Accurately Inferring Ancient Auto and Allopolyploidization Events using Forward-time Simulations

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

[1] Joint Doctoral Program in Evolutionary Biology, San Diego State University and University of California Riverside

[2] Department of Biology, San Diego State University

[*]Corresponding author: Arun Sethuraman

Email: asethuraman@sdsu.edu

Author Contributions: TD developed the *SpecKS* method, performed all simulations, and analyses. TD and AS conceptualized the inference framework, and wrote the manuscript.

Competing Interest Statement: The authors do not have any conflicts of interest to declare.

Classification: Biological Sciences; Evolution

Keywords: Polyploidy; Speciation; Whole Genome Duplication; Angiosperms

Abstract

Multiple rounds of whole genome duplication (WGD) followed by re-diploidization have occurred throughout the evolutionary history of angiosperms. To understand why these cycles occur, much work has been done to model the genomic consequences and evolutionary significance of WGD. Since the machinations of diploidization are strongly influenced by the mode of speciation (whether a lineage was derived from ancient allo or autopolyploid), methods which can classify ancient whole genome duplication events as allo or auto are of great importance. Here we present a forward-time polyploid genome evolution simulator called *SpecKS*. Using extensive simulations, we demonstrate that allo and autopolyploid-derived species exhibit differently shaped Ks histograms. We also demonstrate sensitivity of the Ks histogram to the effective population size (Ne) of the ancestral species. Our findings indicate that 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 results to derive tests that reveal whether a genome is descended from allo or autopolyploidy, and whether the ancestral species had a high or low Ne. We apply this test to transcriptomic data for over 200 species across the plant kingdom, validating the theory that the majority of angiosperm lineages are derived from allopolyploidization events.

Introduction

Multiple rounds of Whole Genome Duplication (**WGD**) followed by re-diploidization have occurred throughout the evolutionary history of angiosperms (1–3). WGD however

presents a great paradox in plant evolutionary biology. On one hand, WGD is considered a major speciation mechanism (4–7). On the other hand, WGD presents a massive "macromutation", throwing a wrench into sexual reproduction, releasing transposons, and unbalancing molecular signaling pathways (8–18). One way to reconcile this paradox is to consider that the hypothetical advantage may not be the multiplicity of genomic material itself, but the period of diploidization that follows (19–30).

The diploidization process depends on whether an ancient WGD event is due to allopolyploidy (the multiplicity of genomes originating from two different species) or due to autopolyploidy (the multiplicity of genomes originating from within the same species). For example, allopolyploids are only capable of recombination between homeologous chromosomes from the same parental lineage, while autopolyploids are capable of recombination across subgenomes (31, 32). Furthermore, the existence of intermediate types of polyploids suggest that polyploidy is best modeled as a continuum (31, 33–35).

The estimation of the relative proportion of speciation by allopolyploidy or autopolyploidy across the plant tree of life has been much debated (36–38). Recent assessments suggest that the majority of sequenced plant genomes are derived from allopolyploid ancestry (21, 36, 38–42). Another critical parameter for assessing the evolutionary significance of WGD is the timing of the WGD event. This is critical for correlating WGD events with ecological and geological factors (43–45), as well as understanding the timescale of the diploidization process (5, 28, 46–50). The dating of WGD events is typically accomplished through genetic analysis. This can be done by comparing the polyploid genetic data to its presumptive parental species (5, 51–53) or by a comparison of duplicated genes within the polyploid genome (54–56).

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 (Figs. 1 and 2), and the mode (the Ks "peak" on the histogram) is used to infer the time of WGD. Univariate mixture models are typically used to fit the peak (56–61).

The Ks-based approach however has several pitfalls. The conversion between Ks and time is not straightforward (5, 43, 62), Ks saturates for Ks > 2, or roughly 200 million years (47, 58, 59), and multivariate fitting methods have been shown to overfit distributions (56, 57, 63). A further drawback is that while Ks-based methods can in theory correctly date the origin of a polyploid lineage which traces back to a single individual (where the (51, 53, 64–66) subgenomes began to diverge at the same instant they duplicated), the same methods may fail for allopolyploids. This is because the peak of the Ks distribution corresponds to the divergence time between the diploid parental species (hereon **SPC time**), not the time of origin of the polyploid (hereon **WGD time**) (5, 61, 65), as shown in (Fig. 1). Since plant genomes can remain compatible for 10 mya or more after the last common ancestor (67, 68) and up to 50 mya in one documented case (69), the difference between SPC time and WGD time can be significant.

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 (51, 53, 64–66). Population-demographic approaches have also been used with recent success to time WGD (31, 70, 70–74).

As part of the 1000 Plants (1KP) initiative (75), transcriptomic data from over 1000 plants spanning the plant kingdom was used to compile Ks histograms. As in other works, Gaussian mixture models were fit and used to detect WGD from Ks peaks (54,

58, 76, 77). But a review of empirical Ks histogram shapes from the 1KP dataset show great complexity beyond the Gaussian (63, 78), which simulations have thus far do not replicate (79).

Here we present the Ks simulator SpecKS. SpecKS simulates the forward evolution of auto and allopolyploid genomes (Fig. 3). SpeckS is highly configurable (Table 1 and S2) and capable of producing a rich array of Ks distribution shapes, including but not limited to the Gaussian. SpecKS demonstrates the sensitivity of the Ks distribution to a number of ancient speciation parameters, such as effective population size, gene shedding rates, and the mode of speciation.

Results

SpecKS accurately recovers key speciation parameters

WGD time and SPC time are key parameters which define an ancient polyploidization event. We show the response of the SpecKS-generated Ks histogram to a range of WGD and SPC times (Figs. 4, 5, S1, S2). SpecKS accurately retrieves the expected linear relationship between the peak of the Ks histogram and SPC time (Fig. S2, left). SpecKS additionally recaptures the expected logarithmic relationship between the number of genes lost and the WGD times (Fig. S2, right). We also show that the whole genome duplication time and speciation time can be accurately recovered from the output histogram (Fig. S2., bottom).

SpecKS demonstrates allo and autopolyploid-derived species exhibit differently shaped Ks histograms.

Since speciation is a gradual process, we theorized that the Ks histograms of allopolyploids would display a greater diversity of Ks between **ohnologs** (the subset of paralogs derived from WGD) than autopolyploids, whose ohnologs may diverge instantaneously (5, 61). We expected that the difference between the Ks distributions would lose their shape over time due to randomness in Ks accumulation and signal attenuation due to gene shedding. Our simulations confirmed the difference between allo and autopolyploid Ks histogram shapes persist for about 50 MY (Fig. 4).

The Ks histogram is sensitive to the Ne of the ancestral parental species

For a panmictic population, the coalescent times of genes are exponentially distributed, with the half life of the exponential distribution being proportional to the product of the effective population size and the generation time (80). Thus we theorized that allopolyploids whose parental diploids had diverged from common ancestral species with large (thus more diverse) genomes would have more variance in their Ks distribution than allopolyploids which are derived from smaller ancestral populations. Furthermore, we expected that the Ks histogram of autopolyploids would be invariant to population size, as the divergence between ohnologs in an autopolyploid lineage would be close to zero at the initiation of the polyploid lineage. Our simulations show that changes in Ne affect the Ks distributions of allopolyploids, while autopolyploid distributions remain invariant (Fig. 5).

Quantifying the inherent error in estimating WGD time from the Ks-histogram mode

While the common method of approximating time of WGD from the peak of Ks histograms is theoretically sound for autopolyploids, it is erroneous for allopolyploids.

Since the estimation error is a result of the difference between WGD and SPC time (Fig. 1), we expected the error to increase proportionally to the difference. Indeed, simulations with SpecKS suggest a linear relationship between estimation error and the degree of allopolyploidy, where the degree of allopolyploidy is measured as the difference in time between true SPC time and true WGD time (Fig. 6, bottom).

An accurate method to infer WGD time

We previously showed how the SpecKS-generated Ks histogram can be used to infer both WGD and SPC times (Fig. S2). Our simulations show that inferring WGD time from the number of genes lost is theoretically more accurate than using the histogram peak (Fig. 6), and has the further benefit of working equally well for allo and autopolyploids.

Simulated Ks histogram shapes match well with experimental observations

Here we present how SpecKS may be used in an inference framework for estimating demographic parameters that generate the theoretical expected Ks histogram using SpecKS. We compare SpecKS-generated Ks histograms to the true Ks histograms generated from transcriptomic data for 3 well-characterized tetraploid species, maize, coffee, and poplar. Here we see that the simulated histogram shapes match well with experimental observations, with RMSE all under 1.8 (Fig. 7). The SpecKS input parameters have been optimized to minimize the RMSE error (Table S4). The WGD and SPC parameters used in the best-fit simulations correspond well with dates from other sources (Table S5) (81–84). We also note that the main difference between the simulated histograms and the true histograms is that true histograms maintain a set of paralogs whose numbers do not decay over time and whose Ks distribution appears flat. One explanation for this phenomena is that these paralogs are maintained by selection, and thus not presently modeled by our system.

Tests for the degree of allopolyploidy and the effects of ancestral Ne

Here we trained two linear regression models, one to discriminate between allo or autopolyploid-derived species, and another to discriminate between whether the ancestral species was a high or low Ne population at the time of divergence.
SpecKS was used to generate a dataset of 160 simulation results (dataset 1, parameters given in (Table 3)). Using a ⅓ test, ⅔ training approach with the data, we trained a linear regression model to discriminate between allo and autopolyploidy events based on the shape of the Ks histogram. Our model achieved 100% accuracy on the test data (Fig. 8). This approach was further extended to categorize genomes as derived from high, medium, and low Ne parental populations, with the autopolyploids being included as a subset of the low Ne group (since genomes start off with essentially no diversity between them). This extended approach gave 94% accuracy for WGD up to 80 MYA (Fig. 9) and 100% accuracy on test data for WGD up to 50 MYA (Fig. S3).

Categorization of polyploidization events across the plant tree of life

A recent review by Wang *et al.* 2019 suggests that the majority of plant genomes (>99.7%) are likely derived directly from allopolyploidization events (36, 41). We were interested to know if the discrimination algorithms trained on our SpecKS simulation data, when applied to real Ks data, might yield the same results. Since our allo-vs-auto polyploid classifier was tightly tied to lineage specific parameters (nominally, the gene shedding rate, which is used to determine WGD time, see Fig. 6), it was not reasonable to apply this classifier to a broad set of lineages. However, since in the SpecKS model,

the skewness of the Ks histogram was directly dependent only on the product of Ne and generation time, we hypothesized that the low-vs-high Ne classifier would be more globally extendable across a broad set of observations across the plant tree of life. Thus we applied the low-vs-high Ne classifier to 1000 histograms drawn from data from 1KP study (75). Using real Ks observations made public by (58), we were able to classify 228 WGD events, resulting in 12, 38 and 178, low-Ne, medium-Ne and high-Ne determinations respectively (Fig. 10). Our results suggest that >94.7% of the WGD in the 1KP dataset are allopolyploid, and the remaining 5.3% are a mixed category. These results are concordant with (41), suggesting that the majority of ancient polyploidization events which contributed to the genetic conduit may be allopolyploidization events.

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 such as the degree of allopolyploidy time of WGD and ancestral Ne. We also develop an accurate discrimination model that classifies angiosperm taxa into allo or autopolyploids with ancestrally diverse or less diverse WGD events. We contend that the Ks distribution is therefore information rich, with the potential to estimate WGD and SPC events across diverse taxa.

We have demonstrated through extensive forward simulations that the commonly-used method for estimating WGD time from the peak of the Ks histogram will be increasingly inaccurate for more allopolyploid derived species (Fig. 6), and here we develop an alternative method that works equally well for allo and polyploid derived species.
Our simulations also show that the shape and skew of the Ks histogram is sensitive to the evolutionary history of the ancestral polyploidization event: allo and autopolyploid derived species yield characteristically different Ks distributions, and these differences persist for up to 50 million years, until the inevitable attenuation due to gene shedding and stochastic processes in accumulation of Ks renders them indistinguishable (Fig. 4). Furthermore, we have shown that the tail of the Ks histogram depends on the diversity between the ancestral subgenomes of the polyploid (Fig. 5). This is to be expected, since these subgenomes represent individuals originally within the same diploid species (autopolyploids) or derived from individuals descended from the same diploid species that subsequently underwent speciation (allopolyploids). Thus the diversity of the diploid ancestors sort into lineages as orthologous pairs going forwards in time. Looking backwards, we demonstrate with simulations that these orthologous pairs can be used to test inferences about the ancestral divergence time (SPC) and ancestral effective population size (Ne) (Fig. 5&6).

We additionally show that SpecKS is capable of generating a range of realistic histogram shapes. For instance, in Fig. 7, 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 (Fig. 2) appear to be well replicated, with overall RMSE less than 2 in all cases (Fig. 7). This demonstrates that the parameters supplied (WGD time, SSD time, ancestral Ne) are 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 > SPC) whose numbers do not decay over time.

In contrast, in our simulation, the number of paralogous pairs retained asymptotically approaches zero over time. The latter behavior is expected under the SpecKS model, because both SSDs and ohnologs are shed genes following a parameterized decay model (described in Methods).

One explanation for this phenomenon is that the ohnolog pairs that persist in the true Ks histogram and not in the simulations may be duplicates that are 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. 7). Since these species have undergone recent domestication (85, 86), it may be possible that domestication has hindered efficient gene shedding or promoted paralog retention (87).

Despite its utility, SpecKS currently represents a "null hypothesis," and does not include selection, population dynamics, the effects of domestication, and other factors which would reasonably impact the Ks histogram. Its internal model of evolution (GY94 with equal equilibrium codon frequencies) is 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.

We also note that many of the parameters that go into SpecKS are lineage specific (ie, the time of WGD, 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*. Similarly, the discrimination models presented were drawn from simulation runs where a small number of parameters were fixed (for example, fixed gene-birth-and-death rate, generation time, and molecular clock). When making lineage-specific inferences, we caution users to parameterize SpecKS with lineage-appropriate values, or if they are unknown, to use SpecKS to model the impact of uncertainty in their value. Our low-vs-high Ne classifier was trained assuming a generation time of 1 year for all species in the simulation, so by extension, our 1KP classifications were made under the assumption that each species had a generation time of 1 year. While this is not true for all species, allopolyploid species which have a longer generation time (10 or 100 years, for example) may have been classified as having a higher Ne than expected. This might reclassify some true autopolyploids out of the low "Low Ne" species into the "Medium Ne" or "High Ne" categories. However, this would not have affected autopolyploids, and thus would not change our assertion that >94.7% of the WGD in the 1KP dataset appear allopolyploid.

Materials and Methods

Simulation Implementation

SpecKS takes as input an XML configuration file listing the simulation parameters (Table 1, S2 & S3) and outputs a text file (.csv) of all pairwise Ks accumulated between all gene pairs (ohnologs and small-scale duplications) 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. 3). The modules functions are (1) species tree generation, (2) gene tree generation, (3) application of a gene birth and death model, (4) application of a post WGD gene shedding model, (5) gene sequence evolution, (6) the Ks calculation between gene pairs, (7) Ks histogrammer and (8) final results assembly. We give details for each module below.

(1) Species tree generation: A newick-formatted species tree is generated for each polyploid. This is aways in the format (O:□□□□, (P1:□□□, P2:□□□):□□□□-□□□), where □□□□ specifies the full simulation time, and □□□ specific the SPC time. P1 and P2 denote the subgenomes of the polyploid, while O denotes the outgroup.

(2) 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 whole genome duplication, or “WGDs”). Gene-divergence times for autopolyploid ohnologs are modeled using a Dirac-delta distribution, such that P1 and P2 diverge at the exact same time for all gene trees. Allopolyploid ohnolog divergence times are modeled based on the idea the variance in ortholog divergence times at speciation time can be traced to the “founder” individuals of the P1 and P2 species. Presuming the founder individuals are chosen at random from within the ancestral species, the divergence time between their respective gene-copies follows an exponential distribution, under the Kingman coalescent (80). Thus the allopolyploid ohnolog divergence times are initialized following an exponential distribution. The distribution parameters are configurable, and in theory should be based on the ancestral Ne and generation time. The number of gene trees are also configurable (88).

(3) Gene birth and death model: The gene birth and death model introduces small-scale duplications (“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 (89). Genes are assigned a death-date at birth, with a life span drawn randomly from an exponential decay distribution (55, 90). Gene birth rate and mean life expectancy are configurable (91, 92). Genes born which will be dead before the simulation conclues are pruned at this time for computational efficiency.

(4) Ohnolog gene shedding model: Genes duplicated by WGD also have a mean life expectancy given by an exponential decay distribution (93). 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 (91, 94).

(5) Gene sequence evolution: For each simulated gene tree, we simulated codon sequences using the EVOLVER program within PAML v4.10.7 (95, 96). PAML was configured as in (57) to simulate codon evolution along each finalized gene tree using a Goldman–Yang (GY94) model of codon evolution (97, 98) with equal equilibrium codon frequencies, a transition/transversion rate ratio of 2 and a global dN/dS of 0.2 as in (57). The number of codons per alignment is set by default to 1000 as in (57), in line with reported plant transcript lengths (99–101), but is configurable.

(6) 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 (95, 96). Ks per million years is 0.01 by default and is configurable (57, 102).

(7) 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 are only meant to give a confirmation of the run success. It is expected that the end user will use their own visualization tools to generate more elegant histograms.

(8) 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 non-synonymous 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 is organized into 4 columns, giving the gene-tree name the pair is from, 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 the Ks value came from, respectively.

Parallelization: For simplicity, SpecKS is not internally parallelized. A single polyploid of about 3000 genes, 1000 codons long, simulated over 100 MY, runs on a typical cluster (e.g. 16-node Dual-10 Xeon CPU, E5 -2630v4 2.20GHz, 12.8GB RAM per core) in under 10 minutes. 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).

Transcriptomic data: Ks histograms in Fig. 7 were generated from transcriptomic data for *Coffea arabic*a, *Zea mays* and *Populus trichocarpa*. Transcriptomic data was sourced from NCBI (NCBI assembly GCF_003713225.1, GCF_902167145.1 and GCF_000002775.5), and the software package KsRates (Sensalari, 2022) was used to calculate the Ks histograms.

Discrimination models: For the allo-vs-auto discrimination model, the model input data was the difference (in MY) between the inferred WGD and SPC time, and the model target was whether the input polyploid was an allo or autopolyploid. For the low-vs-high Ne discrimination model, the model input data was the difference (in Ks space) between histogram peak and center of mass, and the model target was whether the input polyploid had a high, medium or low Ne. We chose three levels of Ne (low, medium and high) to span ranges corresponding to empirical observations (103). Gt is the generation time of the species being simulated. The Low Ne group included simulated autopolyploids (whose genomes start off with essentially no diversity between them) and allopolyploids whose diversity corresponded to Ne*Gt less than or equal to 1 MY. Medium Ne were allopolyploids with Ne*Gt greater than 1 and less than 10. And high-Ne were allopolyploids with Ne*Gt of 10 or more. For both classifiers, two thirds of the data was used for training and one third was kept for testing. The logistic regression function from sklearn.linear_model v1.3.0 was used to build the classifiers (104).

1KP data: In results section 8, the low-vs-high Ne discrimination model is run on histograms generated from Ks data available at (https://gitlab.com/barker-lab/1KP/-/tree/master/1KP_ksplots). This Ks data was generated by (58, 75).

Source code and data availability

SpecKS is available from the github repository https://github.com/tamsen/SpecKS. SpecKS input files used to drive the simulation and the scripts used to generate the figures are at https://github.com/tamsen/SpecKS_paper_scripts.

Acknowledgments

We would like to thank members of the Barker and Sethuraman Labs for valuable inputs during the conceptualization stage of this project. AS and TD were both supported by NSF CAREER: 2147812 and two CSUBIOTECH grants. All computations were performed on the *mesxuuyan* cluster at San Diego State University, which was supported by startup funds and NSF ABI: 1564659 to AS.

References

1. S. P. Otto, J. Whitton, Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
2. J. F. Wendel, The wondrous cycles of polyploidy in plants. *Am J Bot* **102**, 1753–1756 (2015).
3. P. S. Soltis, D. E. Soltis, *Polyploidy and genome evolution* (Springer, 2012).
4. J. W. Clark, P. C. Donoghue, Whole-genome duplication and plant macroevolution. *Trends in Plant Science* **23**, 933–945 (2018).
5. J. J. Doyle, A. N. Egan, Dating the origins of polyploidy events. *New Phytol* **186**, 73–85 (2010).
6. J. F. Wendel, S. A. Jackson, B. C. Meyers, R. A. Wing, Evolution of plant genome architecture. *Genome Biology* **17**, 37 (2016).
7. E. Schranz, S. Mohammadin, P. P. Edger, Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology* **15**, 147–153 (2012).
8. K. Bomblies, G. Jones, C. Franklin, D. Zickler, N. Kleckner, The challenge of evolving stable polyploidy: could an increase in “crossover interference distance” play a central role? *Chromosoma* **125**, 287–300 (2016).
9. L. Yant, K. Bomblies, Genome management and mismanagement--cell-level opportunities and challenges of whole-genome duplication. *Genes Dev* **29**, 2405–2419 (2015).
10. B. McClintock, A cytological and genetical study of triploid maize. *Genetics* **14**, 180 (1929).
11. P. Baduel, S. Bray, M. Vallejo-Marin, F. Kolář, L. Yant, The “polyploid hop”: shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Frontiers in Ecology and Evolution* **6**, 117 (2018).
12. J. Ramsey, D. W. Schemske, Neopolyploidy in flowering plants. *Annu. Rev. Ecol.*

Syst. **33**, 589–639 (2002).
13. L. Comai, *et al.*, Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. *Plant Cell* **12**, 1551 (2000).
14. V. W. Mayer, A. Aguilera, High levels of chromosome instability in polyploids of Saccharomyces cerevisiae. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **231**, 177–186 (1990).
15. S. C. Le Comber, M. L. Ainouche, A. Kovarik, A. R. Leitch, Making a functional diploid: from polysomic to disomic inheritance. *New Phytol* **186**, 113–122 (2010).
16. H. Zhang, *et al.*, Transcriptome shock invokes disruption of parental expression-conserved genes in tetraploid wheat. *Scientific Reports* **6**, 26363 (2016).
17. N. Arrigo, M. S. Barker, Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* **15**, 140–146 (2012).
18. G. L. Stebbins, Cataclysmic Evolution. *Scientific American* **184**, 54–59 (1951).
19. R. J. A. Buggs, *et al.*, Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Curr Biol* **21**, 551–556 (2011).
20. A. Madlung, Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* **110**, 99–104 (2013).
21. A. E. Baniaga, H. E. Marx, N. Arrigo, M. S. Barker, Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecology Letters* **23**, 68–78 (2020).
22. Y. Van de Peer, T.-L. Ashman, P. S. Soltis, D. E. Soltis, Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* (2020). https://doi.org/10.1093/plcell/koaa015.
23. D. E. Soltis, C. J. Visger, D. B. Marchant, P. S. Soltis, Polyploidy: pitfalls and paths to a paradigm. *American Journal of Botany* **103**, 1146–1166 (2016).
24. S. Dodsworth, M. W. Chase, A. R. Leitch, Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* **180**, 1–5 (2016).
25. F. M. Robertson, *et al.*, Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology* **18**, 111 (2017).
26. G. Nieto Feliner, J. Casacuberta, J. F. Wendel, Genomics of evolutionary novelty in hybrids and polyploids. *Frontiers in Genetics* **11**, 792 (2020).
27. P. S. Soltis, X. Liu, D. B. Marchant, C. J. Visger, D. E. Soltis, Polyploidy and novelty: Gottlieb's legacy. *Philos Trans R Soc Lond B Biol Sci* **369**, 20130351 (2014).
28. Z. Li, *et al.*, Patterns and Processes of Diploidization in Land Plants. *Annual Review of Plant Biology* **72**, 387–410 (2021).
29. D. C. Tank, *et al.*, Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* **207**, 454–467 (2015).
30. L. Carretero-Paulet, Y. Van de Peer, The evolutionary conundrum of whole-genome duplication. *Am J Bot* **107**, 1101–1105 (2020).
31. P. D. Blischak, M. Sajan, M. S. Barker, R. N. Gutenkunst, Demographic history inference and the polyploid continuum. *Genetics* **224**, iyad107 (2023).
32. S. K. Deb, P. P. Edger, J. C. Pires, M. R. McKain, Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: a genomic and epigenomic perspective. *New Phytologist* (2023).
33. A. S. Mason, J. F. Wendel, Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Frontiers in Genetics* 1014 (2020).

34. P. Meirmans, P. Van Tienderen, The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity* **110**, 131–137 (2013).
35. C. L. Stebbins, Variation and evolution in plants. *Variation and evolution in plants.* (1950).
36. M. S. Barker, N. Arrigo, A. E. Baniaga, Z. Li, D. A. Levin, On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* **210**, 391–398 (2016).
37. G. L. Stebbins, "Types of polyploids: their classification and significance" in *Advances in Genetics*, (Elsevier, 1947), pp. 403–429.
38. D. E. Soltis, R. J. A. Buggs, J. J. Doyle, P. S. Soltis, What we still don't know about polyploidy. *TAXON* **59**, 1387–1403 (2010).
39. J. Clausen, D. D. Keck, W. M. Hiesey, Experimental studies on the nature of species. 2. Plant evolution through amphiploidy and autoploidy, with examples from the Madiinae. *Experimental studies on the nature of species. 2. Plant evolution through amphiploidy and autoploidy, with examples from the Madiinae.* (1945).
40. C. D. Darlington, Recent advances in cytology. *Recent advances in cytology.* (1937).
41. J. Wang, *et al.*, Polyploidy index and its implications for the evolution of polyploids. *Frontiers in genetics* **10**, 807 (2019).
42. C. Parisod, R. Holderegger, C. Brochmann, Evolutionary consequences of autopolyploidy. *New Phytologist* **186**, 5–17 (2010).
43. J. Barba-Montoya, M. Dos Reis, H. Schneider, P. C. J. Donoghue, Z. Yang, Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. *New Phytol* **218**, 819–834 (2018).
44. J. W. Clark, P. C. Donoghue, "Constraining Whole-Genome Duplication Events in Geological Time" in *Polyploidy: Methods and Protocols*, (Springer, 2023), pp. 139–154.
45. L. Zhang, *et al.*, The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant, Cell & Environment* **43**, 2847–2856 (2020).
46. K. T. David, J. R. Oaks, K. M. Halanych, Patterns of gene evolution following duplications and speciations in vertebrates. *PeerJ* **8**, e8813 (2020).
47. A. R. De La Torre, Z. Li, Y. Van de Peer, P. K. Ingvarsson, Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biology and Evolution* **34**, 1363–1377 (2017).
48. S. Hoegg, H. Brinkmann, J. S. Taylor, A. Meyer, Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal of Molecular Evolution* **59**, 190–203 (2004).
49. S. Laurent, N. Salamin, M. Robinson-Rechavi, No evidence for the radiation time lag model after whole genome duplications in Teleostei. *PLoS One* **12**, e0176384–e0176384 (2017).
50. B. K. Mable, Polyploids and hybrids in changing environments: winners or losers in the struggle for adaptation? *Heredity* **110**, 95–96 (2013).
51. G. C. Conant, POInT: Modeling Polyploidy in the Era of Ubiquitous Genomics. *Methods Mol Biol* **2545**, 77–90 (2023).
52. G. M. Douglas, *et al.*, Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proceedings of the National Academy of Sciences* **112**, 2806–2811 (2015).
53. J. Mccann, *et al.*, Dating the species network: Allopolyploidy and repetitive DNA evolution in American daisies (*Melampodium sect. Melampodium*, Asteraceae). *Syst Biol* **67**, 1010–1024 (2018).

54. J. W. Clark, M. N. Puttick, P. C. J. Donoghue, Origin of horsetails and the role of whole-genome duplication in plant macroevolution. *Proc Biol Sci* **286**, 20191662 (2019).
55. M. Lynch, J. S. Conery, The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
56. K. Vanneste, Y. Van de Peer, S. Maere, Inference of genome duplications from age distributions revisited. *Mol Biol Evol* **30**, 177–190 (2013).
57. G. P. Tiley, M. S. Barker, J. G. Burleigh, Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biology and Evolution* **10**, 2882–2898 (2018).
58. Z. Li, M. S. Barker, Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. *GigaScience* **9**, giaa004 (2020).
59. L. Cui, *et al.*, Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16**, 738–749 (2006).
60. J. A. Schlueter, *et al.*, Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**, 868–876 (2004).
61. H. Chen, A. Zwaenepoel, Inference of ancient polyploidy from genomic data. *Methods Mol Biol* **2545**, 3–18 (2023).
62. K. H. Wolfe, W. H. Li, P. M. Sharp, Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci U S A* **84**, 9054–9058 (1987).
63. A. Zwaenepoel, Y. Van de Peer, WGD—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155 (2019).
64. R. Bouckaert, *et al.*, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019).
65. G. W. C. Thomas, S. H. Ather, M. W. Hahn, Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Systematic Biology* **66**, 1007–1018 (2017).
66. D. Wen, Y. Yu, J. Zhu, L. Nakhleh, Inferring phylogenetic networks using phyloNet. *Systematic Biology* **67**, 735–740 (2018).
67. D. A. Levin, The timetable for allopolyploidy in flowering plants. *Ann Bot* **112**, 1201–1208 (2013).
68. D. S. Senchina, *et al.*, Rate variation among nuclear genes and the age of polyploidy in Gossypium. *Mol Biol Evol* **20**, 633–643 (2003).
69. C. J. Rothfels, *et al.*, Natural hybridization between genera that diverged from each other approximately 60 million years ago. *Am Nat* **185**, 433–442 (2015).
70. W. W. Booker, D. R. Schrider, The genetic consequences of range expansion and its influence on diploidization in polyploids. *bioRxiv* (2023). https://doi.org/10.1101/2023.10.18.562992.
71. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics* **5**, e1000695 (2009).
72. C. Roux, J. R. Pannell, Inferring the mode of origin of polyploid species from next-generation sequence data. *Mol Ecol* **24**, 1047–1059 (2015).
73. J. Roux, J. Liu, M. Robinson-Rechavi, Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. *Molecular biology and evolution* **34**, 2773–2791 (2017).
74. K. R. St Onge, *et al.*, Coalescent-based analysis distinguishes between allo- and autopolyploid origin in Shepherd's Purse (*Capsella bursa-pastoris*). *Mol Biol Evol* **29**,

1721–1733 (2012).
75. J. H. Leebens-Mack, *et al.*, One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
76. X. Guo, *et al.*, Linked by ancestral bonds: Multiple whole-genome duplications and reticulate evolution in a Brassicaceae tribe. *Molecular Biology and Evolution* **38**, 1695–1714 (2020).
77. X. Qiao, *et al.*, Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology* **20**, 38 (2019).
78. C. Sensalari, S. Maere, R. Lohaus, ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics* **38**, 530–532 (2022).
79. B. L. Sutherland, G. P. Tiley, Z. Li, M. T. McKibben, M. S. Barker, SLEDGe: Inference of ancient whole genome duplications using machine learning. *bioRxiv* 2024.01.17.574559 (2024). https://doi.org/10.1101/2024.01.17.574559.
80. J. F. C. Kingman, The coalescent. *Stochastic Processes and their Applications* **13**, 235–248 (1982).
81. J. Salojärvi, *et al.*, The genome and population genomics of allopolyploid Coffea arabica reveal the diversification history of modern coffee cultivars. *Nature genetics* **56**, 721–731 (2024).
82. Q. Yu, *et al.*, Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). *The Plant Journal* **67**, 305–317 (2011).
83. X. Dai, *et al.*, The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research* **24**, 1274–1277 (2014).
84. B. S. Gaut, J. F. Doebley, DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences* **94**, 6809–6814 (1997).
85. N. Yang, *et al.*, Two teosintes made modern maize. *Science* **382**, eadg8940.
86. C. Montagnon, A. Mahyoub, W. Solano, F. Sheibani, Unveiling a unique genetic diversity of cultivated Coffea arabica L. in its main domestication center: Yemen. *Genetic Resources and Crop Evolution* **68**, 2411–2422 (2021).
87. B. S. Gaut, D. K. Seymour, Q. Liu, Y. Zhou, Demography and its effects on genomic variation in crop domestication. *Nature Plants* **4**, 512–520 (2018).
88. L. Sterck, S. Rombauts, K. Vandepoele, P. Rouzé, Y. Van de Peer, How many genes are there in plants (… and why are they there)? *Current Opinion in Plant Biology* **10**, 199–203 (2007).
89. J. Zhao, A. I. Teufel, D. A. Liberles, L. Liu, A generalized birth and death process for modeling the fates of gene duplication. *BMC Evol Biol* **15**, 275 (2015).
90. M. Lynch, M. O'Hely, B. Walsh, A. Force, The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**, 1789–1804 (2001).
91. Y.-L. Guo, Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes. *Plant J* **73**, 941–951 (2013).
92. M. Lynch, J. S. Conery, The evolutionary demography of duplicate genes. *Genome Evolution: Gene and Genome Duplications and the Origin of Novel Gene Functions* 35–44 (2003).
93. R. Ren, *et al.*, Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Molecular Plant* **11**, 414–428 (2018).
94. S. Maere, *et al.*, Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* **102**, 5454–5459 (2005).

95. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).
96. Z. Yang, R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32–43 (2000).
97. N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725–736 (1994).
98. Z. Yang, R. Nielsen, Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**, 409–418 (1998).
99. D. Luo, *et al.*, 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 Biology* **19**, 32 (2019).
100. O. Al-Dossary, *et al.*, Long read sequencing to reveal the full complexity of a plant transcriptome by targeting both standard and long workflows. *Plant Methods* **19**, 112 (2023).
101. S. Zhang, *et al.*, New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. *Nucleic Acids Res* **48**, 7700–7711 (2020).
102. G. Blanc, K. H. Wolfe, Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667 (2004).
103. T. I. Gossmann, *et al.*, Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol* **27**, 1822–1832 (2010).
104. F. Pedregosa, *et al.*, Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

Figures and Tables

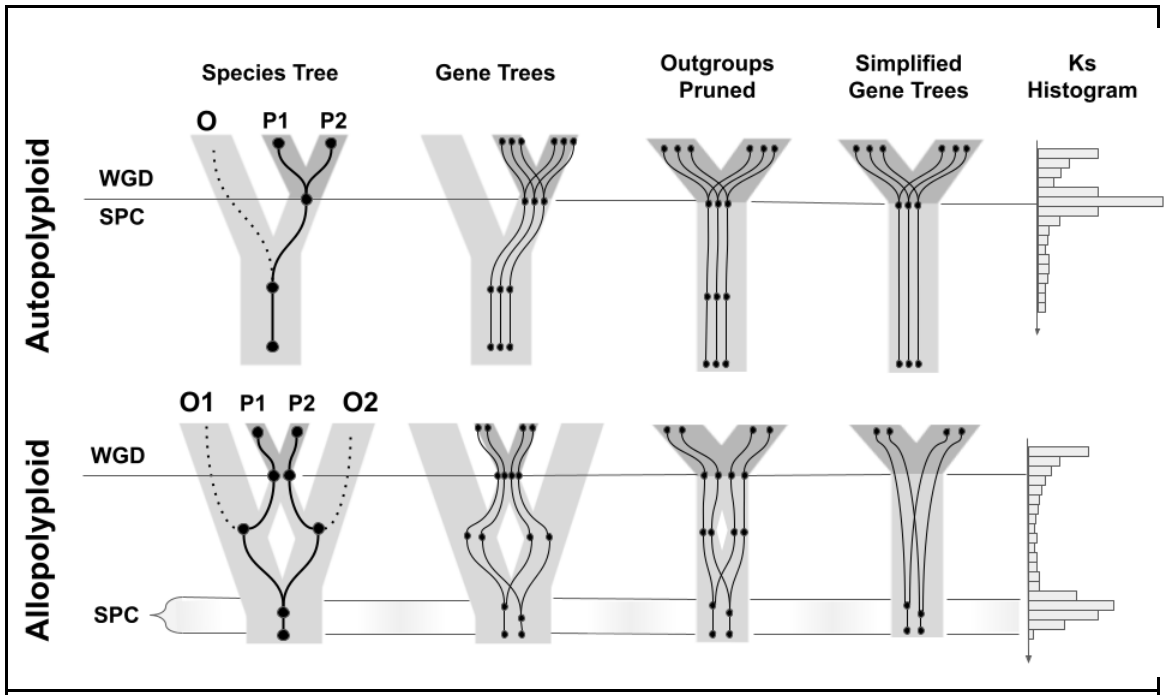


Figure 1. Relationship between the shape of the Ks histogram and the modes of auto versus allopolyploid speciation.
Top: Autopolyploid speciation. Bottom: Allopolyploid speciation. Outgroup (O) and parental species are indicated in light gray. Polyploid species are shown in dark gray with duplicated subgenomes denoted as P1 and P2. The arrow in the Ks histogram points backwards in time along the X axis. This figure illustrates that for autopolyploids the time of WGD is synonymous with subgenome divergence (instantaneous SPC) and thus the Ks peak. For allopolyploids the time of WGD is offset from subgenome divergence (SPC), which is gradual. The Ks peak therefore corresponds to SPC, and not WGD. Figures at the far right and left are derived from (Chen and Zwaenepoel, 2023).

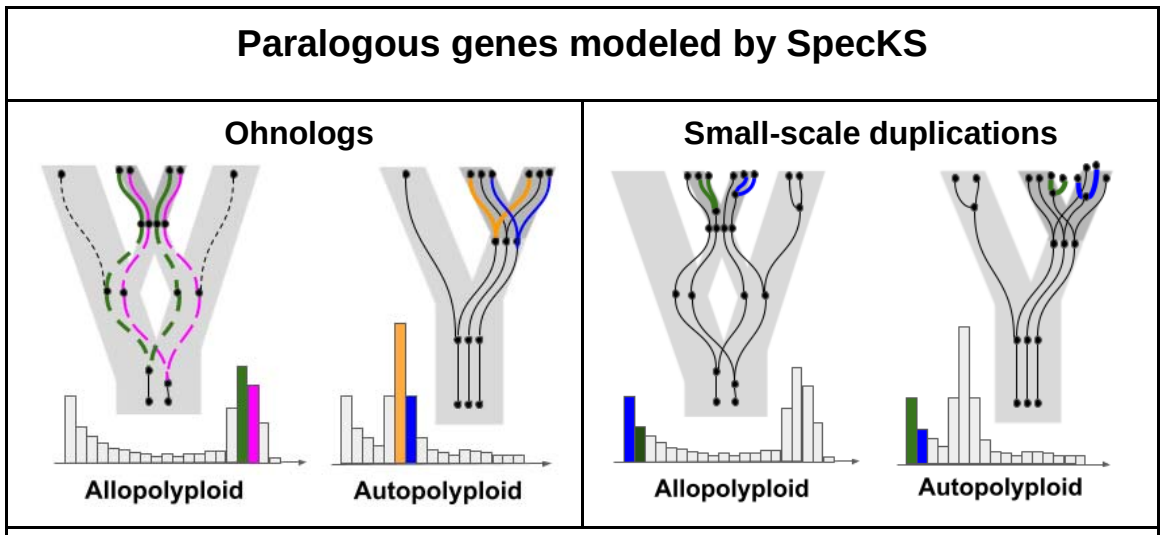


Figure 2. Ohnologs (paralogs generated by WGD) and small-scale duplications ("SSDs", paralogs generated as single or small scale copies) have unique contributions to the Ks histogram. Ohnologs (left panel) are shown for both allo and autopolyploids. Green and pink ohnologs (far left) are attributed to 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.

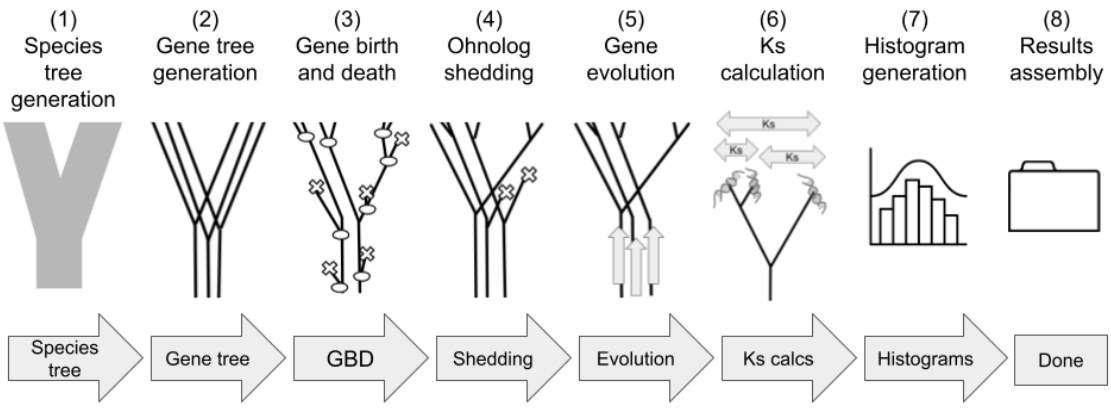


Figure 3. The SpecKS pipeline.
Here we give a cartoon depicting the eight modules that make up the SpecKS pipeline.

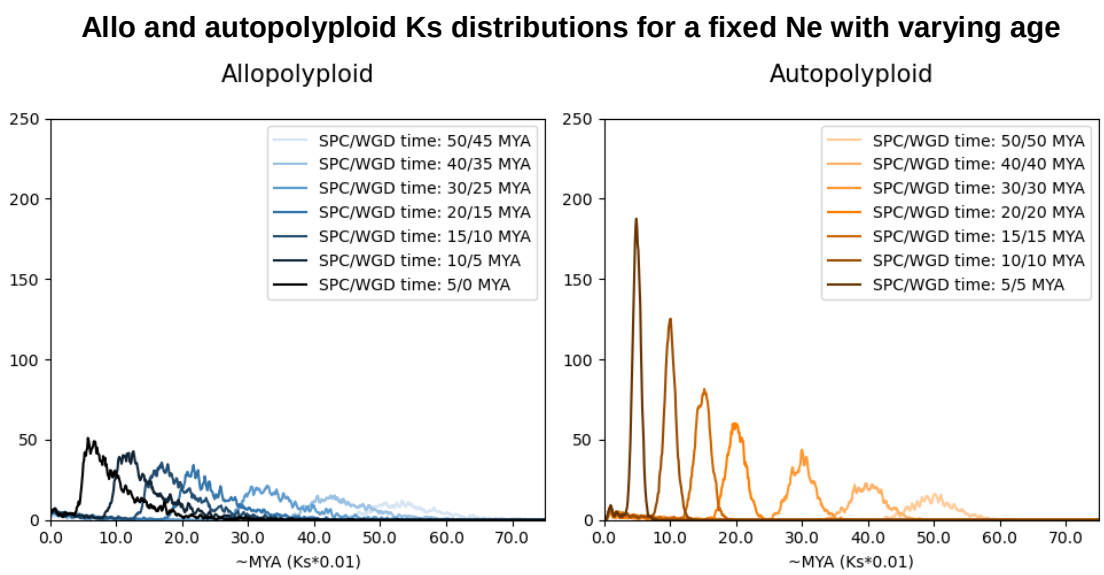


Figure 4. SpecKS demonstrates how the characteristically different shapes of allo and autopolyploid Ks distributions become more similar over time.
(Left) Allopolyploids. (Right) Autopolyploids. Y-axis gives the number of gene pairs per bin. X-axis is the histogram bins, in Ks, converted to the equivalent MYA. SPC times were set from 5 to 50 MYA. Allopolyploids had WGD offset by 5 MY, and autopolyploids had WGD offset by 0 MY. Allopolyploid Ks divergence at SPC time was modeled as an exponential decay with a decay constant of 5 MY.

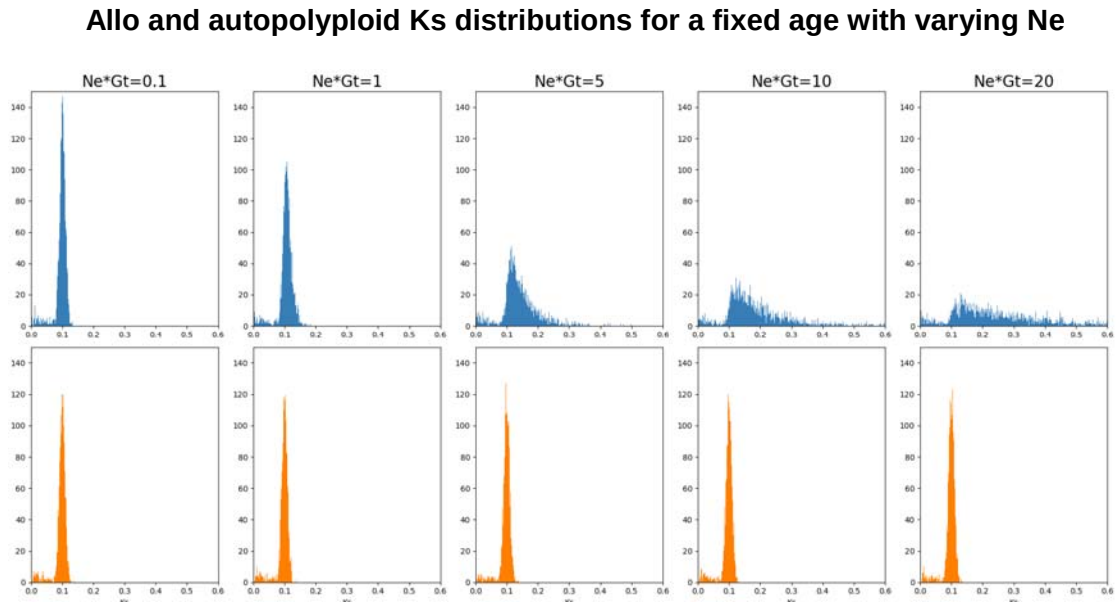


Figure 5. Allopolyploid Ks distributions exhibit characteristically different shapes depending on ancestral Ne, while autopolyploid distributions are similar to low-Ne distributions.

(Top) Allopolyploids. (Bottom) Autopolyploids. Y-axis gives the number of gene pairs per bin. X-axis is the histogram bins, in Ks. SPC time was set to 10 MYA for all simulations. For autopolyploids, WGD time was set to SPC time. For allopolyploids, WGD time was 5 MYA after SPC. Allopolyploid Ks divergence at SPC time was modeled as an exponential decay function with a decay constant K=Ne*Gt, where Ne is the effective population size and Gt is the generation time. Here Ne varies from 0.1 to 20, and Gt is held constant.

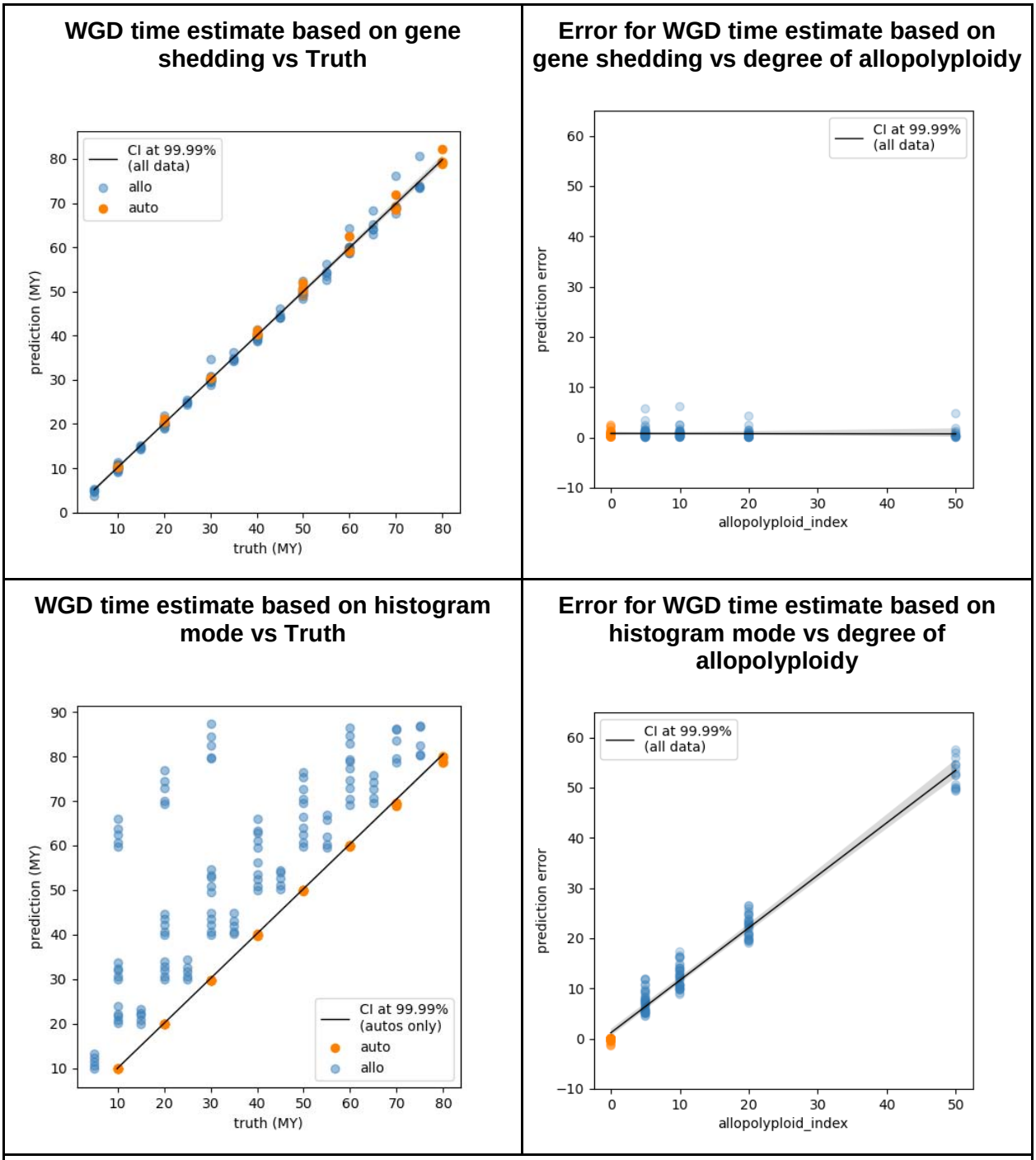
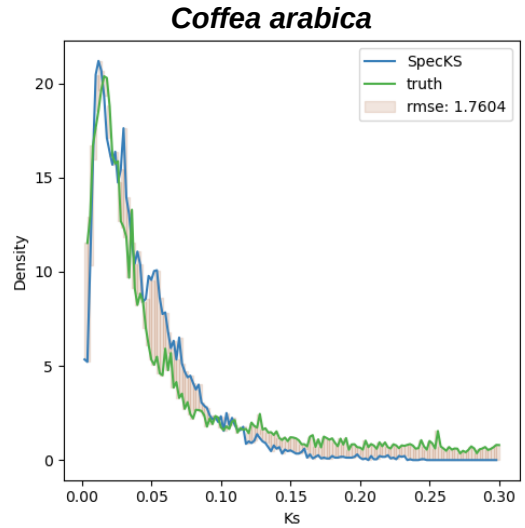


Figure 6. SpecKS shows that estimation of WGD time based number of genes remaining yields accurate results, irrespective of whether a polyploidization event is allo or autopolyploid.

(Left) WGD time as estimated from the number of genes shed. (Top right) Error in the WGD time as predicted from the number of genes shed (y axis) vs the degree of allopolyploidy (x-axis). (Bottom right) Error in the WGD time as estimated from the histogram mode (y axis) vs

the degree of allopolyploidy (x-axis). To make these plots, 160 simulations were run, with input parameters forming a 3-dimensional grid over SPC, WGD and Ne*Gt. Here SPC times range from 10 to 80 MYA, by 10, with WGD offset by 0, 5, 10 and 50 MY. Ne*GT is 0.1,1.0,5.0, 10 and 20 MY. Autopolyploids are red,. Allopolyploids are blue. Line fit and confidence interval is black/gray.



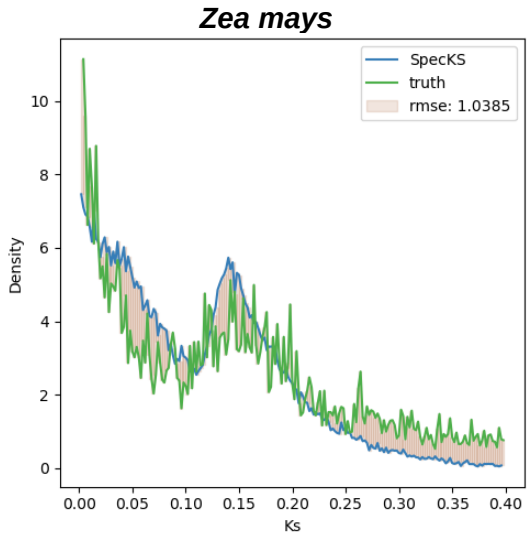
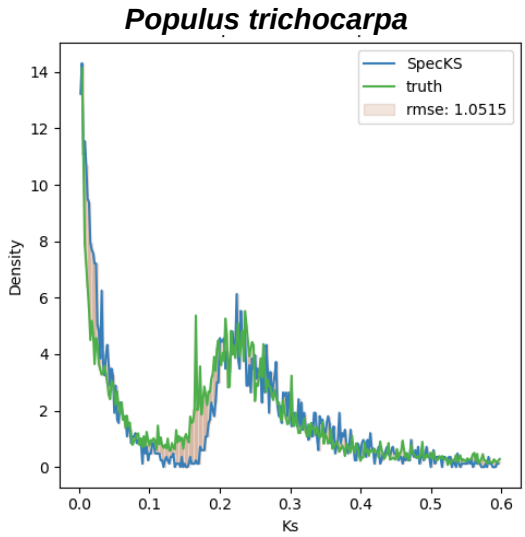


Figure 7. Using SpecKS to reproduce Ks empirical histograms from transcriptomic data.
(top) SpecKS-simulated histogram in blue and the histogram from transcriptomic data in green.
(bottom) The difference between the true and simulated histograms (y-axis), by Ks bin (x-axis).



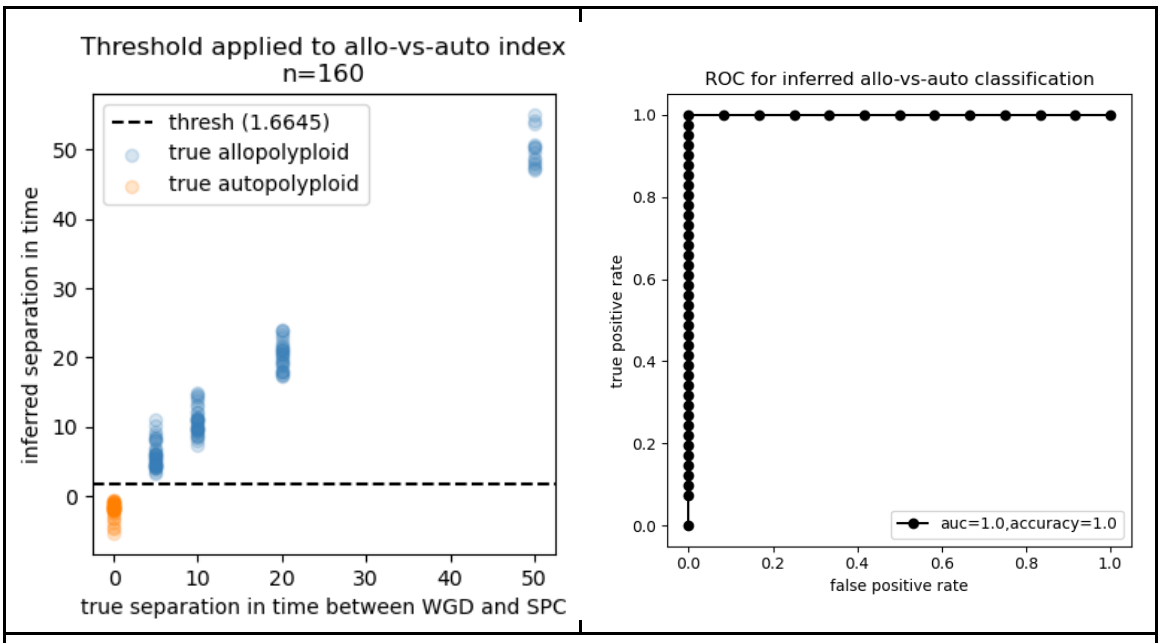


Figure. 8. The allo-vs-auto discrimination model.
(left) The allo-vs-auto discrimination threshold applied to the allo-vs-auto discrimination metric. The x-axis gives the true separation in time between SPC and WGD, as SPC-WGD in MYA. The y-axis gives the inferred time between SPC and WGD (i.e., the discrimination metric). (right) Accuracy of inference of historical allo versus autopolyploidy represented as an ROC plot. WGD times are [80,70, 60, 50, 40, 30, 20,10] as given in Table 3.

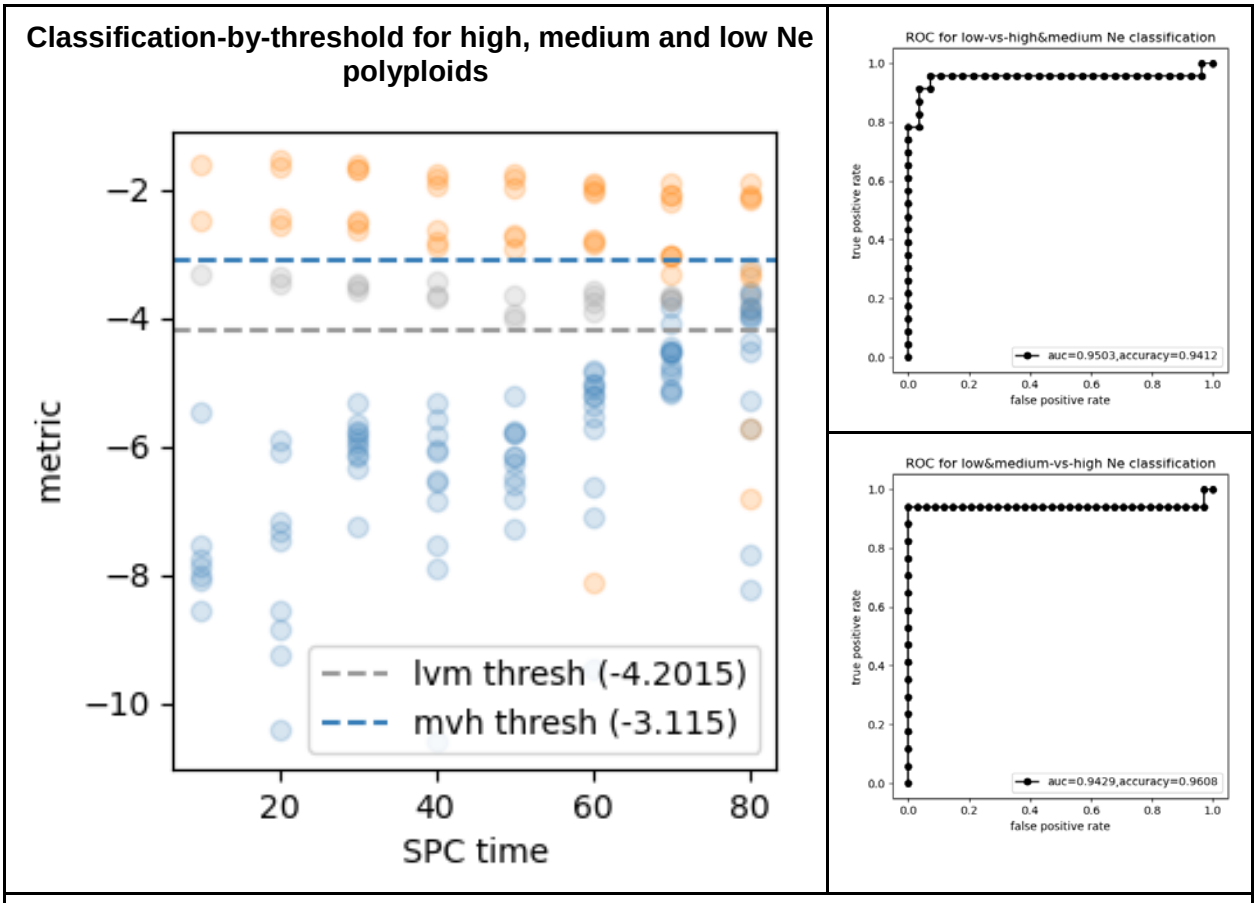


Figure 9. The low-vs-high Ne discrimination model.
(left) The low-vs-high Ne discrimination threshold applied to simulated polyploids. The x-axis gives the SPC time. The y-axis gives the high-vs-low Ne discrimination metric. (right) The ROC plot to assess accuracy of inference into high versus low Ne, across threshold values. Red, gray and blue represent simulations ran with low, medium and high Ne, respectively. Note: Figs. 8&9 use the same simulated dataset, whose input parameters formed a 3-dimensional grid over SPC, WGD and Ne*GT. For fig 7a SPC times are from 10 to 80 MYA, by 10, with WGD offset by 0, 5, 10 and 50 MY. Ne*GT is 0.1, 1.0, 5.0, 10 and 20 MY, totaling 160 runs (Table 3). Fig7b is the same dataset, keeping only runs with spec time < 60, for a total of 85 runs. A 2:1 split was used between training and test data in both cases, and only test datasets were used in accuracy assessments, here shown using ROC plots.

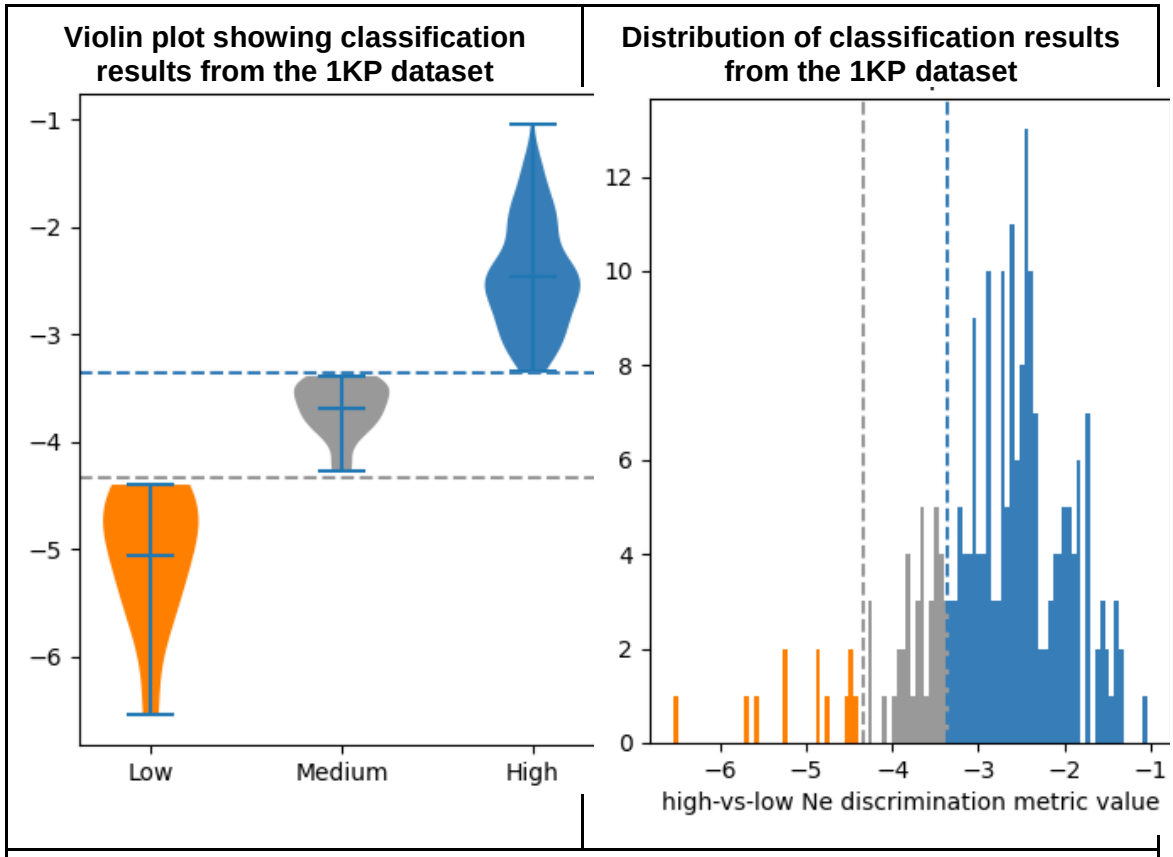


Figure. 10. The Ne-classifier results for the 1KP dataset
(a) Violin plots indicating the number of species in each category.
(b) Histogram giving the number of species in each category.
Red, gray and blue colors represent polyploid species categorized and having a low, medium or high Ne ancestral species at SPC time, respectively. X-axis gives the numerical value of the discrimination metric. Y-axis gives the number of species per bin.

Table 1. Subset of SpecKs configurable parameters		
Polyploid-specific. (Set for each simulated polyploid in the run):		
Parameter & default	**Default**	**Description**
SPC_time_MYA	0 MY	SPC 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 SPC time. For **autopolyploids**, use format: “impulse,1,1”. For **allopolyploids**, 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 (ie, **segmental allopolyploids**), multiple distributions may be given, with the last parameter being the proportion of genes which belong in each distribution. Details in Table S2.
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 small-scale duplications. Our default is chosen from [Lynch and Conery 2003]
WGD_half_life_M Y	31 MY	Half life of ohnologs (whole-genome duplications). Our default is based on [Guo 2013 and Maere 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 [Sterk 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. 1000 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 1996 and Koch 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-evoultionary_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.

Table 2. Summary of 1KP categorization results			
Ne 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

Table 3. Specs parameters used in dataset 1		
Polyploid	SPC_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 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.		