51	4.55	5.91	4.2
	10	3.57	4.63	6.01	4.3
	12.5	3.67	4.76	6.17	4.7
	15	3.75	4.85	6.27	4.8
	17.5	3.79	4.9	6.33	4.6
	20	3.81	4.93	6.36	4.9
	25	3.83	4.95	6.4	5
	1	0.99	1.26	1.59	1.2
	2	1.82	2.35	3.03	2.1
	3	2.3	2.96	3.8	2.8
	4	2.59	3.32	4.27	3
	5	2.78	3.59	4.62	3.6
	6	2.96	3.82	4.92	3.4
	7	3.1	4	5.16	3.8
C. albicans	8	3.23	4.17	5.38	4
	9	3.35	4.33	5.6	4.3
	10	3.45	4.47	5.8	4.4
	12.5	3.68	4.77	6.18	4.7
	15	3.86	5	6.49	4.7
	17.5	3.99	5.18	6.72	5
	20	4.11	5.31	6.85	5.3
	25	4.24	5.49	7.12	5.3
	1	0.9	1.12	1.4	1
	2	1.77	2.31	3.02	2
	3	2.44	3.18	4.14	3
	4	2.89	3.75	4.86	3.4
	5	3.21	4.16	5.39	3.7
	6	3.44	4.45	5.75	4.1
	7	3.63	4.66	5.98	4.3
C. intestinalis	8	3.77	4.84	6.2	4.6
	9	3.89	4.98	6.39	4.7
	10	3.97	5.09	6.52	4.7
	12.5	4.13	5.28	6.74	5.2
	15	4.22	5.38	6.87	5.4
	17.5	4.26	5.44	6.94	5.3
	20	4.32	5.49	6.99	5.1
	25	4.31	5.5	7.02	5.6

Continued on next page

Table E.1 – Continued from previous page

Species	Simulated K_S	Lower SD	Geometric mean	Upper SD	Modus
	1	0.88	1.06	1.29	1
	2	1.65	2.1	2.66	1.9
	3	2.19	2.8	3.58	2.7
	4	2.52	3.23	4.15	3
	5	2.77	3.56	4.57	3.4
	6	2.96	3.81	4.91	3.9
	7	3.1	4	5.16	4
K. lactis	8	3.22	4.16	5.37	4.1
	9	3.29	4.25	5.49	4.3
	10	3.41	4.38	5.64	4
	12.5	3.58	4.61	5.94	4.8
	15	3.72	4.78	6.14	4.5
	17.5	3.81	4.91	6.34	4.7
	20	3.91	5.03	6.47	4.9
	25	4.03	5.19	6.68	5

Table E.2: Detailed synonymous evolution simulation results for *A. thaliana*. Mean and SD of K_S estimates for a simulated K_S of 0.7 and 0.8 are similar to those of the 242 contemporaneously duplicated gene pairs (mean=0.82 and SD=0.36) in Zhang et al.[360].

Simulated K_S	Mean of K_S estimates	SD of K_S estimates
0.1	0.11	0.03
0.2	0.21	0.05
0.3	0.32	0.07
0.4	0.43	0.11
0.5	0.54	0.16
0.6	0.66	0.21
0.7	0.77	0.26
0.8	0.89	0.32
0.9	1.02	0.38
1	1.15	0.53

E.3 Supplementary information

E.3.1 Introduction

The results presented in the main text hinge on the accuracy of the codon model used in the evolutionary simulations. We employed a simplified version of the codon model of Yang and Nielsen[299] that only considers synonymous evolution as proof-of-concept to demonstrate that saturation dynamics for different species lead to diffuse SSD saturation peaks in (K_S-based) age distributions. In support of our approach, the genome-wide saturation dynamics described in the main text are in qualitative agreement with previous smaller-scale empirical examples[309–311]. Additionally, the K_S stochasticity effects observed in our genome-wide simulations for *A. thaliana* at a synonymous age of 0.7-0.8 were in quantitative agreement with an empirical example of 242 simultaneously duplicated gene pairs remaining from the most recent WGD in the *A. thaliana* lineage[360] (see supplementary table E.2).

It could nevertheless be argued that our simulation strategy explores a special case of sequence evolution that seems especially implausible for recently duplicated genes that typically undergo a period of accelerated non-synonymous sequence evolution. Additionally, the space of possible mutations is limited because some changes between synonymous codons of the same codon set are never possible as they

require a non-synonymous intermediate (e.g., Serine), or there are no synonymous partners in the codon set (e.g., Methionine). Here, we consider the more complex scenario where non-synonymous mutations are allowed, corresponding to the full form of the codon model as specified by Yang and Nielsen[299], and demonstrate based on the paranome of *A. thaliana* this does not qualitatively change the observations and interpretations presented in the main text.

E.3.2 Material and methods

Characterization of K_S stochasticity and saturation effects based on evolutionary simulations that incorporate both synonymous and non-synonymous changes

The full form of the codon model of Yang and Nielsen[299] is as follows:

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\ \kappa\pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\ \omega\pi_j & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion} \\ \omega\kappa\pi_j & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition} \end{cases} \quad \text{(E.1)}$$

Values for the equilibrium frequencies π_j and the transition/transversion rate ratio parameter κ were extracted as described in the main text. Parameter ω serves as a measure for selection strength because it relates the number of synonymous and non-synonymous changes (ω=K_N/K_S). We performed non-synonymous evolution for different values of ω: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. We did not consider higher values because a value of one already entails a highly relaxed constraint, which is typically only encountered for a few select sites in genes under strong positive selection, but is unlikely to apply to the full gene sequence[471]. The simulation of evolution using a Markov chain with both the waiting times and jump chain probabilities fully specified by the substitution rate matrix Q=$\{q_{ij}\}$ was performed as described in the main text, except for the total evolutionary time the simulation was run, because the latter now also needs to accommodate for non-synonymous mutations. Based on equation 3.4, and since ω=K_N/K_S, the total evolutionary time becomes:

$$t = K_S \frac{S}{L_c} + \omega K_S \frac{N}{L_c} \quad \text{(E.2)}$$

S/L_c and N/L_c represent the genome-wide average number of synonymous sites per codon and non-synonymous sites per codon, respectively. We performed the simulations with different values of ω for 20,000 randomly chosen genes from the *A. thaliana* genome. We evolved genes in time step equivalents corresponding to an average K_S increase of 0.1, until an average K_S of 20 had been reached, after which K_S estimates between the original and evolved sequences at each time step were re-estimated using CODEML[204]. The incorporation of non-synonymous evolution resulted in a large proportion of obvious outliners at higher synonymous ages, especially for high values of ω, because many of these genes had undergone multiple non-synonymous substitutions per non-synonymous site (see further). We therefore employed the Thompson-Tau method[472] to remove these outliers. The resulting frequency distributions

of K_S estimates for different ω values and synonymous ages are presented in figures E.20 to E.29, on which the frequency distributions of K_S estimates with ω=0, and after Thompson-Tau cleaning, is also indicated to allow comparison with the results of the main text.

The impact of saturation effects on age distributions

We used our duplicate population dynamics model to construct 'real age' SSD distributions as described the main text, equations 3.6 - 3.9, using the same standard model parameters (G_0=10,000, ν=0.03, and α_0=0.80). The transformation of these real age SSD distributions into K_S-based age distributions is however less straightforward under different selection regimes (i.e., values of ω) that are expected for different duplicate pairs in an empirical age distribution. Newly duplicated genes are expected to undergo a period of relaxed selection, characterized by high values of ω, while genes present in the tail of the distribution are expected to be under strong purifying selection, characterized by low values of ω[473]. Converting a real age distribution into a K_S-based distribution based on the K_S estimation biases gathered from evolutionary simulation data for just one value of ω, i.e., equation 3.10 of the main text, therefore does not capture the intricacies present in empirical age distributions.

To characterize how ω varies between duplicate pairs of different age, we plotted the values of ω (calculated using CODEML) between all duplicated gene family members of the *A. thaliana* paranome (identified as described in the main text) against their duplication age, using K_S as a proxy for time since duplication. The results are presented in figure E.30, hereafter referred to as the 'ω paranome plot', and confirm our assumptions about different selection pressures for genes of different ages. Newly duplicated genes ($0<K_S \leq 1$) exhibit a wide variety of associated ω values, with many genes still exhibiting a strong selection pressure, as could be explained by mechanisms such as dosage amplification or strongly conserved evolutionary processes; but many genes also exhibit relaxed selection, as could be explained by duplicate genes that are on route to being pseudogenized[308]. The associated values of ω show a steady decrease for genes of intermediate ages ($1<K_S \leq 2.5$), until they seem to stabilize at an average ω of about 0.1-0.2 for older genes ($2.5<K_S \leq 5$), indicating these genes are under strong purifying selection.

We therefore took the following approach, which is also illustrated in figure E.31 for improved clarity. We know the 'true K_S age' of all genes present in the real age distribution. To get a representative estimate as to which 'observed K_S age' this corresponds after incorporation of K_S noise, we employed the synonymous evolutionary data for which ω=0. More specifically, we sampled the observed K_S age from the range of K_S estimates corresponding to a simulated synonymous age equal to the true K_S age, with a probability proportional to the frequency of individual K_S estimates. We then referred to the ω paranome plot (see figure E.30) to get an estimate of a representative ω value for that observed K_S age. This was done by sampling from a probability distribution where each individual ω value has a probability equal to its frequency in the ω paranome plot for a certain observed K_S age.

This ω estimate is however biased, because the ω paranome plot is not a real age plot, but rather also the result of K_S estimation bias. To remove this bias, we used a stepping bridge rule. Although both the observed K_S ages and ω values in the ω paranome plot are biased through K_S estimation bias, we assume their corresponding K_N values (easily extracted by the relationship $K_N=\omega K_S$) are not biased, or at least to a much lesser extent. Multiple non-synonymous back-substitutions, and their associated

problems with K_N saturation and stochasticity, are expected to start at an average K_N value of 1. This would effectively mean each non-synonymous site in the sequence has undergone one non-synonymous substitution on average. It is however very unlikely such situations are encountered in the ω paranome plot, since such sequences would *a priori* be unidentifiable according to the Li-Rost criterion of 30% sequence identity for gene families[329,474]. We therefore effectively know the real K_S age and have an unbiased estimate for its associated real K_N age. Using the same relationship as before (ω=K_N/K_S), this enables us to calculate an unbiased ω estimate for each duplicate pair present in our real age distribution. Through our saturation dynamics profiles for different synonymous ages and ω values (see figures E.20-E.29), we can thus transform the real age distribution into a K_S-based age distribution. The results are presented in figures E.32-E.33.

E.3.3 Results and discussion

Characterization of K_S stochasticity and saturation effects based on evolutionary simulations that incorporate both synonymous and non-synonymous changes

Figures E.20 to E.29 present the results of the evolutionary simulations for different values of ω in *A. thaliana*. Results on each figure are displayed as the frequency distributions of K_S estimates corresponding to a certain synonymous age for a particular value of ω and are indicated in red, while the frequency distributions of K_S estimates corresponding to the same synonymous age but with ω=0 (i.e., no selection pressure) are also indicated in black to allow comparison. For all values of ω considered, the red distribution of K_S estimates closely follows the black distribution until a synonymous age of 2 to 2.5. Both the modus (serving as a proxy for K_S saturation) and the spread (serving as a proxy for K_S stochasticity) thus follow the characteristics described in the main text where only ω=0 was considered, supporting our claim that K_S estimates generally can be considered trustworthy until a K_S value of 2-2.5.

After this cutoff, the profiles of the K_S estimate frequency distributions however start to vary markedly between different values of ω. At low values of ω (0.1-0.3), the modus of K_S estimates is always shifted to the right in respect to the K_S estimates without selection, especially at intermediate synonymous ages ($<$5-10). Although the spread of these K_S estimates is also larger compared to a scenario without selection, this suggests K_S saturation is of a lesser extent under these circumstances. At higher synonymous ages ($>$10), the spread of the K_S estimates however reaches dramatic proportions. This is to be expected, since a scenario with ω=0.1 effectively entails a K_N of 1 will be reached on average at a synonymous age of 10. Since every non-synonymous site will have undergone one non-synonymous substitution on average, the original and simulated sequence will have diverged to such an extent on an amino-acid level that they would not be recognized as being part of the same gene family following the Li-Rost criterion of 30% sequence identity for members of the same gene family[329,474]. At higher values of ω ($>$0.3), the modus of K_S estimates progressively shifts to the left in respect to the K_S estimates without selection, while their spread also increases drastically, indicating K_S saturation and stochasticity are more pronounced for increasingly younger synonymous ages. Following the above logic, this is also to be expected as a K_N of 1 will be reached on average at synonymous ages of 2.5, 2, 1.67, 1.43, 1.25, 1.12, and 1 for ω = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1; respectively.

In summary, for simulated synonymous ages until 2-2.5 and for different values of ω, distributional trends reported here follow those of the main text where no selection was present, reinforcing our notion that a K_S value of 2-2.5 is a good cut-off boundary. For higher simulated synonymous ages and low values of ω (and hence a slow rate of non-synonymous mutations), K_S saturation seems in fact to be more mitigated at the cost of more K_S stochasticity. For increasingly larger values of ω, the fast rate of non-synonymous mutations entails both drastic K_S saturation and stochasticity, but such scenarios are highly unlikely in real datasets.

The impact of saturation effects on age distributions

Figure E.32 demonstrates the effects of applying a correction for K_S saturation and stochasticity on real age distributions, based on evolutionary simulations that did also incorporate non-synonymous changes. The real age distributions of increasing synonymous age that are depicted in figure E.32a do not incorporate K_S noise, and are exactly the same as figure 3.3a of the main text. The K_S-based age distributions depicted in figure E.32b are the result of incorporating the K_S noise, and demonstrate diffuse SSD saturation peaks are still present. They however display a rugged curve surface as the transformation procedure was based on sampling ω estimates from a finite number of possibilities. Figure E.33 therefore compares the K_S-based age distributions of increasing synonymous ages based on different values of ω depicted in figure E.32b with a scenario where ω=0, but based on the same finite number of genes, the latter corresponding to figure 3.4b of the main text.

Figure E.33 demonstrates that despite the rugged surface curve, the existence of diffuse SSD saturation peaks is still evident when using evolutionary data where non-synonymous changes are also allowed. Moreover, the saturation peak mode across evolutionary timespans is the same as described in the main text. However, the peak amplitude across evolutionary timespans is lower. The inclusion of varying K_N/K_S ratios results in flattening out the SSD saturation peak to some extent, ranging from negligible to moderately evident across evolutionary timespans going from 5 to 20. As mentioned in the main text, it remains very difficult to assess to which corresponding peak amplitude the SSD saturation peak will correspond for empirical age distributions, as this is mainly dependent on how many ancient duplicates can still be identified. Assuming an average synonymous substitution rate in the order of 10 per synonymous site per billion years (from 2.5/ss/BY for mammals to 15/ss/BY for invertebrates[308]), duplicates with a synonymous age of 20 may be well over a billion years old, and therefore not recognizable anymore as such.

E.3.4 Conclusion

A more realistic evolutionary scenario that also incorporate non-synonymous evolution did not qualitatively change the findings described in the main text, namely that K_S estimates are more or less reliable until a synonymous age of 2-2.5, and that SSD age distributions are characterized by a diffuse saturation peak, which could easily be mistaken for a WGD signature. This demonstrates the validity of our approach using the simplified version of the full codon model outlined by Yang and Nielsen[299] that only considers synonymous evolution, and suggests it provides a reasonably good approximation for future work[313].

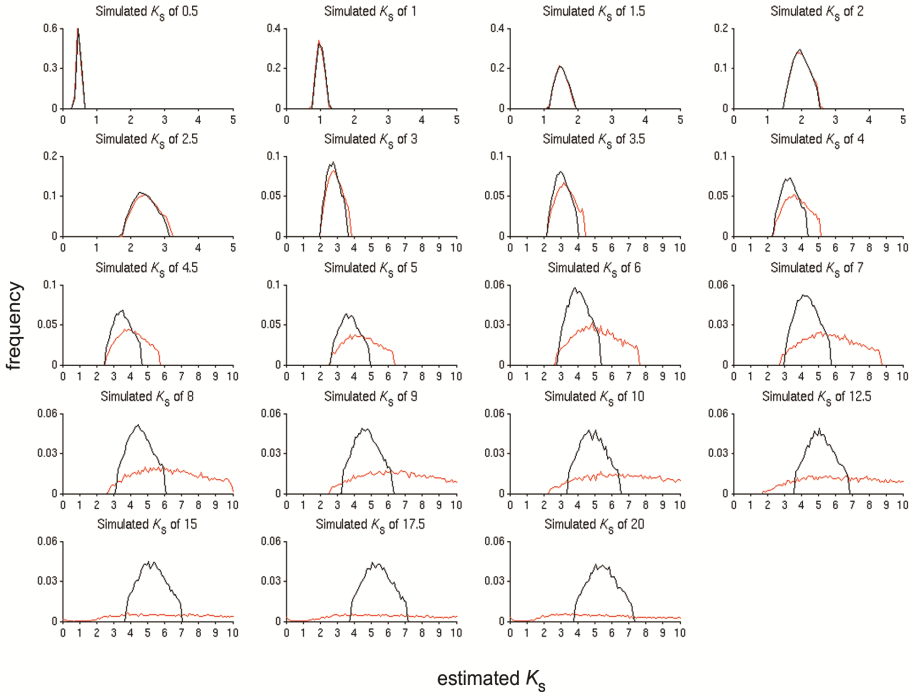


Figure E.20: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.1. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.1 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

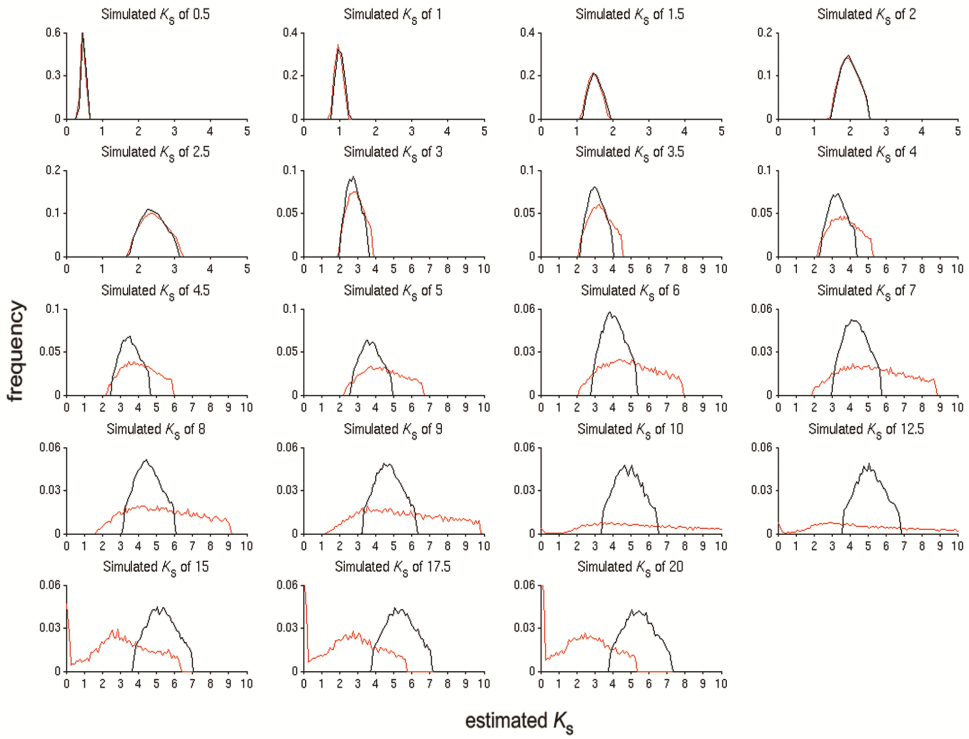


Figure E.21: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.2. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.2 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

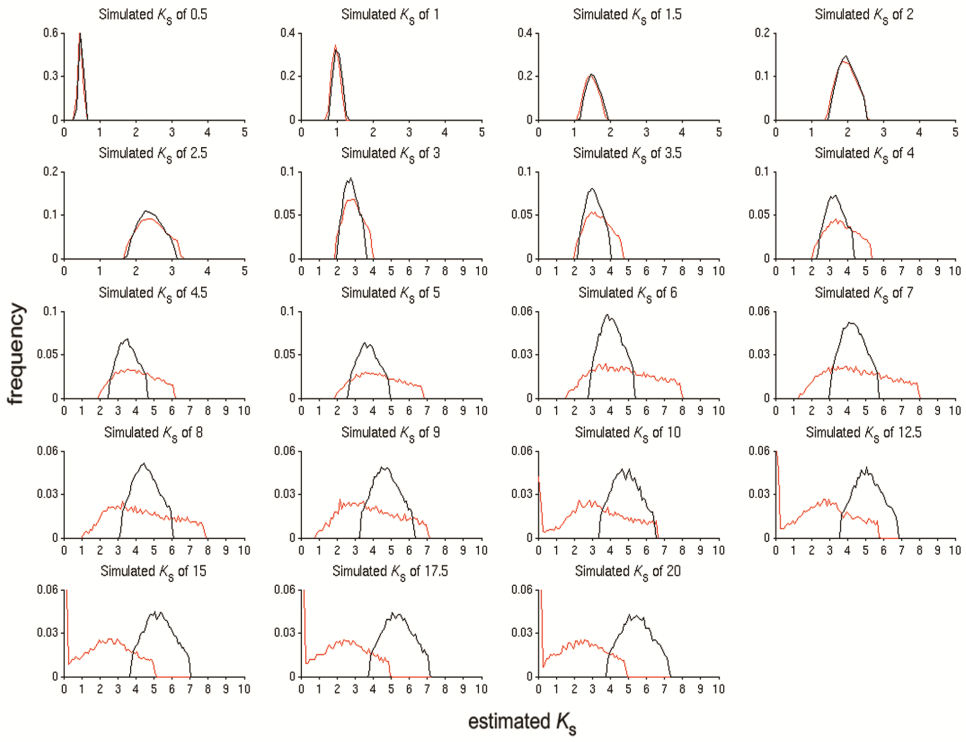


Figure E.22: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.3. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.3 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

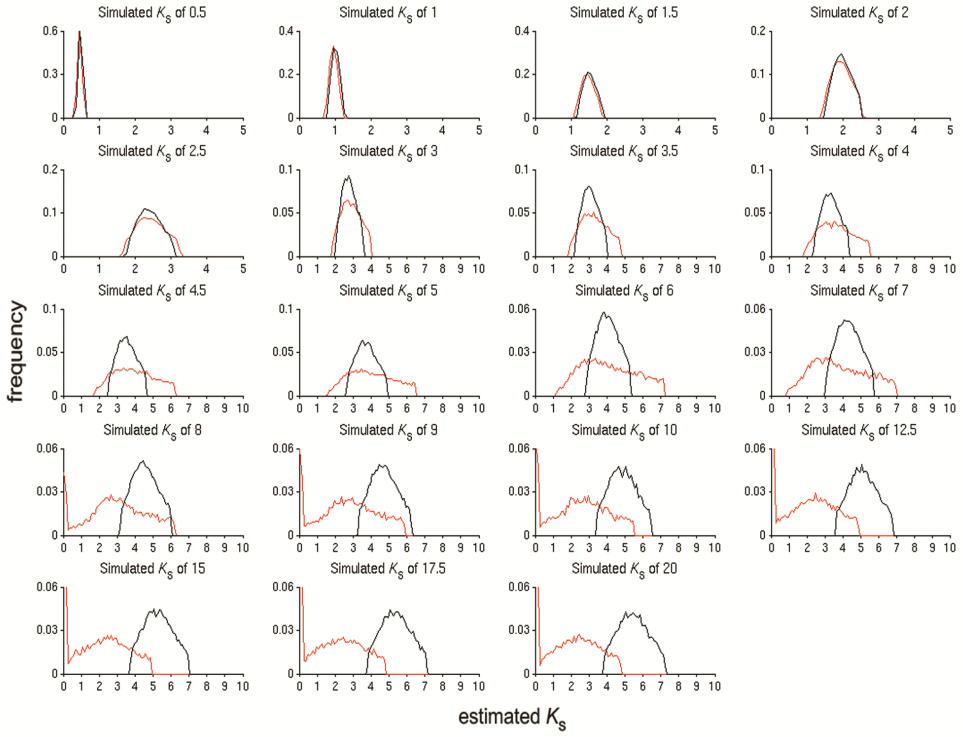


Figure E.23: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.4. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.4 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

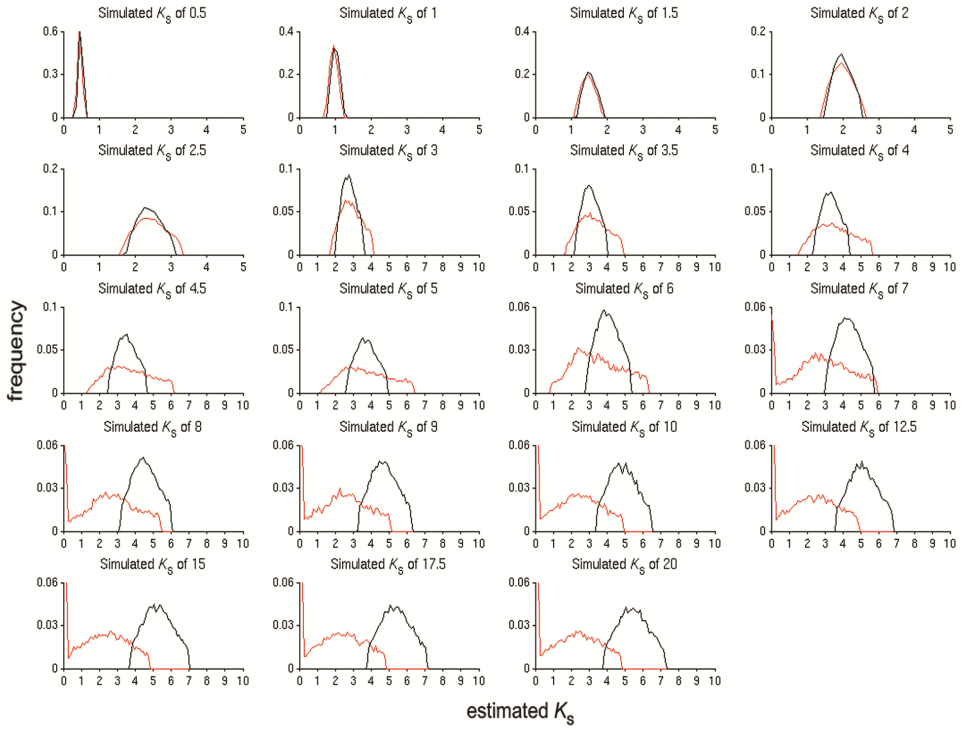


Figure E.24: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.5. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.5 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

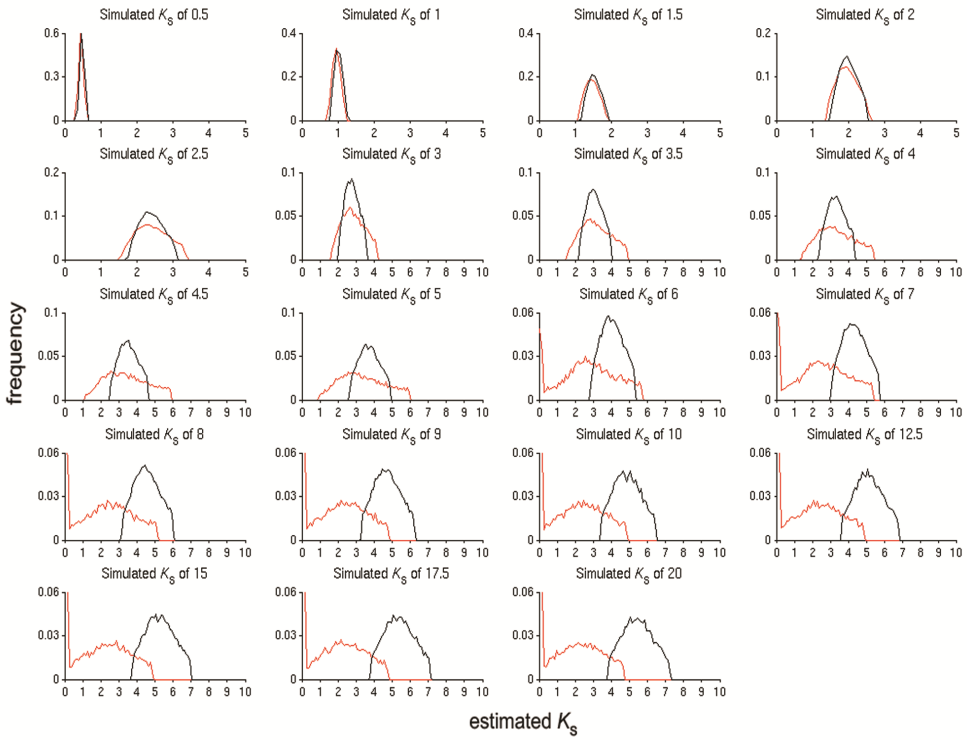


Figure E.25: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.6. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.6 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

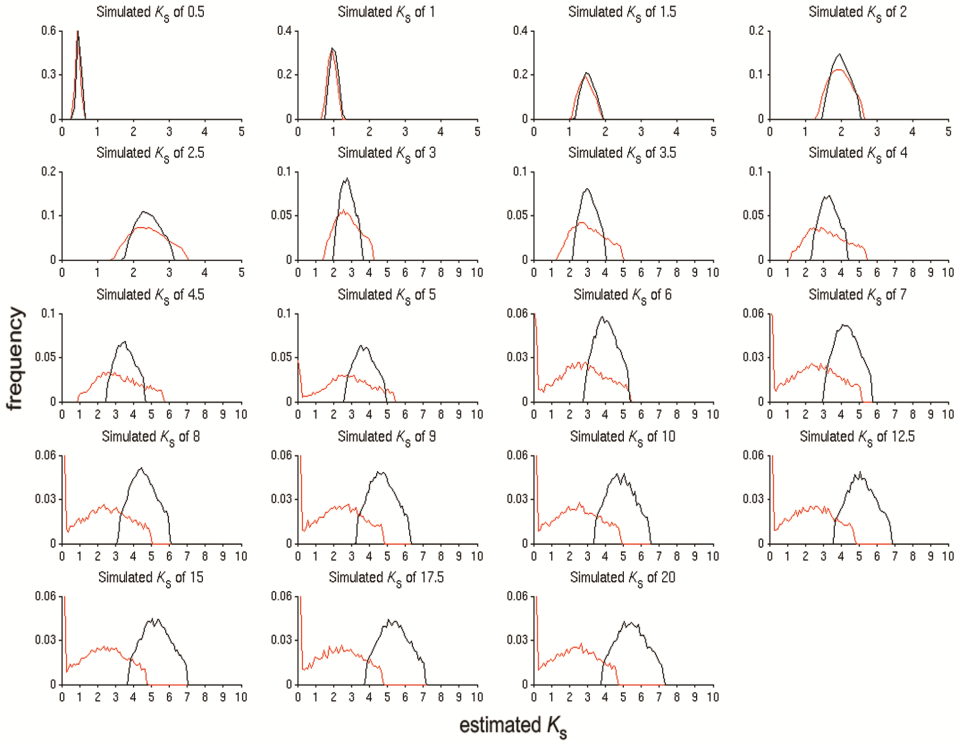


Figure E.26: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.7. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.7 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

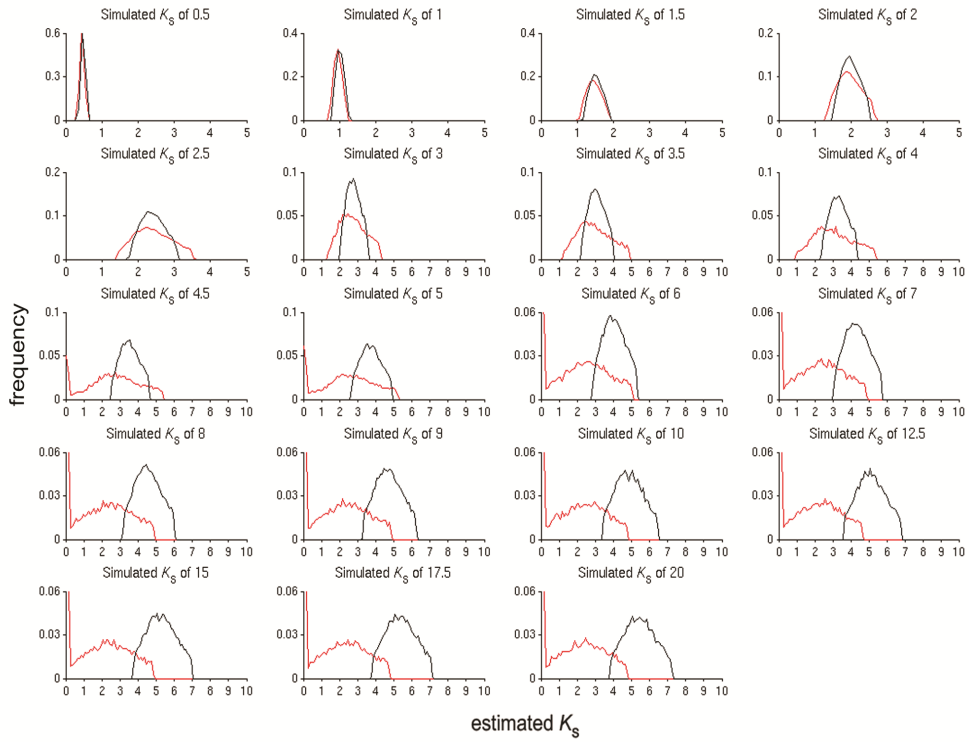


Figure E.27: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.8. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.8 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

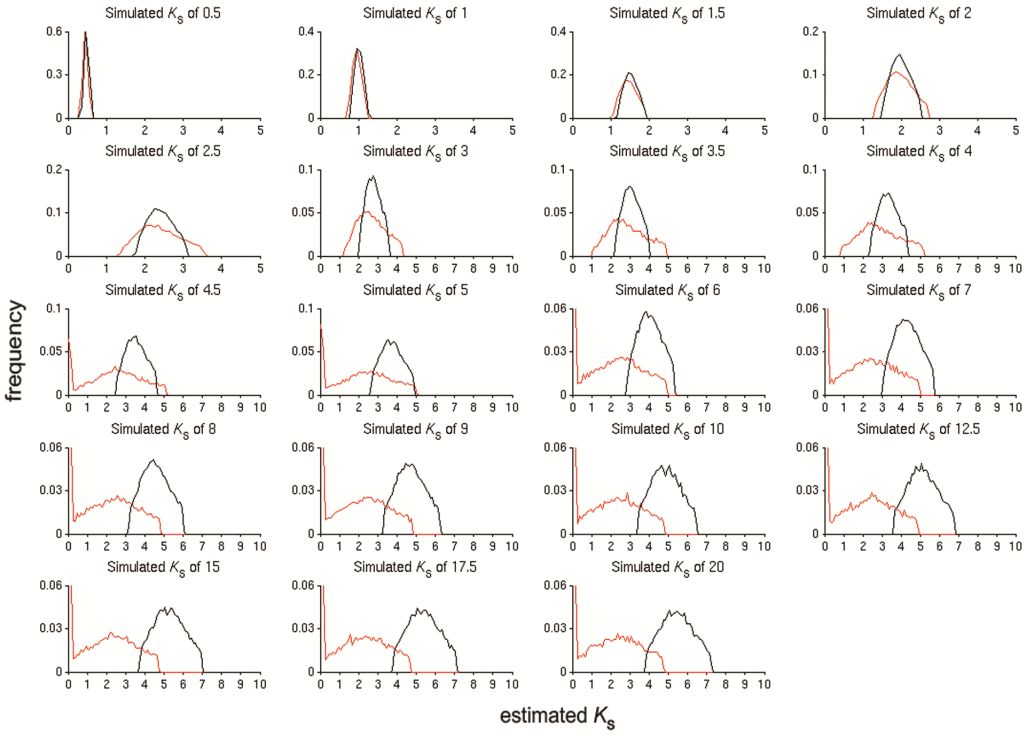


Figure E.28: Detailed results of evolutionary simulations for *A. thaliana* with ω=0.9. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=0.9 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

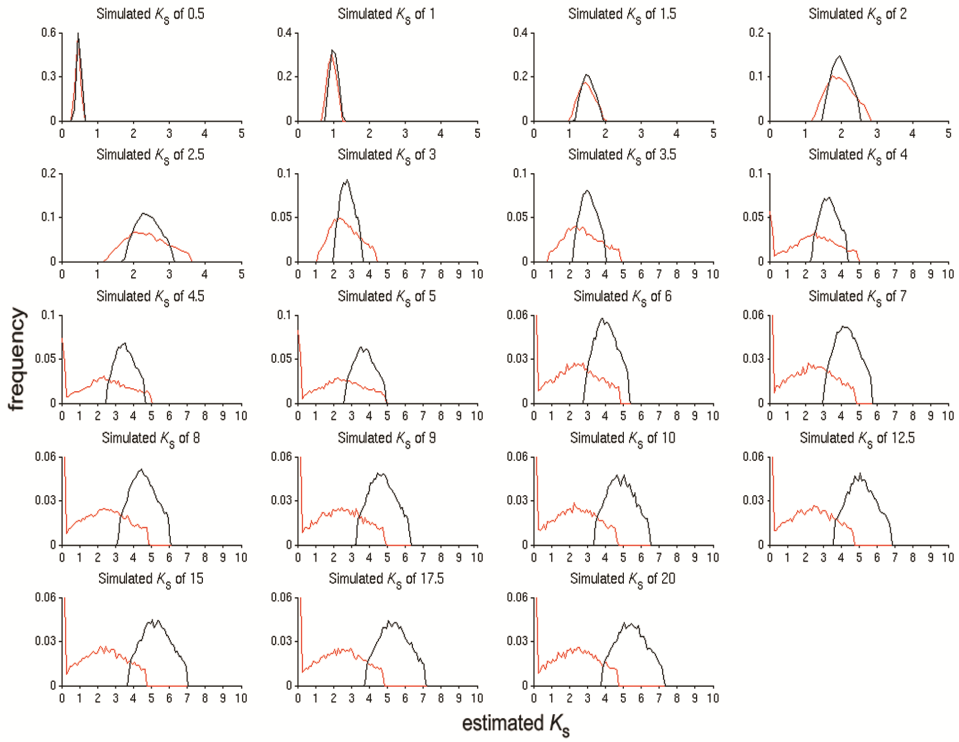


Figure E.29: Detailed results of evolutionary simulations for *A. thaliana* with ω=1.0. In total, 20,000 protein coding genes were randomly selected from the genome and evolved for time equivalents corresponding to predefined synonymous ages, indicated on top of the panels. Afterwards, K_S estimates between the real and simulated sequences were calculated. The panels display the resulting K_S estimate frequency distributions for ω=1.0 in red, while the K_S estimate frequency distributions for ω=0 are indicated in black for comparison. The ordinate scale varies between panels.

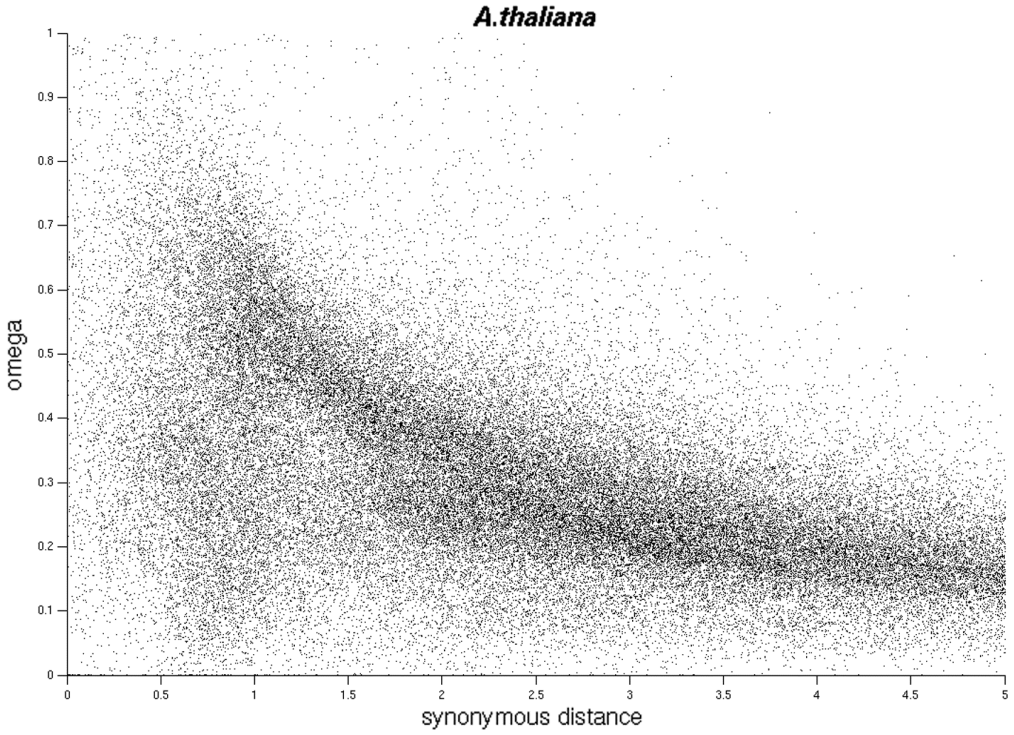


Figure E.30: ω paranome plot of *A. thaliana*. Both the synonymous distance (serving as a proxy for age since duplication) and ω (serving as a measure for selection strength) for all duplicated gene family members of the *A. thaliana* paranome were calculated with CODEML and plotted against each other. Newly duplicated genes display a wide variety in associated selection pressures, with many duplicate pairs exhibiting both high and low values of ω, while older duplicate gene pairs in general show a trend towards more stringent purifying selection.

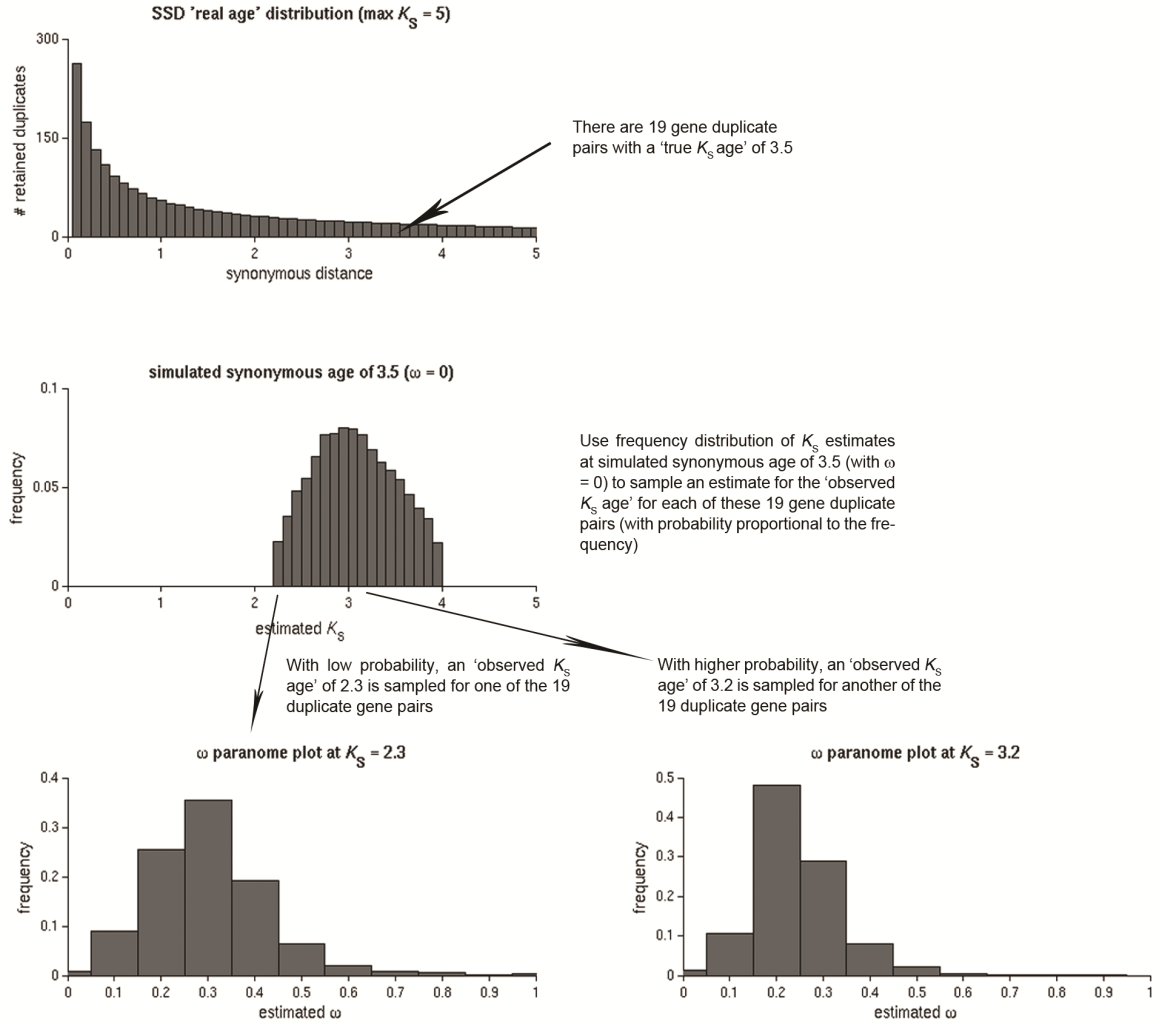


Figure E.31: Illustration of our approach to transform real age SSD distributions into K_S-based age distributions. Our approach accounts for different selection pressures (i.e., values of ω) between duplicate pairs of different age. The example is shown for a real age distribution that considers a timespan corresponding to a synonymous age of 5.

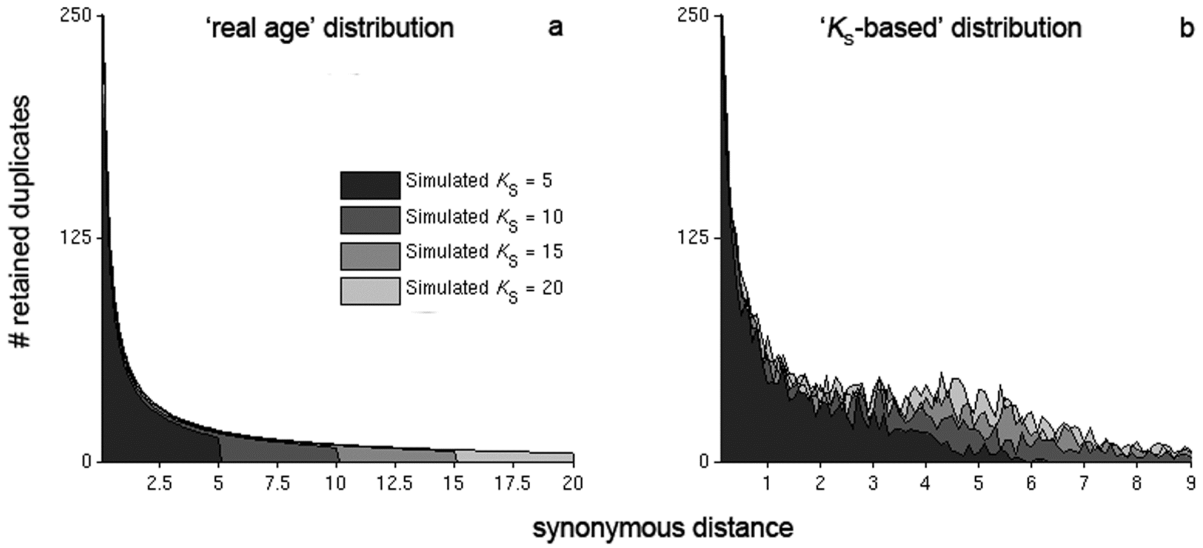


Figure E.32: SSD age distributions are characterized by a saturation peak. (a) SSD real age distributions generated by our population dynamics model under standard parameters over increasing evolutionary timespans without correcting for the effects of K_S saturation and stochasticity. (b) SSD K_S-based transformed age distributions of the real age distributions displayed in panel (a) for *A. thaliana*. Correcting for K_S noise was based on evolutionary simulations that did also consider non-synonymous mutations, and results in a diffuse SSD saturation peak. The surface curve is rugged because the transformation is necessarily based on a finite number of genes.

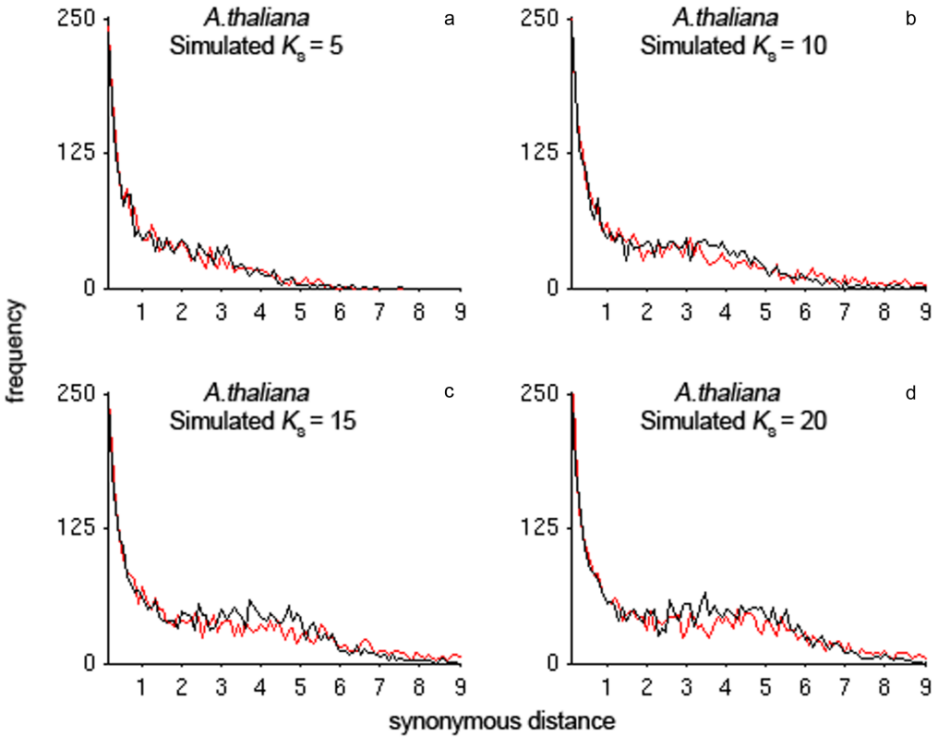


Figure E.33: SSD age distributions are characterized by a saturation peak. Each panel shows a close-up of the K_S-based transformed age distributions for increasing evolutionary timespans presented in figure E.32b, indicated by the red solid line. K_S-based transformed age distributions based on synonymous evolution only (i.e., ω=0) as described in the main text are indicated by the solid black lines, and are based on the same finite number of genes to allow better comparison. For all panels, both scenarios are characterized by a diffuse SSD saturation peak, although its amplitude is lower when varying K_N/K_S ratios are considered, especially for longer evolutionary timespans.

Appendix F

Supplementary material - Dating of genome duplications

F.1 Supplementary figures

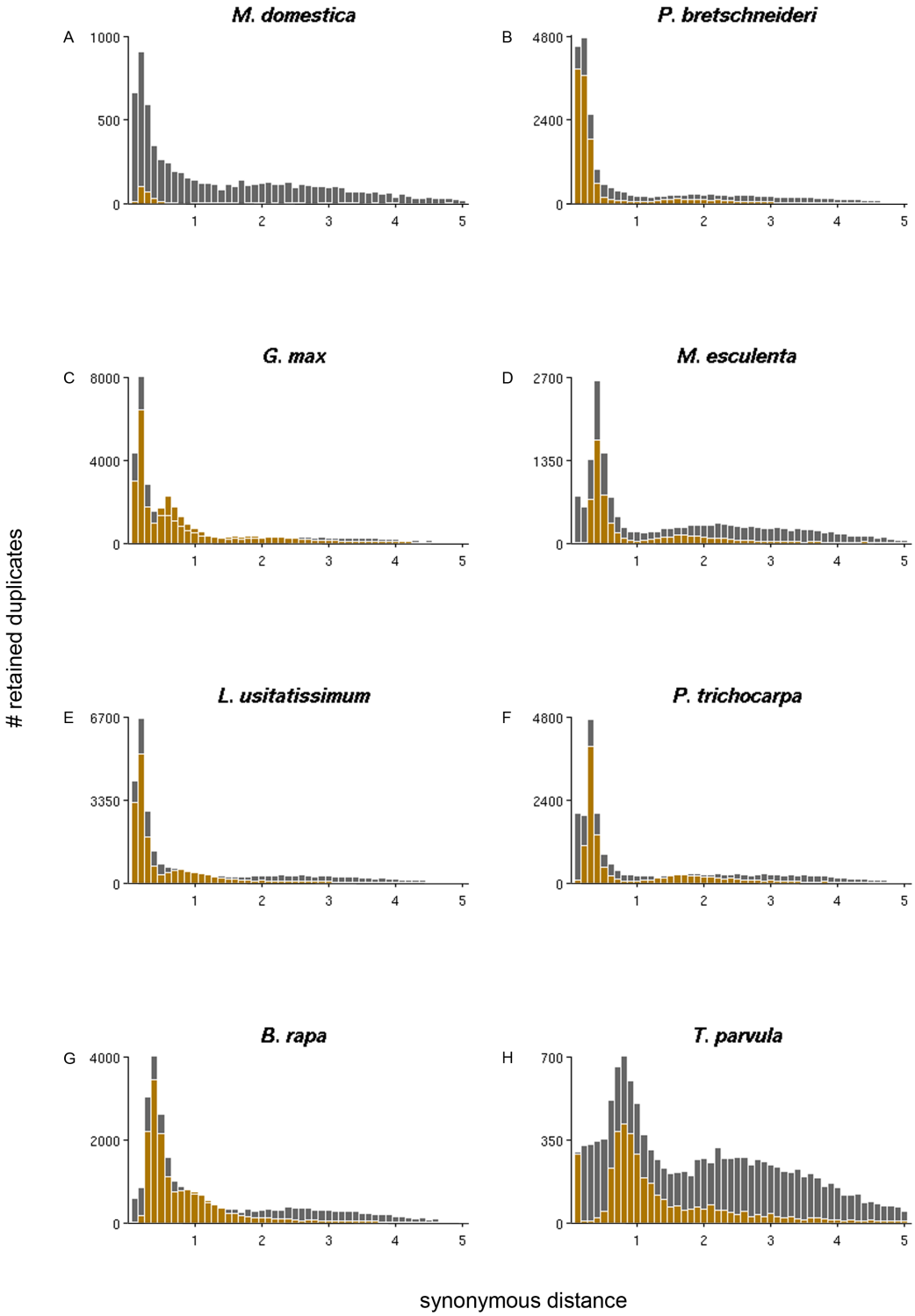


Figure F.1: K_S age distributions for all species. K_S age distributions for (A) *M. domestica*, (B) *P. bretschneideri*, (C) *G. max*, (D) *M. esculenta*, (E) *L. usitatissimum*, (F) *P. trichocarpa*, (G) *B. rapa*, and (H) *T. parvula*. The grey and beige bars represent the distribution of the paranome and duplicated anchors identified with i-ADHoRe, respectively. Anchors and peak-based duplicates used as homeologs for absolute dating were extracted between the WGD peak boundaries (see table 4.1).

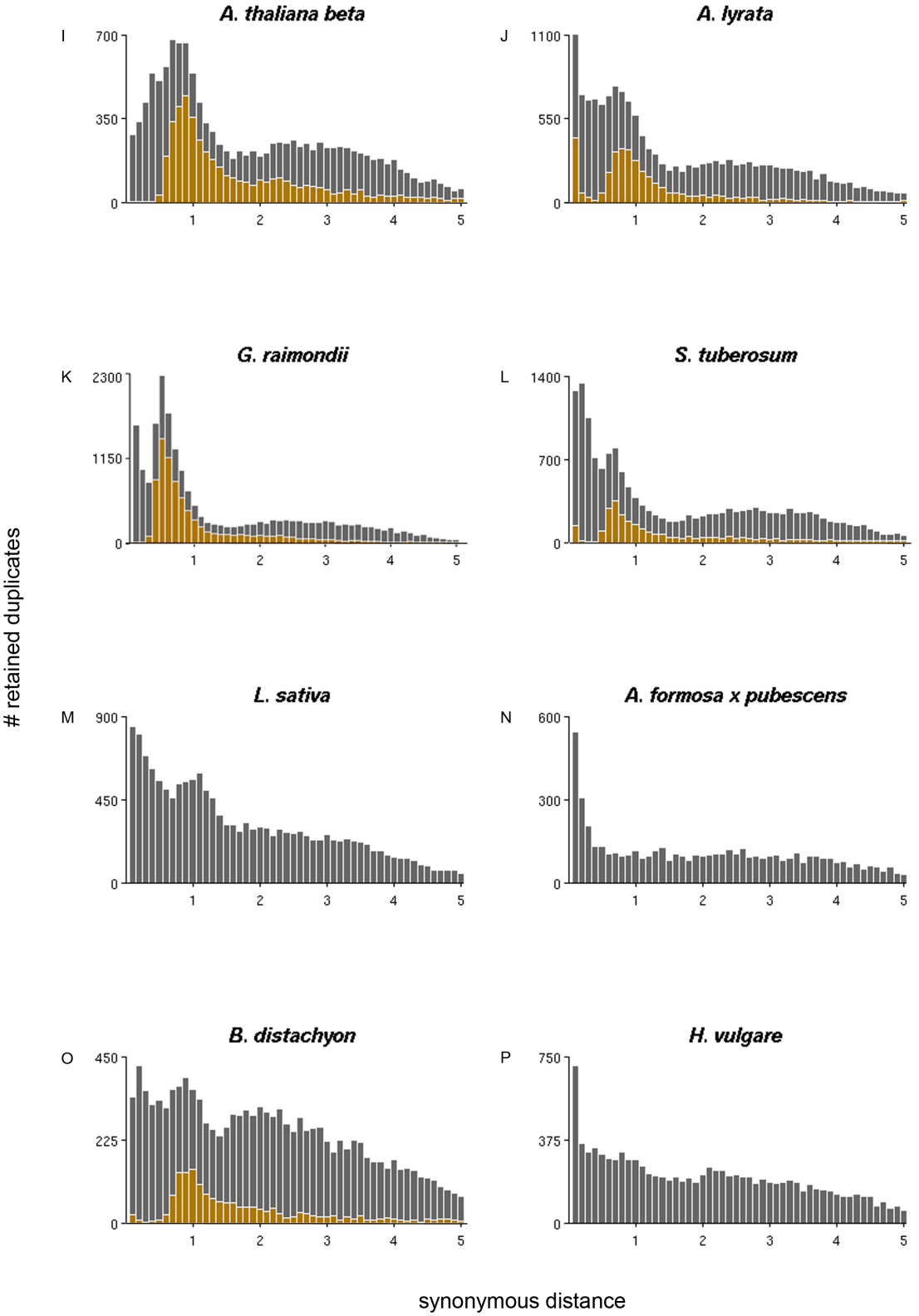


Figure F.1: K_S age distributions for all species - Continued. K_S age distributions for (I) *A. thaliana beta*, (J) *A. lyrata*, (K) *G. raimondii*, (L) *S. tuberosum*, (M) *L. sativa*, (N) *A. formosa x pubescens*, (O) *B. distachyon*, and (P) *H. vulgare*. The grey and beige bars represent the distribution of the paranome and duplicated anchors identified with i-ADHoRe, respectively. Anchors and peak-based duplicates used as homeologs for absolute dating were extracted between the WGD peak boundaries (see table 4.1).

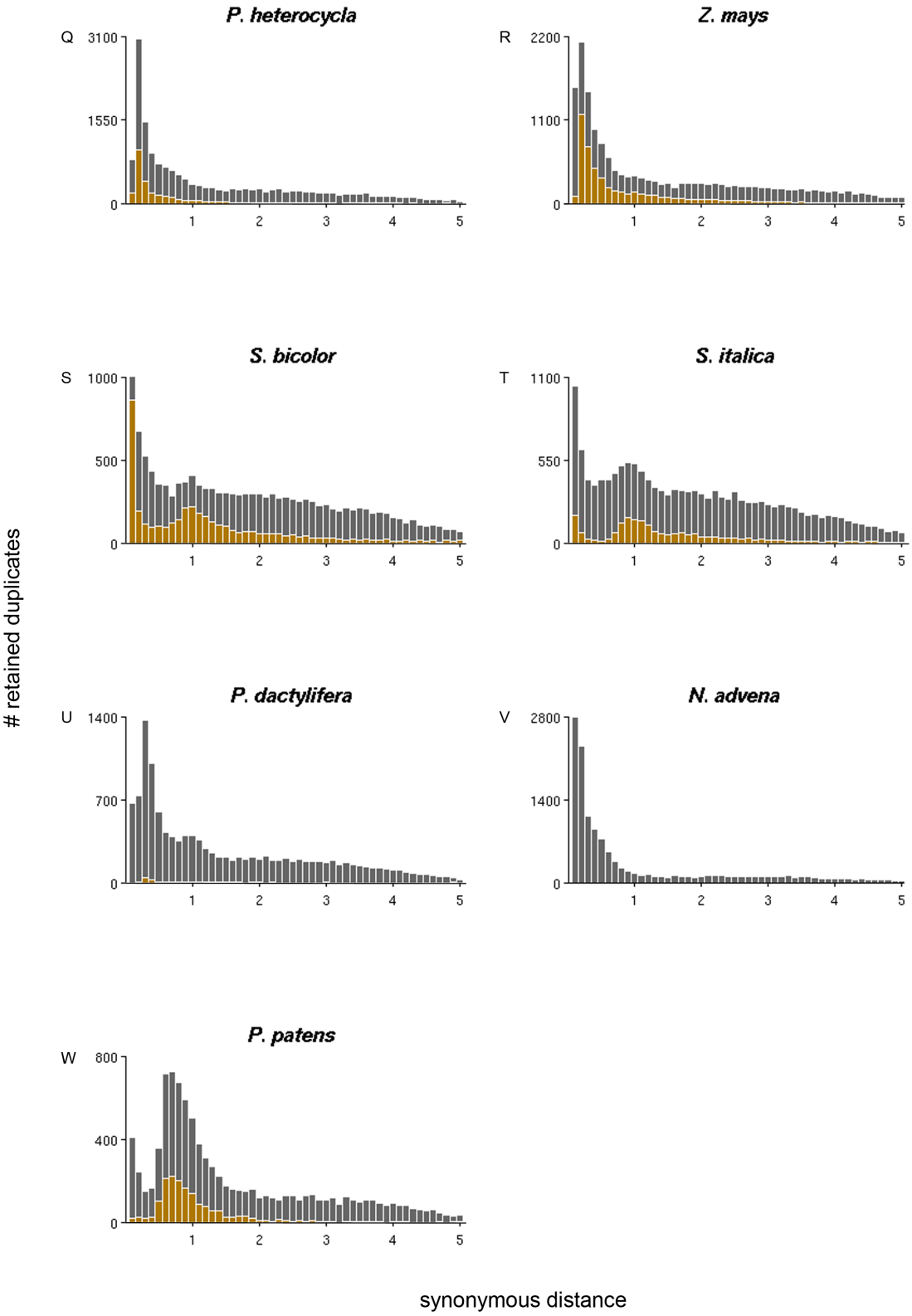


Figure F.1: K_S age distributions for all species - Continued. K_S age distributions for (Q) *P. heterocycla*, (R) *Z. mays*, (S) *S. bicolor*, (T) *S. italica*, (U) *P. dactylifera*, (V) *N. advena*, and (W) *P. patens*. The grey and beige bars represent the distribution of the paranome and duplicated anchors identified with i-ADHoRe, respectively. Anchors and peak-based duplicates used as homeologs for absolute dating were extracted between the WGD peak boundaries (see table 4.1).

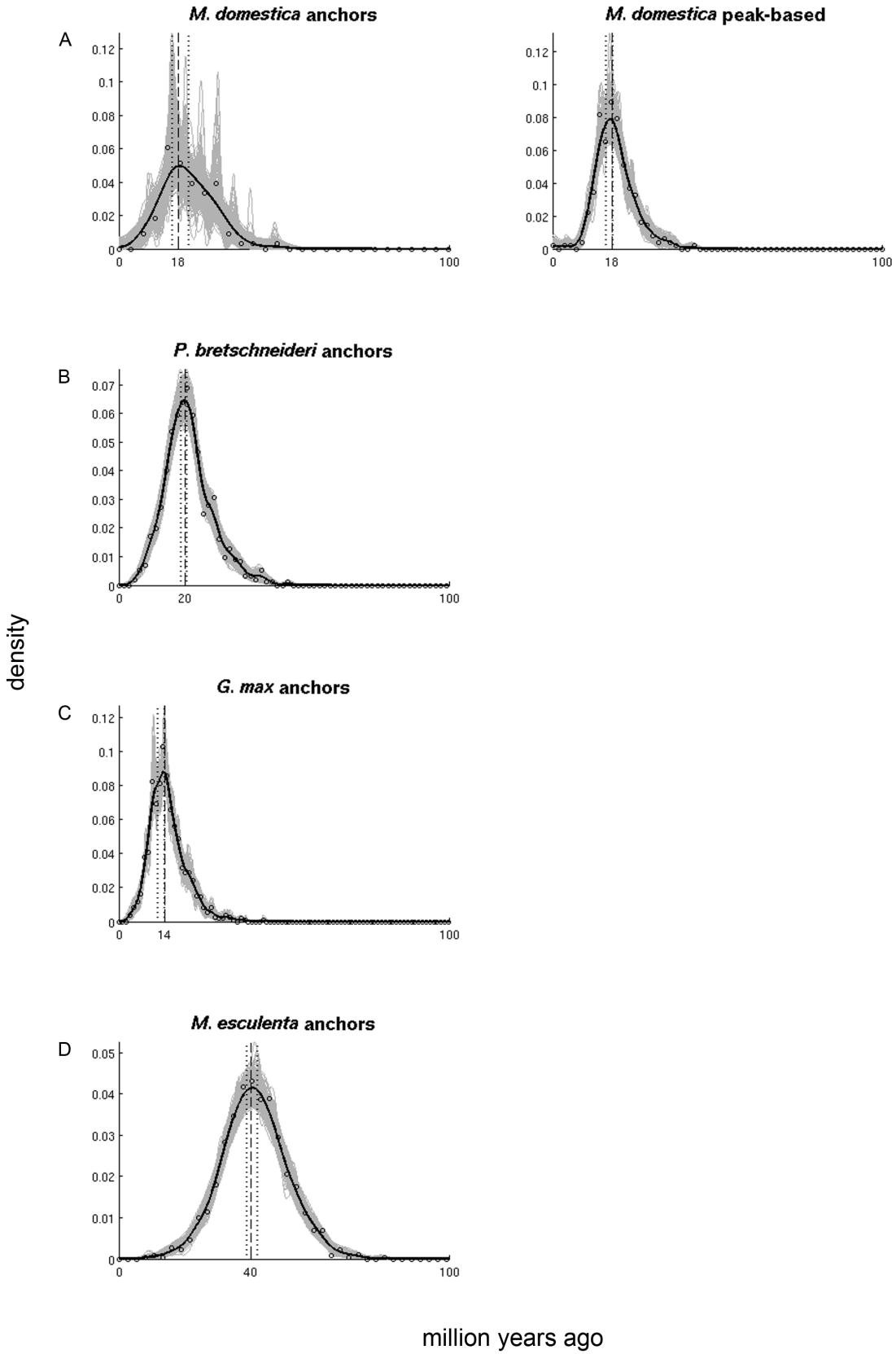


Figure F.2: Absolute age distributions of dated anchors and/or peak-based duplicates. Anchor and/or peak-based duplicate results are listed, where applicable (see table 4.1), for (A) *M. domestica*, (B) *P. bretschneideri*, (C) *G. max*, and (D) *M. esculenta*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries.

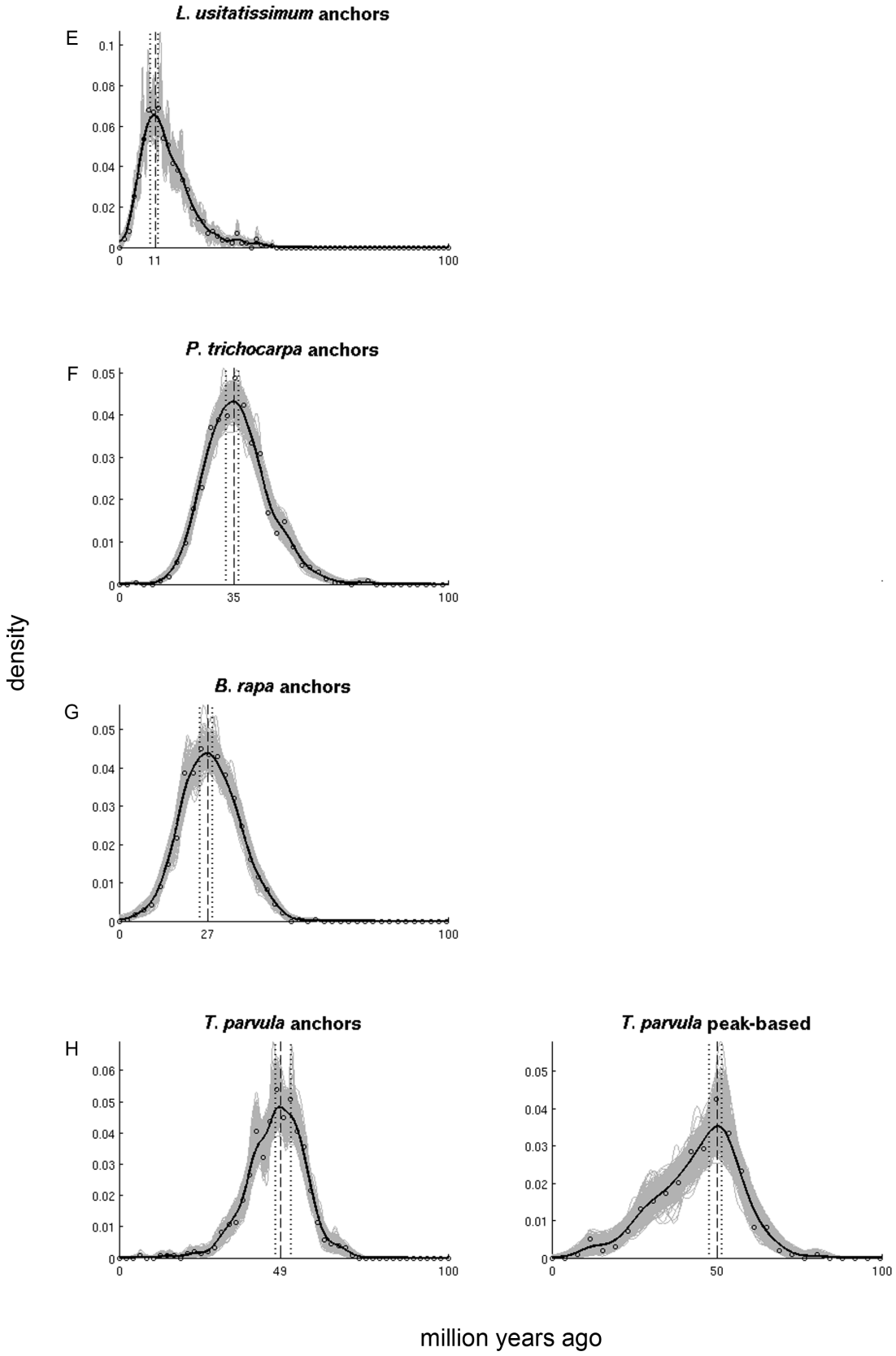


Figure F.2: Absolute age distributions of dated anchors and/or peak-based duplicates - Continued. Anchor and/or peak-based duplicate results are listed, where applicable (see table 4.1), for (E) *L. usitatissimum*, (F) *P. trichocarpa*, (G) *B. rapa*, and (H) *T. parvula*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries.

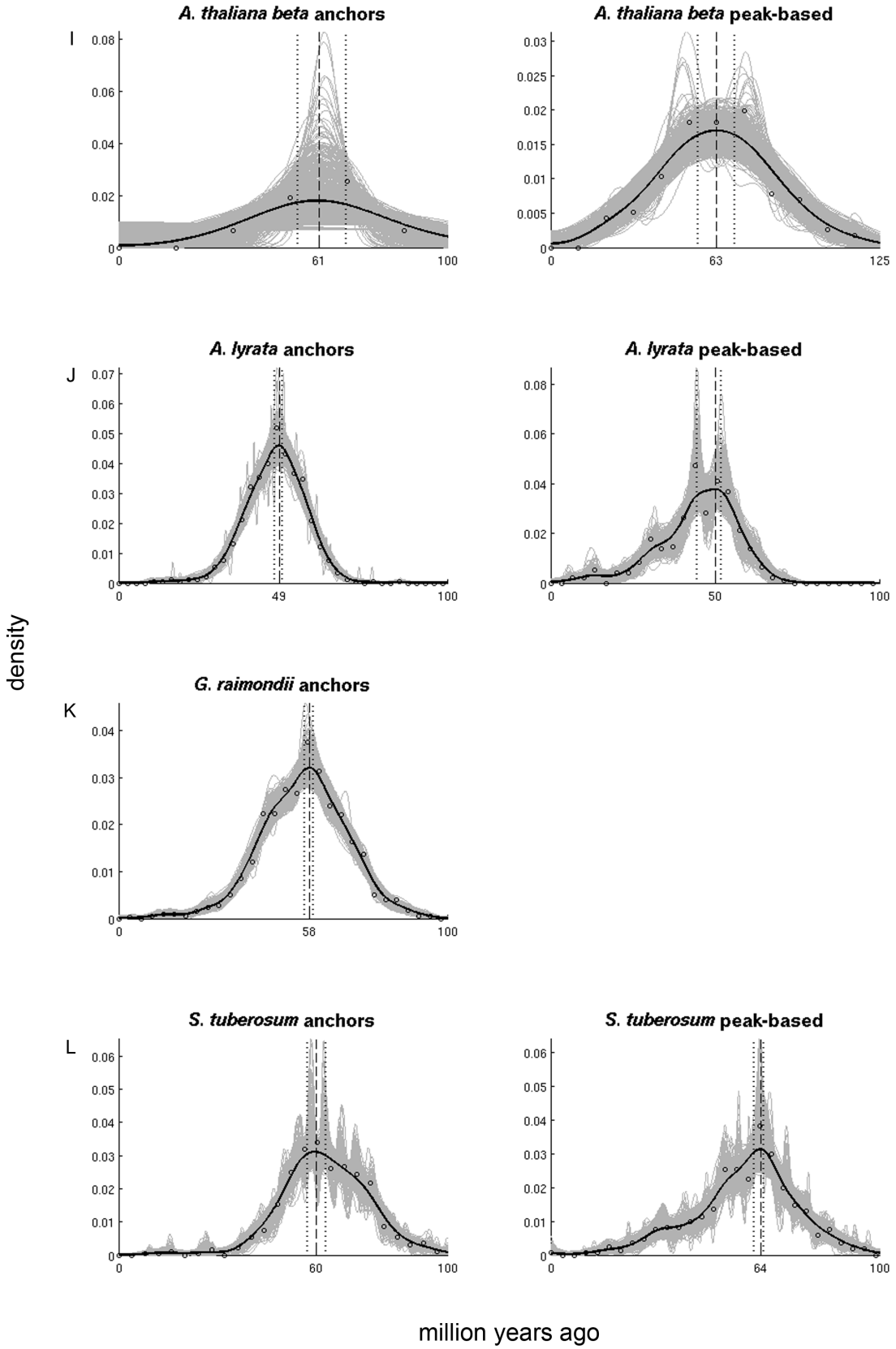


Figure F.2: Absolute age distributions of dated anchors and/or peak-based duplicates - Continued. Anchor and/or peak-based duplicate results are listed, where applicable (see table 4.1), for (I) *A. thaliana beta*, (J) *A. lyrata*, (K) *G. raimondii*, and (L) *S. tuberosum*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries.

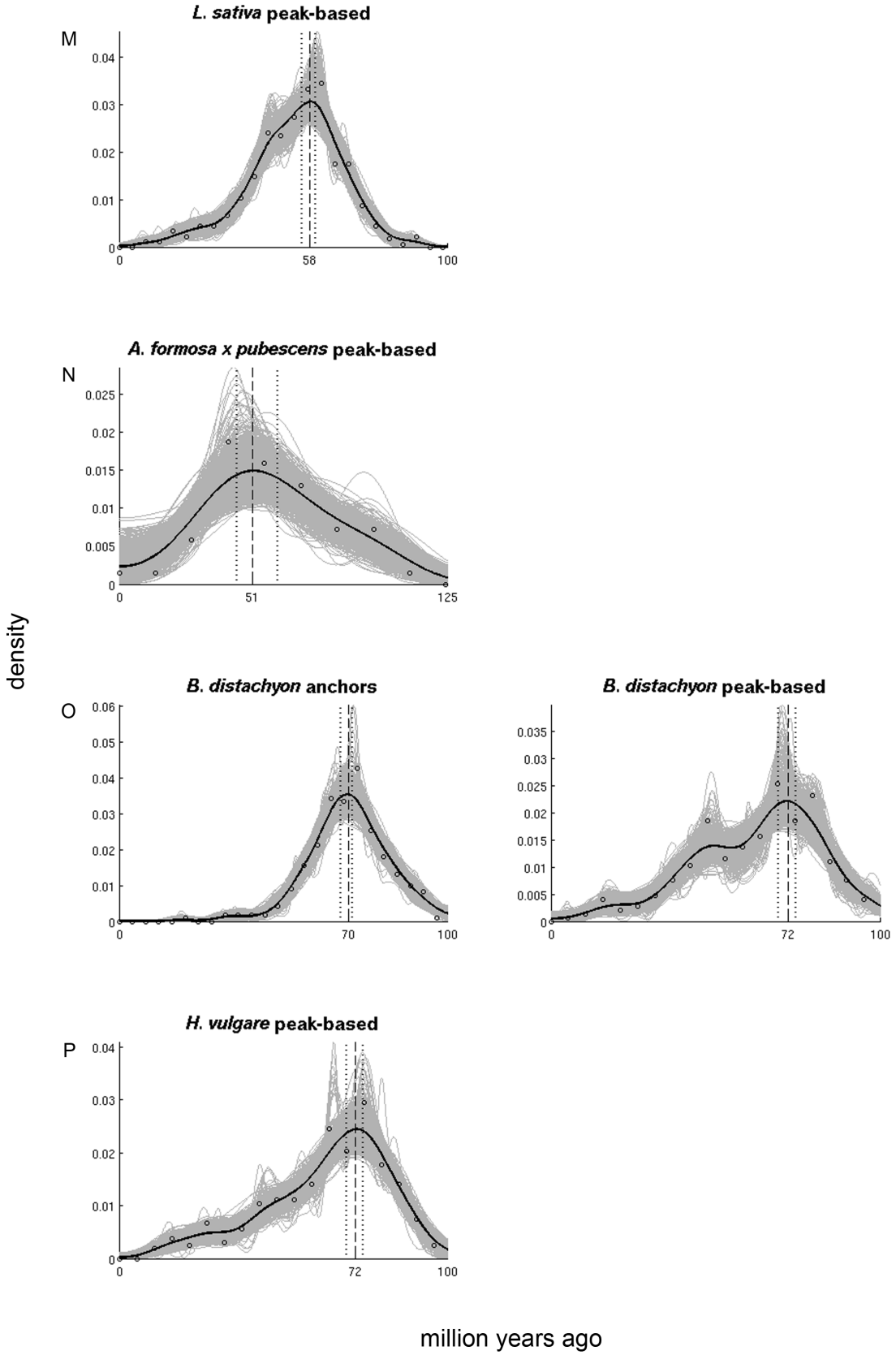


Figure F.2: Absolute age distributions of dated anchors and/or peak-based duplicates - Continued. Anchor and/or peak-based duplicate results are listed, where applicable (see table 4.1), for (M) *L. sativa*, (N) *A. formosa x pubescens*, (O) *B. distachyon*, and (P) *H. vulgare*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries.

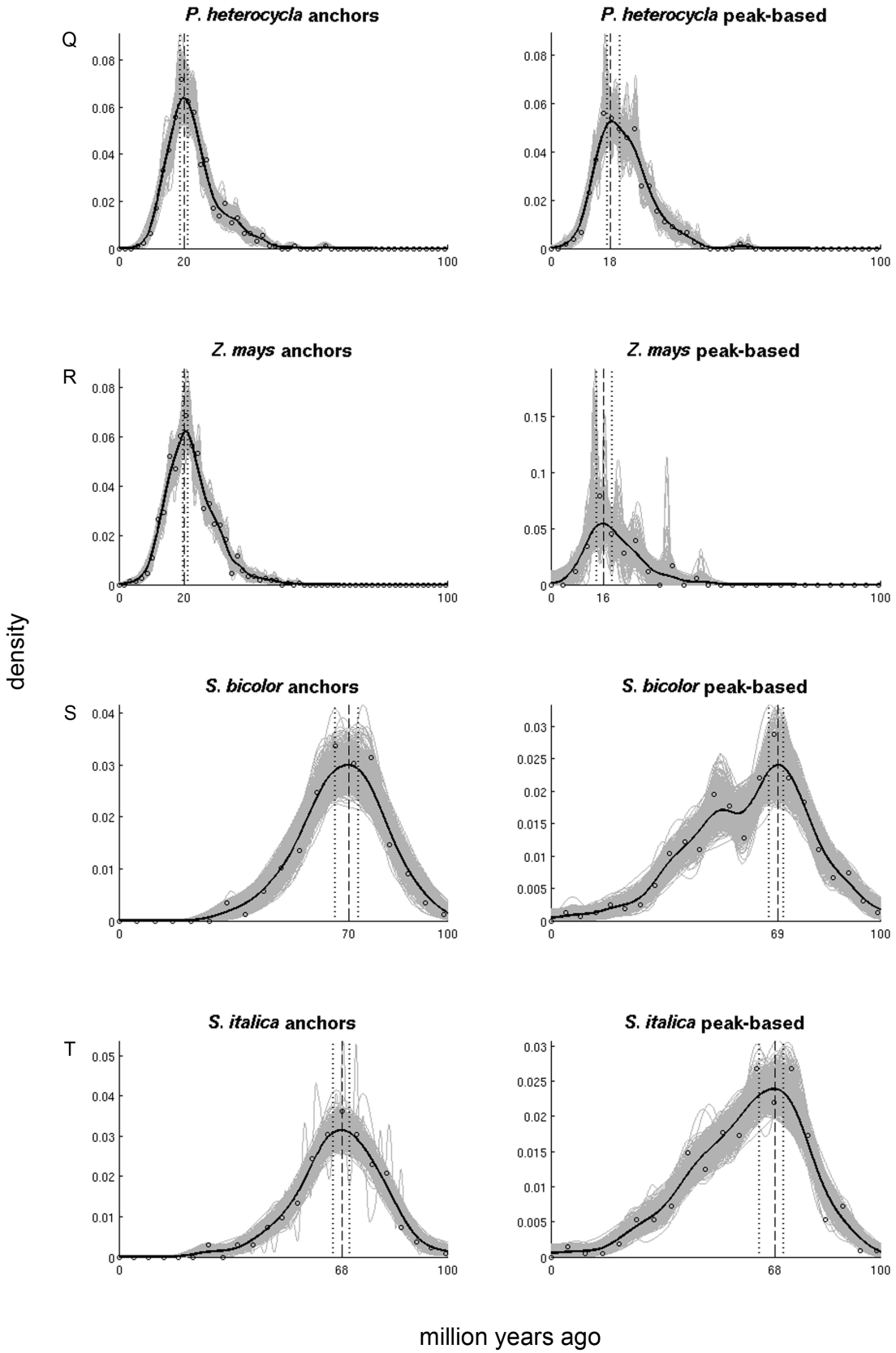


Figure F.2: Absolute age distributions of dated anchors and/or peak-based duplicates - Continued. Anchor and/or peak-based duplicate results are listed, where applicable (see table 4.1), for (Q) *P. heterocycla*, (R) *Z. mays*, (S) *S. bicolor*, and (T) *S. italica*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries.

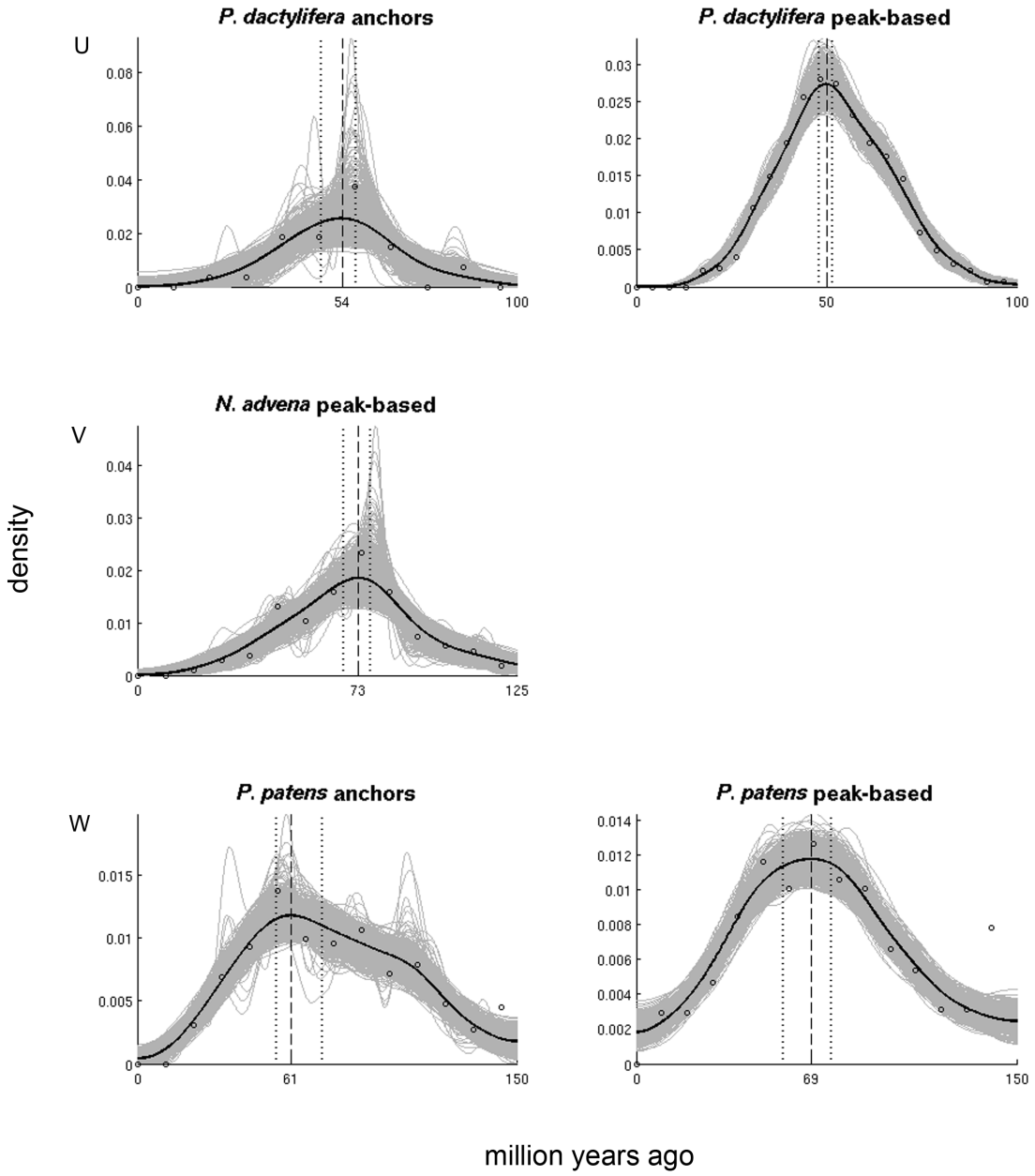


Figure F.2: Absolute age distributions of dated anchors and/or peak-based duplicates - Continued. Anchor and/or peak-based duplicate results are listed, where applicable (see table 4.1), for (U) *P. dactylifera*, (V) *N. advena*, and (W) *P. patens*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table 4.1 for sample sizes and exact confidence interval boundaries.

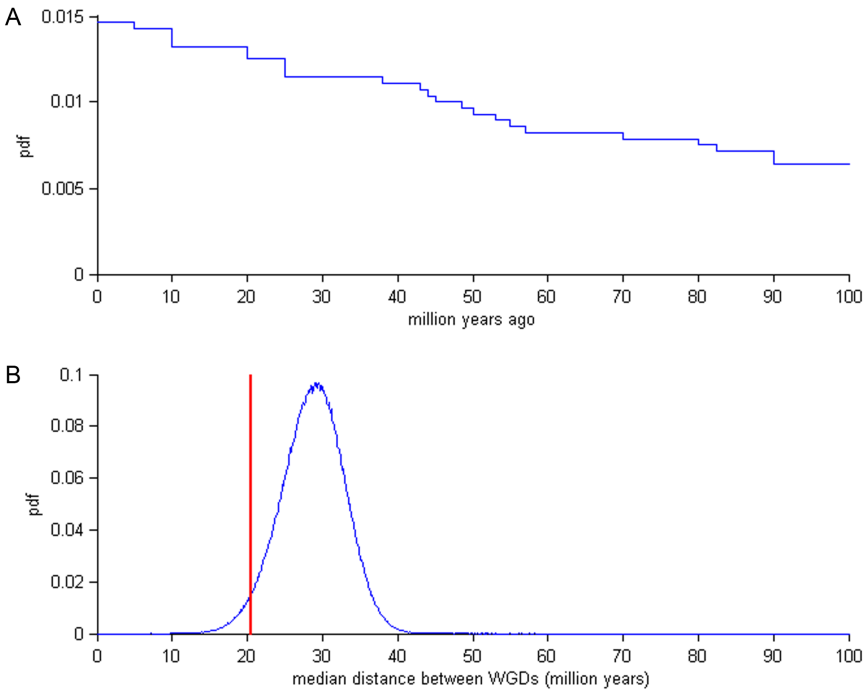


Figure F.3: The dated WGDs cluster statistically significantly in time. (A) Probability density function (pdf) for the null model of random WGD occurrence over time. An interval between 0 and 100 mya is considered. Each discontinuity in the pdf corresponds to a speciation event in figure 4.3, and the probability of WGD occurrence at a certain point in time is proportional to the total number of species present at that time. (B) Assessment of the statistical significance of WGD clustering in time. The true median distance between WGD age estimates presented in table 4.1 is indicated by the vertical red line (true median WGD distance = 20.42 million years). Note that shared WGDs were only counted once by taking the average of anchor-based WGD age estimates, or peak-based WGD age estimates if the former were not available, in their descendant species. The distribution of one million random samples is indicated in blue. Each sample is represented by a median WGD distance that was calculated based on pulling WGD ages randomly from the null model in A (average random median WGD distance = 28.65 million years). The true median WGD distance was significantly lower than expected under the null model (p=0.0301), indicating that plant paleopolyploidizations cluster statistically significantly in time. Exclusion of the *M. acuminata* WGD, because this most likely represents two WGDs in close succession, does not change these results although exclusion of the latter does decrease statistical significance (p=0.0430).

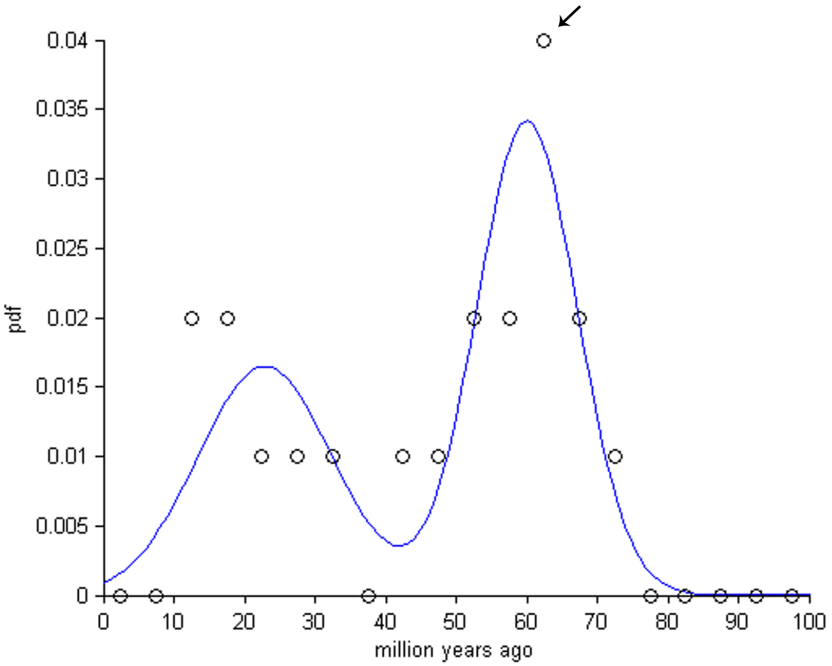


Figure F.4: Distribution of age estimates for all dated WGDs. Probability density function (pdf) of WGD age estimates. The blue curve represents the fit of a mixture of Gaussians that was used to find where WGDs cluster in time (see figure F.3). A mixture of two components was selected according to the AIC criterion (AIC=174.90 compared to AIC=180.33 and 177.96 for a mixture with one and three components, respectively). The total probability of WGD occurrence between 0 and 100 mya is equal to one (i.e., the sum of everything under the blue curve, its integral, sums to one). Note that shared WGDs were only counted once by taking the average of anchor-based WGD age estimates, or peak-based WGD age estimates if the former were not available, in their descendant species. The mixture contains one relatively thin and high component with a peak located at 60.05 mya, corresponding to the clustering of WGDs with the K-Pg boundary, and a broader and lower component with a peak located at 22.91 mya. The raw data is also presented on the figure by open circles. Every circle indicates the relative frequency of WGDs falling within an age bin of 5 million years (i.e., the first circle is located at 2.5 mya and represents the relative frequency of all WGDs falling between 0 and 5 mya etc.). Note that the particular bin size of 5 million years was arbitrarily chosen to allow a visual comparison of the raw data with the estimated fit of the Gaussian mixture, and does not influence the Gaussian mixture model fitting (i.e., the bin size does not have any influence on the shape of the mixture and its peak at 60.05 mya). The mixture demonstrates an overall good fit to the raw data, especially considering the relatively small sample size of only 20 independent WGDs. The open circle indicated with an arrow represents the relative frequency of WGDs falling between an interval of 60 and 65 mya. Exclusion of the *M. acuminata* WGD, because this most likely represents two WGDs in close succession, does not change these results (first and second peak located at 22.47 and 59.21 mya, respectively).

F.2 Supplementary tables

Table F.1: Overview of all employed species and their sequence sources.

Species	Provider	Source
Aquilegia formosa x pubescens	PLANTGDB (v187a)	www.plantgdb.org
Arabidopsis lyrata	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Arabidopsis thaliana	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Brachypodium distachyon	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Brassica rapa	Phytozome (v8)	www.phytozome.net
Cajanus cajan	IIPG (v5)	www.icrisat.org/gt-bt/iipg/Genome_Manuscript.html
Carica papaya	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Cicer arietinum	LIS (v1)	cicar.comparative-legumes.org
Citrullus lanatus	BGI (v1)	www.icugi.org/cgi-bin/ICuGI/index.cgi
Cucumis melo	MELONOMICS (v3.5)	melonomics.net
Cucumis sativus	BGI (v2)	www.icugi.org/cgi-bin/ICuGI/index.cgi
Fragaria vesca	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Glycine max	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Gossypium raimondii	BGI (v1)	cgp.genomics.org.cn
Hordeum vulgare	IBSC (v1)	www.public.iastate.edu/~imagefpc/IBSCWebpage
Jatropha curcas	JGD (v4.5)	www.kazusa.or.jp/jatropha
Lactuca sativa	PLANTGDB (v187a)	www.plantgdb.org
Linum usitatissimum	Phytozome (v8)	www.phytozome.net
Lotus japonicus	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Malus domestica	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Manihot esculenta	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Medicago truncatula	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Musa acuminata	Genoscope (v1)	banana-genome.cirad.fr
Nuphar advena	AAGP (v3)	ancangio.uga.edu/content/nuphar-advena
Oryza sativa	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Phoenix dactylifera	Weill Cornell Medical College (v3)	qatar-weill.cornell.edu/research/datepalmGenome
Phyllostachys heterocycla	ICBR (v1.0)	202.127.18.221/bamboo/index.php
Physcomitrella patens	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Populus trichocarpa	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Prunus mume	BGI (v1)	prunusmumegenome.bjfu.edu.cn
Prunus persica	Phytozome (v8)	www.phytozome.net
Pyrus bretschneideri	BGI (v1)	peargenome.njau.edu.cn
Ricinus communis	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Setaria italica	Phytozome (v8)	www.phytozome.net
Solanum lycopersicum	ITAG (v2.3)	solgenomics.net/organism/Solanum_lycopersicum/genome
Solanum tuberosum	ITAG (v1)	solgenomics.net/organism/Solanum_tuberosum/genome
Sorghum bicolor	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Thellungiella parvula	Thellungiella Consortium (v2)	thellungiella.org
Theobroma cacao	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Vitis vinifera	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza
Zea mays	PLAZA (v2.5)	bioinformatics.psb.ugent.be/plaza

F.3 Supplementary information

F.3.1 Species grouping topology

In order to date the node joining the homeologous pair, orthogroups were constructed consisting of both homeologs and orthologs from other plant species for which full genome sequence information was available. Different plant species were grouped into 'species groups' for which one ortholog was selected and added to the orthogroup, in order to keep the orthogroup topology fixed and to facilitate automation on the one hand, but also to allow enough orthogroups to be constructed on the other hand. Figure F.5 illustrates the employed species grouping topology.

The topology presented in figure F.5 is a trade-off between the total amount of sequence information within each individual orthogroup, and the total number of orthogroups that can be recovered. For instance, in case of the Brassicales, there is ample high-quality sequence information available from multiple genomes, so that splitting this order up in two different species groups (i.e., *A. thaliana* and *A. lyrata* on the one hand, and *T. parvula*, *B. rapa*, and *C. papaya* on the other hand) instead of one single group entails that every orthogroup contains more sequence information (which increases the accuracy in the age estimate of the homeologous pair that is dated in the orthogroup), while the total number of recovered

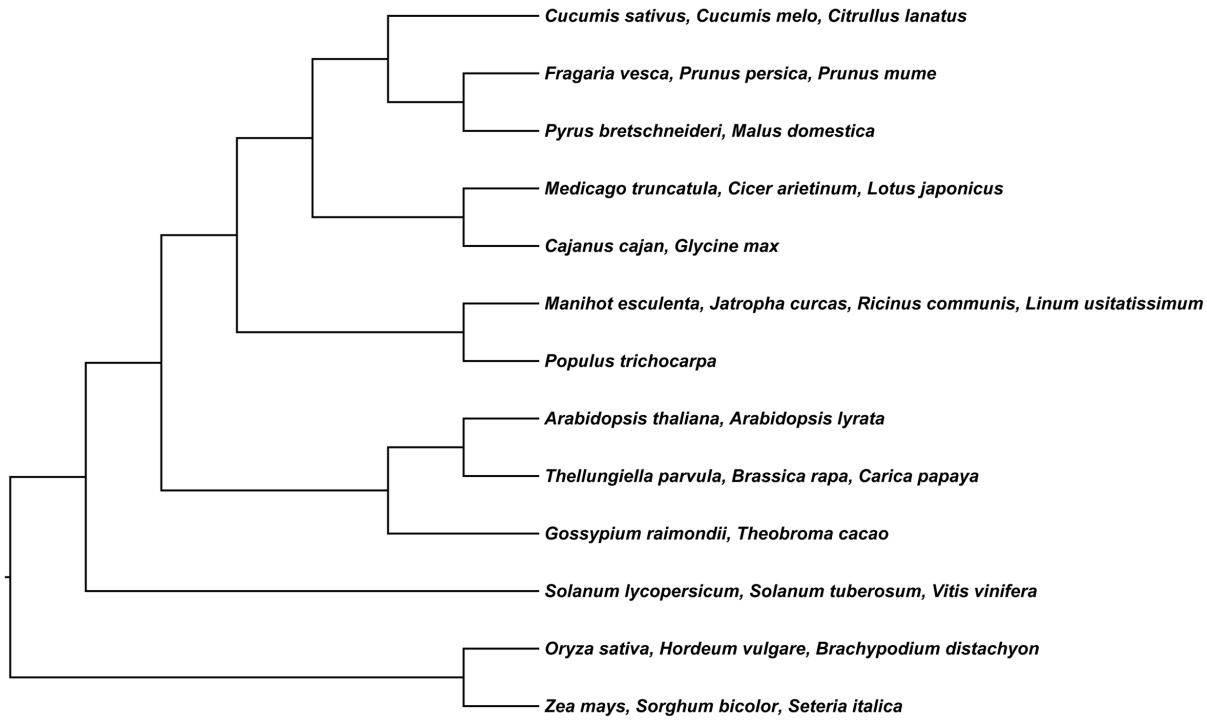


Figure F.5: Employed species grouping topology.

orthogroups also remains adequately high (which increases the total number of homeologous pairs that can be dated). Conversely, *Vitis* and *Solanum* were merged into one species group, because although splitting them would result in more sequence information per individual orthogroup, we found that in most cases not both a *Vitis* and *Solanum* ortholog could be found, drastically decreasing the total number of recovered orthogroups. The topology illustrated in figure F.5 was the result of some 'trial-and-error', i.e., merging and splitting different groupings of species until we found a topology that maximized the total amount of sequence information per individual orthogroup, while still allowing a sufficiently large number of orthogroups to be recovered.

The topology presented in figure F.5 also offers some additional advantages. First, it avoids any phylogenetic uncertainties, as the underlying topology between the different grouped species conforms to the well-accepted current plant phylogeny[61–63,65,475–478], and is in accordance with the Angiosperm Phylogeny Group classification (APGIII)[479]. Second, because most often closely related species were grouped into species groups, the overall phylogenetic coverage remains high through including at least one ortholog for most major plant clades for which full genome sequence information is available. Third, WGDs in species not included in the topology could still be dated by introducing their homeologs at their respective phylogenetic location, after which one ortholog per species group (see figure F.5) was added. This was the case for *L. sativa*, *A. formosa x pubescens*, and *N. advena*, because only a transcriptome assembly was available for these, for *P. heterocycla* because this genome only became available towards the end of this study when dating for the other species was finishing, for *P. patens* because of its very large phylogenetic distance from all the other species, and for *M. acuminata* and *P. dactylifera* because these were used only for dating WGDs in monocot species (see F.3.2). The exact phylogenetic position of these species is indicated on figure 4.3.

F.3.2 Calibrations and constraints

General

Recent molecular dating studies within the angiosperms benefit from a relatively wide array of fossil information that has become available, which typically allows implementing several high-quality primary fossil calibrations in large-scale dating studies where representatives from a large set of taxa are included based on a few high-quality sequenced marker genes[63–67]. However, in our study, the value of any particular calibration is highly dependent on the species sampling in our trees, which is limited by the number of full plant genome sequences that are currently available. Only a small minority of the available fossils can in fact properly describe the divergence events within the species grouping topology (see figure F.5). The majority of fossils routinely used in recent large-scale molecular dating studies cannot be used because no representative orthologs could be included in the orthogroups, due to the lack of a representative sequenced plant genome. For instance, there are several high-quality fossils available within the order Sapindales that could increase dating quality, but no representatives from this clade have been sequenced yet. Similarly, there are several high-quality fossils available within the order Arecales[480,481], but only one representative genome sequence is currently available (*P. dactylifera*) so that all these fossils can only describe the same divergence event in the orthogroups (i.e., the divergence from a *P. dactylifera* ortholog from other monocot species orthologs) and are therefore redundant. In such cases, only the oldest available fossil can be used to describe the divergence event[222].

A considerable body of literature has emerged in the last few years on the proper use of fossil data in molecular dating analysis. It is known that calibration priors in Bayesian time estimation can have a profound impact on posterior time estimates[66,67,198,216,359,365–367]. Point calibrations result in illusionary precision of the posterior time estimate, so that flexible statistical distributions that describe the error associated with the fossil age more realistically are preferred[217]. Early work focused on uniform distributions with hard minimum and maximum boundaries. These are however limited to clearly delineated fossil age boundaries, and can also lead to illusionary precision in the confidence intervals of the resulting posterior time estimate[368]. Such problems are mitigated by the introduction of soft maximum bounds that allow a certain small but nonzero part of the probability distribution, typically 2.5 to 5%, to be outside the maximum bound[198]. The youngest possible age to which a fossil can reliably be attributed (based on radiometric dating, biostratigraphy etc.) still constitutes a hard minimum bound[365]. Soft maximum bounds eliminate the need for arbitrarily ‘safe’ high hard maximum bounds because they allow the sequence signal to overcome and correct poor calibrations by pulling the posterior past the maximum bound[198]. Several flexible statistical distributions are commonly used but the lognormal distribution is particularly useful because of the way it mimics the error associated with estimating the divergence time of lineages from fossil information[67,222]. It has a hard minimum bound but allows placing its peak mass probability anywhere between the minimum and maximum bound. This way, it can accommodate for the lag-phase between the first appearance of a particular fossil and the actual divergence event it documents, a discrepancy that has led to much controversy in the early days of molecular dating[397]. The lognormal distribution also accommodates for soft maximum bounds because it has an infinitely extending horizontal asymptote.

Recent research demonstrates that the use of arbitrary lognormal calibration priors without justification for their shape, perhaps not surprisingly, can however still have a profound impact on the resulting posterior time estimates[367]. Especially the position of the peak mass probability within the calibration boundaries has been demonstrated to pull the posterior time estimates towards its location[66,367]. There is no reason to assume that the lag between lineage origin and first fossil occurrence will be consistent for all calibration points across the tree[482]. Guidelines about the magnitude of the parameters of the lognormal distribution are therefore currently assigned based on rough confidence around prior beliefs, see for instance Magallon et al.[67]. We calibrated any particular divergence by concentrating the prior peak mass probability on the most recent and accurate estimates found in literature (described below in detail for the individual calibrations). Although these literature-based estimates do not necessarily represent the true time of divergence, their effect on posterior time estimates should be less biased compared to a strategy where the peak mass probability is always arbitrarily placed at the beginning, middle, or end of a calibration interval. The proper placement of the calibration priors was always checked by performing a run without data[219] because the marginal calibration prior does not necessarily correspond to the desired calibration density, since the former is combined with the tree prior[483]. A starting tree with branch lengths satisfying all the fossil prior constraints was manually constructed. Figure F.6 represents an overview of both the initial tree branch lengths and all fossil calibrations (initial branch lengths were implemented based on the specific ortholog selected for each species group).

Eudicot calibrations (E1, E2, E3, and E4)

E1 is based on the fossil *Paleoclusia chevalieri*, which is the oldest known fossil we found from the order Malpighiales[484]. This fossil originates from the South Amboy Fire Clay at Old Crossman Clay Pit (New Jersey, USA), with a minimum bound of 82.8 mya[66]. This fossil is a member of the Clusiaceae family, but there exists some uncertainty whether the Clusiaceae split off between the Salicaceae and Euphorbiaceae[485], or if they are rather sister to both of these[68]. We therefore used this fossil to calibrate the divergence of the total group Malpighiales from their nearest sister group for which full genome sequence information was available, namely the remainder of the Eurosids I. The divergence between the former has been estimated at ~122.5 mya[68]. The mode of the lognormal distribution is located at $e^{\mu-\sigma^2}$, with μ and σ the mean and standard deviation of the lognormal distribution, respectively. We therefore specified a lognormal calibration prior with μ=3.9314, σ=0.5, and a minimum bound of 82.8 mya (because the peak of the lognormal calibration prior is hence located at 82.8 + $e^{3.9314-0.5^2}$ = 122.5 mya).

E2 is based on the fossil *Dressiantha bicarpellata*, which is the oldest known fossil from the order Brassicales[486], also originating from the South Amboy Fire Clay at Old Crossman Clay Pit (New Jersey, USA). We used this fossil to calibrate the divergence of the Brassicales from their nearest sister group for which full genome sequence information was available, namely the order Malvales. The divergence between the former has been estimated at ~119.5 mya[69]. We therefore specified a lognormal calibration prior with μ=3.8528, σ=0.5, and a minimum bound of 82.8 mya.

E3 is based on the fossil *Icacinicarya budvarensis*, which is the oldest known fossil from the asterids[487]. This fossil originates from České Budějovice Budvar (Czech Republic), with a minimum bound of 89.3 mya[70]. We used this fossil to calibrate the divergence of the asterids from their nearest

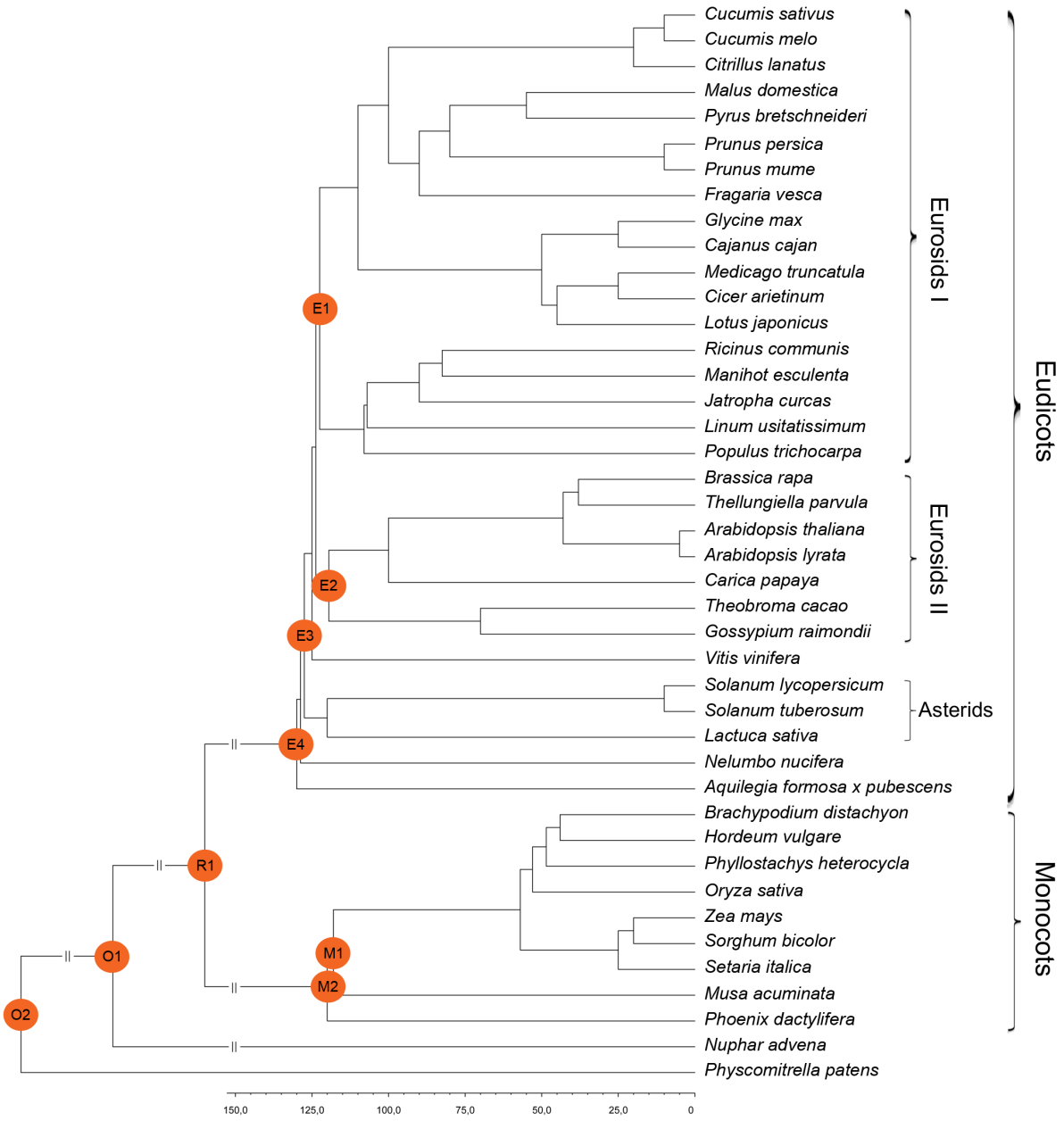


Figure F.6: Tree with initial branch lengths and employed fossil calibrations. Branch lengths are truncated after 150 mya for improved clarity (the initial branch length for the divergence described by O2, O1, and R1, was put at 450 mya, 220 mya, and 170 mya, respectively).

sister group for which full genome sequence information was available, namely the remainder of the rosids. The divergence between the former has been estimated at ~125 mya[63,70]. We therefore specified a lognormal calibration prior with μ=3.8252, σ=0.5, and a minimum bound of 89.3 mya.

E4 is based on the fossil *Leefructus mirus*, which is the oldest known fossil from the order Ranunculales[488]. This fossil originates from the Daxinfangzi Bed at the Yixian Formation (China), with a minimum bound of 123.0 mya. We used this fossil to calibrate the divergence of the Ranunculales from their nearest sister group for which full genome sequence information was available, namely the total group of rosids and asterids. The divergence between the former has been estimated at ~130 mya[63,489]. We therefore specified a lognormal calibration prior with μ=2.1959, σ=0.5, and a minimum bound of 123.0 mya.

Performing a run without data[219,483] indicated however that implementation of all these four calibrations resulted in a situation where the marginal prior calibration distributions did not correspond to their specified calibration densities anymore. Rather, the prior calibration distributions of E1 and E2 pushed away the prior calibration distributions of E3 and E4, most likely because they were located on consecutive nodes (see figure F.6). Calibrations E3 and E4 was therefore only used when dating WGDs in the asterids (i.e., *S. lycopersicum*, *S. tuberosum*, and *L. sativa*), and Ranunculales (i.e., *A. formosa x pubescens*), respectively, while calibrations E1 and E2 were used for dating WGDs in all other species (including non-eudicots). This ensures that always at least one rate-correcting calibration was present between the homeologous pair and root for dating the WGDs in all eudicot species.

Monocot calibrations (M1 and M2)

M1 and M2 were used only when dating WGDs in monocot species (*O. sativa*, *B. distachyon*, *Z. mays*, *S. bicolor*, *M. acuminata*, *S. italica*, *P. heterocycla*, *H. vulgare*, and *P. dactylifera*). This is because monocot calibrations necessitated the inclusion of either *M. acuminata* or *P. dactylifera* into the orthogroups, which led to a drastic drop in orthogroup recovery. This was true especially when dating WGDs in non-monocot species, but also to a large extent for dating WGDs in monocot species themselves, which is why we considered *M. acuminata* and *P. dactylifera* as a single species group and required only one representative ortholog with its corresponding calibration to be present (i.e., there are two possible monocot calibrations that were only implemented when dating WGDs in monocot species to ensure at least one rate-correcting calibration between the root and homeologous pair, but for each orthogroup only one was implemented based on whether a *M. acuminata* or *P. dactylifera* ortholog was added to the orthogroup).

M1 is based on the fossil *Spirematospermum chandlerae*, which is the oldest known fossil from the order Zingiberales[490]. This fossil originates from the Black Creek Formation at Neuse River Cut-Off (North Carolina, USA), with a minimum bound of 83.5 mya. We used this fossil when a *M. acuminata* ortholog was included in the orthogroup to calibrate the divergence of the Zingiberales from their nearest sister group for which full genome sequence information was available, namely the order Poales. The divergence between the former has been estimated at $\sim$118 mya[71,491]. We therefore specified a lognormal calibration prior with μ=3.7910, σ=0.5, and a minimum bound of 83.5 mya.

M2 is based on the fossil *Sabalites carolinensis*, which is the oldest known fossil from the order Arecales[492]. This fossil originates from the Black Creek Formation near Langley (South Carolina, USA), with a minimum bound of 85.8 mya[480]. We used this fossil when a *P. dactylifera* ortholog was included in the orthogroup to calibrate the divergence of the Arecales from their nearest sister group for which full genome sequence information was available, namely the order Poales. The divergence between the former has been estimated at $\sim$120 mya[71,480,481]. We therefore specified a lognormal calibration prior with μ=3.7822, σ=0.5, and a minimum bound of 85.8 mya.

Root calibration (R1)

R1 is based on the sudden abundant appearance of eudicot tricolpate pollen in the fossil record at $\sim$125 mya at several separate geographical localities (Doyle 2005). An error of 1 million year based on

magnetostratigraphic evaluation is associated with the above described estimate of 125 mya, placing its minimum bound effectively at 124.0 mya[66]. We used this fossil information to calibrate the divergence of the eudicots from the monocots, which constitutes the root of orthogroup phylogenies. Selecting an appropriate peak mass probability location for this divergence is however less straightforward because there exists considerable variation in its estimate, ranging from about 140 mya until as old as 200 mya[63–66,493]. We consequently selected a peak mass probability at 170 mya (effectively the middle of these intervals), and therefore specified a lognormal calibration prior with μ=4.0786, σ=0.5, and a minimum bound of 124.0 mya. The more uncertain position of this split, in combination with placing a soft bound on the maximum root age, could place undue weight on the assumption of the age of the root[66]. The effects thereof on our results are however most likely small because for all species, with the exception of *N. advena* and *P. patens* (see below), at least one extra rate-correcting calibration was incorporated between the root and homeologous pair.

N. advena and *P. patens* calibrations (O1 and O2)

N. advena and *P. patens* were not part of the species grouping topology because of their isolated basal position in the plant phylogeny. Applying the same strategy as for other species not part of the species grouping topology, i.e., adding the homeologous pair at its respective phylogenetic location in the orthogroup topology, entails however that a new root is instituted. When dating the WGD in *N. advena* and *P. patens*, we therefore implemented O1 and O2 as new root calibrations, respectively.

O1 is based on the sudden abundant appearance of eudicot tricolpate pollen in the fossil record at 125 mya at several separate geographical localities[56], with a minimum bound of 124.0 mya (see before). We used this fossil information to calibrate the divergence of the *N. advena* homeologous pair from the eudicots and monocots, which constitutes the new root when the *N. advena* WGD was dated. This divergence has been estimated at ~220 mya[65–67]. We therefore specified a lognormal calibration prior with μ=4.8143, σ=0.5, and a minimum bound of 124.0 mya.

O2 is based on the fossil *Cooksonia*, which is the oldest known fossil from the Lycopsida[494]. This fossil originates from the Cloncannon Formation of County Tipperary (Ireland), with a minimum bound of 420.4 mya[66]. We used this fossil to calibrate the divergence of the *P. patens* homeologous pair from the eudicots and monocots, which constitutes the new root when the *P. patens* WGD was dated. This divergence has been estimated at ~450 mya[65–67]. We therefore specified a lognormal calibration prior with μ=3.6378, σ=0.5, and a minimum bound of 420.4 mya.

F.3.3 Alternative calibrations and constraints

General

The set of calibrations used for the WGD age estimates presented in table 4.1 are necessarily limited through the availability of full genome sequences and the species grouping topology. With regard to the remaining fossil calibration options, some of the choices we made may seem suboptimal at first sight. In particular, one may wonder why we did not adopt the eudicot crown node calibration based on eudicot tricolpate fossil pollen, in accordance with its sudden abundant appearance in the fossil record at ~125

mya[56]. The latter has a long history of use in molecular dating studies to enforce a hard maximum bound of 125 mya on the eudicot crown node. The interpretation of this fossil information has however recently been called into question. The earliest tricolpate fossil pollen already displays considerable structural variety and can be found across widespread geographical localities, suggesting that they represent the rise to dominance, rather than the first origin of the eudicots[65]. Additionally, the recent description of a fossil from the early-branching eudicot order Ranunculales estimated at 122.6-125.8 mya, argues that eudicots may have already been present some time before 125 mya[488]. The latter is also supported by several recent clade-specific molecular dating studies that place key divergence events within the eudicots typically very close to 125 mya[68,69,75]. Although it is difficult to explain why eudicots would remain hidden for so long if they had already diversified into clades that rose so rapidly in the mid-Cretaceous, angiosperms possibly originated in isolated freshwater lake-related wetlands from where they later quickly invaded other habitats, which would explain the discrepancy in the molecular record[73].

In light of this recent uncertainty, we preferred avoiding any controversy by not including this fossil calibration in our dating analysis. However, most recent large-scale molecular dating studies of the angiosperms converge mostly on the same age estimates for key divergence events within the eudicots, irrespective of whether this calibration was employed or not[63–67]. Not surprisingly, studies that impose a hard maximum bound of ~125 mya on the eudicot crown typically find age estimates that are somewhat younger than studies that do not impose this constraint, but both nevertheless agree particularly well on most divergence time estimates within the eudicots, despite the fact that both disagree strongly on their estimates for the age of the eudicots themselves. We investigated the effects of including this eudicot crown calibration in our analysis by rerunning a substantial part of the calculations on our dataset with this particular calibration implemented (see below).

Simultaneously, we took advantage of the relatively rich fossil record of the eudicots to investigate how reliable our WGD age estimates are under an alternative calibration set. For instance, the fossil *Dressiantha bicarpellata* was used in our original calibration set to describe the divergence of the order Brassicales, in which it was originally placed based on morphological data[486]. This classification was later challenged by a combined molecular sequence + morphological character analysis[495], but afterwards placed firmly again within the Brassicales based on a more recent combined molecular sequence + morphological character analysis[69]. This fossil has consequently been used in a series of recent molecular dating studies[61,66,69,480]. Here, we studied the effect of omitting this fossil calibration in favor of other calibrations (see below).

The alternative calibration set

Re-dating all constructed orthogroups with an alternative calibration set was computationally prohibitive due to the immense computational resources required for running the MCMC component of the molecular sequence divergence estimation[346,347]. We therefore chose to re-date all orthogroups based on anchors, because these are based on actual duplicated segments, and we only employed orthogroups based on peak-based duplicates if the former were not available (i.e., for *L. sativa*, *A. formosa x pubescens*, *H. vulgare*, and *N. advena*). The analysis methods were exactly the same as described before (see 4.2), with the exception that the original calibration set within the eudicots (i.e., E1, E2, E3, and E4 - see figure

F.6) was replaced in all orthogroups by a new alternative calibration set (i.e., E1', E2', E3', and E4' - see figure F.7), as discussed in the next paragraphs.

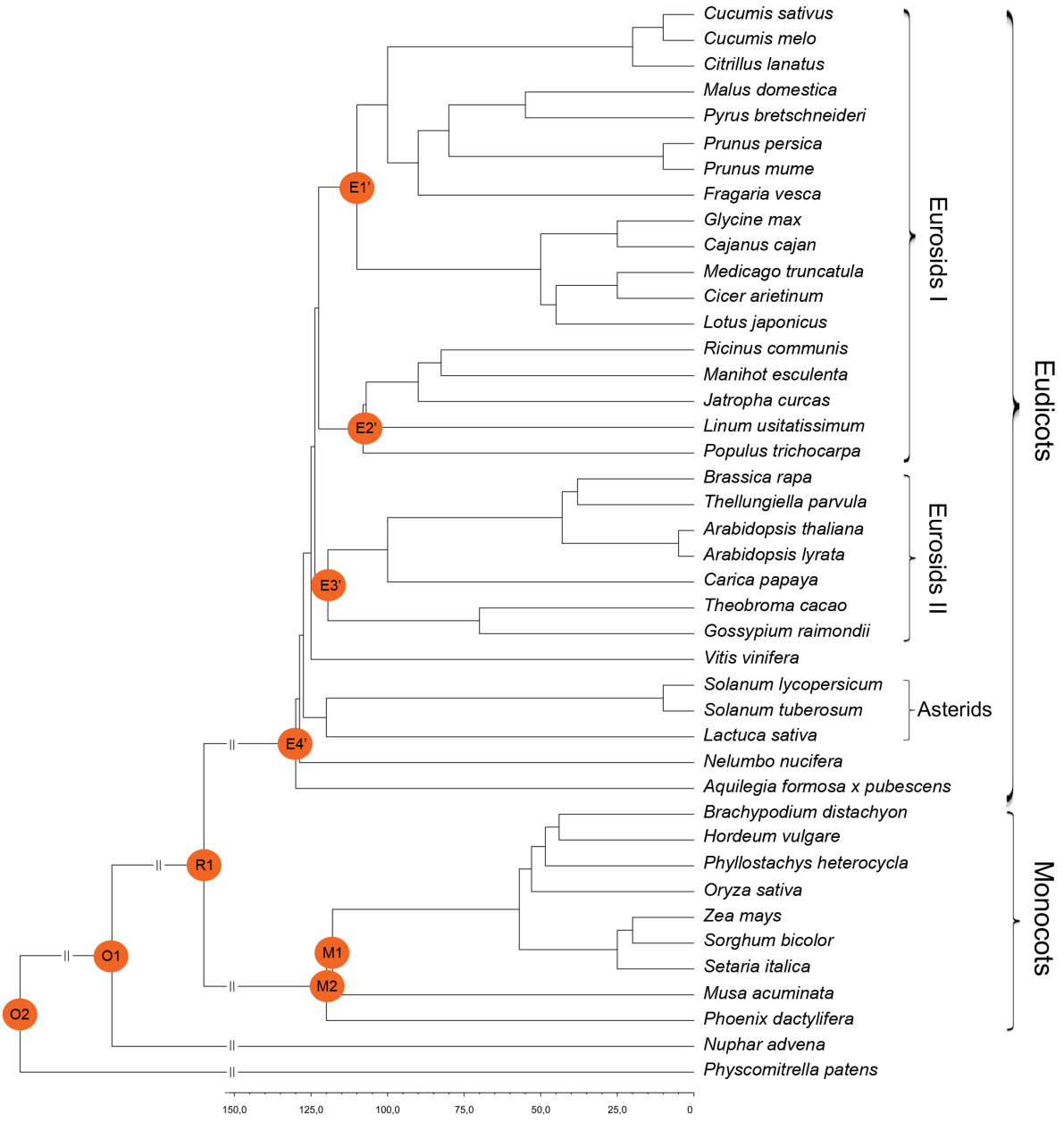


Figure F.7: Tree with initial branch lengths and employed fossil calibrations for the alternative calibration set. Branch lengths are truncated after 150 mya for improved clarity (the initial branch length for the divergence described by O2, O1, and R1, was put at 450 mya, 220 mya, and 170 mya, respectively).

The alternative calibration E1' is based on an unnamed fossil from the order Fabales[496], which is the oldest known fossil we found for this order, with a minimum bound of 59.9 mya. We used this fossil to calibrate the divergence of the Fabales from their nearest sister group for which full genome sequence information was available, namely the total group Rosales + Cucurbitales. The divergence between the former has been estimated at ~120 mya[75]. We therefore specified a lognormal calibration prior with μ=4.3460 (but see below), σ=0.5, and a minimum bound of 59.9 mya.

E2' is based on the fossil *Pseudosalix*, which is the oldest known fossil from the family Salicaceae[375], with a minimum bound of 48.0 mya. We used this fossil to calibrate the divergence of the Salicaceae from

their nearest sister group for which full genome sequence information was available, namely all other representatives from the order Malpighiales. The divergence between the former has been estimated at ~108 mya[68]. We therefore specified a lognormal calibration prior with μ=4.3443 (but see below), σ=0.5, and a minimum bound of 59.9 mya.

E3' is based on the fossil *Parbombacaceoxylon*, which is the oldest known fossil from the order Malvales[497,498], with a minimum bound of 65.5 mya. We used this fossil to calibrate the divergence of the Malvales from their nearest sister group for which full genome sequence information was available, namely the Brassicales. The divergence between the former has been estimated at ~119.5 mya[69]. We therefore specified a lognormal calibration prior with μ=4.2390 (but see below), σ=0.5, and a minimum bound of 65.5 mya.

E4' is based on the aforementioned eudicot tricolpate fossil pollen at ~125 mya[56]. We used this fossil information to constrain the crown group of the eudicots with a maximum age. To accommodate some small margin of error around this boundary, as suggested by recent findings of a fossil from the early-branching eudicot order Ranunculales estimated at 122.6-125.8 mya[488], we imposed a hard bound of 130 mya on the eudicots by implementing a uniform calibration prior between 0 and 130 mya.

We found that when imposing E4' and running a scenario without data[219], the marginal prior calibration distributions of E1', E2', and E3' did not correspond to their specified calibration densities anymore. This type of behavior has been observed before, and has been ascribed to the fact that the marginal prior distribution is the combination of both the specified calibration density and the tree prior[367,483]. In fact, we experienced that implementing calibrations on nodes that were located very close to each other, in particular consecutive nodes, always resulted in a discrepancy between the specified calibration densities and effective marginal prior calibration distributions. We therefore increased parameter μ of calibrations E1', E2', and E3' until their marginal prior calibration distributions corresponded with their specified location at μ=8.0978, μ=4.5675, and μ=5.0703, respectively, as also illustrated in figure F.8.

WGD age estimates under the alternative calibration set

Table F.2 summarizes the WGD age estimates and their 90% CIs, as obtained using the alternative calibration set, while figure F.9 illustrates the resulting absolute age distributions.

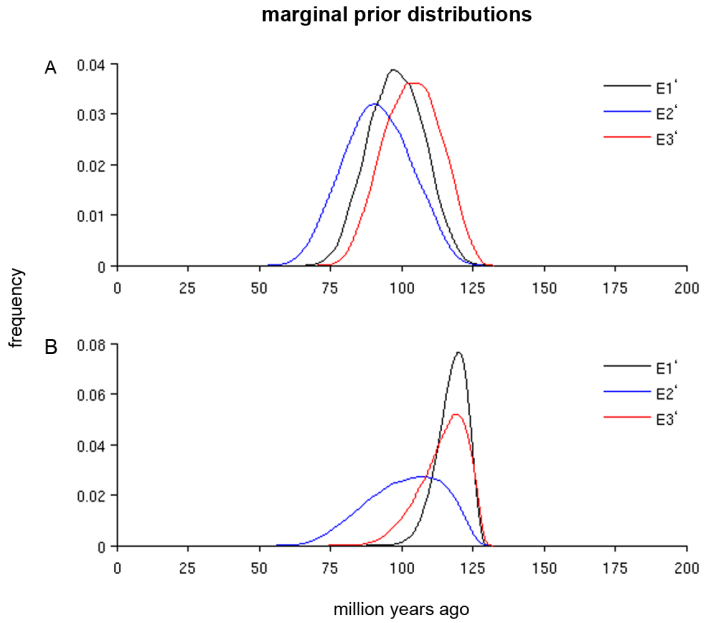


Figure F.8: Marginal prior distributions for the alternative calibration set. Marginal prior distributions for calibrations E1', E2', and E3' when E4' was also implemented with (A) μ=4.3460, μ=4.3443, and μ=4.2390, respectively (B) μ=8.0978, μ=4.5675, and μ=5.0703, respectively.

Table F.2: Overview of WGD age estimates under the alternative calibration set. Overview of the number of dated and accepted (ESS >200 for all statistics) orthogroups per species, and their resulting WGD age estimates with 90% confidence intervals (CIs). All orthogroups are based on anchors, except if indicated otherwise.

Species	# Dated (accepted) orthogroups	WGD age estimate (90% CI)
Malus domestica	99 (90)	17.95 (16.48-20.07)
Pyrus bretschneideri	1,000 (986)	18.53 (17.47-19.45)
Glycine max	1,000 (987)	12.31 (10.33-13.08)
Cajanus cajan	361 (351)	56.41 (53.41-60.26)
Medicago truncatula	79 (77)	64.95 (62.78-66.67)
Cicer arietinum	210 (201)	60.73 (59.01-65.20)
Lotus japonicus	19 (19)	61.87 (56.96-66.26)
Manihot esculenta	1,000 (977)	43.52 (42.45-44.80)
Linum usitatissimum	1,000 (987)	9.67 (8.94-10.62)
Populus trichocarpa	1,000 (983)	35.38 (34.07-36.56)
Brassica rapa	1,000 (975)	24.95 (23.22-26.34)
Thellungiella parvula	779 (758)	46.01 (44.91-47.14)
Arabidopsis thaliana α*	754 (736)	47.58 (45.90-48.75)
Arabidopsis thaliana β*	9 (9)	55.86 (0-65.20)
Arabidopsis lyrata	706 (686)	46.37 (45.13-47.22)
Gossypium raimondii	1,000 (968)	54.36 (53.00-55.49)
Solanum lycopersicum	479 (466)	62.27 (61.01-63.63)
Solanum tuberosum	478 (462)	59.74 (57.77-62.67)
Lactuca sativa †	451 (422)	55.97 (53.70-57.80)
Aquilegia formosa x pubescens †	55 (49)	51.17 (45.82-60.55)
Brachypodium distachyon	319 (300)	66.04 (63.85-68.75)
Hordeum vulgare †	323 (303)	72.93 (70.26-74.49)
Phyllostachys heterocycla	503 (487)	18.53 (17.47-20.11)
Oryza sativa	334 (319)	62.75 (60.37-68.28)
Zea mays	948 (913)	19.30 (18.42-19.93)
Sorghum bicolor	170 (164)	66.08 (63.11-69.96)
Setaria italica	309 (296)	66.15 (64.10-68.75)
*Musa acuminata***	367 (346)	65.27 (61.54-67.73)
Phoenix dactylifera	32 (29)	53.11 (47.66-55.79)
Nuphar advena †	119 (115)	69.23 (63.74-73.15)
Physcomitrella patens	319 (255)	55.79 (51.83-65.79)

† Based on peak-based duplicates.
*α and β refer to the *A. thaliana alpha* and *beta* duplication, respectively [173].
**This event most likely represents 2 separate WGDs in close succession [330].

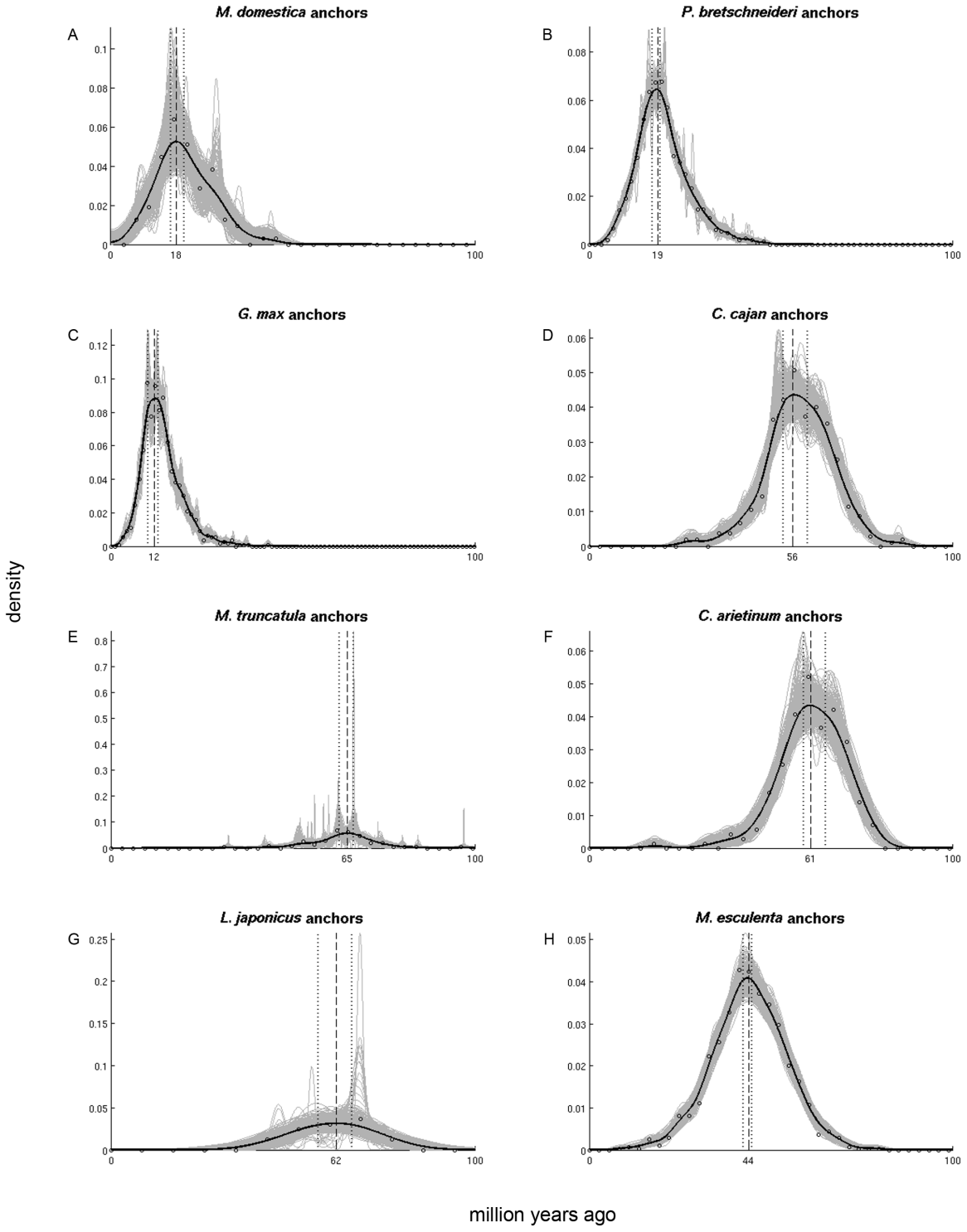


Figure F.9: Absolute age distributions obtained under the alternative calibration set. Absolute age distributions obtained under the alternative calibration set for (A) *M. domestica*, (B) *P. bretschneideri*, (C) *G. max*, (D) *C. cajan*, (E) *M. truncatula*, (F) *C. arietinum*, (G) *L. japonicus*, and (H) *M. esculenta*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table F.2 for sample sizes and exact confidence interval boundaries.

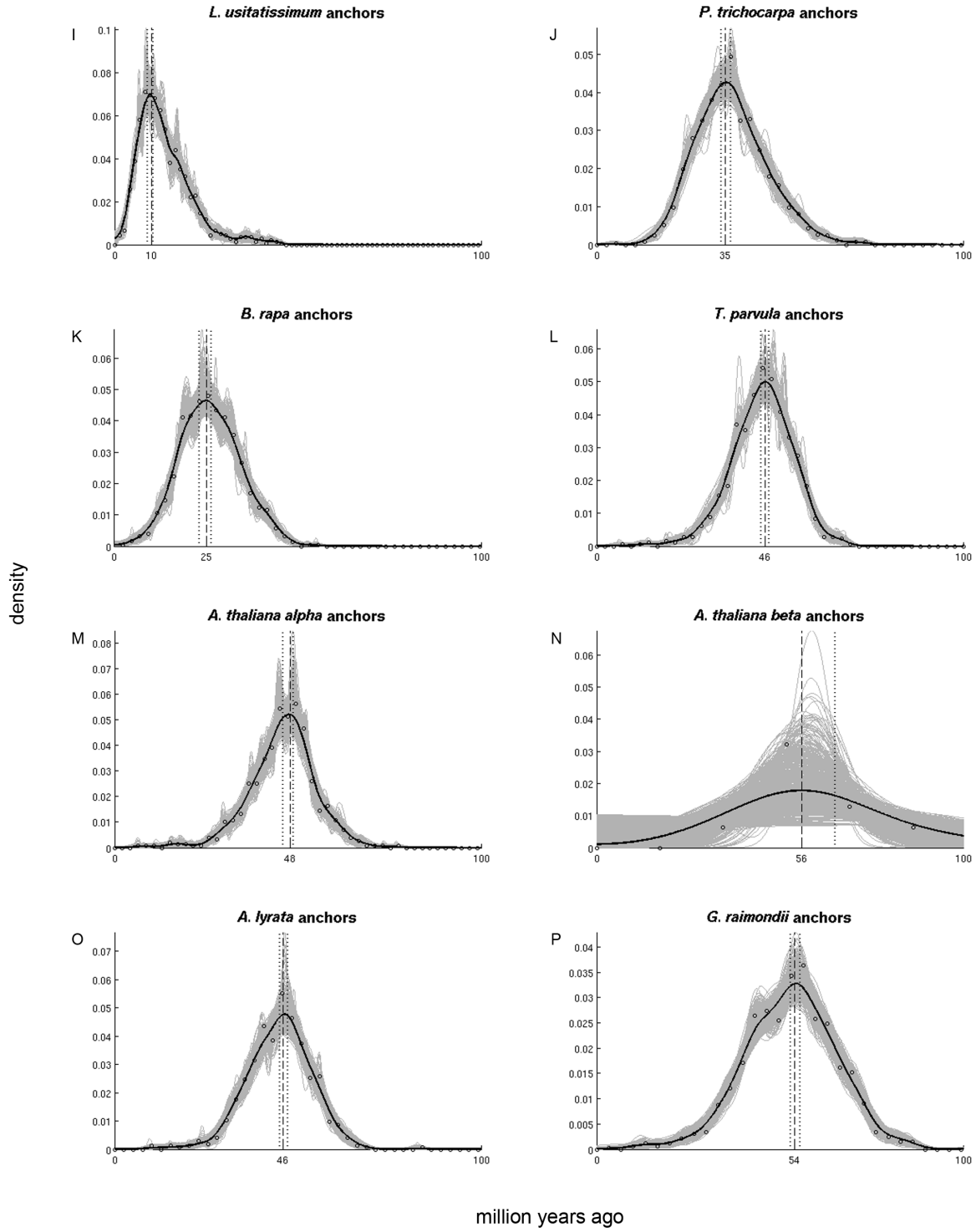


Figure F.9: Absolute age distributions obtained under the alternative calibration set - Continued. Absolute age distributions obtained under the alternative calibration set for (I) *L. usitatissimum*, (J) *P. trichocarpa*, (K) *B. rapa*, (L) *T. parvula*, (M) *A. thaliana alpha*, (N) *A. thaliana beta*, (O) *A. lyrata*, and (P) *G. raimondii*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table F.2 for sample sizes and exact confidence interval boundaries.

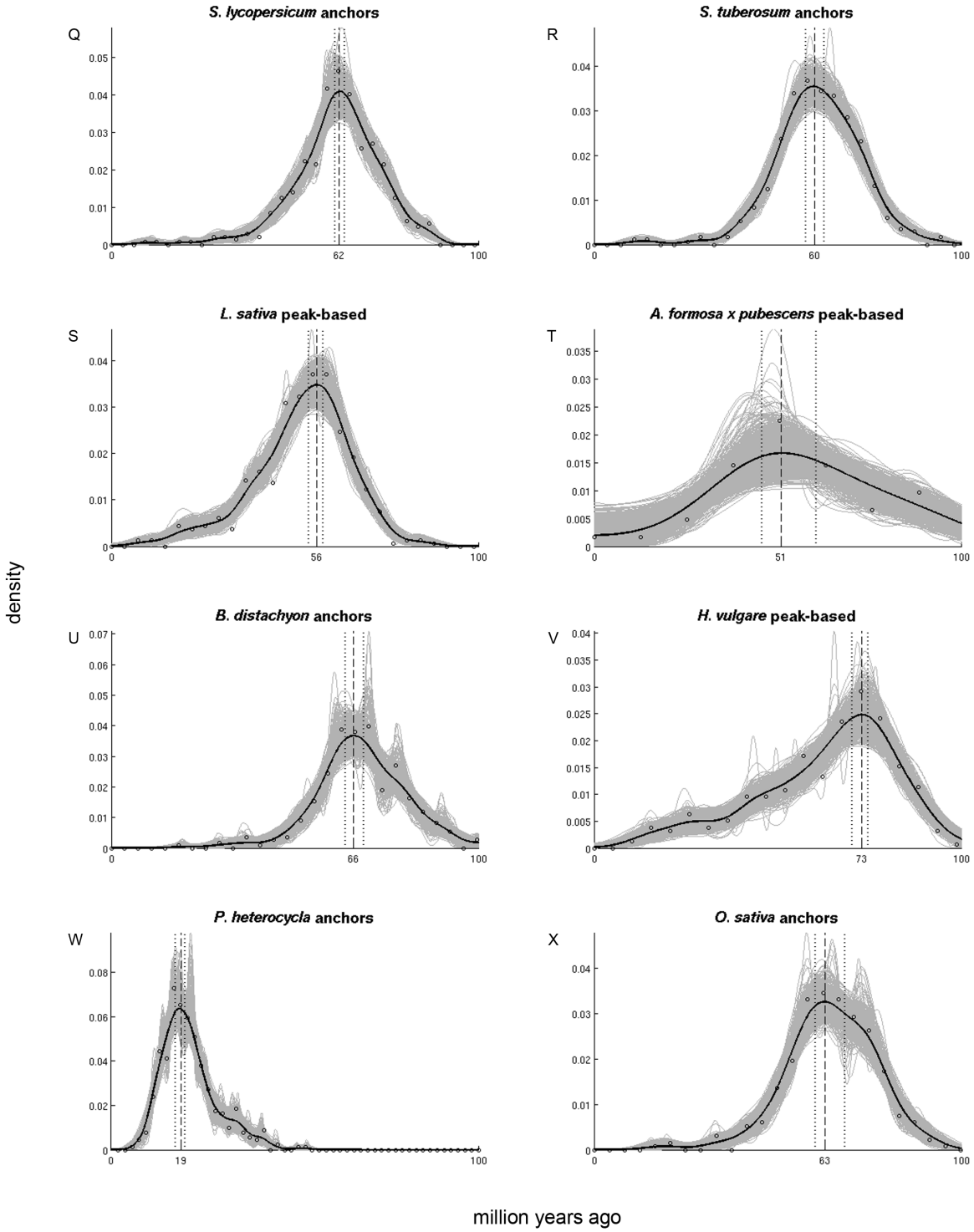


Figure F.9: Absolute age distributions obtained under the alternative calibration set - Continued. Absolute age distributions obtained under the alternative calibration set for (Q) *S. lycopersicum*, (R) *S. tuberosum*, (S) *L. sativa*, (T) *A. formosa x pubescens*, (U) *B. distachyon*, (V) *H. vulgare*, (W) *P. heterocycla*, and (X) *O. sativa*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table F.2 for sample sizes and exact confidence interval boundaries.

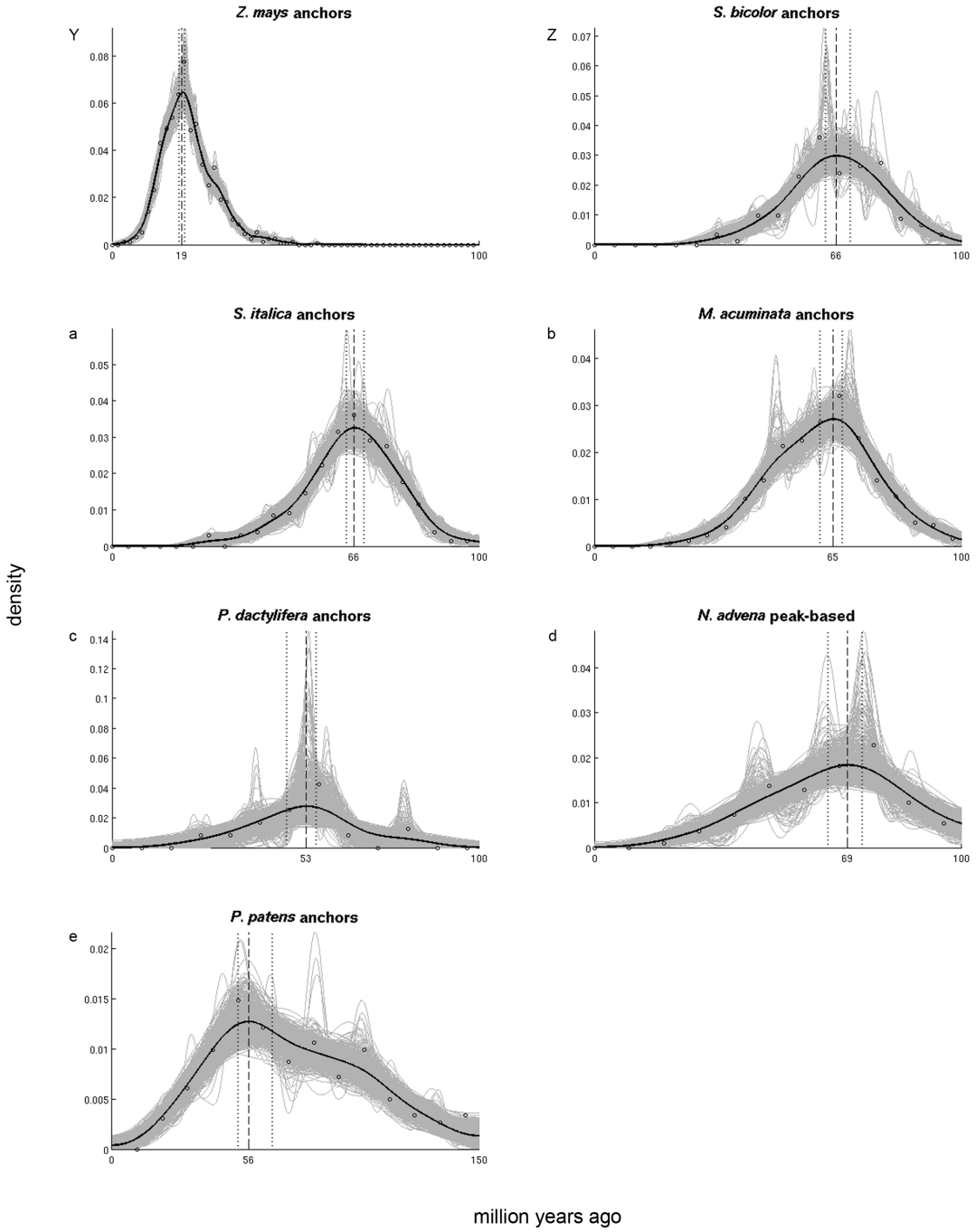


Figure F.9: Absolute age distributions obtained under the alternative calibration set - Continued. Absolute age distributions obtained under the alternative calibration set for (Y) *Z. mays*, (Z) *S. bicolor*, (a) *S. italica*, (b) *M. acuminata*, (c) *P. dactylifera*, (d) *N. advena*, and (e) *P. patens*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See table F.2 for sample sizes and exact confidence interval boundaries.

The WGD age estimates obtained under the alternative calibration set presented in table F.2 generally agree very well with the WGD age estimates obtained under the original calibration set presented in table 4.1. Not surprisingly, implementation of a hard maximum bound on the eudicot crown node results in WGD age estimates and 90% CIs that are slightly younger. A similar shift is also apparent in other large-scale molecular dating studies within the angiosperms where this constraint was implemented[64], compared to studies where this was not the case[65]. However, the 90% CIs obtained under the alternative calibration set overlap in all but two cases with the 90% CIs obtained under the original calibration set, and are on average only 1.57 million years younger for the complete set of all 31 species-specific WGD age estimates presented in table F.2. The *G. raimondii* WGD, and the Brassicaceae *alpha* WGD shared by *A. thaliana*, *A. lyrata*, and *T. parvula*, constitute the only two WGDs where the 90% CIs of WGD age estimates obtained under the alternative calibration set do not overlap with those of the original calibration set. The *G. raimondii* WGD is 3.66 million years younger under the alternative calibration set, while the Brassicaceae alpha WGD is 2.53 million years younger (average of WGD age estimates of *A. thaliana*, *A. lyrata*, and *T. parvula*).

The WGD age estimates obtained under the original calibration set can arguably be considered more reliable for three reasons. First, with regard to the hard maximum bound used for the eudicot fossil pollen calibration, it needs to be remarked that a fossil can in fact only provide unequivocal evidence on a hard minimum bound, but not on a hard maximum bound. A hard minimum bound is provided by the earliest age to which the fossil can reliably be attributed to, whereas a maximum bound always needs to be inferred based on other types of evidence such as older fossils and stratigraphic information. The latter is therefore error-prone, which is exactly why soft maximum bounds were introduced[198]. Recently, it was convincingly demonstrated that when the sequence signal is sufficiently strong and indicates an age different from the one suggested by the fossil calibration prior, soft maximum bounds can indeed allow to overcome a strong calibration prior[67], whereas this evidently is not possible when a hard maximum bound has been imposed. Additionally, it has been suggested that hard maximum bounds result in narrower confidence intervals on the posterior divergence time estimates, which do not represent genuine high precision but rather the conflict between fossil and sequence information[198]. Soft maximum bounds are therefore always preferred, and it was in fact argued that eudicot tricolpate fossil pollen constituted the only exception against these guidelines that was deemed acceptable[222]. In light of later scrutiny of the interpretation of eudicot tricolpate fossil pollen[65,488], a calibration strategy that strictly follows the conservative guidelines detailed above without allowing for any exceptions is preferable. Such a strategy does not question the value of eudicot tricolpate fossil pollen itself, but simply applies the same rules as enforced for all other fossil information.

Second, irrespective of this hard maximum constraint on the eudicot crown node, the fossils employed in the alternative calibration set may also be less optimal in the context of molecular sequence divergence estimation. The alternative calibration set contains calibrations with minimum bounds located more closely to the tips of the orthogroups compared to the original calibrations. It has been demonstrated that an abundance of constraints near the tips can bias the estimates for deeper nodes[499]. Further, because the alternative calibrations have much younger minimum bounds, but necessarily still describe divergence events quite far from these minimum bounds due to a lack of genome sequences for intermediate taxa, the resulting marginal calibration priors are much wider, and hence more diffuse and

uninformative (see figure F.8). Informative calibration priors on these nodes are nevertheless important because they represent a period of angiosperm diversification that is characterized by "layer upon layer of rapid radiation"[63], for which informative calibration priors are most likely imperative to guide the posterior divergence time estimates. Simply combining all calibrations from both the original and alternative calibration set is not a viable option, because this would result in a scenario where the large majority of nodes within the orthogroup topology have a calibration prior imposed. This is problematic because calibrating the large majority of the available nodes can only lead to conclusions compatible with the prior assumptions, since even a very strong sequence signal will not be able to correct posterior divergence time estimates if the majority of the nodes situated close to the divergence of interest (i.e., the homeologous pair) carry a strong prior[482]. Additionally, the effective marginal prior distributions and specified calibration densities will always differ when specified priors on nested clades overlap temporally[367], which is something we noticed in our own dataset as soon as calibration priors were specified on nodes located too close to each other.

Third, evaluation of the resulting absolute age distributions for all species-specific WGDs obtained under the alternative calibration set (see figure F.9), indicates that they become less informative compared to the absolute age distributions obtained under the original calibration set (see figure 4.2 and figure F.2). This is for instance particularly evident for the *A. thaliana beta* absolute age distribution. The original WGD age estimate and 90% CI of 61.21 mya and 54.58 to 69.38 mya, respectively, were necessarily based on only nine dated anchor pairs (see table 4.1). Despite this very low number, we deemed this WGD age estimate fairly reliable because of the relatively strong unimodal pattern of its absolute age distribution (see figure F.2, panel I). Furthermore, this was re-affirmed by its peak-based absolute age distribution that was based on a much larger number of orthogroups, but still arrived at a very similar WGD age estimate and 90% CI of 62.97 mya and 56.04 to 70.01 mya, respectively. Under the alternative calibration set however, a WGD age estimate and 90% CI of 55.86 mya and 0 to 65.20 mya, respectively, were obtained for this WGD (see table F.2). The latter appears a particularly strong shift, but evaluation of the new absolute age distribution indicates that it exhibits a very uninformative shape (see figure F.9, panel N). In particular, its kernel density estimate is very wide with only a poorly supported peak, as also indicated by the bootstrap replicates that reveal a mostly flat surface curve with a very diffuse peak. Consequently, the resulting 90% CI is over 65 million years wide. Although the uninformative shape of this absolute age distribution obtained under the alternative calibration set is not particularly striking, considering that it only consists of nine dated anchors, the drastic difference with the informative shape obtained under the original calibration set is remarkable. This most likely indicates that the new constraints imposed by the alternative calibration set conflict with the sequence signal to some extent.

In conclusion, using an alternative calibration set with in particular a hard maximum constraint on the eudicot crown node, we find that the resulting WGD age estimates are overall in good agreement with those obtained under the original calibration set, being on average only 1.57 mya younger and possessing overlapping 90% CIs for all but two independent WGDs, suggesting that our conclusions are robust against the particular choice of employed calibrations.

F.3.4 Relative rate tests

To obtain a measure for the relative rate at which species used in dating the WGDs evolve, we performed pairwise relative rate tests (RRTs) between the different WGDs. We used *P. patens* as an outgroup, since this allows consistent comparison of all other dated WGDs. Anchors and peak-based duplicates from different species used for dating WGDs were collected and grouped by plant order. Transcriptome assemblies were not considered because no positional information is available for these. Table F.3 lists all employed species.

Table F.3: Overview of species employed for RRT comparisons.

Plant order	Code	Used species
Rosales	ROS	*P. bretschneideri*, *M. domestica*
Fabales	FAB	*M. truncatula*, *C. cajan*, *L. japonicus*, *C. arietinum*
Malpighiales	MAL	*M. esculenta*, *P. trichocarpa*
Brassicales	BRA	*A. thaliana*, *A. lyrata*, *T. parvula*
Malvales	MAV	*G. raimondii*
Solanales	SOL	*S. lycopersicum*, *S. tuberosum*
Poales	POA	*O. sativa*, *B. distachyon*, *S. italica*, *S. bicolor*, *H. vulgare*
Zingiberales	ZIN	*M. acuminata*
Arecales	ARE	*P. dactylifera*

The evolutionary rates between orthologs used in dating the WGDs, grouped by plant order, were then compared in a pairwise fashion. Orthogroups were constructed for each pairwise comparison based on Inparanoid data for *P. patens*, and always included the *P. patens* ortholog as outgroup and two orthologs representing the specific plant orders being compared. We performed the RRTs employing HyPhy (v2.0)[252], using a WAG model of evolution[500] with gamma-distributed rate heterogeneity across sites using four rate categories[341] for all orthogroups. Table F.4 lists the fraction of all orthogroups evolving faster, and the total sample sizes, between all pairwise comparisons of orders. Table F.5 does the same but only considers the orthogroups that were found to evolve significantly faster ($p<0.05$).

Table F.4: Fraction of all orthogroups evolving faster. Fraction of orthogroups evolving faster for the orders listed in the rows compared to the orders listed in the columns. The lower diagonal of the matrix lists the percentages, while the upper diagonal lists the sample sizes upon which these percentages are based.

from/to	ROS	FAB	MAL	BRA	MAV	SOL	POA	ZIN	ARE
ROS	x	438	1129	544	71	460	552	161	303
FAB	0.56	x	660	450	52	406	469	107	200
MAL	0.46	0.38	x	841	120	666	846	216	439
BRA	0.63	0.57	0.66	x	74	503	633	98	252
MAV	0.52	0.42	0.53	0.23	x	55	79	22	27
SOL	0.52	0.46	0.53	0.41	0.56	x	524	99	175
POA	0.62	0.63	0.68	0.51	0.7	0.58	x	120	249
ZIN	0.55	0.55	0.56	0.38	0.5	0.4	0.38	x	97
ARE	0.45	0.43	0.5	0.35	0.56	0.46	0.31	0.41	x

To facilitate evaluation, we scored each comparison binary as either evolving faster (1) or slower (0) depending on the fractions listed in table F.5, using 50% as the cut-off. Since for the comparison between the Malvales and Fabales, no single statistically significant orthogroup was identified, this was scored as 1 based on the comparison of all their orthogroups in table F.4. Similarly, since exactly half of all scored orthogroups evolved slower/faster for the comparison between the Solanales and Malvales, this was scored as 0 based on the comparison of all their orthogroups in table F.4. The resulting binary matrix is listed in table F.6.

Table F.5: Fraction of orthogroups evolving significantly faster ($p<0.05$). Fraction of orthogroups evolving significantly faster ($p<0.05$) for the order listed in the rows compared to the orders listed in the columns. The lower diagonal of the matrix lists the percentages, while the upper diagonal lists the sample sizes upon which these percentages are based.

from/to	ROS	FAB	MAL	BRA	MAV	SOL	POA	ZIN	ARE
ROS	x	49	94	71	4	43	95	14	36
FAB	0.65	x	89	73	n/a	47	83	13	36
MAL	0.44	0.27	x	115	7	77	143	25	57
BRA	0.77	0.67	0.83	x	12	42	73	9	58
MAV	0.25	n/a	0.43	0.17	x	6	10	3	3
SOL	0.58	0.36	0.51	0.31	0.5	x	74	8	21
POA	0.75	0.65	0.87	0.59	1	0.72	x	17	38
ZIN	0.79	0.54	0.72	0.33	0.33	0.63	0.35	x	9
ARE	0.58	0.36	0.65	0.28	1	0.52	0.08	0.44	x

Table F.6: Binary matrix representing the relationships between all considered plant orders. 0 and 1 represent an overall slower or faster evolutionary rate between the orders listed in the rows compared to the orders listed in the columns, respectively.

from/to	ROS	FAB	MAL	BRA	MAV	SOL	POA	ZIN	ARE
ROS	x	0	1	0	1	0	0	0	0
FAB	1	x	1	0	1	1	0	0	1
MAL	0	0	x	0	1	0	0	0	0
BRA	1	1	1	x	1	1	0	1	1
MAV	0	0	0	0	x	0	0	1	0
SOL	1	0	1	0	1	x	0	0	0
POA	1	1	1	1	1	1	x	1	1
ZIN	1	1	1	0	0	1	0	x	1
ARE	1	0	1	0	1	1	0	0	x

Although our current approach is arguably very crude because different species belonging to the same plant order do not necessarily share the same evolutionary rates, similar trends based on similar life history traits are expected[191]. We tried an alternative strategy where individual species instead of plant orders were compared but this led to sample sizes that were too low for statistical evaluation. Despite the fact that our results should therefore be interpreted with due caution, our current approach allows for a rudimentary comparison between the different plant orders. This is supported by the fact that the resulting relationships between the different plant orders in the binary matrix are very consistent, ordered from slowest to fastest as follows:

$$\text{MAV} < \text{MAL} < \text{ROS} < \text{SOL} < \text{ARE} < \text{FAB} < \text{ZIN} < \text{BRA} < \text{POA}$$

The above association represents the most parsimonious relationship between all plant orders. There was only one error in the binary matrix against this relationship, namely the comparison between the Zingiberales and Malvales, which was scored as 0 but should have been scored as 1. This is most likely because of a low sample size, as only three orthogroups were scored as statistically significant. All other comparisons in the binary matrix were consistent according to the relationships listed above.

F.3.5 Re-dating the *Pyrus bretschneideri* WGD

We presented fossil evidence that suggests that the ages of both the *P. trichocarpa* WGD and the WGD shared by *M. domestica* and *P. bretschneideri* are underestimated by our dating approach, most likely because of a drastic rate shift associated with their woody status that could not be completely

corrected for. In case of the *P. trichocarpa* WGD, we quoted fossil information that establishes that the divergence between *Salix* and *Populus* is at least 47.4 million years old[375]. Although there is no genome sequence information available for *Salix*, it is well established that *Salix* and *Populus* shared the WGD in question[372,376]. A calibration on the node joining the *P. trichocarpa* homeologous pair enforcing a minimum age of 47.4 million years could therefore theoretically have been implemented. However, it remains very difficult to decide on a proper shape for the calibration prior that would not inadvertently bias the eventual WGD age estimate. Lognormal calibration priors are preferred[67], but posterior time estimates are pulled to some extent towards their peak mass probability[66]. Incorporating prior information on the location of the peak mass, for which the current best estimate is in fact ∼65 mya[193], would hence be highly undesirable because it entails placing a strong peak mass probability at 65 mya on the node joining the homeologous pair. Alternative shapes for this particular calibration are equally questionable. The most basic form, a uniform calibration prior, requires arbitrarily 'safe' high maximum bounds, since it is very difficult to distinguish proper upper boundaries based on the fossil record[198]. The risk that the sequence signal is not strong enough to overcome poor calibration priors is inherent to all molecular dating[198]. A strategy that avoids placing any *a priori* fossil evidence upon the node joining the homeologous pair is hence preferable because it ensures that the sequence signal of this node will yield the most unbiased age estimate possible, based upon other rate-correcting calibrations in the orthogroup topology.

The same applies to the WGD shared by *M. domestica* and *P. bretschneideri*. There is fossil evidence that indicates that their divergence should be at least 48.7 million years old[378], so that a calibration with this minimum bound could theoretically have been implemented on their homeologous pairs, which is nevertheless undesirable in light of the above. However, because there are more sequenced Rosaceae genomes available, we can break up the long branch leading to the homeologous pair by increasing the taxon sampling around this node, and also introduce a new calibration based on this fossil information closer to, but not on, the homeologous pair. Applying the same strategy for *P. trichocarpa* is impossible because the latter is the only genome available at the moment within the Salicaceae, while the most closely related available genome sequences are situated within other families of the Malpighiales, which all diverged about ∼100 mya[68]. We re-dated the *P. bretschneideri* WGD based on its anchors, because these are based on bona fide duplicated segments and many more anchors were available for this species compared to *M. domestica* (see table 4.1). To break up the long branch leading the *P. bretschneideri* homeologous pairs, we included both one *Fragaria* and *Prunus* ortholog into the orthogroup topology, instead of grouping these together in one species group for which only one ortholog was required (see figure F.5). We inserted a new primary fossil calibration, based on the aforementioned fossil evidence, to calibrate the divergence between the homeologous pair and the *Prunus* ortholog. The divergence between *Pyrus* and *Prunus* has been estimated at ∼73 mya[380]. We therefore specified a lognormal calibration prior with μ=3.4405, σ=0.5, and a minimum bound of 48.7 mya. A run without data[219,483] indicated however that the marginal prior calibration distribution did not correspond to its specified calibration density, and we had to increase μ to a value of 3.7851 so that the marginal prior calibration distribution was located at 73 mya. Apart from this new calibration, calibrations E2 and R1 were also implemented (see figure F.6), while calibration E1 had to be removed because it overlapped temporally on a nested clade with the new calibration[367]. In total, 1,000 orthogroups were constructed

and dated, of which 978 were accepted afterwards (ESS >200 for all statistics). The resulting absolute age distribution is presented in figure F.10.

A new WGD age estimate of 30.1 mya was obtained for the *P. bretschneideri* WGD. This constitutes an increase of more than 10 million years with respect to our original WGD age estimate of 19.85 mya, but is still 18.6 million years short of the previously described minimum fossil bound of 48.7 mya. This confirms that incomplete correction of rate deceleration led to an underestimation of the *P. bretschneideri* WGD, and that breaking up long branches in orthogroup phylogenies through better taxon sampling, in combination with new rate-correcting fossil calibrations, will help to correct for drastic rate shifts when more full plant genome sequences become available in the future.

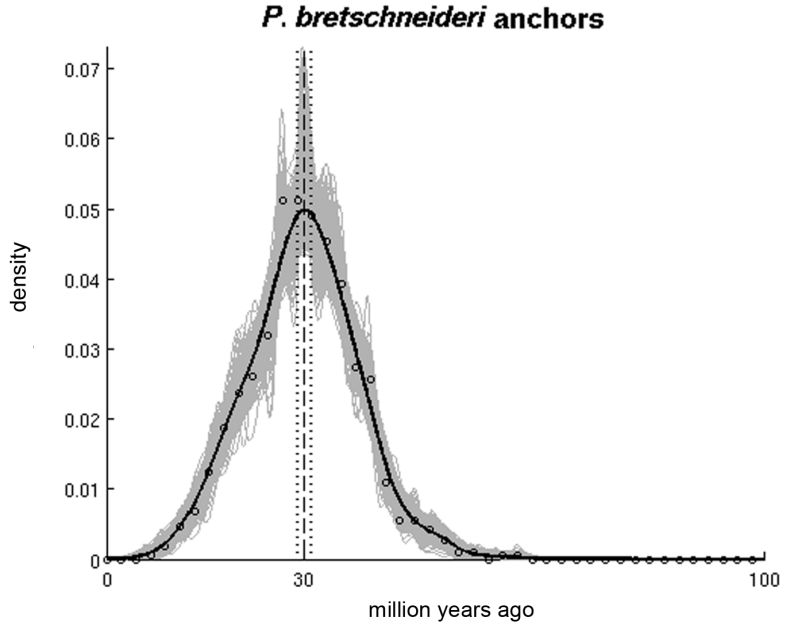


Figure F.10: Re-dating the *Pyrus bretschneideri* WGD. Absolute age distribution of the dated anchors for *P. bretschneideri* with improved taxon sampling and a new primary fossil calibration closer to the homeologous pair. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated by open dots. The WGD age is estimated at 30.15 mya, with a lower and upper 90% confidence interval boundary of 29.23 and 31.14 mya, respectively.

F.3.6 WGD age estimates from literature

The following WGD age estimates, corresponding to the black bars in figure 4.3, were taken from literature. The *N. nucifera* WGD was estimated at 65 mya[459]. The oldest WGD in *M. acuminata* was estimated at 96 mya[330]. The core eudicot shared gamma hexaploidy was estimated somewhere between 117 and 133 mya[138,139]. The oldest shared WGD in the grasses, also referred to as *rho*, was estimated at 130 mya based on the median synonymous substitution rate, which was however close to saturation and therefore should be interpreted with caution[140]. Considering that both the Zingiberales and Arecales, which do not share this event, most likely branched off somewhere around 120 mya[71,480,481,491], we placed this WGD right after the origin of the grasses, but its exact age remains unknown. The angiosperm- and seed plant-wide WGDs were estimated at 192 and 319 mya, respectively[136].

F.3.7 *Eschscholzia californica* and *Acorus americanus*

We originally included all transcriptome assemblies from a previous study[193], including *E. californica* and *A. americanus*, both of which were originally also dated close to the K-Pg boundary. However, in the current study, using the updated approaches, we were unable to obtain unambiguous WGD age estimates for both species. In the case of *E. californica*, only 15 orthogroups based on peak-based duplicates could be constructed, of which 14 were accepted (ESS $>$200 for all statistics). Their resulting absolute age distribution is presented in figure F.11. The mode of the underlying kernel density estimate was located at 58.23 mya, very close to the Gaussian component located at 60.05 mya in association with the K-Pg boundary (see figure F.4). However, our KDE bootstrapping procedure demonstrated the presence of a very strong bimodal underlying shape with one peak located at $\sim$43 mya, and another peak at $\sim$74 mya, as evidenced both by the open dots (representing the raw data) and grey curves (representing the bootstrap samples) on figure F.11. Inclusion of this WGD in our results, represented by a very wide bar on figure 4.3, would however be misleading, as its estimate of 58.23 mya would increase statistical support for the clustering of WGDs with the K-Pg boundary, whereas evaluation of its absolute age distribution demonstrates that this estimate clearly cannot be trusted. This is not necessarily due to the low number of dated homeologs, as other absolute age distributions, such as for instance the absolute age distribution of *L. japonicus* based on anchors (see figure 4.2, panel C), are based on a similar small number of dated homeologs. The latter nevertheless shows strong support for a unimodal distribution, which is reinforced by its peak-based absolute age distribution that is based on a much larger number of homeologous pairs and displays a similar trend. The example of *E. californica* thus demonstrates the strengths of our bootstrapping KDE approach by allowing the exclusion of dubious WGD age estimates. In contrast, fitting a standard parametric distribution, such as a gamma or normal distribution, would forcibly fit a unimodal shape to a bimodal distribution and lead to the inclusion of erroneous data for statistical analysis of clustering.

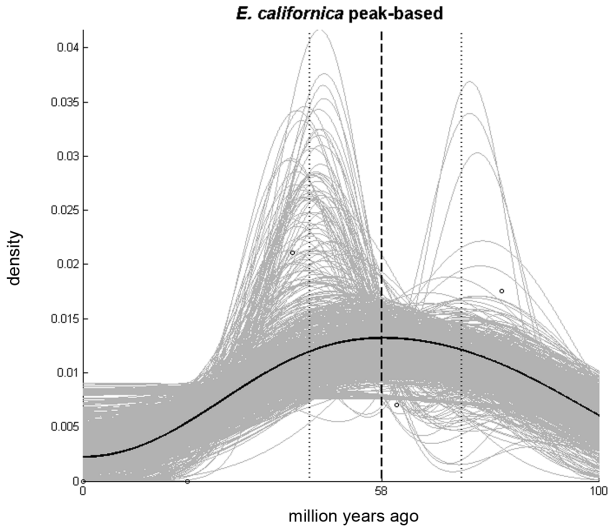


Figure F.11: Absolute age distribution of the dated peak-based duplicates for *E. californica*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated by open dots. The WGD age is estimated at 58.28 mya, with a lower and upper 90% confidence interval boundary of 42.28 and 74.10 mya, respectively.

In the case of *A. americanus*, 35 orthogroups based on peak-based duplicates could be constructed, which were all accepted (ESS >200 for all statistics). Their resulting absolute age distribution is presented in figure F.12. The mode of the underlying kernel density estimate was located at 33.26 mya, very far from the K-Pg boundary. However, our bootstrapping KDE procedure demonstrated a very uninformative shape. In particular, the kernel density estimate is very wide with only a poorly supported peak that barely protrudes above the background, as also indicated by the bootstrap replicates themselves that reveal a mostly flat curve surface. In fact, the bootstrap replicates indicate the presence of a very diffuse peak centered on the 90% confidence interval upper boundary that is masked by the flat left flank, but still evident by the decreasing right flank. A trustworthy estimate for the *A. americanus* WGD, similarly to the *E. californica* WGD, hence remains elusive.

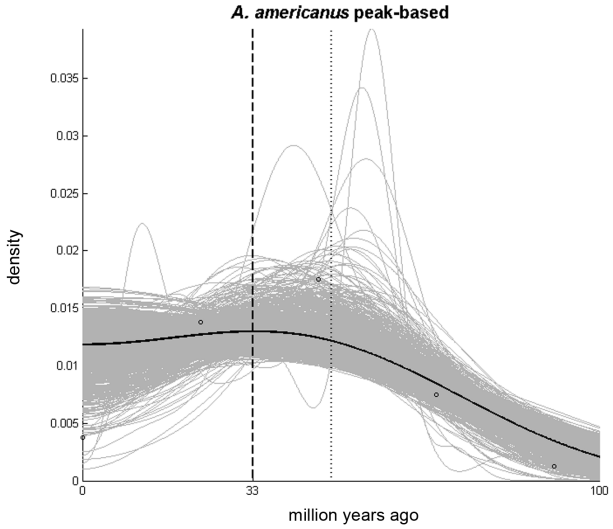


Figure F.12: Absolute age distribution of the dated peak-based duplicates for *A. americanus*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated by open dots. The WGD age is estimated at 33.26 mya, with a lower and upper 90% confidence interval boundary of 0.00 and 48.17 mya, respectively.

Appendix G

Bibliography

[1] Watson JD and Crick FH 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**(4356):737–738.

[2] Luger K, Dechassa ML, and Tremethick DJ 2012. New insights into nucleosome and chromatin structure: An ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology* **13**(7):436–447.

[3] Crick FH 1958. On protein synthesis. *Symposia of the Society for Experimental Biology* **12**:138–163.

[4] Crick F 1970. Central dogma of molecular biology. *Nature* **227**(5258):561–563.

[5] Li WH 1997. *Molecular evolution*. Sinauer Associates.

[6] Shapiro JA 2009. Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences* **1178**:6–28.

[7] House AE and Lynch KW 2008. Regulation of alternative splicing: More than just the ABCs. *The Journal of Biological Chemistry* **283**(3):1217–1221.

[8] Maas S and Rich A 2000. Changing genetic information through RNA editing. *Bioessays* **22**(9):790–802.

[9] Khoury GA, Baliban RC, and Floudas CA 2011. Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database. *Scientific Reports* **1**(90):1–5.

[10] Shapiro JA and von Sternberg R 2005. Why repetitive DNA is essential to genome function. *Biological Reviews of the Cambridge Philosophical Society* **80**(2):227–250.

[11] Cech TR 1989. RNA as an enzyme. *Biochemistry International* **18**(1):7–14.

[12] Temin HM and Mizutani S 1970. Viral RNA-dependent DNA polymerase - RNA-dependent DNA polymerase in virions of rous sarcoma virus. *Nature* **226**(5252):1211–1213.

[13] Mattick JS 2004. RNA regulation: A new genetics? *Nature Reviews Genetics* **5**(4):316–323.

[14] Goodman MF 2002. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annual Review of Biochemistry* **71**:17–50.

[15] Dobzhansky T 1973. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher* **35**(3):125–129.

[16] Darwin C 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.

[17] Pigliucci M 2007. Do we need an extended evolutionary synthesis? *Evolution* **61**(12):2743–2749.

[18] Lyell CS 1830. *Principles of geology : Being an attempt to explain the former changes of the earth's surface, by reference to causes now in operation*. John Murray.

[19] Friedman WE 2009. The meaning of Darwin's 'abominable mystery'. *American Journal of Botany* **96**(1):5–21.

[20] Gould SJ and Lewontin RC 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society B: Biological Sciences* **205**(1161):581–598.

[21] Fisher RA 1919. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**(2):399–433.

[22] Brooks DR 2011. The mastodon in the room: How Darwinian is neo-Darwinism? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **42**(1):82–88.

[23] Gould SJ 1980. Is a new and general theory of evolution emerging? *Paleobiology* **6**(1):119–130.

[24] Rose MR and Oakley TH 2007. The new biology: Beyond the modern synthesis. *Biology Direct* **2**(30):1–17.

[25] Gould SJ 1997. The exaptive excellence of spandrels as a term and prototype. *Proceedings of the National Academy of Sciences USA* **94**(20):10750–10755.

[26] Sheldon MP 2014. Claiming Darwin: Stephen Jay Gould in contests over evolutionary orthodoxy and public perception, 1977-2002. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* .

[27] Mc CB 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences USA* **36**(6):344–355.

[28] Kimura M 1968. Evolutionary rate at the molecular level. *Nature* **217**(5129):624–626.

[29] Knight RD, Freeland SJ, and Landweber LF 2005. *The Genetic Code and the Origin of Life*, chapter Adaptive evolution of the genetic code. Springer.

[30] Ohta T 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **246**(5428):96–98.

[31] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822):860–921.

[32] Lynch M and Conery JS 2003. The origins of genome complexity. *Science* **302**(5649):1401–1404.

[33] Kreitman M 2000. Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics* **1**:539–559.

[34] Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *Plos Biology* **5**(11):2534–2559.

[35] Hahn MW 2008. Toward a selection theory of molecular evolution. *Evolution* **62**(2):255–265.

[36] Nei M, Suzuki Y, and Nozawa M 2010. The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics* **11**:265–289.

[37] West-Eberhard MJ 2003. *Developmental plasticity and evolution*. Oxford University Press.

[38] Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57–74.

[39] Pennisi E 2012. Genomics ENCODE project writes eulogy for junk DNA. *Science* **337**(6099):1159–1161.

[40] Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, et al. 2013. On the immortality of television sets: Function in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution* **5**(3):578–590.

[41] Eldredge N and Gould SJ 1972. Punctuated equilibria: An alternative to phyletic gradualism. *Models in Paleobiology* **82**:115.

[42] Goldschmidt R 1940. *The material basis of evolution*. Yale University Press.

[43] Dietrich MR 2003. Richard Goldschmidt: Hopeful monsters and other 'heresies'. *Nature Reviews Genetics* **4**(1):68–74.

[44] Vergara-Silva F 2003. Plants and the conceptual articulation of evolutionary developmental biology. *Biology & Philosophy* **18**(2):249–284.

[45] Ronshaugen M, McGinnis N, and McGinnis W 2002. Hox protein mutation and macroevolution of the insect body plan. *Nature* **415**(6874):914–917.

[46] Mondragon-Palomino M and Theissen G 2009. Why are orchid flowers so diverse? Reduction of evolutionary constraints by paralogues of class B floral homeotic genes. *Annals of Botany* **104**(3):583–594.

[47] Ohya YK, Kuraku S, and Kuratani S 2005. Hox code in embryos of Chinese soft-shelled turtle *Pelodiscus sinensis* correlates with the evolutionary innovation in the turtle. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **304**(2):107–118.

[48] Theissen G 2006. The proper place of hopeful monsters in evolutionary biology. *Theory in Biosciences* **124**(3-4):349–369.

[49] Theissen G 2009. Saltational evolution: Hopeful monsters are here to stay. *Theory in Biosciences* **128**(1):43–51.

[50] Pennell MW, Harmon LJ, and Uyeda JC 2014. Is there room for punctuated equilibrium in macroevolution? *Trends in Ecology & Evolution* **29**(1):23–32.

[51] Dickins TE and Rahman Q 2012. The extended evolutionary synthesis and the role of soft inheritance in evolution. *Proceedings of the Royal Society B: Biological Sciences* **279**(1740):2913–2921.

[52] Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, et al. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**(1):336–348.

[53] Hickey LJ and Wolfe JA 1975. The bases of angiosperm phylogeny: Vegetative morphology. *Annals of the Missouri Botanical Garden* **62**(3):538–589.

[54] Friis EM, Pedersen KR, and Crane PR 2006. Cretaceous angiosperm flowers: Innovation and evolution in plant reproduction. *Palaeogeography, Palaeoclimatology, Palaeoecology* **232**(2-4):251–293.

[55] Doyle JA 2012. Molecular and fossil evidence on the origin of angiosperms. *Annual Review of Earth and Planetary Sciences* **40**:301–326.

[56] Doyle JA 2005. Early evolution of angiosperm pollen as inferred from molecular and morphological phylogenetic analyses. *Grana* **44**(4):227–251.

[57] Sun G, Dilcher DL, Zheng SL, and Zhou ZK 1998. In search of the first flower: A Jurassic angiosperm, *Archaefructus*, from northeast China. *Science* **282**(5394):1692–1695.

[58] Zhou ZH, Barrett PM, and Hilton J 2003. An exceptionally preserved Lower Cretaceous ecosystem. *Nature* **421**(6925):807–814.

[59] Hochuli P, Heimhofer U, and Weissert H 2006. Timing of early angiosperm radiation: Recalibrating the classical succession. *Journal of the Geological Society* **163**(4):587–594.

[60] Morley RJ 2003. Interplate dispersal paths for megathermal angiosperms. *Perspectives in Plant Ecology, Evolution and Systematics* **6**(1):5–20.

[61] Magallon S and Castillo A 2009. Angiosperm diversification through time. *American Journal of Botany* **96**(1):349–365.

[62] Wang HC, Moore MJ, Soltis PS, Bell CD, Brockington SF, et al. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences USA* **106**(10):3853–3858.

[63] Bell CD, Soltis DE, and Soltis PS 2010. The age and diversification of the angiosperms re-revisited. *American Journal of Botany* **97**(8):1296–1303.

[64] Magallon S 2010. Using fossils to break long branches in molecular dating: A comparison of relaxed clocks applied to the origin of angiosperms. *Systematic Biology* **59**(4):384–399.

[65] Smith SA, Beaulieu JM, and Donoghue MJ 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proceedings of the National Academy of Sciences USA* **107**(13):5897–5902.

[66] Clarke JT, Warnock RC, and Donoghue PC 2011. Establishing a time-scale for plant evolution. *New Phytologist* **192**(1):266–301.

[67] Magallon S, Hilu KW, and Quandt D 2013. 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**(3):556–573.

[68] Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences USA* **109**(43):17519–17524.

[69] Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, and Mathews S 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences USA* **107**(43):18724–18728.

[70] Bremer K, Friis EM, and Bremer B 2004. Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. *Systematic Biology* **53**(3):496–505.

[71] Janssen T and Bremer K 2004. The age of major monocot groups inferred from 800+ *rbcl* sequences. *Botanical Journal of the Linnean Society* **146**(4):385–398.

[72] Soltis DE, Bell CD, Kim S, and Soltis PS 2008. Origin and early evolution of angiosperms. *Annals of the New York Academy of Sciences* **1133**:3–25.

[73] Coiffard C, Gomez B, Daviero-Gomez V, and Dilcher DL 2012. Rise to dominance of angiosperm pioneers in European Cretaceous environments. *Proceedings of the National Academy of Sciences USA* **109**(51):20955–20959.

[74] Ruban DA 2012. Mesozoic mass extinctions and angiosperm radiation: Does the molecular clock tell something new? *Geologos* **18**(1):37–42.

[75] Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, et al. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: The case of *Nothofagus* (Fagales). *Systematic Biology* **61**(2):289–313.

[76] Fiz-Palacios O, Schneider H, Heinrichs J, and Savolainen V 2011. Diversification of land plants: Insights from a family-level phylogenetic analysis. *BMC Evolutionary Biology* **11**(341):1–10.

[77] Friis EM, Pedersen KR, and Crane PR 2010. Diversity in obscurity: Fossil flowers and the early history of angiosperms. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**(1539):369–382.

[78] Rohde RA and Muller RA 2005. Cycles in fossil diversity. *Nature* **434**(7030):208–210.

[79] Raup DM 1994. The role of extinction in evolution. *Proceedings of the National Academy of Sciences USA* **91**(15):6758–6763.

[80] Renne PR, Deino AL, Hilgen FJ, Kuiper KF, Mark DF, et al. 2013. Time scales of critical events around the Cretaceous-Paleogene boundary. *Science* **339**(6120):684–687.

[81] Robertson DS, Lewis WM, Sheehan PM, and Toon OB 2013. K-Pg extinction: Reevaluation of the heat-fire hypothesis. *Journal of Geophysical Research: Biogeosciences* **118**(1):329–336.

[82] McElwain JC and Punyasena SW 2007. Mass extinction events and the plant fossil record. *Trends in Ecology & Evolution* **22**(10):548–557.

[83] Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, et al. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**(6055):521–524.

[84] Wilf P and Johnson KR 2004. Land plant extinction at the end of the Cretaceous: A quantitative analysis of the North Dakota megafloral record. *Paleobiology* **30**(3):347–368.

[85] Vajda V, Raine JI, and Hollis CJ 2001. Indication of global deforestation at the Creataceous-Tertiary boundary by New Zealand fern spike. *Science* **294**(5547):1700–1702.

[86] Anderson JM, Anderson HM, Archangelsky S, Bamford M, Chandra S, et al. 1999. Patterns of Gondwana plant colonisation and diversification. *Journal of African Earth Sciences* **28**(1):145–167.

[87] Ramirez SR, Gravendeel B, Singer RB, Marshall CR, and Pierce NE 2007. Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. *Nature* **448**(7157):1042–1045.

[88] Lavin M, Herendeen PS, and Wojciechowski MF 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Systematic Biology* **54**(4):575–594.

[89] Prasad V, Stromberg CA, Leache AD, Samant B, Patnaik R, et al. 2011. Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. *Nature Communications* **2**:480.

[90] Smith JF, Stevens AC, Tepe EJ, and Davidson C 2008. Placing the origin of two species-rich genera in the late Cretaceous with later species divergence in the Tertiary: A phylogenetic, biogeographic and molecular dating analysis of *Piper* and *Peperomia* (Piperaceae). *Plant Systematics and Evolution* **275**(1-2):9–30.

[91] Ricklefs RE 2007. Estimating diversification rates from phylogenetic information. *Trends in Ecology & Evolution* **22**(11):601–610.

[92] Quental TB and Marshall CR 2010. Diversity dynamics: Molecular phylogenies need the fossil record. *Trends in Ecology & Evolution* **25**(8):434–441.

[93] Ohno S 1970. *Evolution by gene duplication*. Springer-Verlag.

[94] Taylor JS and Raes J 2004. Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics* **38**:615–643.

[95] Katju V and Lynch M 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics* **165**(4):1793–1803.

[96] Bailey JA, Liu G, and Eichler EE 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *The American Journal of Human Genetics* **73**(4):823–834.

[97] Brosius J 1991. Retroposons - seeds of evolution. *Science* **251**(4995):753.

[98] Hahn MW 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity* **100**(5):605–617.

[99] Zhang JZ 2003. Evolution by gene duplication: An update. *Trends in Ecology & Evolution* **18**(6):292–298.

[100] Kondrashov FA and Kondrashov AS 2006. Role of selection in fixation of gene duplications. *Journal of Theoretical Biology* **239**(2):141–151.

[101] O'Hely M 2006. A diffusion approach to approximating preservation probabilities for gene duplicates. *Journal of Mathematical Biology* **53**(2):215–230.

[102] Hurles M 2004. Gene duplication: The genomic trade in spare parts. *PLoS Biology* **2**(7):E206.

[103] Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4):1531–1545.

[104] Proulx SR and Phillips PC 2006. Allelic divergence precedes and promotes gene duplication. *Evolution* **60**(5):881–892.

[105] Otto S and Young P 2002. *Homology effects*, chapter The evolution of gene duplicates. Academic press.

[106] Orgel LE 1977. Gene-duplication and the origin of proteins with novel functions. *Journal of Theoretical Biology* **67**(4):773.

[107] Dykhuizen D and Hartl DL 1980. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics* **96**(4):801–817.

[108] Hartl DL and Dykhuizen DE 1981. Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia Coli*. *Proceedings of the National Academy of Sciences USA* **78**(10):6344–6348.

[109] Clement Y, Tavares R, and Marais GAB 2006. Does lack of recombination enhance asymmetric evolution among duplicate genes? Insights from the *Drosophila melanogaster* genome. *Gene* **385**:89–95.

[110] Hittinger CT and Carroll SB 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**(7163):677–681.

[111] Des Marais DL and Rausher MD 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**(7205):762–765.

[112] Hendrickson H, Slechta ES, Bergthorsson U, Andersson DI, and Roth JR 2002. Amplification-mutagenesis: Evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. *Proceedings of the National Academy of Sciences USA* **99**(4):2164–2169.

[113] Francino MP 2005. An adaptive radiation model for the origin of new gene functions. *Nature Genetics* **37**(6):573–577.

[114] Ramsey J and Schemske DW 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics* **29**:467–501.

[115] Coleman L 1950. Nuclear conditions in normal stem tissue of *Vicia faba*. *Canadian Journal of Research* **28**(3):382–391.

[116] Randolph L 1932. Some effects of high temperature on polyploidy and other variations in maize. *Proceedings of the National Academy of Sciences USA* **18**(3):222.

[117] Grant V 1981. *Plant speciation*. Columbia University Press.

[118] Harlan J and De Wet J 1975. On ö winge and a prayer: The origins of polyploidy. *The Botanical Review* **41**(4):361–390.

[119] De Storme N and Geelen D 2013. Sexual polyploidization in plants - cytological mechanisms and molecular regulation. *New Phytologist* **198**(3):670–684.

[120] Parrott WA and Smith RR 1986. Recurrent selection for 2*n* pollen formation in red clover. *Crop Science* **26**(6):1132–1135.

[121] Tavoletti S, Mariani A, and Veronesi F 1991. Phenotypic recurrent selection for 2*n*-pollen and 2*n*-egg production in diploid alfalfa. *Euphytica* **57**(2):97–102.

[122] Levin D 1975. Minority cytotype exclusion in local plant populations. *Taxon* **24**(1):35–43.

[123] Bretagnolle F, , and Thompson J 1995. Gametes with the somatic chromosome number: Mechanisms of their formation and role in the evolution of autopolyploid plants. *New Phytologist* **129**(1):1–22.

[124] Suda J and Herben T 2013. Ploidy frequencies in plants with ploidy heterogeneity: Fitting a general gametic model to empirical population data. *Proceedings of the Royal Society B: Biological Sciences* **280**(1751):1–11.

[125] Soltis DE, Buggs RJA, Doyle JJ, and Soltis PS 2010. What we still don't know about polyploidy. *Taxon* **59**(5):1387–1403.

[126] Ramsey J and Schemske DW 2002. Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics* **33**:589–639.

[127] Madlung A 2013. Polyploidy and its effect on evolutionary success: Old questions revisited with new tools. *Heredity* **110**(2):99–104.

[128] Van de Peer Y, Maere S, and Meyer A 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10**(10):725–732.

[129] Dehal P and Boore JL 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* **3**(10):1700–1708.

[130] Panopoulou G and Poustka AJ 2005. Timing and mechanism of ancient vertebrate genome duplications - the adventure of a hypothesis. *Trends in Genetics* **21**(10):559–567.

[131] Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**(7198):1064–1071.

[132] Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**(7116):171–178.

[133] Wolfe KH and Shields DC 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**(6634):708–713.

[134] Kellis M, Birren BW, and Lander ES 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**(6983):617–624.

[135] Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**(6):738–749.

[136] Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**(7345):97–100.

[137] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**(7161):463–467.

[138] Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* **13**(1):R3.

[139] Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, et al. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: Significance for MADS-box gene and species diversification. *Molecular Biology and Evolution* **29**(12):3793–3806.

[140] Paterson AH, Bowers JE, and Chapman BA 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences USA* **101**(26):9903–9908.

[141] Van de Peer Y, Fawcett JA, Proost S, Sterck L, and Vandepoele K 2009. The flowering world: A tale of duplications. *Trends in Plant Science* **14**(12):680–688.

[142] Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, et al. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences USA* **106**(33):13875–13879.

[143] Pandit MK, Pocock MJO, and Kunin WE 2011. Ploidy influences rarity and invasiveness in plants. *Journal of Ecology* **99**(5):1108–1115.

[144] te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, et al. 2012. The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany* **109**(1):19–45.

[145] Brochmann C, Brysting AK, Alsos IG, Borgen L, Grundt HH, et al. 2004. Polyploidy in arctic plants. *Biological Journal of the Linnean Society* **82**(4):521–536.

[146] Mable BK 2004. Why polyploidy is rarer in animals than in plants: Myths and mechanisms. *Biological Journal of the Linnean Society* **82**(4):453–466.

[147] Song C, Liu SJ, Xiao J, He WG, Zhou Y, et al. 2012. Polyploid organisms. *Science China - Life Sciences* **55**(4):301–311.

[148] Mallet J 2007. Hybrid speciation. *Nature* **446**(7133):279–283.

[149] Otto SP and Whitton J 2000. Polyploid incidence and evolution. *Annual Review of Genetics* **34**:401–437.

[150] Van de Peer Y, Maere S, and Meyer A 2010. 2R or not 2R is not the question anymore. *Nature Reviews Genetics* **11**(2):166.

[151] Fawcett J and Van de Peer Y 2010. Angiosperm polyploids and their road to evolutionary success. *Trends in Evolutionary Biology* **2**(1):e3.

[152] Stebbins GL 1950. *Variation and evolution in plants*. Columbia University Press.

[153] Wagner WH 1970. Biosystematics and evolutionary noise. *Taxon* **19**:146–151.

[154] Levin DA 1983. Polyploidy and novelty in flowering plants. *American Naturalist* **122**(1):1–25.

[155] Arrigo N and Barker MS 2012. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* **15**(2):140–146.

[156] Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, et al. 2005. Genomic changes in synthetic *Arabidopsis* polyploids. *The Plant Journal* **41**(2):221–230.

[157] Wang YX, Jha AK, Chen RJ, Doonan JH, and Yang M 2010. Polyploidy-associated genomic instability in *Arabidopsis thaliana*. *Genesis* **48**(4):254–263.

[158] Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, et al. 2011. Recently formed polyploid plants diversify at lower rates. *Science* **333**(6047):1257.

[159] Lynch M and Force A 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**(1):459–473.

[160] Freeling M and Thomas BC 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome research* **16**(7):805–814.

[161] Birchler JA and Veitia RA 2010. The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. *New Phytologist* **186**(1):54–62.

[162] De Smet R and Van de Peer Y 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology* **15**(2):168–176.

[163] Fawcett J, Van de Peer Y, and Maere S 2013. *Plant Genome Diversity Volume 2*, chapter Significance and biological consequences of polyploidization in land plant evolution. Springer.

[164] Thompson JD and Lumaret R 1992. The evolutionary dynamics of polyploid plants - origins, establishment and persistence. *Trends in Ecology & Evolution* **7**(9):302–307.

[165] Kron P, Suda J, and Husband BC 2007. Applications of flow cytometry to evolutionary and population biology. *Annual Review of Ecology Evolution and Systematics* **38**:847–876.

[166] Wolfe KH 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* **2**(5):333–341.

[167] Van de Peer Y 2004. Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics* **5**(10):752–763.

[168] Fostier J, Proost S, Dhoedt B, Saeys Y, Demeester P, et al. 2011. A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **27**(6):749–756.

[169] Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, et al. 2012. i-ADHoRe 3.0 - fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research* **40**(2):e11.

[170] Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, et al. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* **166**(2):935–945.

[171] Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, et al. 2009. PLAZA: A comparative genomics resource to study gene and genome evolution in plants. *The Plant Cell* **21**(12):3718–3731.

[172] Vandesteene L, López-Galvis L, Vanneste K, Feil R, Maere S, et al. 2012. Expansive evolution of the *trehalose-6-phosphate phosphatase* gene family in *Arabidopsis*. *Plant Physiology* **160**(2):884–896.

[173] Bowers JE, Chapman BA, Rong J, and Paterson AH 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**(6930):433–438.

[174] Chen K, Durand D, and Farach-Colton M 2000. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology* **7**(3-4):429–447.

[175] Kimura M 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**(5608):275–276.

[176] Blanc G and Wolfe KH 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**(7):1667–1678.

[177] Schlueter JA, Dixon P, Granger C, Grant D, Clark L, et al. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**(5):868–876.

[178] Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, et al. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences USA* **102**(15):5454–5459.

[179] Pfeil BE, Schlueter JA, Shoemaker RC, and Doyle JJ 2005. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. *Systematic Biology* **54**(3):441–454.

[180] Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, et al. 2005. EST data suggest that poplar is an ancient polyploid. *New Phytologist* **167**(1):165–170.

[181] Yu J, Wang J, Lin W, Li S, Li H, et al. 2005. The genomes of *Oryza sativa*: A history of duplications. *PLoS Biology* **3**(2):e38.

[182] Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, et al. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* **25**(11):2445–2455.

[183] Barker MS, Vogel H, and Schranz ME 2009. Paleopolyploidy in the Brassicales: Analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* **1**:391–399.

[184] Shi T, Huang H, and Barker MS 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* **106**(3):497–504.

[185] Tang HB, Bowers JE, Wang XY, and Paterson AH 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences USA* **107**(1):472–477.

[186] McKain MR, Wickett N, Zhang Y, Ayyampalayam S, McCombie WR, et al. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany* **99**(2):397–406.

[187] Chaudhuri P and Marron J 1999. SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**:807–823.

[188] Zuckerkandl E and Pauling L 1965. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* **97**:97–166.

[189] Koch MA, Haubold B, and Mitchell-Olds T 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution* **17**(10):1483–1498.

[190] Lanfear R, Welch JJ, and Bromham L 2010. Watching the clock: Studying variation in rates of molecular evolution between species. *Trends in Ecology & Evolution* **25**(9):495–503.

[191] Smith SA and Donoghue MJ 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* **322**(5898):86–89.

[192] Yang L and Gaut BS 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Molecular Biology and Evolution* **28**(8):2359–2369.

[193] Fawcett JA, Maere S, and Van de Peer Y 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Sciences USA* **106**(14):5737–5742.

[194] Egan AN and Doyle J 2010. A comparison of global, gene-specific, and relaxed clock methods in a comparative genomics framework: Dating the polyploid history of soybean (*Glycine max*). *Systematic Biology* **59**(5):534–547.

[195] Ho SY 2009. An examination of phylogenetic models of substitution rate variation among lineages. *Biology Letters* **5**(3):421–424.

[196] Sanderson MJ 2003. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**(2):301–302.

[197] Sanderson MJ 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution* **19**(1):101–109.

[198] Yang Z and Rannala B 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution* **23**(1):212–226.

[199] Crow KD and Wagner GP 2006. What is the role of genome duplication in the evolution of complexity and diversity? *Molecular Biology and Evolution* **23**(5):887–892.

[200] Comai L 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**(11):836–846.

[201] Hegarty M and Hiscock S 2007. Polyploidy: Doubling up for evolutionary success. *Current Biology* **17**(21):R927–R929.

[202] Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, and Wilson AC 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345**(6270):86–89.

[203] Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, et al. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics* **37**(6):630–635.

[204] Goldman N and Yang Z 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**(5):725–736.

[205] Yang ZH and Nielsen R 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**(4):409–418.

[206] Anisimova M, Bielawski JP, and Yang ZH 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution* **19**(6):950–958.

[207] Zhang J, Nielsen R, and Yang Z 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22**(12):2472–2479.

[208] Hughes AL 2007. Looking for Darwin in all the wrong places: The misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**(4):364–373.

[209] Nozawa M, Suzuki Y, and Nei M 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences USA* **106**(16):6700–6705.

[210] Yang ZH, Nielsen R, and Goldman N 2009. In defense of statistical methods for detecting positive selection. *Proceedings of the National Academy of Sciences USA* **106**(36):E95.

[211] Teste MA, Francois JM, and Parrou JL 2010. Characterization of a new multigene family encoding isomaltases in the yeast *Saccharomyces cerevisiae*, the *IMA* family. *Journal of Biological Chemistry* **285**(35):26815–26824.

[212] Maere S and Van de Peer Y 2010. *Evolution after gene duplication*, chapter Duplicate retention after small- and large-scale duplications. John Wiley & Sons.

[213] Long MY and Thornton K 2001. Gene duplication and evolution. *Science* **293**(5535):1551.

[214] Graur D and Martin W 2004. Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. *Trends in Genetics* **20**(2):80–86.

[215] Soltis DE and Burleigh JG 2009. Surviving the K-T mass extinction: New perspectives of polyploidization in angiosperms. *Proceedings of the National Academy of Sciences USA* **106**(14):5455–5456.

[216] Mulcahy DG, Noonan BP, Moss T, Townsend TM, Reeder TW, et al. 2012. Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. *Molecular Phylogenetics and Evolution* **65**(3):974–991.

[217] Ho SY and Phillips MJ 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology* **58**(3):367–380.

[218] Fritz MHY, Leinonen R, Cochrane G, and Birney E 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research* **21**(5):734–740.

[219] Drummond AJ, Ho SY, Phillips MJ, and Rambaut A 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**(5):e88.

[220] Drummond AJ, Suchard MA, Xie D, and Rambaut A 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**(8):1969–1973.

[221] Anisimova M, Liberles DA, Philippe H, Provan J, Pupko T, et al. 2013. State-of-the-art methodologies dictate new standards for phylogenetic analysis. *BMC Evolutionary Biology* **13**(161):1–8.

[222] Forest F 2009. Calibrating the Tree of Life: Fossils, molecules and evolutionary timescales. *Annals of Botany* **104**(5):789–794.

[223] Hughes AL 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society B: Biological Sciences* **256**(1346):119–124.

[224] Kondrashov FA, Rogozin IB, Wolf YI, and Koonin EV 2002. Selection in the evolution of gene duplications. *Genome Biology* **3**(2):1–9.

[225] Cheung J, Wilson MD, Zhang J, Khaja R, MacDonald JR, et al. 2003. Recent segmental and gene duplications in the mouse genome. *Genome Biology* **4**(8):R47.

[226] Wapinski I, Pfeffer A, Friedman N, and Regev A 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**(7158):54–61.

[227] Lynch M and Conery JS 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494):1151–1155.

[228] Piatigorsky J and Wistow G 1991. The recruitment of crystallins: New functions precede gene duplication. *Science* **252**(5009):1078–1079.

[229] Wistow G and Piatigorsky J 1987. Recruitment of enzymes as lens structural proteins. *Science* **236**(4808):1554–1556.

[230] Conant GC and Wolfe KH 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics* **9**(12):938–950.

[231] Innan H and Kondrashov F 2010. The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics* **11**(2):97–108.

[232] He X and Zhang J 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**(2):1157–1164.

[233] Hughes AL 2005. Gene duplication and the origin of novel proteins. *Proceedings of the National Academy of Sciences USA* **102**(25):8791–8792.

[234] Rueffler C, Hermisson J, and Wagner GP 2012. Evolution of functional specialization and division of labor. *Proceedings of the National Academy of Sciences USA* **109**(6):326–335.

[235] Bergthorsson U, Andersson DI, and Roth JR 2007. Ohno's dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences USA* **104**(43):17004–17009.

[236] Bridgham JT, Carroll SM, and Thornton JW 2006. Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**(5770):97–101.

[237] Carroll SM, Ortlund EA, and Thornton JW 2011. Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. *PLoS Genetics* **7**(6):e1002117.

[238] Chandrasekharan UM, Sanker S, Glynias MJ, Karnik SS, and Husain A 1996. Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science* **271**(5248):502–505.

[239] Gaucher EA, Govindarajan S, and Ganesh OK 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**(7179):704–707.

[240] Jermann TM, Opitz JG, Stackhouse J, and Benner SA 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* **374**(6517):57–59.

[241] Zhang J, Dyer KD, and Rosenberg HF 2002. RNase 8, a novel RNase A superfamily ribonuclease expressed uniquely in placenta. *Nucleic Acids Research* **30**(5):1169–1175.

[242] Wouters MA, Liu K, Riek P, and Husain A 2003. A despecialization step underlying evolution of a family of serine proteases. *Molecular Cell* **12**(2):343–354.

[243] Brown CA, Murray AW, and Verstrepen KJ 2010. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology* **20**(10):895–903.

[244] Yamamoto K, Miyake H, Kusunoki M, and Osaki S 2010. Crystal structures of isomaltase from *Saccharomyces cerevisiae* and in complex with its competitive inhibitor maltose. *FEBS Journal* **277**(20):4205–4214.

[245] Katoh K and Standley DM 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**(4):772–780.

[246] Abascal F, Zardoya R, and Posada D 2005. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* **21**(9):2104–2105.

[247] Huelsenbeck JP and Ronquist F 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8):754–755.

[248] Nylander JA, Wilgenbusch JC, Warren DL, and Swofford DL 2008. AWTY (are we there yet?): A system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* **24**(4):581–583.

[249] Guindon S and Gascuel O 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**(5):696–704.

[250] Yang Z 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(8):1586–1591.

[251] Pond SL, Frost SD, and Muse SV 2005. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**(5):676–967.

[252] Pond SL and Frost SD 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Molecular Biology and Evolution* **22**(3):478–485.

[253] Arnegard ME, Zwickl DJ, Lu Y, and Zakon HH 2010. Old gene duplication facilitates origin and diversification of an innovative communication system - twice. *Proceedings of the National Academy of Sciences USA* **107**(51):22172–22177.

[254] Yang Z and Nielsen R 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**(6):908–917.

[255] Anisimova M and Yang Z 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular Biology and Evolution* **24**(5):1219–1228.

[256] Yang Z, Wong WS, and Nielsen R 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**(4):1107–1118.

[257] Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution* **28**(11):3033–3043.

[258] Brown CA and Brown KS 2010. Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PLoS One* **5**(6):e10779.

[259] Kurtzman CP and Robnett CJ 2003. Phylogenetic relationships among yeasts of the *Saccharomyces cerevisiae* determined from multigene sequence analyses. *FEMS Yeast Research* **3**(4):417–432.

[260] Low N and Sporns P 1988. Analysis and quantitation of minor di- and trisaccharides in honey, using capillary gas chromatography. *Journal of Food Science* **53**(2):558–561.

[261] Yamamoto K, Miyake H, Kusunoki M, and Osaki S 2011. Steric hindrance by 2 amino acid residues determines the substrate specificity of isomaltase from *Saccharomyces cerevisiae*. *Journal of Bioscience and Bioengineering* **112**(6):545–550.

[262] Yamamoto K, Nakayama A, Yamamoto Y, and Tabata S 2004. Val216 decides the substrate specificity of alpha-glucosidase in *Saccharomyces cerevisiae*. *European Journal of Biochemistry* **271**(16):3414–3420.

[263] Suzuki Y and Nei M 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **18**(12):2179–2185.

[264] Suzuki Y and Nei M 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **19**(11):1865–1869.

[265] Mefford HC and Trask BJ 2002. The complex structure and dynamic evolution of human subtelomeres. *Nature Reviews Genetics* **3**(2):91–102.

[266] Whittington A, Gow NAR, and Hube B 2014. *Human Fungal Pathogens*, chapter From commensal to pathogen: *Candida albicans*. Springer.

[267] Byrne KP and Wolfe KH 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research* **15**(10):1456–1461.

[268] Gordon JL, Byrne KP, and Wolfe KH 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *Plos Genetics* **5**(5):e1000485.

[269] Barkman T and Zhang J 2009. Evidence for escape from adaptive conflict? *Nature* **462**(7274):E1–E3.

[270] Zhai W, Nielsen R, Goldman N, and Yang Z 2012. Looking for Darwin in genomic sequences - validity and success of statistical methods. *Molecular Biology and Evolution* **29**(10):2889–2893.

[271] Deng C, Cheng CH, Ye H, He X, and Chen L 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences USA* **107**(50):21593–21598.

[272] Huang R, Hippauf F, Rohrbeck D, Haustein M, Wenke K, et al. 2012. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proceedings of the National Academy of Sciences USA* **109**(8):2966–2971.

[273] Sobreira TJ, Marletaz F, Simoes-Costa M, Schechtman D, Pereira AC, et al. 2011. Structural shifts of aldehyde dehydrogenase enzymes were instrumental for the early evolution of retinoid-dependent axial patterning in metazoans. *Proceedings of the National Academy of Sciences USA* **108**(1):226–231.

[274] Des Marais DL and Rausher M 2008. Reply to: Barkman T. and Zhang J. *Nature* **462**:E2–E3.

[275] Bridgham JT, Ortlund EA, and Thornton JW 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**(7263):515–519.

[276] Khersonsky O and Tawfik DS 2010. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry* **79**:471–505.

[277] Copley SD 2003. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology* **7**(2):265–272.

[278] van Hoof A 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics* **171**(4):1455–1461.

[279] Afriat L, Roodveldt C, Manco G, and Tawfik DS 2006. The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. *Biochemistry* **45**(46):13677–13686.

[280] Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, et al. 2005. The “evolvability” of promiscuous protein functions. *Nature Genetics* **37**(1):73–76.

[281] Bone R, Silen JL, and Agard DA 1989. Structural plasticity broadens the specificity of an engineered protease. *Nature* **339**(6221):191–195.

[282] Friis EM, Pedersen KR, and Crane PR 2005. When Earth started blooming: Insights from the fossil record. *Current Opinion in Plant Biology* **8**(1):5–12.

[283] De Bodt S, Maere S, and Van de Peer Y 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution* **20**(11):591–597.

[284] Hakes L, Pinney JW, Lovell SC, Oliver SG, and Robertson DL 2007. All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biology* **8**(10):1–13.

[285] Freeling M 2009. 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.

[286] Hughes AL and Friedman R 2003. 2R or not 2R: Testing hypotheses of genome duplication in early vertebrates. *Journal of Structural and Functional Genomics* **3**(1-4):85–93.

[287] Abbasi AA 2010. Piecemeal or big bangs: Correlating the vertebrate evolution with proposed models of gene expansion events. *Nature Reviews Genetics* **11**(2):166.

[288] Vandepoele K, De Vos W, Taylor JS, Meyer A, and Van de Peer Y 2004. Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences USA* **101**(6):1638–1643.

[289] Sato Y and Nishida M 2010. Teleost fish with specific genome duplication as unique models of vertebrate evolution. *Environmental Biology of Fishes* **88**(2):169–188.

[290] Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. *Science* **331**(6017):555–561.

[291] Arnaud MB, Costanzo MC, Skrzypek MS, Shah P, Binkley G, et al. 2007. Sequence resources at the *Candida* Genome Database. *Nucleic Acids Research* **35**:D452–456.

[292] Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, et al. 2009. Genolevures: Protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Research* **37**:D550–554.

[293] Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. 2011. Ensembl 2011. *Nucleic Acids Research* **39**:D800–806.

[294] Enright AJ, Van Dongen S, and Ouzounis CA 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**(7):1575–1584.

[295] Edgar RC 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**(5):1792–1797.

[296] Fitch WM 1967. Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *Journal of Molecular Biology* **26**(3):499–507.

[297] Wakeley J 1996. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution* **11**(4):158–162.

[298] Hershberg R and Petrov DA 2008. Selection on codon bias. *Annual Review of Genetics* **42**:287–299.

[299] Yang Z and Nielsen R 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**(1):32–43.

[300] Muse SV and Gaut BS 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**(5):715–724.

[301] Seo TK and Kishino H 2009. Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. *Systematic Biology* **58**(2):199–210.

[302] Miyazawa S 2011. Advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences. *PLoS One* **6**(12):e28892.

[303] Gojobori T 1983. Codon substitution in evolution and the saturation of synonymous changes. *Genetics* **105**(4):1011–1027.

[304] Kristina Strandberg AK and Salter LA 2004. A comparison of methods for estimating the transition:transversion ratio from DNA sequences. *Molecular Phylogenetics and Evolution* **32**(2):495–503.

[305] Rosenberg MS, Subramanian S, and Kumar S 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Molecular Biology and Evolution* **20**(6):988–993.

[306] Yang Z 2006. *Computational molecular evolution*. Oxford University Press.

[307] Morrison DA 2008. How to summarize estimates of ancestral divergence times. *Evolutionary Bioinformatics Online* **4**:75–95.

[308] Lynch M and Conery JS 2003. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics* **3**(1-4):35–44.

[309] Brown WM, George M, and Wilson AC 1979. Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences USA* **76**(4):1967–1971.

[310] Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, et al. 1980. The evolution of genes - the chicken preproinsulin gene. *Cell* **20**(2):555–566.

[311] Smith JM and Smith NH 1996. Synonymous nucleotide divergence: What is "saturation"? *Genetics* **142**(3):1033–1036.

[312] Berg OG 1999. Synonymous nucleotide divergence and saturation: Effects of site-specific variations in codon bias and mutation rates. *Journal of Molecular Evolution* **48**(4):398–407.

[313] Anisimova M and Kosiol C 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution* **26**(2):255–271.

[314] Naik PA, Shi P, and Tsai CL 2007. Extending the Akaike Information Criterion to mixture regression models. *Journal of the American Statistical Association* **102**(477):244–254.

[315] Yoo MJ, Chanderbali AS, Altman NS, Soltis PS, and Soltis DE 2010. Evolutionary trends in the floral transcriptome: Insights from one of the basalmost angiosperms, the water lily *Nuphar advena* (Nymphaeaceae). *The Plant Journal* **64**(4):687–698.

[316] Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, and Van de Peer Y 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences USA* **99**(21):13627–13632.

[317] Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**(7011):946–957.

[318] Meyer A and Van de Peer Y 2005. From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD). *BioEssays* **27**(9):937–945.

[319] Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, et al. 2004. Genome evolution in yeasts. *Nature* **430**(6995):35–44.

[320] Conant GC and Wolfe KH 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology* **3**:129.

[321] Gao LZ and Innan H 2004. Very low gene duplication rate in the yeast genome. *Science* **306**(5700):1367–1370.

[322] Sugino RP and Innan H 2006. Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends in Genetics* **22**(12):642–644.

[323] Lin YS, Byrnes JK, Hwang JK, and Li WH 2006. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. *Proceedings of the National Academy of Sciences USA* **103**(39):14412–14416.

[324] Schranz ME, Mohammadin S, and Edger PP 2012. Ancient whole genome duplications, novelty and diversification: The WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology* **15**(2):147–153.

[325] Meyers LA and Levin DA 2006. On the abundance of polyploids in flowering plants. *Evolution* **60**(6):1198–1206.

[326] Gouzy J, Carrere S, and Schiex T 2009. FrameDP: Sensitive peptide detection on noisy matured sequences. *Bioinformatics* **25**(5):670–671.

[327] Bairoch A, Boeckmann B, Ferro S, and Gasteiger E 2004. Swiss-Prot: Juggling between evolution and stability. *Briefings in Bioinformatics* **5**(1):39–55.

[328] Vanneste K, Van de Peer Y, and Maere S 2013. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* **30**(1):177–190.

[329] Li WH, Gu Z, Wang H, and Nekrutenko A 2001. Evolutionary analyses of the human genome. *Nature* **409**(6822):847–849.

[330] D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**(7410):213–217.

[331] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* **59**(3):307–321.

[332] Hess PN and De Moraes Russo CA 2007. An empirical test of the midpoint rooting method. *Biological Journal of the Linnean Society* **92**(4):669–674.

[333] Altenhoff AM, Studer RA, Robinson-Rechavi M, and Dessimoz C 2012. Resolving the Ortholog Conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PloS Computational Biology* **8**(5):e1002514.

[334] Gabaldon T and Koonin EV 2013. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* **14**(5):360–366.

[335] Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, et al. 2010. Inparanoid 7: New algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research* **38**:D196–203.

[336] Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, et al. 2011. Orthology prediction methods: A quality assessment using curated protein families. *Bioessays* **33**(10):769–780.

[337] Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, and Gabaldon T 2010. 2x genomes - depth does matter. *Genome Biology* **11**(2):R16.

[338] Koonin EV 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* **39**:309–338.

[339] Brysting AK, Oxelman B, Huber KT, Moulton V, and Brochmann C 2007. Untangling complex histories of genome mergings in high polyploids. *Systematic Biology* **56**(3):467–476.

[340] Le SQ and Gascuel O 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**(7):1307–1320.

[341] Yang Z 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* **11**(9):367–372.

[342] Baele G, Li WL, Drummond AJ, Suchard MA, and Lemey P 2013. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* **30**(2):239–243.

[343] Yule 1925. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis., F.R.S. *Philosophical Transactions of the Royal Society of London. Series B* **213**:21–87.

[344] Lewis PO 2001. Phylogenetic systematics turns over a new leaf. *Trends in Ecology & Evolution* **16**(1):30–37.

[345] Sanderson MJ, Thorne JL, Wikstrom N, and Bremer K 2004. Molecular evidence on plant divergence times. *American Journal of Botany* **91**(10):1656–1665.

[346] Suchard MA and Rambaut A 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**(11):1370–1376.

[347] Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, et al. 2012. BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology* **61**(1):170–173.

[348] Battistuzzi FU, Billing-Ross P, Paliwal A, and Kumar S 2011. Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. *Molecular Biology and Evolution* **28**(9):2439–2442.

[349] Moret BME, Bader DA, and Warnow T 2002. High-performance algorithm engineering for computational phylogenetics. *Journal of Supercomputing* **22**(1):99–110.

[350] Botev ZI, Grotowski J, and Kroese DP 2010. Kernel density estimation via diffusion. *Annals of Statistics* **38**(5):2916–2957.

[351] Hall P and Kang KH 2001. Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *Annals of Statistics* **29**(5):1443–1468.

[352] Akaike H 1974. A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers Transactions on Automatic Control* **19**(6):716–723.

[353] Doyle JJ 2012. *Polyploidy and Genome Evolution*, chapter Polyploidy in legumes. Springer.

[354] Doyle JJ and Egan AN 2010. Dating the origins of polyploidy events. *New Phytologist* **186**(1):73–85.

[355] Yang S, Yuan Y, Wang L, Li J, Wang W, et al. 2012. Great majority of recombination events in *Arabidopsis* are gene conversion events. *Proceedings of the National Academy of Sciences USA* **109**(51):20992–20997.

[356] Rannala B and Yang ZH 2007. Inferring speciation times under an episodic molecular clock. *Systematic Biology* **56**(3):453–466.

[357] Holder M and Lewis PO 2003. Phylogeny estimation: Traditional and Bayesian approaches. *Nature Reviews Genetics* **4**(4):275–284.

[358] Thorne JL, Kishino H, and Painter IS 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**(12):1647–1657.

[359] Yang ZH and Yoder AD 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology* **52**(5):705–716.

[360] Zhang L, Vision T, and Gaut B 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **19**(9):1464–1473.

[361] Scannell DR and Wolfe KH 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research* **18**(1):137–147.

[362] Lepage T, Bryant D, Philippe H, and Lartillot N 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* **24**(12):2669–2680.

[363] Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, et al. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* **29**(9):2157–2167.

[364] Baele G and Lemey P 2013. Bayesian evolutionary model testing in the phylogenomics era: Matching model complexity with computational efficiency. *Bioinformatics* **29**(16):1970–1979.

[365] Hug LA and Roger AJ 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. *Molecular Biology and Evolution* **24**(8):1889–1897.

[366] Inoue J, Donoghue PC, and Yang Z 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Systematic Biology* **59**(1):74–89.

[367] Warnock RC, Yang Z, and Donoghue PC 2012. Exploring uncertainty in the calibration of the molecular clock. *Biology Letters* **8**(1):156–159.

[368] Benton MJ and Donoghue PC 2007. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution* **24**(1):26–53.

[369] Lanfear R 2011. The local-clock permutation test: A simple test to compare rates of molecular evolution on phylogenetic trees. *Evolution* **65**(2):606–611.

[370] Lanfear R, Ho SY, Jonathan Davies T, Moles AT, Aarssen L, et al. 2013. Taller plants have lower rates of molecular evolution. *Nature Communications* **4**:1879.

[371] Korall P, Schuettpelz E, and Pryer KM 2010. Abrupt deceleration of molecular evolution linked to the origin of arborescence in ferns. *Evolution* **64**(9):2786–2792.

[372] Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**(5793):1596–1604.

[373] Manchester SR, Dilcher DL, and Tidwell WD 1986. Interconnected reproductive and vegetative remains of *Populus* (Salicaceae) from the Middle Eocene Green River Formation, Northeastern Utah. *American Journal of Botany* **73**(1):156–160.

[374] Manchester SR, Judd WS, and Handley B 2006. Foliage and fruits of early poplars (Salicaceae: *Populus*) from the Eocene of Utah, Colorado, and Wyoming. *International Journal of Plant Sciences* **167**(4):897–908.

[375] Boucher LD, Manchester SR, and Judd WS 2003. An extinct genus of Salicaceae based on twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of Utah and Colorado, USA. *American Journal of Botany* **90**(9):1389–1399.

[376] Berlin S, Lagercrantz U, von Arnold S, Ost T, and Ronnberg-Wastljung AC 2010. High-density linkage mapping and evolution of paralogs and orthologs in *Salix* and *Populus*. *BMC Genomics* **11**(129):1–14.

[377] Wu J, Wang ZW, Shi ZB, Zhang S, Ming R, et al. 2013. The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Research* **23**(2):396–408.

[378] Wehr WC and Hopkins DQ 1994. The Eocene Orchards and Gardens of Republic, Washington. *Washington Geology* **22**(3):27–34.

[379] DeVore ML and Pigg KB 2007. A brief review of the fossil history of the family Rosaceae with a focus on the Eocene Okanogan Highlands of eastern Washington State, USA, and British Columbia, Canada. *Plant Systematics and Evolution* **266**(1-2):45–57.

[380] Lo EY and Donoghue MJ 2012. Expanded phylogenetic and dating analyses of the apples and their relatives (Pyreae, Rosaceae). *Molecular Phylogenetics and Evolution* **63**(2):230–243.

[381] Paun O, Bateman RM, Fay MF, Luna JA, Moat J, et al. 2011. Altered gene expression and ecological divergence in sibling allopolyploids of *Dactylorhiza* (Orchidaceae). *BMC Evolutionary Biology* **11**(113):1–14.

[382] Hahn MA, van Kleunen M, and Muller-Scharer H 2012. Increased phenotypic plasticity to climate may have boosted the invasion success of polyploid *Centaurea stoebe*. *Plos One* **7**(11):e50284.

[383] Voss N, Eckstein RL, and Durka W 2012. Range expansion of a selfing polyploid plant despite widespread genetic uniformity. *Annals of Botany* **110**(3):585–593.

[384] Henery ML, Bowman G, Mraz P, Treier UA, Gex-Fabry E, et al. 2010. Evidence for a combination of pre-adapted traits and rapid adaptive change in the invasive plant *Centaurea stoebe*. *Journal of Ecology* **98**(4):800–813.

[385] Chao DY, Dilkes B, Luo H, Douglas A, Yakubova E, et al. 2013. Polyploids exhibit higher potassium uptake and salinity tolerance in *Arabidopsis*. *Science* **341**(6146):658–659.

[386] Rebernig CA, Weiss-Schneeweiss H, Schneeweiss GM, Schonswetter P, Obermayer R, et al. 2010. Quaternary range dynamics and polyploid evolution in an arid brushland plant species (*Melampodium cinereum*, Asteraceae). *Molecular Phylogenetics and Evolution* **54**(2):594–606.

[387] De Storme N and Geelen D 2014. The impact of environmental stress on male reproductive development in plants: Biological processes and molecular mechanisms. *Plant, Cell & Environment* **37**(1):1–18.

[388] Pecrix Y, Rallo G, Folzer H, Cigna M, Gudin S, et al. 2011. Polyploidization mechanisms: Temperature environment can induce diploid gamete formation in *Rosa* sp. *Journal of Experimental Botany* **62**(10):3587–3597.

[389] De Storme N, Copenhaver GP, and Geelen D 2012. Production of diploid male gametes in *Arabidopsis* by cold-induced destabilization of postmeiotic radial microtubule arrays. *Plant Physiology* **160**(4):1808–1826.

[390] Mason AS, Nelson MN, Yan GJ, and Cowling WA 2011. Production of viable male unreduced gametes in *Brassica* interspecific hybrids is genotype specific and stimulated by cold temperatures. *BMC Plant Biology* **11**(103):1–13.

[391] Kurschner WM, Batenburg SJ, and Mander L 2013. Aberrant *Classopollis* pollen reveals evidence for unreduced (2*n*) pollen in the conifer family Cheirolepidiaceae during the Triassic-Jurassic transition. *Proceedings of the Royal Society B: Biological Sciences* **280**(1768):1–8.

[392] Foster CB and Afonin SA 2005. Abnormal pollen grains: An outcome of deteriorating atmospheric conditions around the Permian–Triassic boundary. *Journal of the Geological Society* **162**:653–659.

[393] Visscher H, Looy CV, Collinson ME, Brinkhuis H, van Konijnenburg-van Cittert JH, et al. 2004. Environmental mutagenesis during the end-Permian ecological crisis. *Proceedings of the National Academy of Sciences USA* **101**(35):12952–12956.

[394] Shen SZ, Crowley JL, Wang Y, Bowring SA, Erwin DH, et al. 2011. Calibrating the end-Permian mass extinction. *Science* **334**(6061):1367–1372.

[395] Oswald BP and Nuismer SL 2011. A unified model of autopolyploid establishment and evolution. *American Naturalist* **178**(6):687–700.

[396] Vanneste K, Baele G, Maere S, and Van de Peer Y 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications at the Cretaceous-Paleogene boundary. *Genome Research* (accepted, pending revisions).

[397] Hedges SB and Kumar S 2004. Precision of molecular time estimates. *Trends in Genetics* **20**(5):242–247.

[398] Donoghue PC and Purnell MA 2005. Genome duplication, extinction and vertebrate evolution. *Trends in Ecology & Evolution* **20**(6):312–319.

[399] Macqueen DJ and Johnston IA 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences* **281**(1778):20132881.

[400] Howe HF and Smallwood J 1982. Ecology of seed dispersal. *Annual Review of Ecology and Systematics* **13**:201–228.

[401] Givnish TJ 2010. Ecology of plant speciation. *Taxon* **59**(5):1326–1366.

[402] Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**(7400):635–641.

[403] Kiers ET, Rousseau RA, West SA, and Denison RF 2003. Host sanctions and the legume-rhizobium mutualism. *Nature* **425**(6953):78–81.

[404] Tian CF, Zhou YJ, Zhang YM, Li QQ, Zhang YZ, et al. 2012. Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. *Proceedings of the National Academy of Sciences USA* **109**(22):8629–8634.

[405] Young ND, Debelle F, Oldroyd GED, Geurts R, Cannon SB, et al. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**(7378):520–524.

[406] Oldroyd GED and Downie JM 2008. Coordinating nodule morphogenesis with rhizobial infection in legumes. *Annual Review of Plant Biology* **59**:519–546.

[407] Middleton PH, Jakab J, Penmetsa RV, Starker CG, Doll J, et al. 2007. An ERF transcription factor in *Medicago truncatula* that is essential for nod factor signal transduction. *The Plant Cell* **19**(4):1221–1234.

[408] Op den Camp R, Streng A, De Mita S, Cao Q, Polone E, et al. 2011. LysM-type mycorrhizal receptor recruited for rhizobium symbiosis in nonlegume *Parasponia*. *Science* **331**(6019):909–912.

[409] Cannon SB, Ilut D, Farmer AD, Maki SL, May GD, et al. 2010. Polyploidy did not predate the evolution of nodulation in all legumes. *PLoS One* **5**(7):e11630.

[410] Soltis DE, Soltis PS, Morgan DR, Swensen SM, Mullin BC, et al. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen-fixation in angiosperms. *Proceedings of the National Academy of Sciences USA* **92**(7):2647–2651.

[411] Li QG, Zhang L, Li C, Dunwell JM, and Zhang YM 2013. Comparative genomics suggests that an ancestral polyploidy event leads to enhanced root nodule symbiosis in the papilionoideae. *Molecular Biology and Evolution* **30**(12):2602–2611.

[412] Schaefer H and Renner SS 2011. Phylogenetic relationships in the order Cucurbitales and a new classification of the gourd family (Cucurbitaceae). *Taxon* **60**(1):122–138.

[413] Borsch T, Hilu KW, Wiersema JH, Lohne C, Barthlott W, et al. 2007. Phylogeny of *Nymphaea* (Nymphaeaceae): Evidence from substitutions and microstructural changes in the chloroplast trnT-trnF region. *International Journal of Plant Sciences* **168**(5):639–671.

[414] Funk V, Susanna A, Stuessy T, and Bayer R 2009. *Systematics, evolution, and biogeography of Compositae*. International Association for Plant Taxonomy.

[415] Werth CR and Windham MD 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *American Naturalist* **137**(4):515–526.

[416] Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, et al. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proceedings of the National Academy of Sciences USA* **104**(20):8397–8402.

[417] Semon M and Wolfe KH 2007. Reciprocal gene loss between *Tetraodon* and zebrafish after whole genome duplication in their ancestor. *Trends in Genetics* **23**(3):108–112.

[418] Schnable JC, Freeling M, and Lyons E 2012. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biology and Evolution* **4**(3):265–277.

[419] Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, et al. 2012. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biology* **10**(12):e1001446.

[420] Postlethwait J, Amores A, Cresko W, Singer A, and Yan YL 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends in Genetics* **20**(10):481–490.

[421] Volff JN 2005. Genome evolution and biodiversity in teleost fish. *Heredity* **94**(3):280–294.

[422] Ehrendorfer F 1980. *Polyploidy:Biological Relevance*, chapter Polyploidy and distribution. Plenum Press.

[423] Parisod C, Holderegger R, and Brochmann C 2010. Evolutionary consequences of autopolyploidy. *New Phytologist* **186**(1):5–17.

[424] Chen ZJ 2013. Genomic and epigenetic insights into the molecular bases of heterosis. *Nature Reviews Genetics* **14**:471–482.

[425] Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, et al. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* **15**(2):131–139.

[426] Goff SA 2011. A unifying theory for general multigenic heterosis: Energy efficiency, protein metabolism, and implications for molecular breeding. *New Phytologist* **189**(4):923–937.

[427] Chen ZJ 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annual Review of Plant Biology* **58**:377–406.

[428] Wright KM, Pires JC, and Madlung A 2009. Mitotic instability in resynthesized and natural polyploids of the genus *Arabidopsis* (Brassicaceae). *American Journal of Botany* **96**(9):1656–1664.

[429] Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, et al. 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics* **19**(3):141–147.

[430] Adams KL 2007. Evolution of duplicate gene expression in polyploid and hybrid plants. *Journal of Heredity* **98**(2):136–141.

[431] Wang JL, Tian L, Lee HS, Wei NE, Jiang HM, et al. 2006. Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics* **172**(1):507–517.

[432] Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, et al. 2006. Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. *Current Biology* **16**(16):1652–1659.

[433] Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, et al. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**(5637):1211–1216.

[434] Buggs RJA and Pannell JR 2007. Ecological differentiation and diploid superiority across a moving ploidy contact zone. *Evolution* **61**(1):125–140.

[435] Husband BC, Ozimec B, Martin SL, and Pollock L 2008. Mating consequences of polyploid evolution in flowering plants: Current trends and insights from synthetic polyploids. *International Journal of Plant Sciences* **169**(1):195–206.

[436] Hanada K, Kuromori T, Myouga F, Toyoda T, Li WH, et al. 2009. Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. *Genome Biology and Evolution* **1**:409–414.

[437] Rodriguez DJ 1996. A model for the establishment of polyploidy in plants. *American Naturalist* **147**(1):33–46.

[438] Felber F and Bever JD 1997. Effect of triploid fitness on the coexistence of diploids and tetraploids. *Biological Journal of the Linnean Society* **60**(1):95–106.

[439] Husband BC 2004. The role of triploid hybrids in the evolutionary dynamics of mixed-ploidy populations. *Biological Journal of the Linnean Society* **82**(4):537–546.

[440] Belling J 1925. The origin of chromosomal mutations in *Uvularia*. *Journal of Genetics* **15**(3):245 –266.

[441] Mchale NA 1983. Environmental induction of high-frequency 2*n* pollen formation in diploid *Solanum*. *Canadian Journal of Genetics and Cytology* **25**(6):609–615.

[442] Felber F 1991. Establishment of a tetraploid cytotype in a diploid population - effect of relative fitness of the cytotypes. *Journal of Evolutionary Biology* **4**(2):195–207.

[443] Mraz P, Spaniel S, Keller A, Bowmann G, Farkas A, et al. 2012. Anthropogenic disturbance as a driver of microspatial and microhabitat segregation of cytotypes of *Centaurea stoebe* and cytotype interactions in secondary contact zones. *Annals of Botany* **110**(3):615–627.

[444] Pysek P, Jarosik V, Pergl J, Randall R, Chytry M, et al. 2009. The global invasion success of Central European plants is related to distribution characteristics in their native range and species traits. *Diversity and Distributions* **15**(5):891–903.

[445] Henry IM, Dilkes BP, Tyagi AP, Lin HY, and Comai L 2009. Dosage and parent-of-origin effects shaping aneuploid swarms in *A. thaliana*. *Heredity* **103**(6):458–468.

[446] Li Y, Shen Y, Cai C, Zhong C, Zhu L, et al. 2010. The type II *Arabidopsis* Formin14 interacts with microtubules and microfilaments to regulate cell division. *The Plant Cell* **22**(8):2710–2726.

[447] d'Erfurth I, Jolivet S, Froger N, Catrice O, Novatchkova M, et al. 2008. Mutations in *AtPS1* (*Arabidopsis thaliana parallel spindle 1*) lead to the production of diploid pollen grains. *PLoS Genetics* **4**(11):e1000274.

[448] Humayun MZ 1998. SOS and Mayday: Multiple inducible mutagenic pathways in *Escherichia coli*. *Molecular Microbiology* **30**(5):905–910.

[449] Radman M, Taddei F, and Matic I 2000. Evolution-driving genes. *Research in Microbiology* **151**(2):91–95.

[450] Aertsen A and Michiels CW 2005. Diversify or die: Generation of diversity in response to stress. *Critical Reviews in Microbiology* **31**(2):69–78.

[451] Arber W 2000. Genetic variation: Molecular mechanisms and impact on microbial evolution. *FEMS Microbiology Reviews* **24**(1):1–7.

[452] Presser A, Elowitz MB, Kellis M, and Kishony R 2008. The evolutionary dynamics of the *Saccharomyces cerevisiae* protein interaction network after duplication. *Proceedings of the National Academy of Sciences USA* **105**(3):950–954.

[453] Bekaert M and Conant GC 2011. Copy number alterations among mammalian enzymes cluster in the metabolic network. *Molecular Biology and Evolution* **28**(2):1111–1121.

[454] Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, et al. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research* **22**(1):95–105.

[455] Evangelisti AM and Conant GC 2010. Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biology and Evolution* **2**:826–834.

[456] Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, et al. 2005. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309**(5736):938–940.

[457] Braasch I, Schartl M, and Volff JN 2007. Evolution of pigment synthesis pathways by gene and genome duplication in fish. *BMC Evolutionary Biology* **7**(74):1–18.

[458] Veron AS, Kaufmann K, and Bornberg-Bauer E 2007. Evidence of interaction network evolution by whole-genome duplications: A case study in MADS-box proteins. *Molecular Biology and Evolution* **24**(3):670–678.

[459] Ming R, Vanburen R, Liu Y, Yang M, Han Y, et al. 2013. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biology* **14**(5):R41.

[460] Peng ZH, Lu Y, Li LB, Zhao Q, Feng Q, et al. 2013. The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nature Genetics* **45**(4):456–461.

[461] Li WLS and Drummond AJ 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* **29**(2):751–761.

[462] Nichols HW 1980. *Polyploidy - Biological Relevance Volume 13*, chapter Polyploidy in algae. Springer.

[463] Jeffroy O, Brinkmann H, Delsuc F, and Philippe H 2006. Phylogenomics: The beginning of incongruence? *Trends in Genetics* **22**(4):225–231.

[464] Kurtzman CP 2003. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Research* **4**(3):233–245.

[465] Fitzpatrick DA, Logue ME, Stajich JE, and Butler G 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology* **6**:99.

[466] Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, et al. 2009. Comparative genomics of protoploid Saccharomycetaceae. *Genome Research* **19**(10):1696–1709.

[467] Bergsten J 2005. A review of long-branch attraction. *Cladistics* **21**:163–193.

[468] Drummond AJ and Rambaut A 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**:214.

[469] Massey SE, Moura G, Beltrao P, Almeida R, Garey JR, et al. 2003. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in *Candida* spp. *Genome Research* **13**(4):544–557.

[470] Magallon S 2004. Dating lineages: Molecular and paleontological approaches to the temporal framework clades. *International Journal of Plant Sciences* **165**:S7–S21.

[471] Yang Z and Bielawski JP 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* **15**(12):496–503.

[472] Thompson WR 1935. On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics* **6**(4):214–219.

[473] Hurst LD 2002. The ka/ks ratio: Diagnosing the form of sequence evolution. *Trends in Genetics* **18**(9):486.

[474] Rost B 1999. Twilight zone of protein sequence alignments. *Protein Engineering* **12**(2):85–94.

[475] Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences USA* **104**(49):19369–19374.

[476] Smith SA, Beaulieu JM, Stamatakis A, and Donoghue MJ 2011. Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany* **98**(3):404–414.

[477] Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* **98**(4):704–730.

[478] Leitch IJ, Greilhuber J, Dolezel J, and Wendel J 2013. *Plant Genome Diversity Volume 2*. Springer-Verlag.

[479] Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, et al. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161**(2):105–121.

[480] Couvreur TLP, Forest F, and Baker WJ 2011. Origin and global diversification patterns of tropical rain forests: Inferences from a complete genus-level phylogeny of palms. *BMC Biology* **9**(44):1–12.

[481] Baker WJ and Couvreur TLP 2013. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. I. Historical biogeography. *Journal of Biogeography* **40**(2):274–285.

[482] Hugall AF, Foster R, and Lee MSY 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene *RAG-1*. *Systematic Biology* **56**(4):543–563.

[483] Heled J and Drummond AJ 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology* **61**(1):138–149.

[484] Crepet W and Nixon K 1998. Fossil Clusiaceae from the late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. *American Journal of Botany* **85**(8):1122.

[485] Davis CC, Webb CO, Wurdack KJ, Jaramillo CA, and Donoghue MJ 2005. Explosive radiation of Malpighiales supports a mid-Cretaceous origin of modern tropical rain forests. *American Naturalist* **165**(3):E36–E65.

[486] Gandolfo MA, Nixon KC, and Crepet WL 1998. A new fossil flower from the Turonian of New Jersey: *Dressiantha bicarpellata* gen. et sp. nov. (Capparales). *American Journal of Botany* **85**(7):964–974.

[487] Pigg KB, Manchester SR, and Devore ML 2008. Fruits of Icacinaceae (tribe Iodeae) from the Late Paleocene of western North America. *American Journal of Botany* **95**(7):824–832.

[488] Sun G, Dilcher DL, Wang HS, and Chen ZD 2011. A eudicot from the Early Cretaceous of China. *Nature* **471**(7340):625–628.

[489] Anderson CL, Bremer K, and Friis EM 2005. Dating phylogenetically basal eudicots using *rbcl* sequences and multiple fossil reference points. *American Journal of Botany* **92**(10):1737–1748.

[490] Friis EM 1988. *Spirematospermum chandlerae* sp. nov., an extinct species of Zingiberaceae from the North American Cretaceous. *Tertiary Research* **9**:7–12.

[491] Kress W 2006. The evolutionary and biogeographic origin and diversification of the tropical monocot order Zingiberales. *Aliso* **22**:621–632.

[492] Berry EW 1914. The Upper Cretaceous and Eocene floras of South Carolina, Georgia. *US Geological Survey* **84**:1–200.

[493] Wikstrom N, Savolainen V, and Chase MW 2001. Evolution of the angiosperms: Calibrating the family tree. *Proceedings of the Royal Society B : Biological Sciences* **268**(1482):2211–2220.

[494] Edwards D and Feehan J 1980. Records of *Cooksonia*-type sporangia from Late Wenlock strata in Ireland. *Nature* **287**(5777):41–42.

[495] De Craene LPR and Haston E 2006. The systematic relationships of glucosinolate-producing plants and related families: A cladistic investigation based on morphological and molecular characters. *Botanical Journal of the Linnean Society* **151**(4):453–494.

[496] Herendeen P and Crane P 1992. *Advances in Legume Systematics: Part 4 The Fossil Record*. Royal Botanical Gardens.

[497] Wheeler EF, Lee M, and Matten LC 1987. Dicotyledonous woods from the Upper Cretaceous of Southern Illinois. *Botanical Journal of the Linnean Society* **95**(2):77–100.

[498] Wheeler RJ, Lecroy SR, Whitlock CH, Purgold GC, and Swanson JS 1994. Surface characteristics for the Alkali Flats and dunes regions at White-Sands-Missile-Range, New-Mexico. *Remote Sensing of Environment* **48**(2):181–190.

[499] Bell CD, Soltis DE, and Soltis PS 2005. The age of the angiosperms: A molecular timescale without a clock. *Evolution* **59**(6):1245–1258.

[500] Whelan S and Goldman N 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**(5):691–699.