A Ks Distribution-Based Study on the Timescale of Angiosperm Evolution (Postprint)

Authors: Jiao Beibei, Wang Xiyin

Date: 2022-04-29T22:11:47Z

Abstract

Estimating species evolutionary timescales constitutes a crucial component of research on life’ s evolution. In recent years, numerous studies have revealed significant disparities in evolutionary rates among different genes and species, necessitating novel approaches to re-estimate the timing of evolutionary events. To re-estimate the evolutionary timeline of angiosperms, we established an evolutionary rate correction model based on genomic data and founded on the principle that shared polyploidization or divergence events should exhibit common Ks peaks. The results are as follows: (1) A comparative analysis of three common methods for obtaining Ks distributions identified the extraction of median Ks values from syntenic blocks as the optimal approach. (2) Simulations of Ks distribution under varying temporal accumulation coefficients v revealed that when v is assumed to follow a normal distribution, the Ks distribution displays a long-tail phenomenon. (3) Application of the correction method to angiosperms demonstrated that different angiosperm lineages exhibit synchronous patterns of radiation and adaptive evolution. Furthermore, although evolutionary rates among angiosperms vary significantly, partial consistency in rates persists across different clades. For instance, magnoliids exhibit the slowest evolutionary rate, followed by eudicots, while monocots display the fastest rate. Ultimately, this yielded a relatively reliable evolutionary timeline for species, providing phylogenetic and evolutionary support for plant research.

Full Text

Preamble

Time Scale of Angiosperm Evolution Based on Ks Distribution

Beibei Jiao[1], Xiyin Wang[1]*
[1]College of Life Sciences, North China University of Science and Technology, Tangshan 063210, Hebei, China

Abstract

Estimating evolutionary timescales is a critical component of life evolution research. Recent studies have revealed significant variation in evolutionary rates among different genes and species, necessitating novel approaches to re-estimate the timing of evolutionary events. To re-evaluate angiosperm evolutionary timescales, we developed an evolutionary rate correction model based on genomic data, founded on the principle that shared polyploidy or divergence events should exhibit common Ks peaks. Our results demonstrate: (1) Among three common methods for obtaining Ks distributions, extracting median Ks values from collinear blocks proves optimal; (2) Simulating Ks distribution evolution under varying time-accumulation coefficients (v) reveals long-tail phenomena when v follows a normal distribution; (3) Applying this correction method to angiosperms reveals synchronous radiative and adaptive evolution across different lineages. Although evolutionary rates vary significantly among angiosperms, partial consistency exists between different branches: Magnoliids exhibit the slowest rate, followed by eudicots, with monocots showing the fastest rate. Ultimately, we established a relatively reliable evolutionary timeline for species, providing phylogenetic and evolutionary support for plant research.

Keywords: Ks distribution, angiosperms, time correction, phylogenetic tree, evolutionary rate

Introduction

The origin and early rapid evolution of angiosperms and their timing have long been focal points in biological research. Current methods for estimating species evolutionary timescales primarily rely on the molecular clock hypothesis, which uses fossil times from specific taxa as calibration points and assumes uniform or similar evolutionary rates across species based on sequence similarity of selected genes to estimate node ages in phylogenetic trees (Tang et al., 2002; Donoghue & Yang, 2016; Luo et al., 2020). However, recent research demonstrates that molecular clocks vary significantly among species, indicating substantial differences in evolutionary rates (Wang et al., 2015; Wang et al., 2017; Wang et al., 2018; Wang et al., 2019), with different epochs exhibiting distinct evolutionary speeds (Luo & Zhang, 2000; Smith & Donoghue, 2008). Moreover, estimates of molecular evolutionary rates show considerable variation across studies (Lanfear et al., 2010).

Additionally, the fossil times used for calibration substantially impact estimated timescales, which consequently shift as more fossils are discovered and more accurate dating becomes available (Hug & Roger, 2007; Wang et al., 2015; Silvestro et al., 2021). Genome sequencing has revealed recurrent historical polyploidization events (Ren et al., 2018). These events duplicate all genes in the genome, and ancient homologous regions often retain substantial numbers of duplicated genes, forming intragenomic or intergenomic collinear homologs (Jiao et al., 2011). Analysis of these collinear homologs provides a crucial approach for revealing ancient polyploidization or species divergence events and determining their timing and scale. Following polyploidization, plant genomes often become highly unstable, with evolutionary rates varying significantly. These duplicated genes typically evolve more rapidly due to reduced selective constraints (Wang et al., 2016). For instance, studies of Cucurbitaceae genomes revealed that melon evolves most slowly, while watermelon and cucumber evolve 23.6% and 27.4% faster, respectively (Wang et al., 2018).

Synonymous substitution rate (Ks) generally does not alter amino acid composition and is considered unaffected by natural selection, making Ks distributions a common criterion for identifying historical polyploidization or divergence events (Vanneste et al., 2013). Based on the principle that shared evolutionary events should exhibit identical Ks peaks, Wang et al. first proposed a Ks peak-based correction method for estimating evolutionary timescales, which has gained recognition and wide application (Zhuang et al., 2019; Shang et al., 2020; Song et al., 2020; Yang et al., 2020; Song et al., 2021; Wang et al., 2021). For example, two teams independently analyzed water lily (Zhang et al., 2020a) and fox nut (Yang et al., 2020) genomes, with Yang et al.' s Ks peak-corrected estimate for an ancient polyploidization event (later confirmed as shared across Nymphaeales) aligning closely with another team' s transcriptome-based timescale for Nymphaeales. Accurate Ks peak extraction is critical for precise timescale estimation, yet current methods for obtaining Ks distributions lack standardization and frequently exhibit long-tail phenomena (Tang et al., 2008). The causes of these long tails and their impact on Ks peaks remain unclear.

Currently, over 400 angiosperm genomes have been sequenced at various levels, enabling genome-wide understanding of angiosperm evolution (Kress et al., 2022). Whole-genome data can effectively mitigate effects of horizontal gene transfer and rate variation among lineages on phylogenetic inference. Therefore, new methods are urgently needed to re-estimate angiosperm evolutionary timescales using whole-genome data. This study compares three methods for obtaining Ks distributions to identify which yields peaks closest to reality, investigates the causes of long-tail phenomena through simulation, and develops a whole-genome-based Ks distribution correction model that distinguishes between shared polyploidy and shared early divergence. We re-estimated timescales for evolutionary events in 44 representative angiosperm genomes to establish a relatively reliable angiosperm evolutionary timeline, facilitating deeper understanding of angiosperm diversity, phylogeny, and genome evolution patterns.

Materials and Methods

Genome Data Collection

We collected 44 high-quality chromosome-level angiosperm genomes, comprising 43 families and 39 orders, primarily from NCBI and Phytozome (Table 1).

Collinearity and Ks Distribution Analysis

Collinearity Analysis. We performed collinearity analysis using WGDI v0.5.3 (Sun et al., 2021). First, BLASTP identified intragenomic or intergenomic gene similarities. Subsequently, WGDI' s `-d` submodule generated dot plots of homologous genes, and the `-icl` submodule extracted collinear genes.

Ks Distribution Calculation. Ks distributions were primarily generated using WGDI. The `-ks` submodule called PAML (Yang, 2007) to calculate Ks values for collinear gene pairs. The `-bi` submodule integrated collinearity and Ks results, while `-bk` visualized Ks distributions as dot plots (Figure 1A). Based on known polyploidy or divergence events within or between species, WGDI's `-c` submodule filtered collinear fragments to retain only those generated by target events. The `-kp` submodule then obtained Ks distributions (Figure 1B), and the `-pf` submodule performed separate fitting for different events (Figure 1C).

Results

Analysis of Ks Distributions and Long-Tail Phenomena

Ks distributions commonly serve as evidence for historical polyploidization or divergence events. Three primary methods exist for obtaining Ks distributions: (1) Using clustering software like OrthoMCL (Li et al., 2003) to identify paralogous gene pairs, then calculating and plotting their Ks values; (2) Performing genome collinearity analysis first, then calculating and plotting Ks values for collinear gene pairs; (3) Extracting median Ks values from collinear blocks before plotting. Method 1 lacks collinearity analysis and often includes numerous tandem duplicates that distort Ks distributions. Methods 2 and 3 both employ collinearity analysis. When plotting Ks values from collinear blocks (length >5) as dot plots (Figure 1A, using rice as an example), most fragments appear as green dots (e.g., between chromosomes 8 and 9), consistent with rice' s recent polyploidization event. The similar coloring of most dots indicates minimal Ks variation. Normal distribution fitting (bandwidth=0.01, homo range 0.3-1) of median Ks values (Method 3), mean values, and all gene pairs (Method 2) from collinear regions (Figure 2B) shows that Method 2 produces no clear peak and exhibits long tails. Methods 3 and the block mean approach yield distinct peaks with more concentrated data. Since the median robustly estimates central tendency and its peak color closely matches the dot plot, median Ks values better approximate true Ks peaks. We therefore fitted normal distributions to Method 3 Ks distributions to extract peaks (Figure 1C).

To further investigate long-tail phenomena, we simulated Ks distribution evolution under varying evolutionary rates. We initially assumed Ks distributions followed a normal distribution $X \sim N(\mu, \sigma)$, where mean μ (peak) and standard deviation σ were constants. Molecular clock theory suggests relatively constant gene evolutionary rates, so we defined v ($v > 0$) as the time-accumulation coefficient representing Ks value accumulation over time, simulating constant evolutionary rates. However, other studies indicate molecular clocks are not strictly isochronous, so we also assumed v follows a normal distribution $N(\mu_v, \sigma_v)$ and simulated both scenarios. Ks values were iteratively updated as X', with iteration count n.

When v was constant, setting the initial Ks distribution as $X \sim N(\mu, \sigma)$ with $\mu = 0.2$, $\sigma = 0.05$, $v = 1.02$, and $n = 100$, we plotted distributions every 10 iterations (Figure 2A). Ks peaks gradually increased while distributions remained perfectly normal without long tails.

When v followed a normal distribution, with initial settings $X \sim N(\mu, \sigma)$, $\mu = 0.2$, $\sigma = 0.05$, $\mu_v = 1.02$, $\sigma_v = 0.02$, and $n = 100$, we plotted results every 10 iterations (Figure 2B). Ks peaks increased over time, but distributions deviated from normality and exhibited pronounced long tails. This scenario more closely resembles reality, suggesting evolutionary rates are not constant but vary across epochs, potentially following a normal distribution. When we re-extracted peaks from simulated distributions via Gaussian fitting, no significant differences were found compared to constant-rate simulations (Table 2), indicating that long-tail phenomena minimally affect Ks peak extraction. However, previous research shows that Ks values >1 suffer from saturation effects that intensify with increasing Ks (Vanneste et al., 2013). Our simulated peaks approached 1, suggesting saturation effects may influence estimated peaks as Ks values increase.

Ks Distribution Correction Method

Angiosperm genomes often experience multiple polyploidization events. Significant rate variation among species leads to different Ks peaks for shared polyploidy events. The core principle of our correction method is to align these shared event peaks. The approach differs for shared polyploidy versus shared early divergence.

For shared polyploidy events between species A and B, the event should have occurred simultaneously, yielding equal Ks peaks (Figure 3A). Yellow blocks represent shared polyploidy, with corresponding time ranges from the event to the present (green brackets). Due to different evolutionary rates, actual AA_{Ks} and BB_{Ks} values differ. Assuming post-polyploidy rates v_A and v_B, and ancestral rate v from the polyploidy event to divergence node O, species A' s correction coefficient is $\lambda_A = v/v_A$, and species B' s is $\lambda_B = v/v_B$. Thus, corrected divergence AB_{Ks} becomes $AB_{Ks\_correction} = \lambda_A \lambda_B AB_{Ks}$ (Yang et al., 2020).

For species A and B lacking shared polyploidy but sharing early divergence, we use outgroups for correction (Figure 3B). Outgroup species C, D, and E diverged from the A-B ancestor at point P, so Ks peaks between C and A/B, and between D and A/B should be equal: $CA_{Ks} = CB_{Ks}$ and $DA_{Ks} = DB_{Ks}$. Due to rate variation, these are typically unequal. Following previous assumptions, we derive correction coefficients: $CA_{Ks\_correction} = \lambda_C \lambda_A CA_{Ks}$, $CB_{Ks\_correction} = \lambda_C \lambda_B CB_{Ks}$, $DA_{Ks\_correction} = \lambda_D \lambda_A DA_{Ks}$, $DB_{Ks\_correction} = \lambda_D \lambda_B DB_{Ks}$.

Using multiple outgroups yields more accurate λ_B relationships, which we average: $\lambda_B = \text{mean}(\lambda_{B1}, \lambda_{B2}, ...)$.

Timescale Correction of the Angiosperm Phylogenetic Tree

Many phylogenetic studies place angiosperm origins in the Triassic, 225-240 million years ago (Magallón, 2010), coinciding with the origin of core pollinating lepidopteran insects (~230 Mya) (Li et al., 2019). The relationships among Amborellales, Nymphaeales, and the five major core angiosperm clades remain unresolved, with evidence suggesting rapid radiative divergence of the core angiosperm ancestor (Yang et al., 2020). Therefore, we used Amborellales as a reference without resolving its relationship with Nymphaeales, and assumed the five major clades diverged within the same timeframe. Using the core eudicot-shared γ event (115-130 million years ago, Mya) as calibration, we corrected timescales for 44 angiosperm genomes (Table 1, Figure 4). The corrected timescale shows rapid radiative evolution in monocot, eudicot, and magnoliid ancestors around 130 Mya, consistent with previous findings (Zhang et al., 2020b). Additionally, numerous polyploidization events occurred in the Early Cretaceous (130 Mya), at the Cretaceous-Paleogene boundary (66 Mya), and in the Miocene (20 Mya, near glaciation periods), supporting the non-random distribution of WGD events across angiosperm phylogeny (Wu et al., 2020).

Despite significant rate variation among species, evolutionary rates show partial consistency within clades. Since corrected Ks peaks should be equal, larger Ks peaks indicate faster rates. Comparing Ks peaks between magnoliids, eudicots, and monocots relative to Amborella reveals magnoliids (mostly woody) evolve slowest, followed by eudicots (mostly shrubs), with monocots (mostly herbaceous) fastest (Table 3), consistent with findings that perennial woody plants have slower molecular evolutionary rates than herbs (Lanfear et al., 2013). Comparing polyploidization event times with pre- and post-correction Ks peaks (Figure 5) shows pre-correction peaks are not linear with time—more ancient events do not correspond to larger Ks peaks. Post-correction peaks are directly proportional to time, demonstrating the necessity of Ks peak correction for accurate timescale estimation.

Discussion and Conclusion

Traditional angiosperm timescale estimation relies heavily on the molecular clock hypothesis, but widespread evolutionary rate heterogeneity severely compromises its accuracy. Wang et al.' s Ks distribution-based correction method yields compelling timescales. This study compared three Ks distribution extraction methods, establishing that median Ks values from collinear blocks best represent true Ks peaks. We further investigated long-tail phenomena through simulation, demonstrating that evolutionary rates are not constant but vary across epochs. When rates follow a normal distribution, Ks distributions develop pronounced long tails, though this minimally affects peak extraction accuracy. Previous studies indicate Ks values >1 suffer saturation effects that intensify with increasing Ks (Vanneste et al., 2013). Our simulated peaks approached 1, suggesting potential saturation effects on estimated peaks as Ks values increase.

We also detailed the Ks peak-based correction procedure. While previous studies described shared polyploidy and shared early divergence separately, this comprehensive description facilitates deeper understanding and broader application. Applying this method to 44 high-quality angiosperm genomes yielded timescales consistent with recent publications (Li et al., 2019; Wu et al., 2020). Results demonstrate significant rate variation among angiosperm genomes but partial consistency between branches, with synchronous radiative and adaptive evolution across lineages. As more high-quality angiosperm genomes become available and fossil dates are more accurately determined, angiosperm evolutionary timescales will become increasingly refined, benefiting phylogenetic reconstruction and deeper understanding of species evolution.

References

Donoghue PC, Yang Z, 2016. The evolution of methods for establishing evolutionary timescales. *Phil Trans Roy Soc B: Biol Sci*, 371(1699).

Hug LA, Roger AJ, 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. *Mol Biol Evol*, 24(8): 1889-1897.

Jiao Y, Wickett NJ, Ayyampalayam S, et al., 2011. Ancestral polyploidy in seed plants and angiosperms. *Nat*, 473(7345): 97-100.

Kress WJ, Soltis DE, Kersey PJ, et al., 2022. Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proc Nat Acad Sci*, 119(4).

Lanfear R, Ho SYW, Jonathan Davies T, et al., 2013. Taller plants have lower rates of molecular evolution. *Nat Comm*, 4(1): 1879.

Lanfear R, Welch JJ, Bromham L, 2010. Watching the clock: studying variation in rates of molecular evolution between species. *Trend Ecol Evol*, 25(9): 495-503.

Li HT, Yi TS, Gao LM, et al., 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat Plant*, 5(5): 461-470.

Li L, Stoeckert CJ, Roos DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9): 2178-2189.

Luo A, Duchêne DA, Zhang C, et al., 2020. A simulation-based evaluation of tip-dating under the fossilized birth–death process. *Syst Biol*, 69(2): 325-344.

Luo J, Zhang YP, 2000. Molecular clock and its existing problems. *Acta Anthropol Sinica*, 19(2): 151-159.

Magallón S, 2010. Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. *Syst Biol*, 59(4): 384-399.

Ren R, Wang H, Guo C, et al., 2018. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol Plant*, 11(3): 414-428.

Shang J, Tian J, Cheng H, et al., 2020. The chromosome-level wintersweet (*Chimonanthus praecox*) genome provides insights into floral scent biosynthesis and flowering in winter. *Genome Biol*, 21(1): 1-28.

Silvestro D, Bacon CD, Ding W, et al., 2021. Fossil data support a pre-Cretaceous origin of flowering plants. *Nat Ecol Evol*, 5(4): 449-457.

Smith SA, Donoghue MJ, 2008. Rates of molecular evolution are linked to life history in flowering plants. *Sci*, 322(5898): 86-89.

Song X, Sun P, Yuan J, et al., 2021. The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. *Plant Biotechnol J*, 19(4): 731-744.

Song X, Wang J, Li N, et al., 2020. Deciphering the high-quality genome sequence of coriander that causes controversial feelings. *Plant Biotechnol J*, 18(6): 1444-1456.

Sun P, Jiao B, Yang Y, et al., 2021. WGDI: a user-friendly toolkit for evolutionary analyses of ancestral whole-genome duplications and karyotypes. *BioRxiv*. https://doi.org/10.1101/2021.04.29.441969.

Tang H, Wang X, Bowers JE, et al., 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res*, 18(12): 1944-1954.

Tang XH, Lai XL, Zhong Y, 2002. Molecular clock hypothesis and fossil record. *Earth Sci Front*, 9(2): 465-474.

Vanneste K, Van de Peer Y, Maere S, 2013. Inference of genome duplications from age distributions revisited. *Mol Biol Evol*, 30(1): 177-190.

Wang J, Sun P, Li Y, et al., 2018. An overlooked paleotetraploidization in Cucurbitaceae. *Mol Biol Evol*, 35(1): 16-26.

Wang J, Sun P, Li Y, et al., 2017. Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol*, 174(1): 284-300.

Wang J, Yuan J, Yu J, et al., 2019. Recursive Paleohexaploidization shaped the durian genome. *Plant physiol*, 179(1): 209-219.

Wang S, Xiao Y, Zhou ZW, et al., 2021. High-quality reference genome sequences of two coconut cultivars provide insights into evolution of monocot chromosomes and differentiation of fiber content and plant height. *Genome Biol*, 22(1): 1-25.

Wang X, Guo H, Wang J, et al., 2016. Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. *New Phytol*, 209(3): 1252-1263.

Wang X, Wang J, Jin D, et al., 2015. Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Mol Plant*, 8(6): 885-898.

Wu S, Han B, Jiao Y, 2020. Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol Plant*, 13(1): 59-71.

Yang Y, Sun P, Lv L, et al., 2020. Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nat Plant*, 6(3): 215-222.

Yang Z, 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8): 1586-1591.

Zhang L, Chen F, Zhang X, et al., 2020a. The water lily genome and the early evolution of flowering plants. *Nat*, 577(7788): 79-84.

Zhang L, Wu S, Chang X, et al., 2020b. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant Cell Environ*, 43(12): 2847-2856.

Zhuang W, Chen H, Yang M, et al., 2019. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet*, 51(5): 865-876.

Table 1. List of the 44 angiosperms and genome data sources

Table 2. Ks peaks under simulations at different evolution rates

Table 3. Ks peaks between some species of Mesangiospermae and *Amborella trichopoda*

Figure 1. Ks distribution. (A) Synteny blocks of the *Oryza sativa* genome; (B) Fitted distribution of Ks values for synteny blocks; (C) Kernel density of Ks values for synteny blocks.

Figure 2. Simulation results of Ks distribution at different evolution rates. (A) Simulation of Ks distribution at a constant evolution rate; (B) Simulation of Ks distribution under a normal distribution of evolution rates.

Figure 3. Principle of the Ks distribution correction method. (A) Shared polyploidy events; (B) Shared early divergence.

Figure 4. Angiosperm phylogenetic tree after time correction.

Figure 5. Relationship between Ks peaks and time before and after correction.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.