Accurate Inference of the Polyploid Continuum Using Forward-Time Simulations

Tamsen Dunn [1,2] Arun Sethuraman [*,1]

[1]Department of Biology, San Diego State University, San Diego, CA, USA
[2]Department of Evolution, Ecology, and Organismal Biology, University of California Riverside, Riverside, CA, USA

***Corresponding author:** E-mail: asethuraman@sdsu.edu.

Associate editor: Emily Josephs

Abstract

Multiple rounds of whole-genome duplication (WGD) followed by diploidization have occurred throughout the evolutionary history of angiosperms. Much work has been done to model the genomic consequences and evolutionary significance of WGD. While researchers have historically modeled polyploids as either allopolyploids or autopolyploids, the variety of natural polyploids span a continuum of differentiation across multiple parameters, such as the extent of polysomic versus disomic inheritance, and the degree of genetic differentiation between the ancestral lineages. Here we present a forward-time polyploid genome evolution simulator called SpecKS. SpecKS models polyploid speciation as originating from a 2D continuum, whose dimensions account for both the level of genetic differentiation between the ancestral parental genomes, as well the time lag between ancestral speciation and their subsequent reunion in the derived polyploid. Using extensive simulations, we demonstrate that changes in initial conditions along either dimension of the 2D continuum deterministically affect the shape of the *Ks* histogram. Our findings indicate that the error in the common method of estimating WGD time from the *Ks* histogram peak scales with the degree of allopolyploidy, and we present an alternative, accurate estimation method that is independent of the degree of allopolyploidy. Lastly, we use SpecKS to derive tests that infer both the lag time between parental divergence and WGD time, and the diversity of the ancestral species, from an input *Ks* histogram. We apply the latter test to transcriptomic data from over 200 species across the plant kingdom, the results of which are concordant with the prevailing theory that the majority of angiosperm lineages are derived from diverse parental genomes and may be of allopolyploid origin.

Key words: **forward-time simulations, polyploidy, genomics.**

Introduction

Multiple rounds of whole-genome duplication (WGD) followed by diploidization have occurred throughout the evolutionary history of angiosperms (Otto and Whitton 2000; Soltis and Soltis 2012; Wendel 2015). WGD is considered a major speciation mechanism (Doyle and Egan 2010; Schranz et al. 2012; Wendel et al. 2016; Clark and Donoghue 2018), presenting a massive "macromutation," potentially interfering with sexual reproduction, releasing transposons, and unbalancing molecular signaling pathways (McClintock 1929; Stebbins 1951; Mayer and Aguilera 1990; Comai et al. 2000; Ramsey and Schemske 2002; Le Comber et al. 2010; Arrigo and Barker 2012; Yant and Bomblies 2015; Bomblies et al. 2016; Zhang et al. 2016; Baduel et al. 2018). It has been theorized that the advantage of WGD lies not in the multiplicity of genomic material itself, but the intense period of genomic reorganization and gene shedding that follows, known as diploidization (Buggs et al. 2011; Madlung 2013; Soltis et al. 2014; Tank et al. 2015; Dodsworth et al. 2016; Soltis et al. 2016; Robertson et al. 2017; Baniaga et al. 2020; Carretero-Paulet and Van de Peer 2020; Nieto Feliner et al. 2020; Li et al. 2021; Van de Peer et al. 2021).

Waves of contemporaneous WGD events across multiple plant lineages are observed throughout plant evolutionary history, with the largest cluster reported as roughly contemporaneous with the Cretaceous—Tertiary (K-T) extinction event 65 million years ago (Soltis and Burleigh 2009; Van de Peer et al. 2021; Vanneste et al. 2013). This K-T WGD cluster is associated with developments in stress tolerance in plants, such as heat shock transcription factors and light, drought, and temperature stress response regulators (Fawcett et al. 2009; Zhang et al. 2020; Van de Peer et al. 2021). A more recent wave of polyploidy is associated with the rapid glacial cycling of the quaternary period, and it has been suggested that the ecological upheaval associated with the Anthropocene may yet bring about further waves of WGD (Levin 2020).

Accurate estimation of timing is critical for correlating WGD events with ecological and geological factors

Received: August 15, 2024. **Revised:** October 22, 2024. **Accepted:** November 05, 2024

(Barba-Montoya et al. 2018; Clark and Donoghue 2019). It is also crucial for our understanding of the role and processes of diploidization. For example, errors in the estimation of WGD timing could lead to inaccuracies in estimations of rates of gene mutation, loss and retention and homeologous exchange. An accumulation of systematic errors in assessing the timing of WGD events may bias our understanding of the significance of WGD, both in comparison to competing evolutionary theories (Hoegg et al. 2004: 20; Doyle and Egan 2010; De La Torre et al. 2017; Laurent et al. 2017; Mable et al. 2018; David et al. 2020; Li et al. 2021) and in correlation with major ecological and geoclimatic events (Fawcett et al. 2009; Lohaus and Van de Peer 2016; Levin 2020; Van de Peer et al. 2021).

There are many methods to date WGD events. One method is to use genetic data to place the WGD event on a phylogenetic tree, if enough is known about its presence from multiple lineages (Bowers et al. 2003; Li et al. 2015; Li et al. 2018; Li and Barker 2020; Parey et al. 2022), or its presumptive parental species (Lott et al. 2009; Doyle and Egan 2010; Estep et al. 2014; Douglas et al. 2015; Thomas et al. 2017; Mccann et al. 2018; Wen et al. 2018; Yan et al. 2022; Conant 2023). If only the polyploid genome is available, a comparison of duplicated genes within the polyploid genome may be made to estimate their likely time of divergence (Lynch and Conery 2000; Blanc and Wolfe 2004; Cui et al. 2006; Vanneste et al. 2013; Clark et al. 2019; Chen and Zwaenepoel 2023). Syntenic relationships (self-synteny and with related species) may also be used to corroborate WGD (Vandepoele et al. 2002; Hampson et al. 2003; Wang et al. 2006; Parey et al. 2020).

Historically, the most common method to date a WGD event is to use a molecular clock to calibrate the divergence times between paralogs in a polyploid genome. The number of synonymous substitutions per synonymous site in protein-coding genes (*Ks*) is calculated between all paralogous pairs, and if there is a peak at $Ks > 0$, the position of this peak is used to infer the time of WGD (Blanc and Wolfe 2004; Chen and Zwaenepoel 2023) and see Figs. 1 and 2.

Since paralogs can arise from a number of evolutionary processes, *Ks* histogram shapes may be complex, and their interpretation may be challenging. *Ks* histograms generally have a first peak at $Ks = 0$, due to the constant birth and death of small-scale duplications (SSDs). These SSD paralogs have recently arisen but have not yet been shed, and their lifespan follows an exponential decay model (Lynch and Conery 2000). If a species has undergone one or more WGD events, additional peaks will occur in the histogram for $Ks > 0$, and these peaks may even be overlapping. These additional peaks are due to ohnologs (paralogs formed by WGD). The naive interpretation is that the mode in ohnolog divergence times represents the time of the WGD event (Ohno 1970; Blanc and Wolfe 2004). Univariate mixture models are used to empirically fit the peaks (Schlueter et al. 2004; Cui et al. 2006; Vanneste et al. 2013; Tiley et al. 2018; Li and Barker 2020; Chen and Zwaenepoel 2023).

However, the *Ks*-based approach has several pitfalls. The conversion between *Ks* and time is not straightforward (Wolfe et al. 1987; Doyle and Egan 2010; Barba-Montoya et al. 2018), *Ks* saturates for $Ks > 2$, or roughly 200 million years (Cui et al. 2006; De La Torre et al. 2017; Li and Barker 2020), and multivariate fitting methods have been shown to overfit distributions (Vanneste et al. 2013; Tiley et al. 2018; Zwaenepoel and Van de Peer 2019). Restricting the *Ks* histogram to ohnologs may improve resolution (Van de Peer 2004; Zwaenepoel and Van de Peer 2019; Sensalari et al. 2022; Sutherland et al. 2024), but there remains significant methodological concerns.

While *Ks*-based methods can in theory correctly date the origin of a polyploid lineage, which traces back to a single individual (assuming the genomes began to diverge at the same instant they duplicated; Doyle and Egan 2010), the methods may fail for allopolyploids (polyploids derived from distinct species) (Thomas et al. 2017; Mccann et al. 2018; Wen et al. 2018; Bouckaert et al. 2019; Conant 2023). This is because for allopolyploids the peak of the *Ks* distribution corresponds to the divergence time between the diploid parental species (hereon T_{DIV}), not the time of origin of the polyploid (hereon T_{WGD}) (Doyle and Egan 2010; Thomas et al. 2017; Chen and Zwaenepoel 2023), as shown in Fig. 1. Since plant genomes can remain compatible for 10 mya or more after the last common ancestor (Senchina et al. 2003; Levin 2013) and up to 50 mya in one documented case (Rothfels et al. 2015), the difference between T_{DIV} and T_{WGD} can be significant. Confounding the *Ks* peak with T_{WGD} may also be problematic for autopolyploids (polyploids derived from within a species), whose ohnologs may show complex patterns of divergence post WGD (Gaeta and Pires 2010; Parey et al. 2022; Lv et al. 2024).

Researchers have historically treated allopolyploids and autopolyploids as separate idealized cases. However, in practice the multiplicity of traits, which have been used to distinguish between auto and allopolyploids (for example, polysomic vs. disomic inheritance, levels of genetic differentiation between the diploid progenitor, cytology, and taxonomic assignment) can lead to conflicting classifications. Furthermore, none of these traits lend themselves to binary categorization. In truth, the variety of natural polyploids span a continuum of differentiation across multiple parameters (Stebbins 1950; Ramsey and Schemske 2002; Meirmans and Van Tienderen 2013; De Storme and Mason 2014; Mason and Wendel 2020; Blischak et al. 2023). Because of this complexity, while it is difficult to quantify the extent and ramifications of errors in Ks-based estimates of T_{WGD}, it is clear that in the case of allopolyploids, historical methods will miss the mark.

Recently, several methods have been developed that are capable of dealing with the timing problems unique to allopolyploidy. These methods rely on additional data from the parental species or from broader sampling of the polyploid taxa (Thomas et al. 2017; Mccann et al. 2018; Wen et al. 2018; Bouckaert et al. 2019; Conant 2023). Population-demographic approaches have also

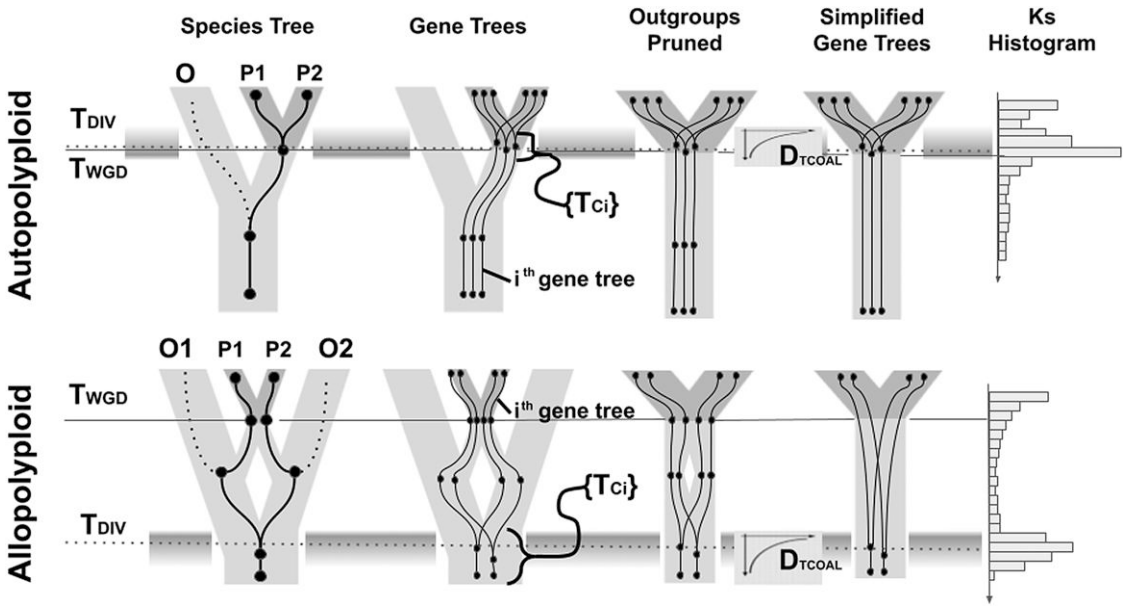

Fig. 1. Relationship between the shape of the *Ks* histogram and the modes of auto versus allopolyploid speciation. Top: Autopolyploid speciation. Bottom: Allopolyploid speciation. Outgroup (O) and parental species are indicated in light gray. Polyploid species are shown in dark gray with duplicated genomes denoted as P1 and P2. The arrow in the *Ks* histogram points backwards in time. The set of gene tree coalescents $\{T_{Ci}\}$ are shown as nodes on the gene trees, and their distribution in time (D_{TCOAL}) is indicated by the gray shaded bands. For allopolyploids the offset between T_{WGD} and T_{DIV} represents the lag time between separation of the diploid parental species (T_{DIV}) and their later conjunction by polyploidization (T_{WGD}). For autopolyploids, there may be instances where the gene tree divergence may begin before T_{WGD} or after it. In both cases, T_{WGD} and T_{DIV} are not synonymous, and D_{TCOAL} may be complex. Figures at the far right and left are derived from (Chen and Zwaenepoel 2023).

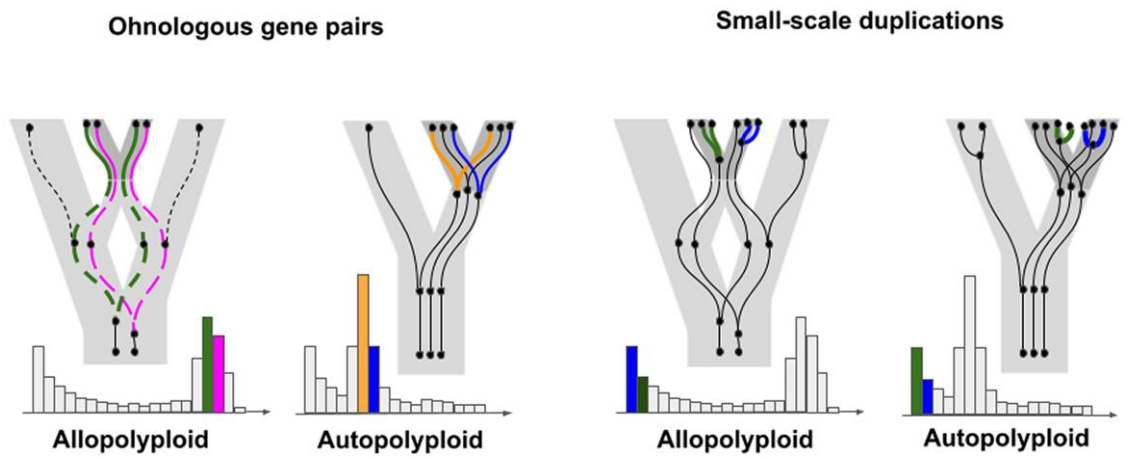

Fig. 2. Ohnologs (paralogs generated by WGD) and "SSDs" (paralogs generated as single or small-scale copies) have unique contributions to the *Ks* histogram. Ohnologs (left panel) are shown for both allo and autopolyploids. Green and pink ohnologs (far left) are attributed to allopolyploidy and are born as orthologs (genes duplicated by speciation, dashed lines). Gold and blue ohnologs (center left) are attributed to autopolyploidy and are born by WGD. SSDs (right panel) follow the same birth and death process for both allo and autopolyploids and are shown here in green and blue. Unless maintained by selection, SSDs tend to be shed rapidly and are thus found at the far left of the *Ks* histograms. As in Fig. 1, the polyploid species is shown in dark gray while the parental and outgroups are shown in light gray. The arrow of time points into the past along the X axis of the *Ks* histogram.

been used with recent success to time WGD (Gutenkunst et al. 2009; St Onge et al. 2012; Roux and Pannell 2015; Roux et al. 2017; Blischak et al. 2023; Booker and Schrider 2024).

As part of the 1,000 Plants (1KP) initiative (Leebens-Mack et al. 2019), transcriptomic data from over 1,000 plants spanning the plant kingdom was used to compile *Ks* histograms (Table 1). As in other works, Gaussian mixture models were fit and used to detect WGD from *Ks* peaks (Clark et al. 2019; Qiao et al. 2019; Guo et al. 2020; Li and Barker 2020). But a review of empirical *Ks* histogram shapes from the 1KP dataset show great

Table 1 Summary of 1KP categorization results

N_e categorization	Number in category	Percent of total	Metric mean value
Low	12	5.26	−5.05
Medium	38	16.67	−3.69
High	178	78.07	−2.45
Total	228	100	−2.80

complexity beyond the Gaussian distribution (Zwaenepoel and Van de Peer 2019; Sensalari et al. 2022), which simulations have thus far have not been able to replicate (Sutherland et al. 2024).

Here, we present the *Ks* simulator SpecKS. SpecKS simulates the forward evolution of polyploid genomes whereby the mode of speciation is not starkly allo or autopolyploid, but instead a point in a 2D continuum model where the parental species' divergence in time and genetic space may vary independently (Fig. 3), which may better align with our evolving and more nuanced understanding of polyploidy (Parisod et al. 2010; Doyle and Sherman-Broyles 2017). Thus, SpecKS is highly configurable (supplementary tables S2 and S3, Supplementary Material online) and capable of producing a rich array of *Ks* distribution shapes, modeling both the SSD and ohnolog components. Here, we use SpecKS to demonstrate the sensitivity of the *Ks* distribution to a number of ancient speciation parameters, such as effective population sizes (N_e), gene shedding rates, and the separation in time between T_{DIV} and T_{WGD}.

Results

Methods Overview

Ks histograms across the tree of life show a great diversity of distribution shapes, with some showing a high degree of symmetry while others appearing more skewed (Li and Barker 2020), suggesting the *Ks* histogram may bear signatures of evolutionary parameters or events beyond the presence or absence of WGD. Thus, we sought to build a simulation engine to investigate the potential effects of a variety of ancient polyploid speciation parameters on the distribution of *Ks* histogram shapes.

We developed a novel simulation-engine, SpecKS, which models polyploid speciation and evolution as a reticulate process. The simulation follows the evolution of an initial ancestral genome, which diverges at a given time (T_{DIV}) into two sister diploid species. These sister species later recombine at T_{WGD}, and the resulting polyploid continues to evolve to the present day. A set of *i* gene-trees $\{G_i\}$ are embedded in this reticulate topology with a configurable distribution (D_{TCOAL}) of coalescent times $\{T_{ci}\}$, which the user might base on ancestral diversity (Π), effective population size (N_e), or some other relationship. Genes (as a set of random strings of nucleotides, $\{N_i\}$) are evolved along $\{G_i\}$ using PAML v4.10.7 (Yang 2007) under neutrality to the present time. *Ks* is then calculated between the resulting paralogs, which are the leaves of the gene trees (Fig. 4).

Implicit in this model is the concept that the polyploid continuum has more than 1D. We allow the degree of allopolyploidy to be a function of both length of the time the parental species were separated (ΔT) and the ancestral diversity (Π) of the subgenomes at T_{DIV}. This allows us to separate the effects of these critical speciation parameters on the *Ks* histogram. We additionally disambiguate T_{DIV} and T_{WGD}, which has confounded estimates of T_{WGD} in the past.

SpecKS Demonstrates that Changes Along Either Dimension of the 2D Continuum Will Deterministically Affect the Shape of the *Ks* Histogram

The initial distribution of the divergence times for the gene trees D_{TCOAL} is supplied by the user as an input parameter. This is to allow the user as much flexibility as possible with regard to modeling their system. To test if differences in these initial distribution shapes can be theoretically detected even after WGD and thus potentially affect the *Ks* histogram of extant polyploids, we simulated polyploids with modes of speciation from all four corners of our 2D continuum (Fig. 3). Specifically, we tested sets of allopolyploids derived from (A) low-N_e ancestral species and a large ΔT between T_{DIV} and T_{WGD}, (B) high-N_e ancestral species and a large ΔT, (C) low-N_e ancestral species and a small ΔT, and (D) high-N_e ancestral species and a small ΔT. Set (B) comprises canonical allopolyploids whose parental species were highly differentiated and have spent several million years apart before hybridization. Set (C) represents polyploids whose parental species were minimally differentiated and had no significant time between divergence and conjunction. The off-diagonals (A and D) represent plausible, but less intuitive polyploids. Specifically, set (A) are derived from parental species with little diversity between them at speciation, but a great amount of time between parental divergence and WGD. Set (D) are polyploids derived from parental species with a greater degree of diversity between them at speciation (highly differentiated subpopulations leading to separate species, but hybridization via polyploidy soon followed parental divergence). For all four sets, we simulated a range of WGD times, from ancient to relatively recent (80–0 MYA) (Fig. 5).

With regard to the diversity dimension of the 2D continuum, our results showed that the polyploids derived from high-N_e ancestral species (B&D) show a more skewed, fat-tailed WGD component in the *Ks* distribution compared to the low-N_e polyploids (Fig. 5). We also see that these differences persist for about 50 MY (Fig. 6, top). Along the ΔT dimension, we see that differences in ΔT had little effect on the skew but will affect the relative height differential between the SSD peak height and the WGD peak height (Figs. 5 and 6 bottom).

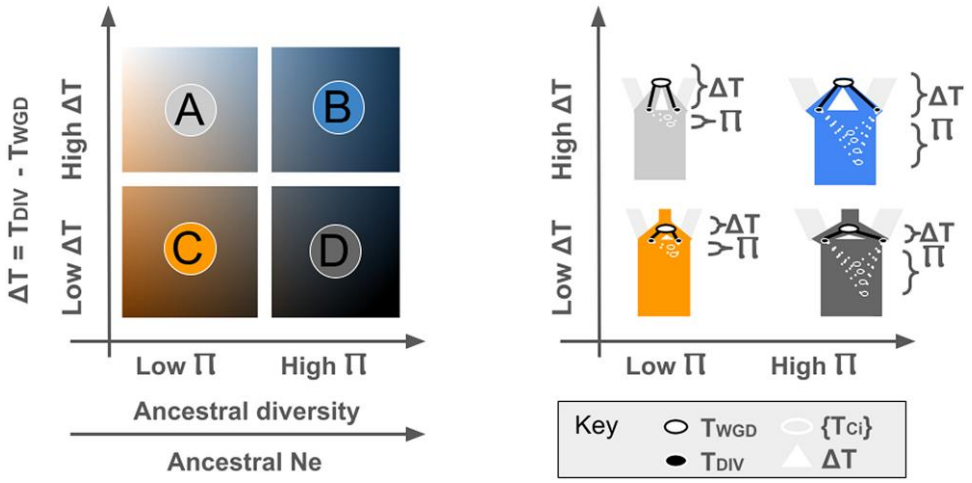

Fig. 3. The SpecKS 2-dimensional model parameterizing allopolyploid speciation. A graphical representation of the 2D polyploid continuum modeled by SpecKS, showing ancestral diversity (Π), the length of time the parental species were separated (ΔT), the parental divergence time (T_{DIV}), the time of whole genome duplication (T_{WGD}), and the set of gene-tree coalescent times $\{T_{COAL\_i}\}$ in the SpecKS model. On the left, we show the 2D continuum, with the x-axis denoting low to high Π, and the y-axis denoting low to high ΔT. More allopolyploid species generally have greater Π and ΔT, but they are not necessarily 1 to 1. Allopolyploid-derived species can fall anywhere on this continuum, and we give examples with A) high ΔT and low Π, B) high ΔT and high Π, C) low ΔT and low Π, and D) low ΔT and high Π. On the right, we give example species-level topologies for each of these situations (A-D). Each of the four diagrams on the right shows two parental diploids (light gray) diverging from a common diploid ancestor (the "trunk" of the tree), and the emergent polyploid (the reticulation). The T_{WGD} is the white circle with the black boundary, and T_{DIV} is the black circle with the white boundary. For each polyploid, ΔT is the vertical distance between T_{WGD} and T_{DIV}, thus the top two polyploids (A and B) have greater ΔT than the bottom two polyploids (C and D), as indicated by the solid black lines. For each polyploid, ancestral Π is indicated by the width of the ancestral trunk, thus the left two polyploids (A and C) have smaller Π than the right two polyploids (B and D). Lastly, the dashed lines indicate the coalescent times between orthologous genes, as measured (vertically) between each parental genomes at T_{DIV}. Thus, when the parental species originate from more diverse populations, the coalescent times reach further back time (right side polyploids, B and D). When the parental species originate from more less-diverse populations, the coalescent times are smaller (left side polyploids, A and C).

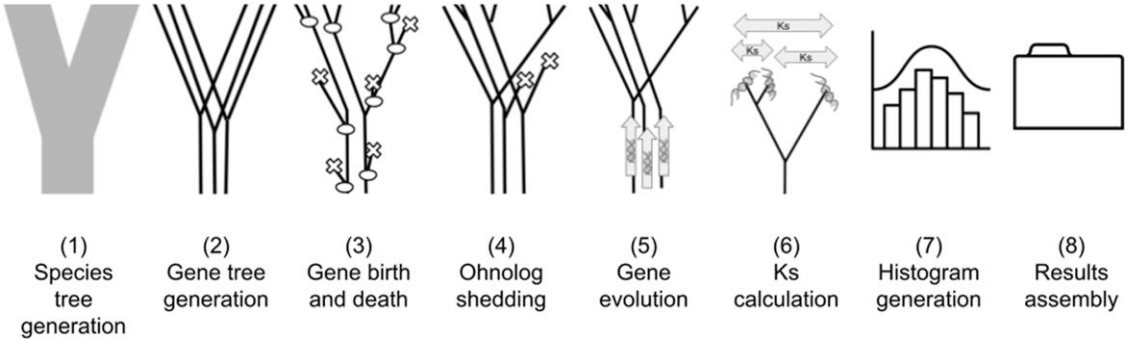

Fig. 4. The eight modules of the SpecKS polyploidy simulation pipeline. (1) The input species tree (gray), indicating ancestral taxa and parental divergence. (2) The gene-trees (black lines) generated within the species tree, according to the requested ortholog distribution model. (3) Within each gene tree, genes are born (white circles) and die (white X's) according to the input gene birth and death rates. (4) Ohnologs are shed (white X's) according to a parameterized exponential decay model. (5) Genes are evolved along each gene tree topology using EVOLVER. (6) *Ks* calculations are made between paralogs using CODEML. (7) Histogram generation. (8) Results are compiled in a single output folder.

The Error in the Common Method of Estimating WGD Time From the *Ks* Histogram Peak Scales With the Degree of Allopolyploidy

The "risk" of using the *Ks* peak to determine the timing of WGD, particularly for allopolyploids, has been well described (Thomas et al. 2017; Chen and Zwaenepoel 2023). However, the *Ks* peak continues to be used to determine the timing of WGD. We were curious to see if SpecKS could be used to quantify the expected error in estimating T_{WGD}, and potentially empirically relate the magnitude of T_{WGD} error to the degree of allopolyploidy along a continuum. We were further interested to test if an alternative method using the inferred start time of ohnolog-shedding would yield more accurate estimates of T_{WGD}.

As expected, our simulations demonstrate that inferring T_{WGD} from the *Ks* peak may be quite problematic (off by millions of years). Since the *Ks* peak gives the T_{DIV}, not the T_{WGD}, it is no surprise that the error in estimates of T_{WGD}

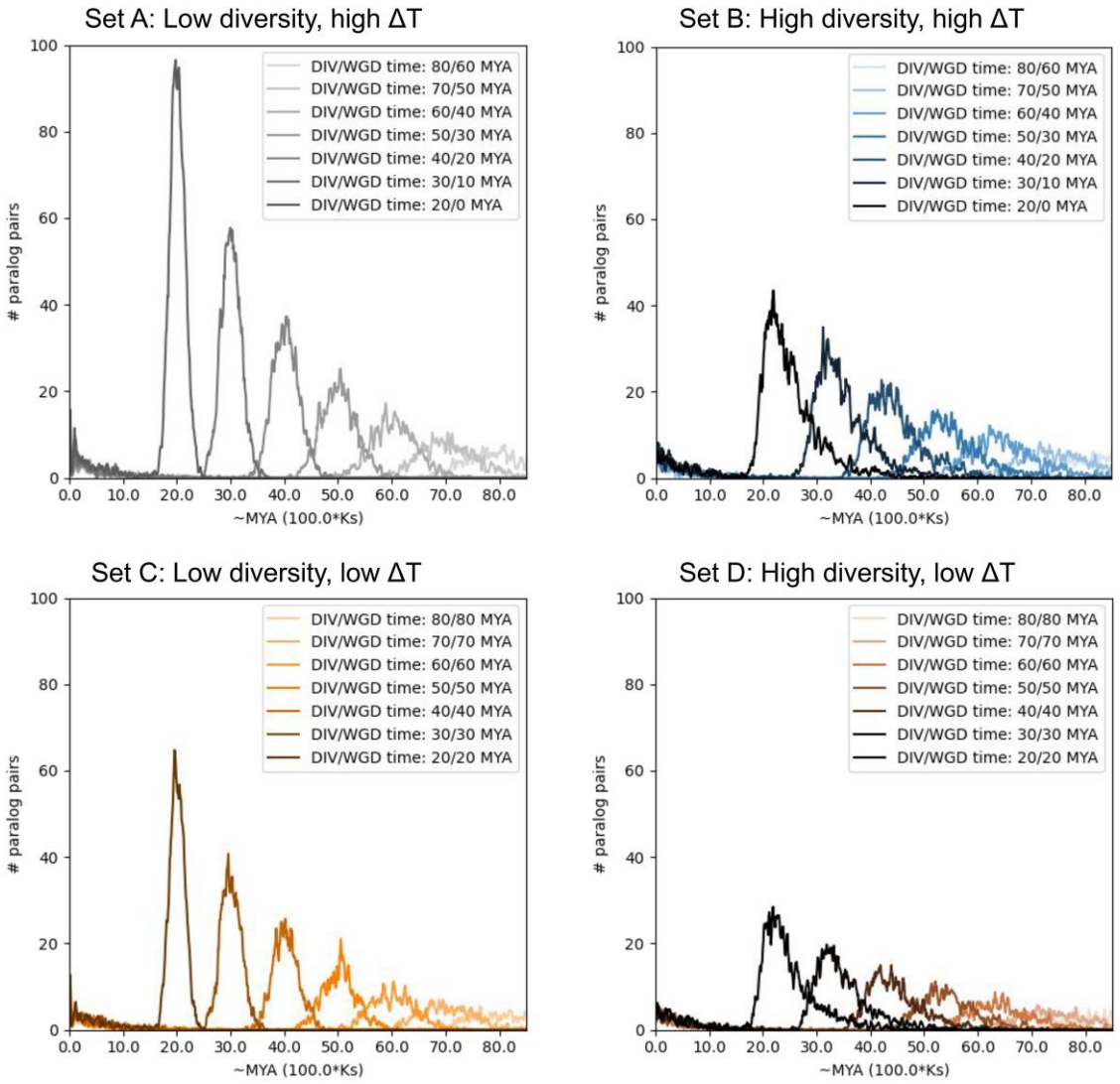

Fig. 5. SpecKS demonstrates how the shapes of the *ks* distributions attenuate over time. Ks histograms (above) are given for simulated polyploids of varying ages, from the four corners of the 2D continuum (left). Older WGD are lighter colored and more recent WGD are darker. A) (gray): low N_e ($1 \times 10\text{^}5$) ancestral species and a large ΔT (20 MY). B) (blue): high N_e ($5 \times 10\text{^}6$) ancestral species and a large ΔT (20 MY). C) (yellow): low N_e ($1 \times 10\text{^}5$) ancestral species and a small ΔT (0 MY). D) (black): high N_e ($5 \times 10\text{^}6$) ancestral species and a small ΔT (0 MY).

linearly scales with their difference. Thus, the error in inferring T_{WGD} from the *Ks* peak is proportional to ΔT, the degree of allopolyploidy as measured along the time axis of the 2D continuum (Fig. 7, top).

An Alternative, Accurate T_{WGD} Estimation Method that is Independent of the Degree of Allopolyploidy

We tested the hypothesis that, since ohnolog-shedding can only begin after WGD irrespective of the mode of polyploid speciation, using the proportion of duplicate genes remaining may be a "less-risky" metric for inference, especially if the mode of speciation is unclear. In practice, the rates of gene shedding would vary by lineage and gene family, but for the purpose of this theoretical test, we assume the gene shedding rate is constant and known a priori for our theoretical species. We thus simulated a range of *Ks* histograms parameterized with a set gene shedding rate for a range of T_{DIV} and T_{WGD}, for polyploids across a 2D polyploid continuum, thus recreating the range of possible *Ks* histograms for this hypothetical polyploid species, under a range of speciation scenarios. The simulations revealed a clear logarithmic relationship between the number of genes shed and T_{WGD}, invariant to the mode of speciation (supplementary fig. S2, Supplementary Material online, left). We were thus able to use the logarithmic relationship between T_{WGD} and the proportion of ohnologs remaining to determine T_{WGD} from the *Ks* histogram. In Fig. 7, bottom left, we demonstrate the accuracy of this method, by comparing the input (true) T_{WGD} to the recovered T_{WGD}, revealing a high accuracy, with r-value of 0.998 and a standard error of 0.004 MY. In Fig. 7, bottom right, we show that the error (the difference between the true and inferred T_{WGD}) does not increase with degree of allopolyploidy, as measured along the ΔT dimension.

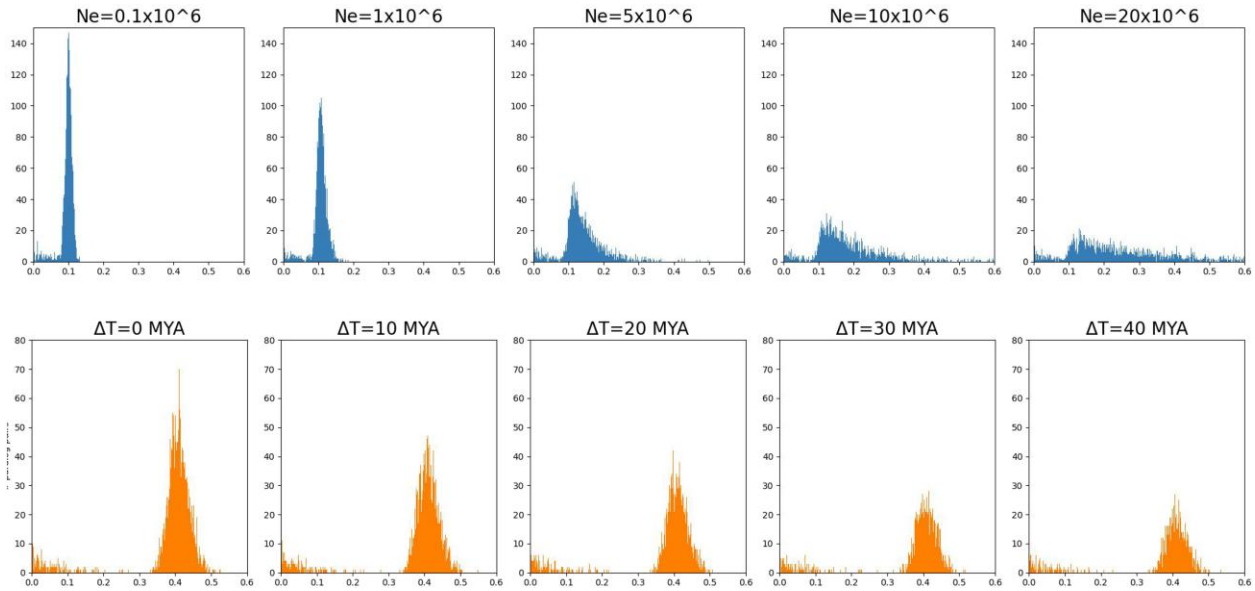

Fig. 6. SpecKS demonstrates how the shapes of the *Ks* distribution vary with respect to ancestral diversity and ΔT. Top: Ancestral N_e varies from 10^5 to 20 × 10^6 (blue). T_{WGD} is fixed at 5 MYA and T_{DIV} is fixed at 10 MYA. Bottom: T_{WGD} varies from 0 to 40 MY, T_{DIV} is fixed at 40 MYA, and N_e is fixed at 1*10^6 (yellow). X-axis is Ks and y-axis is the number of paralogs.

SpecKS Recreates the Main Features of Empirical *Ks* Histograms

Next, we wanted to know if the evolutionary models incorporated in SpecKS would sufficiently reproduce the exponential decay curves typical of the primary "SSD" peak of the *Ks* histogram, as well as the secondary "WGD" peak, which contemporary works have had difficulty to replicate (Tiley et al. 2018; Sutherland et al. 2024). We selected three tetraploid species, *Coffea arabica*, *Zea mays*, and *Populus trichocarpa* to demonstrate this. *Coffea arabica* is a relatively recent allopolyploid (WGD ~10 to 600KYA). *Zea mays* was formed by polyploidy ~14 MYA, and *Populus* ~56 MYA (Gaut and Doebley 1997; Yu et al. 2011; Dai et al. 2014; Salojarvi et al. 2023). We show in Fig. 8 that SpecKS can well replicate both the SSD and ohnolog components of the *Ks* histograms for all three species. It is particularly interesting to note that the simulated *Ks* plots match the exponential-lognormal mixture models which have historically had the greatest success fitting observed *Ks* distributions. We note that this is an emergent property of our simulation, and no lognormal distributions were input. Furthermore, to achieve the best fit, the SpecKS input parameters were optimized to minimize the Root mean squared error (RMSE) error (supplementary table S4, Supplementary Material online), yielding T_{WGD} and T_{DIV} which correspond well with estimated dates from other sources (supplementary table S5, Supplementary Material online) (Gaut and Doebley 1997; Yu et al. 2011; Dai et al. 2014; Salojärvi et al. 2024). We also note that the main difference between the simulated histograms and the true histograms is that the true histograms maintain a set of paralogs whose numbers do not decay over time and whose *Ks* distribution appears flat. One explanation for this phenomenon is that these remaining paralogs are maintained by selection, and thus not presently modeled by our system.

Using SpecKS to Estimate the Time Elapsed Between Divergence to Duplication ($\Delta T = T_{DIV} - T_{WGD}$) and Ancestral Diversity (Π) From the *Ks* Histogram

As we demonstrated previously, the changes along either dimension of the 2D continuum deterministically affect the shape of the *Ks* histogram. Thus, we wanted to know if, given the *Ks* histogram, could the initial parameters of polyploid speciation be recovered? Since we have already demonstrated that T_{WGD} can be recovered (Fig. 7), what remains is to demonstrate the recovery of T_{DIV} and Π.

The peak of the *Ks* histogram theoretically corresponds to T_{DIV} (as in Thomas et al. 2017; Chen and Zwaenepoel 2023). Indeed, SpecKS corroborates this expectation (supplementary fig. S2, Supplementary Material online, right), and we demonstrate the accuracy of this method by comparing the input (true) T_{DIV} to the recovered T_{DIV}, revealing a high accuracy, with r-value of 0.995 and a standard error of 0.008 MY. We note the caveat that this accuracy is contingent on the accuracy of the conversion factor between time and *Ks*, which is a configurable parameter in our simulation. Since $\Delta T = T_{DIV} - T_{WGD}$, the ability to estimate both T_{DIV} and T_{WGD} yields the estimated ΔT.

To test the recovery of ΔT, we generated a dataset of 160 simulations (parameters described in Table 2) spanning a variety of modes of speciation across the 2D continuum, with a range of T_{DIV}, T_{WGD}, and Π. With these simulated datasets, using a ⅓ test, ⅔ training approach

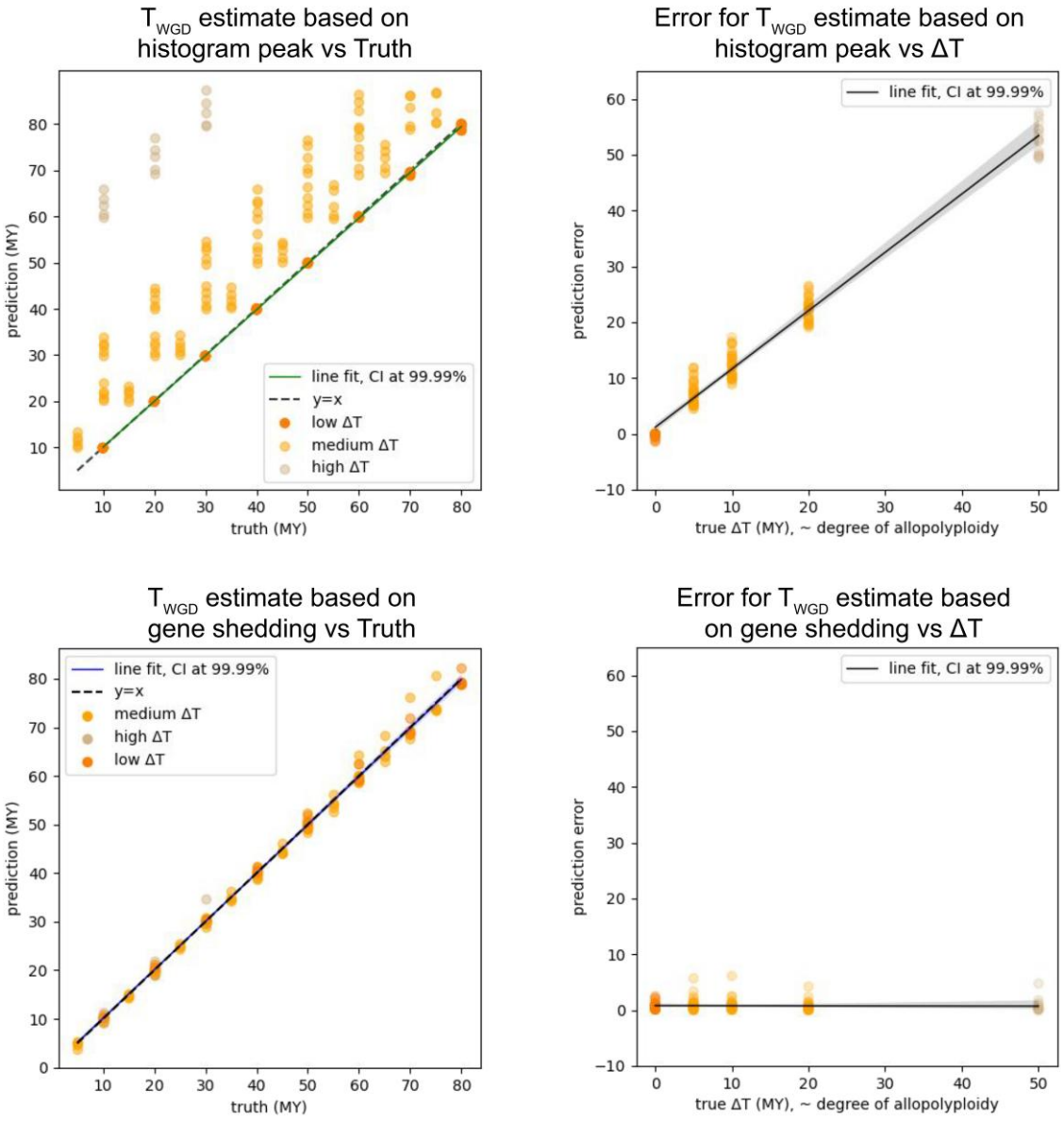

Fig. 7. SpecKS shows that estimation of WGD time based on the number of genes remaining yields accurate results, irrespective of ΔT. (Top left) T_{WGD} estimated from histogram peak. (Top right) Error in T_{WGD} as estimated from the histogram peak (y axis) versus the degree of allopolyploidy (x-axis). (Bottom left) T_{WGD} as estimated from the number of genes shed. (Bottom right) Error in the T_{WGD} as predicted from the number of genes shed (y axis) versus the degree of allopolyploidy (x-axis). Each data point represents results for a given simulation from the full set of simulations described in Table 2, and data from all simulations are given in each figure. T_{DIV} times range from 10 to 80 MYA, by 10, with WGD offset (ΔT) by 0, 5, 10 and 50 MY. $K = N_e{*}G_t$ is 0.1, 1.0, 5.0, 10, and 20 MY. The line fit is made to all points in each figure, with the exception of the top left figure. In the top left, the green line is only fit to the "low ΔT" simulation runs, to highlight that only when T_{DIV} is close to T_{WGD}, does the Ks peak provide a good estimate for T_{WGD} (see conformity with y = x for only these points).

with the data, we trained a logistic regression classification model to discriminate between large, medium or small ΔT. We selected these three categories to represent three biologically distinct t patterns of polyploid speciation: (i) a small ΔT, where WGD follows parental species divergence by 0 to 5 MY, and thus includes parental species who are only recently diverged, where the distinction between autopolyploid (WGD from within the same species) and allopolyploid (WGD derived from different species) might be poorly resolved; (ii) a medium ΔT, between 5 and 30 MY, which would have clearer resolution between parental species and includes most canonical autopolyploids; and (iii) a large ΔT (>30 MY), which constitutes the outer boundary of empirical observations (see Gaut 2002; Senchina et al. 2003; Zeng et al. 2012; Levin 2013; Estep et al. 2014; Rothfels et al. 2015; Barker et al. 2016; Mccann et al. 2018 and more) for estimates of the range of plausible ΔT). Within the context of our simulated results, this approach was able to correctly classify small, medium, and large ΔT-derived polyploids, with 100% accuracy for WGD up to 80 MYA (Fig. 9).

To test the recovery of Π, we chose a model (described in Methods) that would relate Π to the distribution of ortholog divergence times, since it is that distribution,

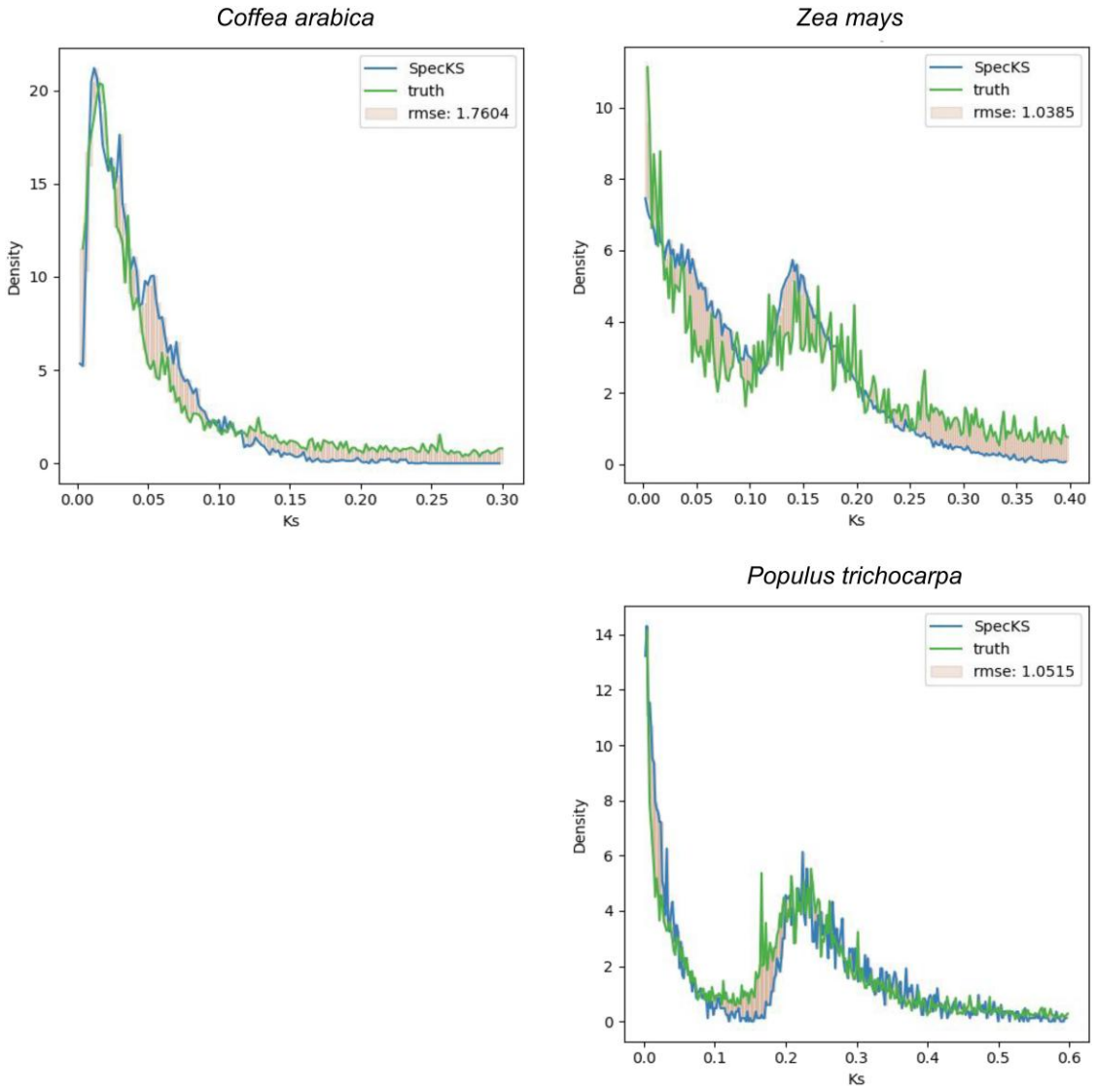

Fig. 8. Using SpecKS to reproduce *Ks* empirical histograms from transcriptomic data. SpecKS-simulated histogram in blue and the empirical *Ks* histogram from transcriptomic data in green. X-axis is the *Ks* value and Y-axis is the paralog density rather than total number of paralogs, since the totals were different between the two datasets. For the SpecKS results, we simulated a genome of 20,000 representative paralogs. The empirical data are the full transcriptomic dataset. RMSE is given in the figure legend.

which is input to SpecKS, not Π directly. Based on this model, we were able to use a range of N_e to generate a set of initial gene-tree divergence distributions as input to SpecKS. Using a ⅓ test, ⅔ training approach with the data, we trained a logistic regression classification model on the set of simulations described in Table 2 to discriminate between high, medium or low ancestral N_e. We selected these three categories (small, medium, and large N_e) to span the levels of N_e that are empirically observed for plant populations at the species level. Our "small N_e" category comprises ancestral species with diversity commensurate with an effective population size of 0 to 1 million (most plant species). Our "medium N_e" comprises N_e between 1 million and 5 million, rarely observed in plants. The "high Ne" category comprises ancestral species with N_e >5 million, surpassing the outer boundary of empirical observations (see Szoveny et al. 2008; Gossmann et al. 2010; Slotte et al. 2010; Strasburg et al. 2011; Ai et al. 2012; Gargiulo et al. 2024) and more for empirical studies of N_e in plants. Thus, the diversity observed in ortholog coalescent times in the "medium N_e" and "high N_e" categories might be poorly explained by the null model (the Kingman coalescent, panmixia), and may better corroborate alternative hypotheses, for example that barriers to mating were already in place at the time of parental divergence. Thus, parentals with these higher levels of divergence may bear the signatures of gradual speciation, which would be strongly concordant with allopolyploidy. In contrast, parentals with the lowest levels of genetic differentiation may only be recently diverged, and the distinction between auto and allopolyploidy may be less clear. Within the context of our simulated results, this

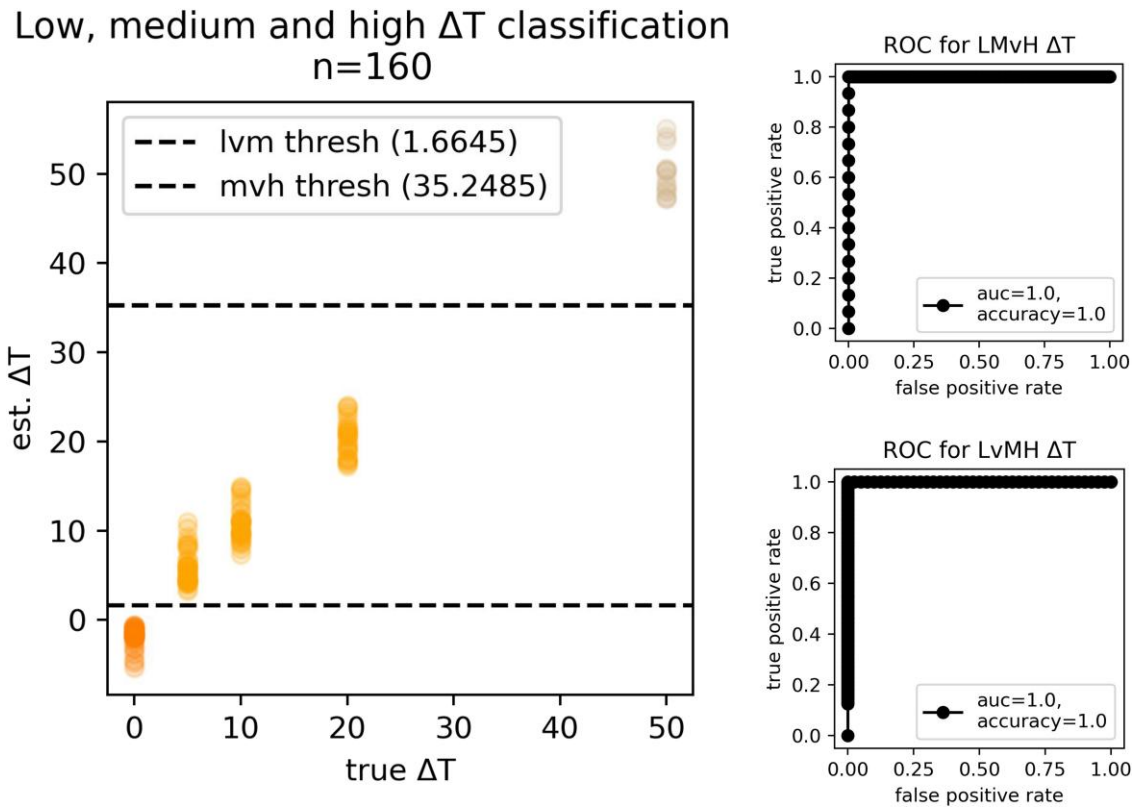

Fig. 9. The low, medium, and high ΔT discrimination model. (left) The ΔT discrimination threshold applied to the estimated ΔT. (right) Accuracy of inference of estimation ΔT, represented as an ROC plot. Each data point represents results for a given simulation from the set of simulations described in Table 2. Lighter colors denote higher ΔT. Orange, yellow and beige correspond to parameters ΔT ≤5, ΔT≤30, ΔT > 30.

approach was able to correctly classify small, medium, and large N_e-derived polyploids, with < 95% accuracy for WGD up to 80 MYA (Fig. 10).

Note, our aim with both the logistic regression classification models above is to demonstrate that SpecKS can be used to derive inference models to address very broad biological questions (hence the categorical results, not point estimates). We do not suggest that SpecKS can be used in isolation to provide precise estimates of ancestral N_e or ΔT. As SpecKS matures as a modeling platform, these inference models may be extended to better capture underlying biological complexity, but this is not currently the case. We binned our inference results into three categories to represent the high and low extremes observed in empirical studies and allowed an intermediary category as "buffer region" between the two. For best practices, if the user is working with a particular lineage, we would suggest (i) building a lineage-specific test dataset where all biological input parameters are tuned to that lineage and using replicates to define confidence intervals. Or (ii) if biological input parameters are not known, a likely range of parameters might be selected and used to build several different sets of simulation settings. Data from these sets of runs could be used to establish the upper and lower bounds of various inferences or bootstrapped to achieve a confidence metric.

Using SpecKS to Infer the Polyploid Continuum With *Ks* Distributions From Transcriptomic Data From >200 Angiosperms

Recent work has suggested that the majority of polyploid lineages may be derived from allopolyploidy rather than autopolyploidy (Wang et al. 2019), although also see (Soltis et al. 2007; Barker et al. 2016). Thus, we were interested to test if the logistical regression model discussed previously, when applied to empirical *Ks* data from species across the plant tree of life, would suggest high or low diversity between ancient subgenomes. Thus we applied our low-versus-high N_e classifier to over 200 real *Ks* observations made public by (Li and Barker 2020), derived from transcriptomic data from the 1KP study (Leebens-Mack et al. 2019). We were able to classify 228 WGD events, resulting in 12, 38, and 178, low-N_e, medium-N_e and high-N_e determinations, respectively (Table 1, and should refer to the "Summary of 1KP categorization" results, Fig. 11).

Our results indicate that >94.7% of the WGD in the 1KP dataset are in the medium-N_e and high-N_e categories, with the remaining 5.3% in the low-N_e category. Given our simplified model relating N_e to the divergence patterns of ancestral orthologs, this suggests that >94.7% of the lineages analyzed had high levels of genetic diversity between the ancestral parental diploid progenitors (equivalent to two random draws from an effective population size of 5 million individuals or

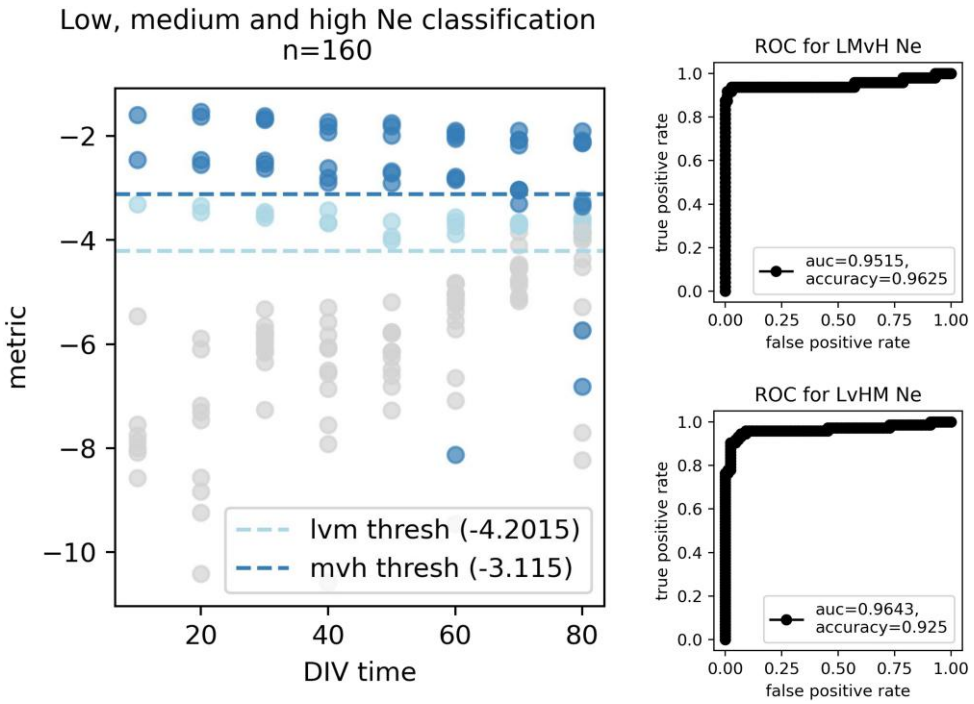

Fig. 10. The low-versus-high n_e discrimination model. (left) The low-versus-high diversity discrimination threshold applied to simulated polyploids. The x-axis gives the T_{DIV} time. The y-axis gives the value of the high-versus-low diversity discrimination metric. (right) The ROC plot to assess accuracy of inference into high versus low N_e, across threshold values. Darker colors denote higher N_e. Gray, light blue and dark blue correspond to K≤1, K≤5, K > 5. Each data point represents results for a given simulation from the set of simulations described in Table 2.

Table 2 SpecKS parameters used in Figs. 7, 9, and 10

Polyploid	**DIV_time_MYA**	**[80,70,60,50,40,30,20,10]**
Polyploid	**WGD_time_MYA**	**[0,5,10,20,50]**
Polyploid	**gene_div_time_distribution_parameters**	**"Impulse,1,1" and "expo,0,K" Where K varied from [0.01,0.1,1.0,5.0,10.0,20.0] To span (Gossmann et al. 2010)**
Species tree	**full_sim_time**	**100 MY**
Gene tree	**mean_gene_birth_rate_GpMY**	**0.001359 genes per MY**
Gene tree	**SSD_half_life_MY**	**4 MY**
Gene tree	**WGD_half_life_MY**	**31 MY**
Gene tree	**num_gene_trees_per_species_tree**	**3000**
Sequence evolution	**num_replicates_per_gene_tree**	**1**
Sequence evolution	**num_codons**	**1000**
Sequence evolution	**Ks_per_Myr**	**0.01**
Sequence evolution	**per_site_evolutionary_distance**	**0.01268182**

All other parameters are default.

more), which may suggest significant population structure emerging at the time of parental divergence, which might be concordant with allopolyploid origins. Of the remaining 5.3% of analyzed lineages, we see lower levels of parental diversity, suggesting more closely related progenitors. This might indicate allopolyploids at the low-N_e end of the continuum and/or autopolyploids. These results are concordant with (Wang et al. 2019), suggesting that the majority of ancient polyploidization events, which contributed to the genetic conduit may be derived from allopolyploidization.

Discussion

Here, we present the polyploid genome evolution simulator SpecKS and use it to demonstrate the dependency of the shape of the *Ks* histogram on critical polyploid speciation parameters.

Our simulations also show that the shape and skew of the *Ks* histogram is sensitive to the evolutionary history of the ancestral polyploidization event. We have shown that different levels of ancestral genetic divergence, as well as the relative timing of T_{WGD} and T_{DIV}, yield characteristically different *Ks* distributions. We also demonstrated that these differences persist for 10 s of millions of years (Fig. 5). Furthermore, we have shown that the skewness of the tail of "WGD peak" on the *Ks* histogram depends on the divergence between the ancestral subgenomes of the polyploid (Fig. 6) and we demonstrate with simulations that the *Ks* distribution can be used to

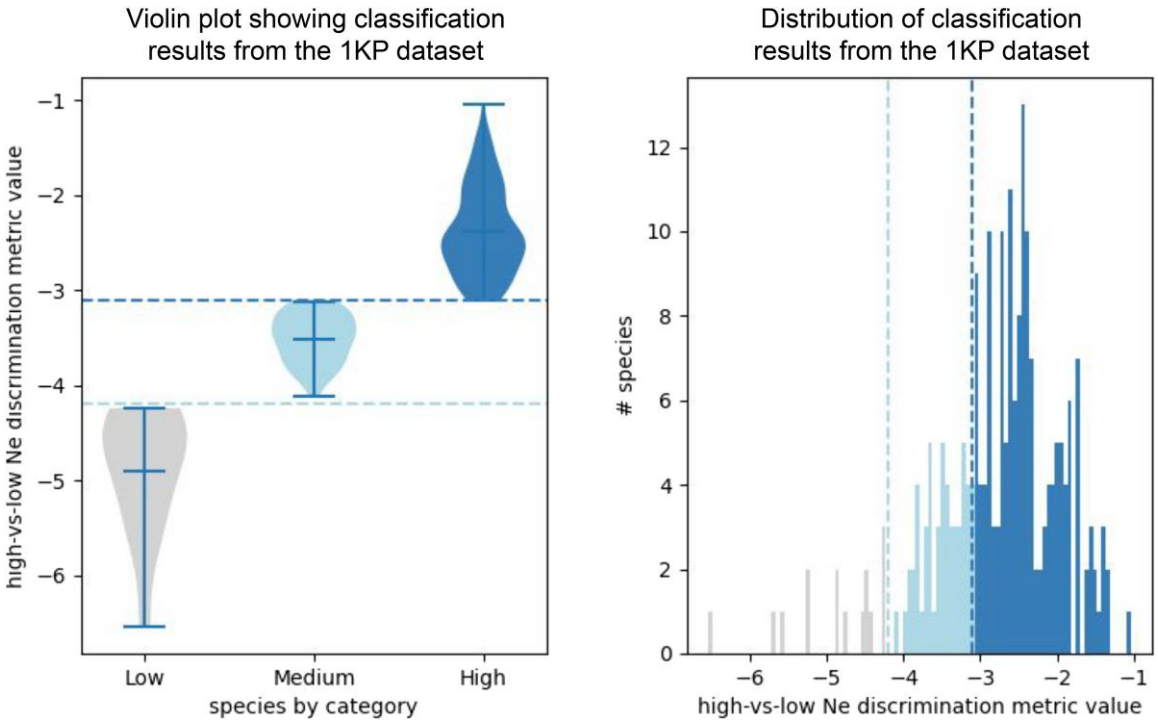

Fig. 11. The Ne-classifier results for the 1KP dataset. a) Violin plots indicating the number of species in each category. b) Histogram giving the number of species in each category. Darker blue color represents higher-N_e ancestral species at SPC time, respectively. X-axis gives the numerical value of the discrimination metric on a log scale. The discrimination metric is the skew of the Ks distribution for each species, measured as the difference between the distribution center of mass and mode. Y-axis gives the number of species per bin. Gray, light blue, and dark blue correspond to parameters $K \leq 1$, $K \leq 5$, $K > 5$. See Table 2 for the full set of parameters.

test inferences linking N_e, Π, T_{WGD}, T_{DIV}, and distributions of ohnolog divergence times (Figs. 9 and 10). Specifically, we demonstrate that the commonly used method for estimating T_{WGD} from the peak of the *Ks* histogram will be increasingly inaccurate for greater differences in time between T_{WGD} and T_{DIV} (Fig. 7), and we develop an alternative method that works equally well for a variety of polyploid speciation scenarios. We contend that the *Ks* distribution is therefore information rich, with the potential to aid in the estimation of a variety of polyploid speciation parameters and scenarios.

We additionally show that SpecKS-generated histograms realistically capture the main features of empirical *Ks* histograms. For instance, in Fig. 8, we parameterized SpecKS to fit the observed *Ks* histograms of three well-studied species, coffee, maize and poplar. In all cases, the histogram features corresponding to the SSDs and ohnolog peaks appear to be well replicated, with overall RMSE <2 in all cases (Fig. 8). This demonstrates that the parameters supplied may be reasonably estimable under the SpecKS model. Interestingly, we note that the main difference between the simulated histograms and the true histograms, is that the true histograms maintain very old paralogous pairs ($Ks > T_{DIV}$) whose numbers do not decay over time. In contrast, in our simulation, the number of paralogous pairs retained asymptotically approaches zero over time. The behavior of our simulation is expected under the SpecKS model, because both SSDs and ohnologs shed genes following a parameterized decay model (described in Methods), thus the number of ancient paralogs retained go to zero over time.

One explanation for the departure between the SpecKS results and the empirical observations is that the SpecKS parameterized decay model does not yet allow for the selective retention of advantageous duplicates. For example, paralogous pairs that persist in the true *Ks* histogram may be maintained due to selection or potentially under linked selection (hitchhiking) and thus not presently modeled by our system. These persistent paralogs are more apparent in *Coffea arabica* and *Zea mays* (Fig. 8). Since these species have undergone recent domestication (Gaut and Doebley 1997; Salojarvi et al. 2023; Yang et al. 2023), it may be possible that domestication has hindered efficient gene shedding or promoted paralog retention (Gaut et al. 2018). We also note that SpecKS currently only models ohnologs and SSDs; some paralogs of a different origin, like tandem duplications, may be retained for longer or shorter periods of time, and thus would be evident in the empirical histograms but not in the modeled histograms.

We have additionally demonstrated that simulated datasets derived from SpecKS may be used to train inference models to gain insight into the interpretation of empirical datasets. For example, we give a machine-learning model trained on a broad range of simulated allopolyploid speciation scenarios. We then apply the model to empirical *Ks* histograms to a diverse set of angiosperms. Using data from the 1KP dataset as input, we are able to infer that the majority of ancient polyploidization events cataloged

appear to be derived from parental diploids with a high level of diversity between them, and thus might be allopolyploid. These results are concordant with (Wang et al. 2019).

However, we note that both our method and that of Wang et al. (2019) may under-report autopolyploids. One concern applicable to both methods is that for autopolyploids, gene copies may not have diverged sufficiently to form true paralogs, thus the gene-pairs may not be detected as a multiplicity in transcriptomic analysis (Mayfield-Jones et al. 2013; Garsmeur et al. 2014). Furthermore, our logistical regression model was trained on simulations drawn from a continuum of allopolyploidy, thus may not be applicable to autopolyploids whose speciation parameters lie outside the scope of the SpecKS implementation. Inherent to our chosen input distribution (see Setting the gene-tree divergence time distribution "D_{TCOAL}" in the Methods section) is the assumption that N_e is related to the divergence time of orthologs via the Kingman coalescent (Kingman 1982). While much work exists describing the complexity of post-WGD divergence of ohnologs in specific autopolyploid systems, there currently exists no general, empirically validated model (Parisod et al. 2010; Spoelhof et al. 2017; Parey et al. 2022; Lallemand et al. 2023.

For some (largely angiosperm) autopolyploid lineages, it has been reported that diploidization and ortholog divergence may be rapid (Santos et al. 2003; Liu et al. 2017; Morgan et al. 2021; Gonzalo 2022; Bomblies 2023; Shi et al. 2023; Zhang et al. 2023) In this case, our 2D continuum model might be extensible to autopolyploids, which would occupy the low Ne, low ΔT portion of the continuum. However, for other ancient autopolyploids, for example salmonids and Acipenseriformes (Robertson et al. 2017; Gundappa et al. 2021; Parey et al. 2022) it seems that tetrasomic inheritance may be maintained for many millions of years, and ortholog divergence may be delayed, protracted or even saltational. In such cases, the pattern of ortholog divergence may be much more complex than predicted by the Kingman coalescent. While the derivation of a generalized model for autopolyploids is beyond the scope of this paper, the SpecKS architecture is easily extensible to more complex ortholog divergence patterns and may prove to be well-suited to the incorporation and assessment of a variety of divergence models. In summary, while SpecKS internal 2D continuum model of polyploidy is currently best suited to model varying degrees of allopolyploidy, we make no general claims regarding autopolyploidy.

We also note that many of the parameters incorporated into SpecKS are lineage specific (most significantly: rates of gene loss, the *Ks* distribution at the time of ancestral divergence, the rate of the molecular clock) and are difficult to ascertain a priori. Furthermore, the simplistic relationship between N_e and the gene tree coalescent distribution we selected as input (described in more detail in the Methods section) might not be appropriate for many use cases, including autopolyploids, or if complex population structure existed within the ancestral species. For these situations, we recommend the user consider the D_{TCOAL} most appropriate to their effort.

For simplicity, the discrimination models presented in our analysis were drawn from simulation runs where a small number of parameters were fixed (for example, fixed gene-birth-and-death rate and molecular clock). These models are presented as proof of concept and not meant to analyze any particular lineage. When making lineage-specific inferences, we caution users to parameterize SpecKS with lineage-appropriate values, or if they are unknown, to use SpecKS over a range of likely values, to model the impact of uncertainty in their value.

Lastly, the SpecKS pipeline closely follows the architecture put forward in Ks simulators by (Sutherland et al. 2024) and (Tiley et al. 2018). Like (Tiley et al. 2018) SpecKS uses EVOLVER's GY94 model of sequence evolution to accrue genetic distance between orthologs, while (Sutherland et al. 2024) generates branch lengths based on distributions fit to empirical observations. All three methods of *Ks* simulation show that in a simulated environment, the time of WGD can be accurately recovered from the *Ks* histogram. However, in prior works, polyploids are generalized such that T_{DIV} and T_{WGD} are equivalent, i.e. $\Delta T = 0$. Thus, the issue of potentially misplacing T_{WGD} by millions of years is unaddressed. In this paper, we show that the magnitude of the error scales linearly with ΔT. SpecKS resolves this issue by making T_{DIV} and T_{WGD} separate input parameters, and our accuracy assessments integrate across a range of ΔT. SpecKS also differs from previous works in that we allow diploid parental speciation parameters to affect the patterns of ortholog coalescence: the user may input an initial distribution based on theorized biological processes. In our simulations, we use the Kingman equation to derive the initial input pattern of ortholog coalescence, but this is not required. Because of these extra dimensions (the 2D continuum), SpecKS is uniquely able to emulate a rich variety of *Ks* histograms with greater fidelity to underlying biological processes.

However, despite its utility, SpecKS remains simplistic. SpecKS' model of genome evolution currently does not include selection, population dynamics, the effects of domestication, and other factors which would reasonably impact the *Ks* histogram. SpecKS' internal model of sequence evolution (GY94 with equal equilibrium codon frequencies) is also simplistic, and the rate of *Ks* accumulation is assumed to be constant over time. Furthermore, our ability to test SpecKS against observed *Ks* histograms is limited to tetraploids whose WGD events are well-separated in time. This is because SpecKS does not yet model multiple superimposed rounds of WGD. An appropriate next step would be to iteratively include more evolutionary complexity, thus making the simulation both more realistic and more testable.

Materials and Methods

Data

Transcriptomic Data

Ks histograms in Fig. 8 were generated from transcriptomic data for *Coffea arabic*a, *Zea mays* and *Populus trichocarpa*. Transcriptomic data were sourced from NCBI (NCBI

assembly GCF_003713225.1, GCF_902167145.1, and GCF_000002775.5), and the software package KsRates (Sensalari et al. 2022) was used to calculate the *Ks* histograms for each species.

1KP Data

The classification of WGD events from the 1KP datset were made by running the low-versus-high N_e discrimination model on histograms generated from *Ks* data from the 1KP dataset. The raw, empirical *Ks* data is available at (https://gitlab.com/barker-lab/1KP/-/tree/master/1KP_ksplots). These *Ks* data were generated by (Leebens-Mack et al. 2019; Li and Barker 2020). *Ks* histograms from this data were plotted using matplotlib (Hunter 2007), and the skew of the *Ks* histogram was measured using the same feature-extraction method used in the derivation of the N_e discrimination model, and then input directly to the N_e discrimination model.

Logistic Regression Models

An example use-case for SpecKS is the provision of simulated truth and training data for the derivation of inference models. Here, we describe two discrimination models used in this paper, one to infer N_e, and another to infer ΔT. For both of these models, we used the logistic regression package from scikit-learn (Pedregosa et al. 2011). SpecKS-simulated *Ks* histograms were used to derive features, and SpecKS input parameters were used to derive target variables, which we organized into three categories (low, medium, and high) for each classifier. With respect to target variable categories, for the ΔT discrimination model, the target variable was based on whether $\Delta T = T_{DIV}\_T_{WGD} < 5, < 30$, or ≥ 30 MY. For the N_e discrimination model, the target variable was based on whether $N_e < 5$, < 10, or ≥ 10 * 10^6. These levels were selected to span realistic values (Gossmann et al. 2010). With respect to feature data, for the ΔT discrimination model, we used the difference between estimated T_{WGD} and T_{DIV} as our single feature, deriving both from the *Ks* histogram directly, using algorithms for estimating T_{WGD} and T_{DIV} as discussed previously in the results section (Fig. 7 and supplementary fig. S2, Supplementary Material online). For the N_e discrimination model, we used the natural logarithm of the skew of the *Ks* histogram (measured as the difference between the x-value of the *Ks* peak and the x-value of the *Ks* center of mass, in *Ks*-space) as the single key feature. The models were trained with a ⅓ test, ⅔ training approach, splitting datasets as appropriate.

Setting the Gene-tree Divergence Time Distribution "D_{TCOAL}"

In the SpecKS simulator, the initial distribution of the divergence times for the gene trees is supplied by the user as an input parameter. This allows the user as much flexibility as possible with regard to modeling their system. Throughout this paper, a simplistic input distribution D_{TCOAL} was used, which relates ancestral genetic diversity (Π) to the distribution of ortholog divergence times, based on the assumption of some level of allopolyploidy. We assumed that the ancestral diploid population was in panmixia, and we used the Kingman coalescent (Kingman 1982) to derive a diversity of initial gene-tree divergence times.

Under this model, the ancestral diploid species exists with some standing genetic variation, which scales with population size, such that N_e (effective population size) is proportional Π. The ancestral species subsequently (at time T_{DIV}) diverges to give rise to two diploid sister species, which evolve forward for a given amount of time ($\Delta T = T_{DIV}\_T_{WGD}$), before conjoining to form the diploid lineage at T_{WGD}. Prior to T_{DIV}, the coalescence times for gene copies between two random individuals in the ancestral population can be approximated by the Kingman coalescent, given Π. If these two ancestral individuals found new species, their individual sets of gene copies are now separated by speciation and are redefined as orthologs. The diversity of coalescent times becomes the initial diversity in nodes for the bifurcating gene trees at T_{DIV}, which follows an exponential distribution, under the Kingman coalescent (Kingman 1982).

Under this simple one-genome-one-species model, for more diverse ancestral species, we see more initial diversity in node times, and a greater skew in the *Ks* distribution toward the past. Mathematically, this is because the decay constant in the Kingman coalescent is inversely proportional to N_e. Note that this model may not be appropriate for autopolyploids, which is beyond the scope of this paper.

In order to obtain the Kingman coalescent from N_e, it is necessary to assume a generation time (G_t). In all our simulations, for simplicity, we assumed at $G_t = 1$. Since the *Ks* distributions generated and the inferences made from them were done in *Ks* space, the exact G_t is immaterial. It only matters that the *Ks*-to-My conversion factor, which is a configurable input, properly factors in the G_t at the initiation of the simulation, and that this same conversion factor is taken into account by the end-user when relating the output *Ks* plots to chronological time.

SpecKS Implementation

SpecKS is implemented as a pipeline application in python. SpecKS takes as input an XML configuration file listing the simulation parameters (Table 3, supplementary S2 and S3, Supplementary Material online) and outputs a text file (.csv) of all pairwise *Ks* accumulated between all gene pairs (ohnologs and SSDs) for each simulated genome. Architecturally, SpecKS is designed as a pipeline with eight modules, which are executed sequentially for each polyploid in the simulation (Fig. 4). The modules functions are (i) species tree generation, (ii) gene tree generation, (iii) application of a gene birth and death model, (iv) application of a post WGD gene shedding model, (v) gene sequence evolution, (vi) the *Ks* calculation between gene pairs, (vii) *Ks* histogrammer, and (viii) final results assembly. We give details for each module below.

Species tree generation: A Newick-formatted species tree is generated for each polyploid. This is always in

Table 3 Subset of SpecKS configurable parameters

Parameter and default	Default	Description
Polyploid-specific. (Set for each simulated polyploid in the run):		
DIV_time_MYA	0 MY	DIV time in MY. Time of subgenome divergence. For gradual speciation, this will be the mode of the gene tree divergence times.
WGD_time_MYA	0 MY	WGD time in MY. This will be the start time of ohnolog shedding, which will continue until the present time.
Gene_div_time_ distribution_ parameters	impulse,1,1	Sets the distribution of divergence times for gene trees at DIV time. For instantaneous divergence, use format: "impulse,1,1". For divergence patterns derived from the Kingman Coalescent, use the format "expon,0,K", where K is the exponential decay constant. (We suggest K = Ne*Gt). For polyploids whose gene tree divergence might be a mix of distributions (i.e. segmental allopolyploids), multiple distributions may be given, with the last parameter being the proportion of genes, which belong in each distribution. Details in supplementary table S2, Supplementary Material online.
General. (The same for all polyploids in the run):		
Full_sim_time	100 MY	The length of the time period to simulate. Note that since speciation is a gradual process, it may be necessary to start the simulation well in advance of the SPC time, in order to fully capture the distribution of gene trees.
mean_gene_birth_rate_GpMY	0.001359 genes per MY	Gene birth rate. Note this may be lineage specific. Our default is chosen from (Guo 2013)
SSD_half_life_MY	4 MY	Half-life of SSDs. Our default is chosen from (Lynch and Conery 2003)
WGD_half_life_MY	31 MY	Half-life of ohnologs (WGDs). Our default is based on (Guo 2013 and Maere et al. 2005)
Num_gene_trees_per_species_tree	3000	Default is set to give a well-supported histogram without taking too long to run. If set too low, the final histogram will look too sparse. Higher numbers may be more realistic (Sterck et al. 2007) and result in smoother histograms but have a longer run time. Since gene trees are simulated independently, the number does not affect the general histogram shape, merely the number of samples in it, so high numbers are not always necessary.
Num_replicates_per_gene_tree	1	SpecKS can automatically run replicates for a given simulation, randomizing appropriately.
Num_codons	1000	The number of codons in each gene to be simulated. All genes in the sim have the same length. 1,000 was chosen in agreement with (Tiley 2018).
Ks_per_Myr	0.01	*Ks* per million years. This number is lineage and gene family specific, and may need to change depending on user needs (Gaut et al. 1996 and Koch et al. 2000). The default of 0.01 was chosen in agreement with (Tiley 2018), and in range with Blanc and Wolfe 2004.
Per_site_evolutionary_distance Default: 0.01268182	0.01268182	The per_site_evolutionary_distance is used to calculate the total tree length per gene tree input to evolver. The default setting was derived by (Tiley 2018) for a Ks_per_Myr of 0.01 and the evolutionary GY94 model.

the format (O:T_{SIM}, (P1:T_{DIV}, P2:T_{DIV}): T_{SIM}-T_{DIV}), where T_{SIM} specifies the full simulation time. P1 and P2 denote the subgenomes of the polyploid, while O denotes the outgroup.

Gene tree generation: Newick-formatted gene trees are generated from the simulated species tree. Random variations in gene divergence times are introduced according to the distribution specified in the configuration file, allowing for the introduction of a range of divergence times for the ohnologs (genes duplicated by WGD). The distribution is configurable and might in theory be selected based on the ancestral genetic diversity (Π), generation time, and evolutionary model. The number of gene trees is also configurable, with a default set at 3,000. The distribution selected for our simulations was based on the Kingman coalescent, and we give details in the methods section.

Gene birth and death model: The gene birth and death model introduces "SSDs" into the simulation. For each gene tree, genes are randomly born at a rate specified in the configuration file, modeled as a Poisson process (Zhao et al. 2015). Genes are assigned a death-date at birth, with a life span drawn randomly from an exponential decay distribution (Lynch and Conery 2000; Lynch et al. 2001). Gene birth rate and mean life expectancy are configurable (Lynch and Conery 2003; Guo 2013). Genes born, which will be dead before the simulation, concludes are pruned at this time for computational efficiency.

Ohnolog gene shedding model: Genes duplicated by WGD also have a mean life expectancy given by an exponential decay distribution (Ren et al. 2018). The proportion of WGD duplicates, which will be dead before the simulation concludes, are calculated based on WGD time and pruned for computational efficiency. For simplicity and traceability, ohnolog duplicates are always removed from the P1 branch. Ohnolog life expectancy is configurable and based on (Maere et al. 2005; Guo 2013).

Gene sequence evolution: For each simulated gene tree, we simulated codon sequences using the EVOLVER program within PAML v4.10.7 (Yang and Nielsen 2000: 200; Yang 2007). PAML was configured as in (Tiley et al. 2018) to simulate codon evolution along each finalized gene tree using a Goldman–Yang (GY94) model of codon evolution (Goldman and Yang 1994; Yang and Nielsen 1998) with equal equilibrium codon frequencies, a transition/transversion rate ratio of 2 and a global dN/dS of 0.2 as in (Tiley et al. 2018). The number of codons per alignment is set by default to 1,000 as in (Tiley et al. 2018), in line with reported plant transcript lengths (Luo et al. 2019; Zhang et al. 2020; Al-Dossary et al. 2023), but is configurable.

Ks calculation: *Ks* is calculated between all pairs of terminal genes in a given gene tree, across all gene trees, using the CODEML program within PAML v4.10.7. Codon frequencies, transition/transversion rate and *dN/dS* were set to match those used in EVOLVER (Yang and Nielsen 2000; Yang 2007). *Ks* per million years is 0.01 by default and is configurable (Blanc and Wolfe 2004; Tiley et al. 2018).

Ks histogrammer: SpecKS thereon generates two default histograms (one with a configurable maximum *Ks*, and one with no maximum *Ks*). The existence of these histograms is only meant to give a confirmation of the run success. It is expected that the end users will use their own visualization tools to generate more elegant histograms.

Final results assembly: The final results assembler collates the CODEML results, producing a single.csv file containing the *Ks* results for each polyploid. Additional products of the simulation include *Kn* results (histogram of accumulated nonsynonymous mutations between paralogs) for the polyploid, and *Ks* and *Kn* results for the outgroup.

Running SpecKS

SpecKS is run calling the "SpecKS.py" function from python and passing it the config xml file. For example, "python3 SpecKS.py myconfig.xml"

SpecKS Output

The SpecKS output file is a.csv file, with each line giving information for a unique gene-pair. The data are organized into four columns, giving the gene-tree name of the pair, the names of the two genes in the pair (the leaves of the tree), the *Ks* value, and the path to the CODEML output file from which the value originated, respectively.

Parallelization

For simplicity, SpecKS is not internally parallelized. A single polyploid of about 3,000 genes, 1,000 codons long, simulated over 100 MY, runs on a typical cluster (e.g. 16-node Dual-10 Xeon CPU, E5 −2630v4 2.20 GHz, 12.8GB RAM per core) in under 10 min. To simulate batches of polyploids, we recommend submitting each polyploid as a separate batch job to a job scheduler such as SGE (Oracle) or SLURM (SchedMD).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank members of the Barker and Sethuraman Labs for valuable inputs during the conceptualization stage of this project. We would like to acknowledge Matthew Hahn, Robert Williamson, and four anonymous reviewers for valuable suggestions on the first version of this manuscript.

Funding

This work was supported by NSF CAREER: 214812 and USDA-HSI:2022-77040-38529 to A.S. T.D. was supported by a CSUBIOTECH grant. All computations were performed on the *mesxuuyan* HPC at San Diego State University, which was supported by NSF ABI: 1564659 and startup funds to A.S.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

SpecKS is available from the Github repository https://github.com/tamsen/SpecKS. SpecKS v1.5.0.0 was used in this analysis, and input files used to drive the simulation and the scripts used to generate the figures are at https://github.com/tamsen/SpecKS_paper_scripts.

References

Ai B, Wang Z-S, Ge S. Genome size is not correlated with effective population size in the oryza species. Evolution. 2012:**66**(10):3302–3310. https://doi.org/10.1111/j.1558-5646.2012.01674.x.

Al-Dossary O, Furtado A, KharabianMasouleh A, Alsubaie B, Al-Mssallem I, Henry RJ. Long read sequencing to reveal the full complexity of a plant transcriptome by targeting both standard and long workflows. Plant Methods. 2023:**19**(1):112. https://doi.org/10.1186/s13007-023-01091-1.

Arrigo N, Barker MS. Rarely successful polyploids and their legacy in plant genomes. Curr Opin Plant Biol. 2012:**15**(2):140–146. https://doi.org/10.1016/j.pbi.2012.03.010.

Baduel P, Bray S, Vallejo-Marin M, Kolář F, Yant L. The "polyploid hop": shifting challenges and opportunities over the evolutionary lifespan of genome duplications. Front Ecol Evol. 2018:**6**:117. https://doi.org/10.3389/fevo.2018.00117.

Baniaga AE, Marx HE, Arrigo N, Barker MS. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. Ecol Lett. 2020:**23**(1):68–78. https://doi.org/10.1111/ele.13402.

Barba-Montoya J, Dos Reis M, Schneider H, Donoghue PCJ, Yang Z. Constraining uncertainty in the timescale of angiosperm

evolution and the veracity of a cretaceous terrestrial revolution. New Phytol. 2018:**218**(2):819–834. https://doi.org/10.1111/nph.15011.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. On the relative abundance of autopolyploids and allopolyploids. New Phytol. 2016:**210**(2):391–398. https://doi.org/10.1111/nph.13698.

Blanc G, Wolfe KH. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 2004:**16**(7):1667. https://doi.org/10.1105/tpc.021345.

Blischak PD, Sajan M, Barker MS, Gutenkunst RN. Demographic history inference and the polyploid continuum. Genetics. 2023:**224**(4):iyad107. https://doi.org/10.1093/genetics/iyad107.

Bomblies K. Learning to tango with four (or more): the molecular basis of adaptation to polyploid meiosis. Plant Reprod. 2023:**36**(1):107–124. https://doi.org/10.1007/s00497-022-00448-1.

Bomblies K, Jones G, Franklin C, Zickler D, Kleckner N. The challenge of evolving stable polyploidy: could an increase in "crossover interference distance" play a central role? Chromosoma. 2016:**125**(2):287–300. https://doi.org/10.1007/s00412-015-0571-4.

Booker WW, Schrider DR. The genetic consequences of range expansion and its influence on diploidization in polyploids. Am Nat. 2024. https://doi.org/10.1086/733334.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019:**15**(4):e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 2003:**422**(6930):433–438. https://doi.org/10.1038/nature01521.

Buggs RJA, Zhang L, Miles N, Tate JA, Gao L, Wei W, Schnable PS, Barbazuk WB, Soltis PS, Soltis DE. Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol. 2011:**21**(7):551–556. https://doi.org/10.1016/j.cub.2011.02.016.

Carretero-Paulet L, Van de Peer Y. The evolutionary conundrum of whole-genome duplication. Am J Bot. 2020:**107**(8):1101–1105. https://doi.org/10.1002/ajb2.1520.

Chen H, Zwaenepoel A. Inference of ancient polyploidy from genomic data. Methods Mol Biol. 2023:**2545**:3–18. https://doi.org/10.1007/978-1-0716-2561-3_1.

Clark JW, Donoghue PC. Whole-genome duplication and plant macroevolution. Trends Plant Sci. 2018:**23**(10):933–945. https://doi.org/10.1016/j.tplants.2018.07.006.

Clark JW, Puttick MN, Donoghue PCJ. Origin of horsetails and the role of whole-genome duplication in plant macroevolution. Proc Biol Sci. 2019:**286**(1914):20191662. https://doi.org/10.1098/rspb.2019.1662.

Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B. Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. Plant Cell. 2000:**12**(9):1551. https://doi.org/10.1105/tpc.12.9.1551.

Conant GC. POInT: modeling polyploidy in the era of ubiquitous genomics. Methods Mol Biol. 2023:**2545**:77–90. https://doi.org/10.1007/978-1-0716-2561-3_4.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al. Widespread genome duplications throughout the history of flowering plants. Genome Res. 2006:**16**(6):738–749. https://doi.org/10.1101/gr.4825606.

Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, Milne R, Chen Y, Wan Z, Wang Z, et al. The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res. 2014:**24**(10):1274–1277. https://doi.org/10.1038/cr.2014.83.

David KT, Oaks JR, Halanych KM. Patterns of gene evolution following duplications and speciations in vertebrates. PeerJ. 2020:**8**:e8813. https://doi.org/10.7717/peerj.8813.

De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. Mol Biol Evol. 2017:**34**(6):1363–1377. https://doi.org/10.1093/molbev/msx069.

De Storme N, Mason A. Plant speciation through chromosome instability and ploidy change: cellular mechanisms, molecular factors and evolutionary relevance. Curr Plant Biol. 2014:**1**:10–33. https://doi.org/10.1016/j.cpb.2014.09.002.

Dodsworth S, Chase MW, Leitch AR. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? Bot J Linn Soc. 2016:**180**(1):1–5. https://doi.org/10.1111/boj.12357.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Ågren JA, Hazzouri KM, Wang W. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. Proc Natl Acad Sci U S A. 2015:**112**(9):2806–2811. https://doi.org/10.1073/pnas.1412277112.

Doyle JJ, Egan AN. Dating the origins of polyploidy events. New Phytol. 2010:**186**(1):73–85. https://doi.org/10.1111/j.1469-8137.2009.03118.x.

Doyle JJ, Sherman-Broyles S. Double trouble: taxonomy and definitions of polyploidy. New Phytol. 2017:**213**(2):487–493. https://doi.org/10.1111/nph.14276.

Estep MC, McKain MR, Diaz DV, Zhong J, Hodge JG, Hodkinson TR, Layton DJ, Malcomber ST, Pasquet R, Kellogg EA. Allopolyploidy, diversification, and the Miocene grassland expansion. Proc Natl Acad Sci U S A. 2014:**111**(42):15149–15154. https://doi.org/10.1073/pnas.1404177111.

Fawcett JA, Maere S, Van de Peer Y. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. Proc Natl Acad Sci U S A. 2009:**106**(14):5737–5742. https://doi.org/10.1073/pnas.0900906106.

Gaeta RT, Pires CJ. Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytol. 2010:**186**(1):18–28. https://doi.org/10.1111/j.1469-8137.2009.03089.x.

Gargiulo R, Decroocq V, González-Martínez SC, Paz-Vinas I, Aury J-M, Lesur Kupin I, Plomion C, Schmitt S, Scotti I, Heuertz M. Estimation of contemporary effective population size in plant populations: limitations of genomic datasets. Evol Appl. 2024:**17**(5):e13691. https://doi.org/10.1111/eva.13691.

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol. 2014:**31**:448–454. https://doi.org/10.1093/molbev/mst188.

Gaut BS. Evolutionary dynamics of grass genomes. New Phytol. 2002:**154**(1):15–28. https://doi.org/10.1046/j.1469-8137.2002.00352.x.

Gaut BS, Doebley JF. DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci U S A. 1997:**94**:6809–6814. https://doi.org/10.1073/pnas.94.13.6809.

Gaut BS, Morton BR, Mccaig BC, Clegg MT. Substitution rate comparisons 839 between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel 840 rate differences at the plastid gene rbcL. Proc Natl Acad Sci U S A. 1996:**93**:842. https://doi.org/10.1073/pnas.93.19.10274.

Gaut BS, Seymour DK, Liu Q, Zhou Y. Demography and its effects on genomic variation in crop domestication. Nat Plants. 2018:**4**(8):512–520. https://doi.org/10.1038/s41477-018-0210-1.

Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994:**11**(5):725–736. https://doi.org/10.1093/oxfordjournals.molbev.a040153.

Gonzalo A. All ways lead to Rome—meiotic stabilization can take many routes in nascent polyploid plants. Genes (Basel). 2022:**13**(1):147. https://doi.org/10.3390/genes13010147.

Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol. 2010:**27**(8):1822–1832. https://doi.org/10.1093/molbev/msq079.

Gundappa MK, To T-H, Grønvold L, Martin S, Lien S, Geist J, Hazlerigg D, Sandve S, Macqueen D. Genome-Wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. Mol Biol Evol. 2021:**39**(1):msab310. https://doi.org/10.1093/molbev/msab310.

Guo X, Mandáková T, Trachtová K, Özüdoğru B, Liu J, Lysak MA. Linked by ancestral bonds: multiple whole-genome duplications and reticulate evolution in a Brassicaceae tribe. Mol Biol Evol. 2020:**38**(5):1695–1714. https://doi.org/10.1093/molbev/msaa327.

Guo Y-L. Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. Plant J. 2013:**73**(6):941–951. https://doi.org/10.1111/tpj.12089.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009:**5**(10):e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Hampson S, McLysaght A, Gaut B, Baldi P. Lineup: statistical detection of chromosomal homology with application to plant comparative genomics. Genome Res. 2003:**13**(5):999–1010. https://doi.org/10.1101/gr.814403.

Hoegg S, Brinkmann H, Taylor JS, Meyer A. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J Mol Evol. 2004:**59**:190–203. https://doi.org/10.1007/s00239-004-2613-z.

Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007:**9**(3):90–95. https://doi.org/10.1109/MCSE.2007.55.

Kingman JFC. The coalescent. Stoch Process Their Appl. 1982:**13**(3):235–248. https://doi.org/10.1016/0304-4149(82)90011-4.

Koch MA, Haubold B, Mitchell-Olds T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol. 2000:**17**:1483–1498. https://doi.org/10.1093/oxfordjournals.molbev.a026248.

Lallemand T, Leduc M, Desmazières A, Aubourg S, Rizzon C, Landès C, Celton J-M. Insights into the evolution of ohnologous sequences and their epigenetic marks post-WGD in *Malus domestica*. Genome Biol Evol. 2023:**15**:evad178. https://doi.org/10.1093/gbe/evad178.

Laurent S, Salamin N, Robinson-Rechavi M. No evidence for the radiation time lag model after whole genome duplications in Teleostei. PLoS One. 2017:**12**:e0176384. https://doi.org/10.1371/journal.pone.0176384.

Le Comber SC, Ainouche ML, Kovarik A, Leitch AR. Making a functional diploid: from polysomic to disomic inheritance. New Phytol. 2010:**186**(1):113–122. https://doi.org/10.1111/j.1469-8137.2009.03117.x.

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019:**574**(7780):679–685. https://doi.org/10.1038/s41586-019-1693-2.

Levin DA. The timetable for allopolyploidy in flowering plants. Ann Bot. 2013:**112**(7):1201–1208. https://doi.org/10.1093/aob/mct194.

Levin DA. Has the polyploid wave ebbed? Front Plant Sci. 2020:**11**:251–251. https://doi.org/10.3389/fpls.2020.00251.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. Early genome duplications in conifers and other seed plants. Sci Adv. 2015:**1**(10):e1501084. https://doi.org/10.1126/sciadv.1501084.

Li Z, Barker MS. Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. GigaScience. 2020:**9**(2):giaa004. https://doi.org/10.1093/gigascience/giaa004.

Li Z, McKibben MTW, Finch GS, Blischak PD, Sutherland BL, Barker MS. Patterns and processes of diploidization in land plants. Annu Rev Plant Biol. 2021:**72**(1):387–410. https://doi.org/10.1146/annurev-arplant-050718-100344.

Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc Natl Acad Sci U S A. 2018:**115**(18):4713–4718. https://doi.org/10.1073/pnas.1710791115.

Liu S, Yang Y, Wei F, Duan J, Braynen J, Tian B, Cao G, Shi G, Yuan J. Autopolyploidy leads to rapid genomic changes in *Arabidopsis thaliana*. Theory Biosci. 2017:**136**(3-4):199–206. https://doi.org/10.1007/s12064-017-0252-3.

Lohaus R, Van de Peer Y. Of dups and dinos: evolution at the K/Pg boundary. Curr Opin Plant Biol. 2016:**30**:62–69. https://doi.org/10.1016/j.pbi.2016.01.006.

Lott M, Spillner A, Huber KT, Moulton V. PADRE: a package for analyzing and displaying reticulate evolution. Bioinformatics. 2009:**25**(9):1199–1200. https://doi.org/10.1093/bioinformatics/btp133.

Luo D, Zhou Q, Wu Y, Chai X, Liu W, Wang Y, Yang Q, Wang Z, Liu Z. Full-length transcript sequencing and comparative transcriptomic analysis to evaluate the contribution of osmotic and ionic stress components towards salinity tolerance in the roots of cultivated alfalfa (*Medicago sativa* L.). BMC Plant Biol. 2019:**19**(1):32. https://doi.org/10.1186/s12870-019-1630-4.

Lv Z, Addo Nyarko C, Ramtekey V, Behn H, Mason AS. Defining autopolyploidy: cytology, genetics, and taxonomy. Am J Bot. 2024:**111**(8):e16292. https://doi.org/10.1002/ajb2.16292.

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000:**290**(5494):1151–1155. https://doi.org/10.1126/science.290.5494.1151.

Lynch M, Conery JS. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 2003:**3**(1/4):35–44. https://doi.org/10.1023/A:1022696612931.

Lynch M, O'Hely M, Walsh B, Force A. The probability of preservation of a newly arisen gene duplicate. Genetics. 2001:**159**(4):1789–1804. https://doi.org/10.1093/genetics/159.4.1789.

Mable BK, Brysting AK, Jørgensen MH, Carbonell Akz, Kiefer C, Ruiz-Duarte P, Lagesen K, Koch MA. Adding complexity to complexity: gene family evolution in polyploids. Front Ecol Evol. 2018:**6**:114. https://doi.org/10.3389/fevo.2018.00114.

Madlung A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity (Edinb). 2013:**110**(2):99–104. https://doi.org/10.1038/hdy.2012.79.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 2005:**102**(15):5454–5459. https://doi.org/10.1073/pnas.0501102102.

Mason AS, Wendel JF. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. Front Genet. 2020:**11**:1014. https://doi.org/10.3389/fgene.2020.01014.

Mayer VW, Aguilera A. High levels of chromosome instability in polyploids of Saccharomyces cerevisiae. Mutat Res. 1990:**231**(2):177–186. https://doi.org/10.1016/0027-5107(90)90024-X.

Mayfield-Jones D, Washburn JD, Arias T, Edger PP, Pires JC, Conant GC. Watching the grin fade: tracing the effects of polyploidy on different evolutionary time scales. Semin Cell Dev Biol. 2013:**24**:320–331. https://doi.org/10.1016/j.semcdb.2013.02.002.

Mccann J, Jang T-S, Macas J, Schneeweiss GM, Matzke NJ, Novák P, Stuessy TF, Villaseñor JL, Weiss-Schneeweiss H. Dating the species network: allopolyploidy and repetitive DNA evolution in American daisies (*Melampodium sect. Melampodium*, Asteraceae). Syst Biol. 2018:**67**(6):1010–1024. https://doi.org/10.1093/sysbio/syy024.

McClintock B. A cytological and genetical study of triploid maize. Genetics. 1929:**14**(2):180. https://doi.org/10.1093/genetics/14.2.180.

Meirmans P, Van Tienderen P. The effects of inheritance in tetraploids on genetic diversity and population divergence. Heredity (Edinb). 2013:**110**(2):131–137. https://doi.org/10.1038/hdy.2012.80.

Morgan C, White MA, Franklin FCH, Zickler D, Kleckner N, Bomblies K. Evolution of crossover interference enables stable autopolyploidy by ensuring pairwise partner connections in *Arabidopsis arenosa*. Curr Biol. 2021:**31**(21):4713–4726.e4. https://doi.org/10.1016/j.cub.2021.08.028.

Nieto Feliner G, Casacuberta J, Wendel JF. Genomics of evolutionary novelty in hybrids and polyploids. Front Genet. 2020:**11**:792. https://doi.org/10.3389/fgene.2020.00792.

Ohno S. Evolution by gene duplication. Berlin, New York: Springer-Verlag; 1970.

Otto SP, Whitton J. Polyploid incidence and evolution. Annu Rev Genet. 2000:**34**(1):401–437. https://doi.org/10.1146/annurev.genet.34.1.401.

Parey E, Louis A, Cabau C, Guiguen Y, Roest Crollius H, Berthelot C. Synteny-guided resolution of gene trees clarifies the functional impact of whole-genome duplications. Mol Biol Evol. 2020:**37**(11):3324–3337. https://doi.org/10.1093/molbev/msaa149.

Parey E, Louis A, Montfort J, Guiguen Y, Crollius HR, Berthelot C. An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication. Genome Res. 2022:**32**(9):1685–1697. https://doi.org/10.1101/gr.276953.122.

Parisod C, Holderegger R, Brochmann C. Evolutionary consequences of autopolyploidy. New Phytol. 2010:**186**(1):5–17. https://doi.org/10.1111/j.1469-8137.2009.03142.x.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in python. J Mach Learn Res. 2011:**12**:2825–2830. https://doi.org/10.48550/arXiv.1201.0490.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol. 2019:**20**(1):38. https://doi.org/10.1186/s13059-019-1650-2.

Ramsey J, Schemske DW. Neopolyploidy in flowering plants. Annu Rev Ecol Syst. 2002:**33**(1):589–639. https://doi.org/10.1146/annurev.ecolsys.33.010802.150437.

Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, Ma H, Qi J. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol Plant.2018:**11**(3):414–428. https://doi.org/10.1016/j.molp.2018.01.002.

Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, Martin SAM, Holland PWH, Sandve SR, Macqueen DJ. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. 2017:**18**(1):111. https://doi.org/10.1186/s13059-017-1241-z.

Rothfels CJ, Johnson AK, Hovenkamp PH, Swofford DL, Roskam HC, Fraser-Jenkins CR, Windham MD, Pryer KM. Natural hybridization between genera that diverged from each other approximately 60 million years ago. Am Nat. 2015:**185**(3):433–442. https://doi.org/10.1086/679662.

Roux C, Pannell JR. Inferring the mode of origin of polyploid species from next-generation sequence data. Mol Ecol. 2015:**24**(5):1047–1059. https://doi.org/10.1111/mec.13078.

Roux J, Liu J, Robinson-Rechavi M. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. Mol Biol Evol. 2017:**34**(11):2773–2791. https://doi.org/10.1093/molbev/msx199.

Salojärvi J, Rambani A, Yu Z, Guyot R, Strickler S, Lepelley M, Wang C, Rajaraman S, Rastas P, Zheng C, et al. The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. Nat Genet.2024:**56**(4):721–731. https://doi.org/10.1038/s41588-024-01695-w.

Santos JL, Alfaro D, Sanchez-Moran E, Armstrong SJ, Franklin FCH, Jones GH. Partial diploidization of meiosis in autotetraploid *Arabidopsis thaliana*. Genetics. 2003:**165**(3):1533–1540. https://doi.org/10.1093/genetics/165.3.1533.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. Mining EST databases to resolve evolutionary events in major crop species. Genome. 2004:**47**(5):868–876. https://doi.org/10.1139/g04-047.

Schranz E, Mohammadin S, Edger PP. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. Curr Opin Plant Biol. 2012:**15**(2):147–153. https://doi.org/10.1016/j.pbi.2012.03.011.

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF. Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol Biol Evol. 2003:**20**(4):633–643. https://doi.org/10.1093/molbev/msg065.

Sensalari C, Maere S, Lohaus R. Ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. Bioinformatics. 2022:**38**(2):530–532. https://doi.org/10.1093/bioinformatics/btab602.

Shi Q, Guo X, Su H, Zhang Y, Hu Z, Zhang J, Han F. Autoploid origin and rapid diploidization of the tetraploid *Thinopyrum elongatum* revealed by genome differentiation and chromosome pairing in meiosis. Plant J. 2023:**113**(3):536–545. https://doi.org/10.1111/tpj.16066.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. Genome-Wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant Species with a large effective population size. Mol Biol Evol. 2010:**27**(8):1813–1821. https://doi.org/10.1093/molbev/msq062.

Soltis DE, Burleigh JG. Surviving the KT mass extinction: new perspectives of polyploidization in angiosperms. Proc Natl Acad Sci. 2009:**106**:5455–5456.

Soltis DE, Soltis PS, Schemske DW, Hancock JF, Thompson JN, Husband BC, Judd WS. Autopolyploidy in angiosperms: have we grossly underestimated the number of species? Taxon. 2007:**56**:13–30. https://doi.org/10.2307/25065732.

Soltis DE, Visger CJ, Marchant DB, Soltis PS. Polyploidy: pitfalls and paths to a paradigm. Am J Bot. 2016:**103**(7):1146–1166. https://doi.org/10.3732/ajb.1500501.

Soltis PS, Liu X, Marchant DB, Visger CJ, Soltis DE. Polyploidy and novelty: gottlieb's legacy. Philos Trans R Soc Lond B Biol Sci. 2014:**369**(1648):20130351. https://doi.org/10.1098/rstb.2013.0351.

Soltis PS, Soltis DE. Polyploidy and genome evolution. Berlin (Heidelberg): Springer; 2012. https://doi.org/10.1007/978-3-642-31442-1.

Spoelhof JP, Soltis PS, Soltis DE. Pure polyploidy: closing the gaps in autopolyploid research. J Syst Evol. 2017:**55**:340–352. https://doi.org/10.1111/jse.12253.

Stebbins GL. Variation and evolution in plants. New York, Chichester (West Sussex): Columbia University Press; 1950. p. 314. https://doi.org/10.7312/steb94536.

Stebbins GL. Cataclysmic evolution. Sci Am. 1951:**184**(4):54–59. https://doi.org/10.1038/scientificamerican0451-54.

Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van De Peer Y. How many genes are there in plants (... and why are they there)? Curr Opin Plant Biol. 2007:**10**(2):199–203. https://doi.org/10.1016/j.pbi.2007.01.004.

St Onge KR, Foxe JP, Li J, Li H, Holm K, Corcoran P, Slotte T, Lascoux M, Wright SI. Coalescent-based analysis distinguishes between allo- and autopolyploid origin in Shepherd's Purse (*Capsella bursa-pastoris*). Mol Biol Evol. 2012:**29**(7):1721–1733. https://doi.org/10.1093/molbev/mss024.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol. 2011:**28**(5):1569–1580. https://doi.org/10.1093/molbev/msq270.

Sutherland BL, Tiley GP, Li Z, McKibben MT, Barker MS. SLEDGe: Inference of ancient whole genome duplications using machine learning. bioRxiv 2024.01.17.574559. https://doi.org/10.1101/2024.01.17.574559, 2024, preprint: not peer reviewed.

Szoveny P, Terracciano S, Ricca M, Giordano S, Shaw AJ. Recent divergence, intercontinental dispersal and shared polymorphism

are shaping the genetic structure of amphi-atlantic peatmoss populations. Mol Ecol. 2008:**17**(24):5364–5377. https://doi.org/10.1111/j.1365-294X.2008.04003.x.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytol. 2015:**207**(2):454–467. https://doi.org/10.1111/nph.13491.

Thomas GWC, Ather SH, Hahn MW. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst Biol. 2017:**66**(6):1007–1018. https://doi.org/10.1093/sysbio/syx044.

Tiley GP, Barker MS, Burleigh JG. Assessing the performance of Ks plots for detecting ancient whole genome duplications. Genome Biol Evol. 2018:**10**(11):2882–2898. https://doi.org/10.1093/gbe/evy200.

Van de Peer Y. Computational approaches to unveiling ancient genome duplications. Nat Rev Genet. 2004:**5**(10):752–763. https://doi.org/10.1038/nrg1449.

Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. Polyploidy: an evolutionary and ecological force in stressful times. Plant Cell. 2021:**33**(1):11–26. https://doi.org/10.1093/plcell/koaa015.

Vandepoele K, Saeys Y, Simillion C, Raes J, Van De Peer Y. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. Genome Res. 2002:**12**(11):1792–1801. https://doi.org/10.1101/gr.400202.

Vanneste K, Van de Peer Y, Maere S. Inference of genome duplications from age distributions revisited. Mol Biol Evol. 2013:**30**(1):177–190. https://doi.org/10.1093/molbev/mss214.

Wang J, Qin J, Sun P, Ma X, Yu J, Li Y, Sun S, Lei T, Meng F, Wei C. Polyploidy index and its implications for the evolution of polyploids. Front Genet. 2019:**10**:807. https://doi.org/10.3389/fgene.2019.00807.

Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J. Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. BMC Bioinformatics. 2006:**7**(1):447. https://doi.org/10.1186/1471-2105-7-447.

Wen D, Yu Y, Zhu J, Nakhleh L. Inferring phylogenetic networks using phyloNet. Syst Biol. 2018:**67**(4):735–740. https://doi.org/10.1093/sysbio/syy015.

Wendel JF. The wondrous cycles of polyploidy in plants. Am J Bot. 2015:**102**(11):1753–1756. https://doi.org/10.3732/ajb.1500320.

Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016:**17**(1):37. https://doi.org/10.1186/s13059-016-0908-1.

Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 1987:**84**(24):9054–9058. https://doi.org/10.1073/pnas.84.24.9054.

Yan Z, Cao Z, Liu Y, Ogilvie HA, Nakhleh L. Maximum parsimony inference of phylogenetic networks in the presence of polyploid complexes. Syst Biol. 2022:**71**(3):706–720. https://doi.org/10.1093/sysbio/syab081.

Yang N, Wang Y, Liu X, Jin M, Vallebueno-Estrada M, Calfee E, Chen L, Dilkes BP, Gui S, Fan X, et al. Two teosintes made modern maize. Science. 2023:**382**(6674):eadg8940. https://doi.org/10.1126/science.adg8940.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007:**24**(8):1586–1591. https://doi.org/10.1093/molbev/msm088.

Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998:**46**(4):409–418. https://doi.org/10.1007/PL00006320.

Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000:**17**(1):32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236.

Yant L, Bomblies K. Genome management and mismanagement–cell-level opportunities and challenges of whole-genome duplication. Genes Dev. 2015:**29**(23):2405–2419. https://doi.org/10.1101/gad.271072.115.

Yu Q, Guyot R, de Kochko A, Byers A, Navajas-Pérez R, Langston BJ, Dubreuil-Tranchant C, Paterson AH, Poncet V, Nagai C, et al. Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (coffea). Plant J. 2011:**67**(2):305–317. https://doi.org/10.1111/j.1365-313X.2011.04590.x.

Zeng X, Yuan Z, Tong X, Li Q, Gao W, Qin M, Liu Z. Phylogenetic study of oryzoideae species and related taxa of the Poaceae based on atpB-rbcL and ndhF DNA sequences. Mol Biol Rep. 2012:**39**(5):5737–5744. https://doi.org/10.1007/s11033-011-1383-0.

Zhang H, Gou X, Zhang A, Wang X, Zhao N, Dong Y, Li L, Liu B. Transcriptome shock invokes disruption of parental expression-conserved genes in tetraploid wheat. Sci Rep. 2016:**6**(1):26363. https://doi.org/10.1038/srep26363.

Zhang S, Li R, Zhang L, Chen S, Xie M, Yang L, Xia Y, Foyer CH, Zhao Z, Lam H-M. New insights into *Arabidopsis* transcriptome complexity revealed by direct sequencing of native RNAs. Nucleic Acids Res. 2020:**48**(14):7700–7711. https://doi.org/10.1093/nar/gkaa588.

Zhang X, Meng Z, Han J, Khurshid H, Esh A, Hasterok R, Wang K. Characterization of meiotic chromosome behavior in the autopolyploid *Saccharum spontaneum* reveals preferential chromosome pairing without distinct DNA sequence variation. The Crop Journal. 2023:**11**(5):1550–1558. https://doi.org/10.1016/j.cj.2023.02.008.

Zhao J, Teufel AI, Liberles DA, Liu L. A generalized birth and death process for modeling the fates of gene duplication. BMC Evol Biol. 2015:**15**(1):275. https://doi.org/10.1186/s12862-015-0539-2.

Zwaenepoel A, Van de Peer Y. WGD—simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics. 2019:**35**(12):2153–2155. doi:10.1093/bioinformatics/bty915.