Chapter 7

Constraining Whole-Genome Duplication Events in Geological Time

James W. Clark and Philip C. J. Donoghue

Abstract

The timing of whole-genome duplication (WGD) events is crucial to understanding their role in evolution and underpins many hypotheses linking WGD to increased diversity and complexity. As such, means of estimating the timing of the WGD events relative to their macroevolutionary outcomes are of considerable importance. Molecular clock methods facilitate direct estimation of the absolute timing of WGD events, integrating information on the rate of sequence evolution between species while accommodating the uncertainty inherent to the fossil record. We present an explanation of the best practice for constructing fossil calibrations and estimating the age of WGD events via molecular clock methods in the program MCMCtree, with an example dataset based on a well-characterized WGD event within the flowering dogwoods (*Cornus*). The approach presented herein allows for the estimation of the age of WGD events and subsequent speciation events, allowing the relationship between WGD and the macroevolutionary outcomes to be explored. In our example, we show that in the case of flowering dogwoods, the WGD event long predates the end-Cretaceous mass extinction and that the two events may be independent.

Key words Molecular clock, Polyploidy, Whole-genome duplication, Fossil calibration, Cornus

1 Introduction

Whole-genome duplication (WGD; Polyploidy) has occurred in the evolutionary history of several major land plant lineages, as well as in fungi and animals. These events are often invoked as agents of macroevolutionary change [1], and instances of WGD have been linked to morphological innovations [2], biogeographic shifts [3], lineage longevity [4], and increased rates of species diversification [5]. Each of these evolutionary hypotheses depends on an estimate of the timing of the WGD event to justify a correlation, let alone causation. The timing of WGD events can be considered in both relative and absolute terms. The relative (or phylogenetic) timing of a WGD event identifies the lineage (branch on a phylogenetic tree) in which the WGD event occurred, based on identifying which species do and do not exhibit genomic evidence of that event.

Yves Van de Peer (ed.), *Polyploidy: Methods and Protocols*,
Methods in Molecular Biology, vol. 2545, https://doi.org/10.1007/978-1-0716-2561-3_7,

However, many hypotheses that relate macroevolutionary consequences to WGD events, such as increased rates of diversification or morphological innovation, are also dependent on the absolute (geological) timing of the event. Following WGD, the process of diploidization or fractionation is believed to result in a time lag between the duplication itself and any proposed macroevolutionary outcomes [6]. The extent of this lag is important, since a drawn-out period of diploidization may explain the disparate outcomes of WGD across sister lineages [7].

Methods of dating WGD events can be categorized as phylogenetic and nonphylogenetic. The primary nonphylogenetic method is to derive the rate of nonsynonymous substitutions (*Ks*) among paralogous gene pairs [8, 9]. Across all paralogous pairs, the distribution of *Ks* values should exhibit a peak if multiple pairs duplicated at the same time, as would be the case in a WGD event. Once identified, this peak can be converted into units of geological time by selecting an external calibration. However, converting the substitution rate using a single-point calibration is problematic since it a) assumes a strict rate of molecular evolution among loci and b) places excessive confidence in the calibration. Further, these methods are known to fail when dating increasingly ancient WGD events since saturation in the rate of substitutions can obscure signal [9, 10].

Phylogenetic approaches rely on the reconciliation of the evolutionary history of genes and species. The most straightforward phylogenetic approach is to bracket the age of gene duplication events between relevant species divergences. The WGD must be older than all of the lineages that underwent the event, but younger than the divergence of those that did not (Fig. 1). Thus, the simplest way of estimating the age of a WGD event is to provide a range between these two ages. However, in order to provide a reasonable estimate of the age of the WGD, this approach relies on a few assumptions. First, that the ages of the species divergences are known and reliably estimated. When characterizing novel WGD events in lineages that previously lacked genomic resources, it is possible that the evolutionary timeline will be poorly understood. Second, that the time between the two species divergences is small relative to their geologic age. Since this method does not directly estimate the timing of WGD, species that sit on long evolutionary branches will only be able to constrain the age of WGD events to unhelpfully broad intervals. Given these considerable shortfalls, we outline a means of directly estimating the age of WGD events. This approach incorporates both phylogenomic data and information from the fossil record in a molecular clock analysis, capable of precisely and accurately coestimating the timing of WGD and subsequent species divergence [11, 12]. To demonstrate this approach, we provide theoretical and practical examples, including the estimation of the timing of a WGD event shared by extant

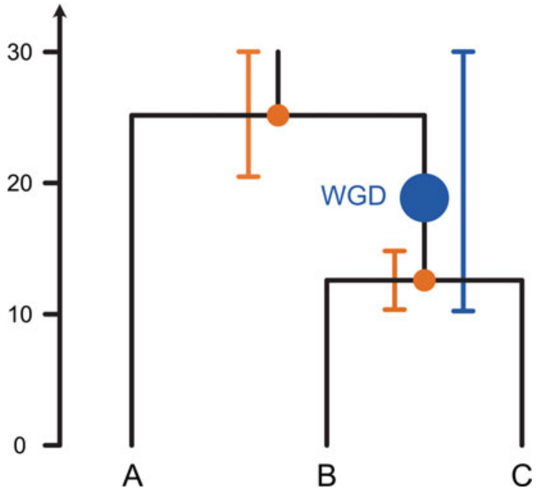

Fig. 1 Constraining the age of whole-genome duplication (WGD) events. Species B and C have undergone a WGD event (blue dot) following their divergence from species A. The timing of species divergence (orange dots) between species B and C and B+C and A is shown as orange bars representing the confidence interval. The WGD event is thus constrained by the minimum divergence time between B and C (10 Ma) and also the maximum divergence time between B+C and A (30 Ma)

members of the flowering dogwood genus, *Cornus* [13]. To perform the analyses described in this chapter, a dataset and set of control files can be found at https://doi.org/10.6084/m9.figshare.16867108.

2 Materials

2.1 Required Data

Molecular clock methods integrate the rate of molecular evolution, measured in substitutions per site of amino acid or nucleotide sequences, with an external calibration. The external calibration may vary, but it is typically provided in the form of a temporal constraint informed by fossil evidence [14]. Therefore, a minimum requirement is molecular sequence data from all species in question that contain a signal of the WGD event and a set of fossil calibrations that inform the divergence times of said species.

Molecular sequence data comes in the form of individual gene families. The most important criterion in selecting a gene family for the analysis is that it contains a clear signal of the WGD event and that the species relationships it predicts do not differ overly from the known species tree. This can be investigated using high-throughput bioinformatic methods, such as gene tree–species tree reconciliation programs, or by manually reconstructing the gene tree and visually inspecting the results.

Beyond this, the quality of the molecular data present in a gene family is measured using the same criterion that is typical of phylogenetic approaches. Each gene family should ideally contain one sequence for each species or paralog in the analysis, though due to processes of gene loss or transcriptomes with lower coverage, some degree of missing data may be acceptable. The length of the gene is not always a reliable indicator of quality, but, in general, longer genes are more likely to contain useful signal over longer evolutionary distances.

Fossil calibrations are established based on the oldest robustly evidenced member of a clade (*see* Subheading 2.3). The selection of species with molecular sequence data and suitable calibrations should not be considered in isolation; taxonomic sampling that maximizes information from the fossil record will result in more accurate estimates of divergence times and, in turn, WGD. Finally, a phylogeny of species relationships is required. This should represent the current best hypothesis of relationships among all species as it is not estimated during these analyses.

2.2 Annotating Sequences

Unlike in a typical phylogenomic or molecular clock analysis, each species may be represented more than once in the tree as a result of a duplication event, forming two or more paralogy groups. Species need to be labeled so as to identify the species or taxon but also the paralogy group to which it belongs (Fig. 2). Paralogy groups must be consistently labeled *within* gene families, but if multiple gene families are being used, then the assignment to either paralogy group *between* gene families is arbitrary (Fig. 2).

2.3 Constructing Calibrations

Several criteria are required to construct a fossil calibration following best practice [15]. These are outlined in the example below:

1. *Node:* Cornaceae + Alangiaceae – Curtisiaceae.
2. *Fossil taxon: Eydeia jerseyensis* [CUPC-1601, Cornell University Palaeobotanical Collection, Cornell University, Ithaca] from the South Amboy Fire Clay Member of the Raritan Formation (Turonian), New Jersey, USA [16].
3. *Phylogenetic justification:* Phylogenetic analysis of morphological characters placed the extinct genus *Eydeia* on the stem of the NMD Group (Nyssaceae, Davidiaceae, Mastixiaceae), in turn sister to Cornaceae + Alangiaceae [16].
4. *Minimum age:* 89.37 Ma.
5. *Maximum age:* 115 Ma.
6. *Age justification:* The age of the Raritan formation has been derived from palynology [17]. The South Amboy Fire Clay member correlates with the *Complexipollenites exigua–Santalacites minor* palynological zone [18, 19], which is considered middle to late Turonian. We follow Atkinson et al. [16] and

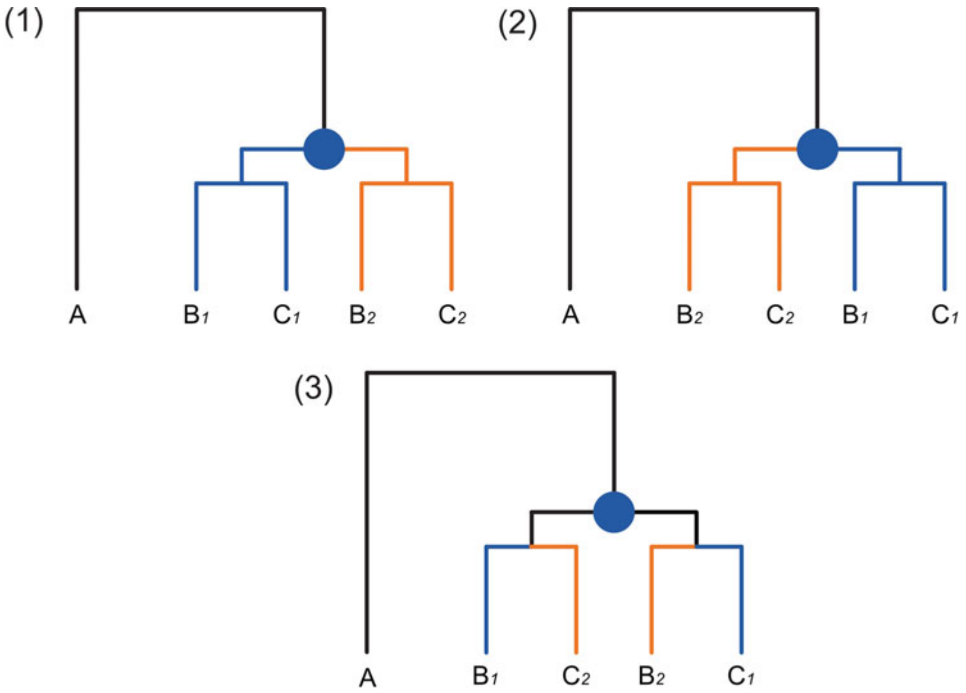

Fig. 2 Annotating sequences across paralogy groups. After duplication (blue dot), each species is represented more than once in the gene tree. Species should be consistently labeled *within* paralogy groups (1 and 2), though between gene families the assignment to either paralogy group is not important (1 vs. 2). However, inconsistent labeling within paralogy groups (3) is incorrect

conservatively consider it latest Turonian, and thus establish a minimum age based on the age of the Turonian-Coniacian boundary of the Turonian, 89.75 ± 0.38 [20].

The maximum age is derived from the most comprehensive analysis of angiosperm divergence times to date [21], which estimated a maximum age for crown group Cornales of 115 Ma.

First, the node on the phylogeny which the fossil is calibrating is clearly stated, in this case, the divergence of Cornaceae and Alangiaceae from Curtisiaceae. Second, the specimen on which the calibration is based, along with where it was collected and where it is housed, is provided. Next, the fossil's inclusion within the stated clade is justified, often based on a formal phylogenetic analysis or the presence of unambiguous synapomorphies established in a previous phylogenetic analysis considering phenotypic data. Fossils are usually used to inform clade-age minimum constraints. Their ages are rarely known directly but, rather, established through correlations between the rock strata in which they are found and strata that have been dated directly. This indirect approach leads to minimum and maximum age interpretations of the fossil, the youngest of which is taken to establish minimum clade-age constraints [14]. These are not sufficient and must be supplemented by maximum clade-age constraints that are commonly established by fitting an arbitrary mathematical distribution to the minimum clade-age constraint to express some visceral perception of how well the fossil on which it is based approximates the true clade age [22]. Alternatively, maximum clade-age constraints can be informed based on evidence of the absence of fossil representatives of a clade constraints [23]. In either instance, justification must be provided for how these ages were derived. Fossil calibrations should be carefully evaluated, and if a full justification is not presented, then a citation to a relevant paper with such a justification should be provided [15].

Different software packages have different means of implementing fossil calibrations, but they are commonly specified as a probability distribution. Different probability distributions reflect different interpretations of the fossil record. A uniform distribution between a minimum and maximum provides a conservative constraint, since no greater probability is assigned to any age [24]. Other probability distributions, such as exponential or Cauchy, may specify a greater probability of a certain node age. Such distributions are commonly applied, yet there is often little justification for weighting the probability toward a particular age and so, though less informative, a uniform distribution is preferable [24]. The R package "MCMCTreeR" allows the specification and visualization of multiple different probability distributions and can be used to produce an input phylogeny with fossil calibrations annotated to the relevant nodes [25].

When assigning fossil calibrations, it is important to also consider the uncertainty of our assessment of the fossil record. Though many fossils have been assigned as members of extant lineages, there is always the possibility of error, especially as hypotheses of clade membership are not always tested. In MCMCTree, this uncertainty can be modeled in the form of "soft" constraints, where a probability tail reflects the possibility that a node may be younger or older than the provided minimum or maximum [26]. These are specified as follows:

```
B(0.8937,1.15,0.01,0.01)
```

The "B" specifies that this calibration contains both a minimum and a maximum constraint. When both a minimum and a maximum are provided, the distribution between them is always uniform. The first number, 0.8937, specifies the minimum constrain, in hundreds of millions of years, followed by the maximum. The final two numbers represent the probability that either the minimum and maximum age can be exceeded, in this case 1%. These soft bounds are especially important when specifying maximum ages.

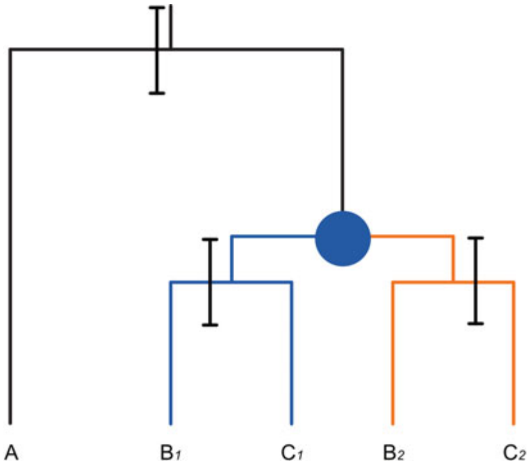

Fig. 3 Cross-calibration. Black bars represent fossil calibrations that can constrain the divergence of B from C and B+C from A. The calibration that constrains the divergence of B from C is present twice in the gene tree after the duplication (blue dot). If this calibration is modeled identically in each paralogy group, then they are cross-calibrated

When dating duplications within gene families, the same speciation node can be represented more than once within a tree in different paralogy groups (Fig. 3). In these instances, "cross-calibration" should be employed, where the same calibration with the same probability distribution is applied to the equivalent speciation nodes in both paralogy groups [27]. A more powerful implementation of this approach, "cross-bracing," has also been proposed, whereby the equivalent speciation nodes are constrained to be exactly the same age [27]. To date, this has only been implemented in the software package BEAST2 [27, 28].

2.4 Examining the Prior

Within MCMCTree, it is possible to examine the combined prior probabilities of the underlying tree and the fossil constraints. This is known as the effective, or joint, time prior, and it can be estimated by running the molecular clock analysis without the sequence data. This is a fast and important step in the analysis, since the tree and fossil priors can interact. For example, the tree topology can truncate the prior for the node ages of some nodes to achieve the expectation that ancestral nodes are older than their descendants. It is important to see the effective prior and to make sure that it does not conflict with the priors that you specified [29]. To run this analysis, we can run the prior control file (`prior.ctl`), where MCMCTree is told to ignore the sequence data:

```
useData = 0
```

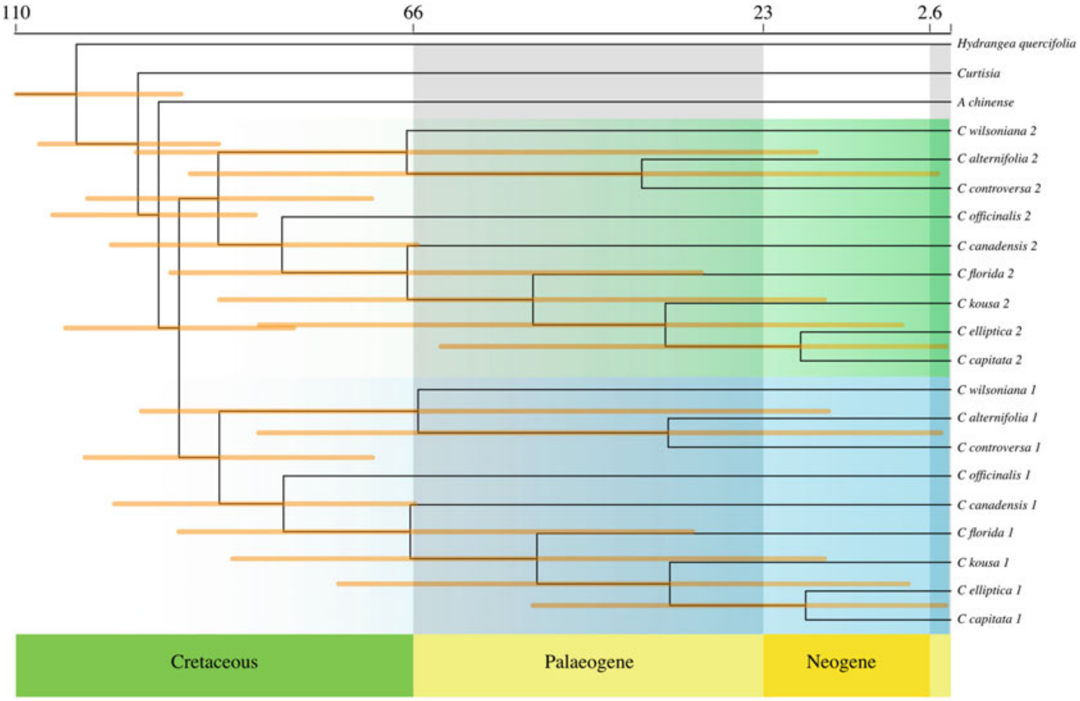

Fig. 4 The effective prior, estimated by running the analysis without molecular sequence data. Divergence times are in millions of years before the presence, as indicated by the top bar. Orange bars represent the 95% highest posterior density (HPD) for each node, with each paralogy group colored blue or green

You will often observe a truncation of the prior that was specified. In our example, the node age prior assigned to the Cornaceae +Alangiaceae node was between 89.37 and 115 Ma, yet the 95% highest posterior density (HPD) for the effective prior is between 90 and 112 Ma (Fig. 4). This is caused by the interaction between the node age prior and the underlying tree model. This is also an opportunity to assess the effect of cross calibration (*see* Subheading 2.3). On each side of the duplication event, the effective priors on divergence times should be very similar; otherwise, a calibration may have been incorrectly specified.

2.5 Running an Analysis

We will describe a clock analysis using the normal approximation method in MCMCTree [30–32]. This is a two-step dating approach, that first estimates branch lengths and allows an approximation of the likelihood surface in the program codeml or baseml [30, 32], and then runs the clock model in MCMCTree. This method is relatively fast and tractable for large datasets. It also allows a high degree of control over all model parameters, all specified in a control file. The sequence data is contained in a Phylip file, consisting of 15 individual gene families that have been concatenated.

The first step in running the normal approximation method is to generate a set of temporary files for the program *codeml* to work with. This is done in the MCMCTree control file (`step_one.ctl`), where the usedata = 3 tells the program to begin the normal approximation method.

```
seed = -1
seqfile = alignment.phy
treefile = calibration_tree.txt
mcmcfile = step_1.txt
outfile = step_1.out
ndata = 1
seqtype = 2
usedata = 3
clock = 2
RootAge = U(1.15,0.001)
```

This will generate a set of files named "`tmp0001.*`". These are the temporary files that are provided to codeml. The most important is `tmp001.ctl`, which is a control file. By default, the file contains instructions to estimate branch lengths according to only a simple model of molecular evolution. Instead, we must change it to describe the appropriate model:

```
seqfile = tmp0001.txt
treefile = tmp0001.trees
outfile = tmp0001.out
noisy = 3
seqtype = 2
aaRatefile = jones.dat
fix_alpha = 0
alpha = 0.5
ncatG = 4
Small_Diff = 0.1e-6
getSE = 2
method = 1
```

Here, we have specified a model file "`jones.dat`" (the JTT model [33]) and have allowed the rate to vary across sites, according to a discrete gamma distribution with four categories and a shape parameter of alpha = 0.5 (a JTT+G4 model). Running codeml will produce a file (named "`rst2`") containing branch length information and the Hessian matrix required for the normal approximation method. To run the normal approximation, we must first rename "`rst2`" to "`in.BV`." Then, in the MCMCtree control file "`step_two.ctl`" change:

```
usedata = 3 to usedata = 2
```

This is now set up to run the analysis with the branch length information estimated from the sequence data. We must also specify some properties of the MCMC, such as the length, sampling frequency, and amount of burn-in.

```
burnin = 2000
sampfreq = 100
nsample = 15000
```

The overall length of the chain is the product of the sample frequency and the number of samples. So, above a sample frequency of every 100 and a total number of samples of 15,000 would specify a chain of length 1,500,000. The burn-in line specifies the number of samples to run the chain for before being discarded. Here, 2000 samples would require 200,000 generations, so the total chain length, including burn-in, is 1,700,000.

2.6 Interpreting and Visualizing Results

The output of the analysis is contained in two files. The treefile (“`FigTree.tre`) contains a Nexus format, and the file “mcmc.txt” contains information from each sampled generation of the Markov chain Monte Carlo (MCMC). First, we can check that the MCMC has run for a sufficient number of generations. The “mcmc.txt” file can be loaded directly into Tracer [34]. This provides the effective sample size (ESS) for each parameter in the analysis. Generally, ESS values greater than 200 are sufficient. The MCMC files from multiple independent runs can be loaded to further compare the posteriors. If each run has converged, the posterior distributions between runs should be the same.

A more intuitive way of looking at the results is to plot the time-scaled phylogeny. This can be done quickly in any tree visualizing program. Publication-ready figures can be produced directly from the MCMCtree output in MCMCtreeR with a set of simple commands [25]:

```
dated.tree <- MCMCtreeR::readMCMCtree(“FigTree.tre”)
```

```
 MCMC.tree.plot(dated.tree, analysis.type = "MCMCtree", time.
correction = 100, plot.type = "phylogram",lwd.bar=5, scale.res
= c("Period"), node.method = "bar")
```

Here, the nodes within the tree are scaled to the *mean* estimates, with the option of plotting bars to represent the 95% HPD. Note that while mean ages are useful for visualizing the tree (the nodes must be positioned somewhere), they may be a misleading (inaccurately precise) interpretation of the results that is likely to exclude the true divergence time [35]. Instead, it is essential to always report the 95% HPD intervals when presenting divergence

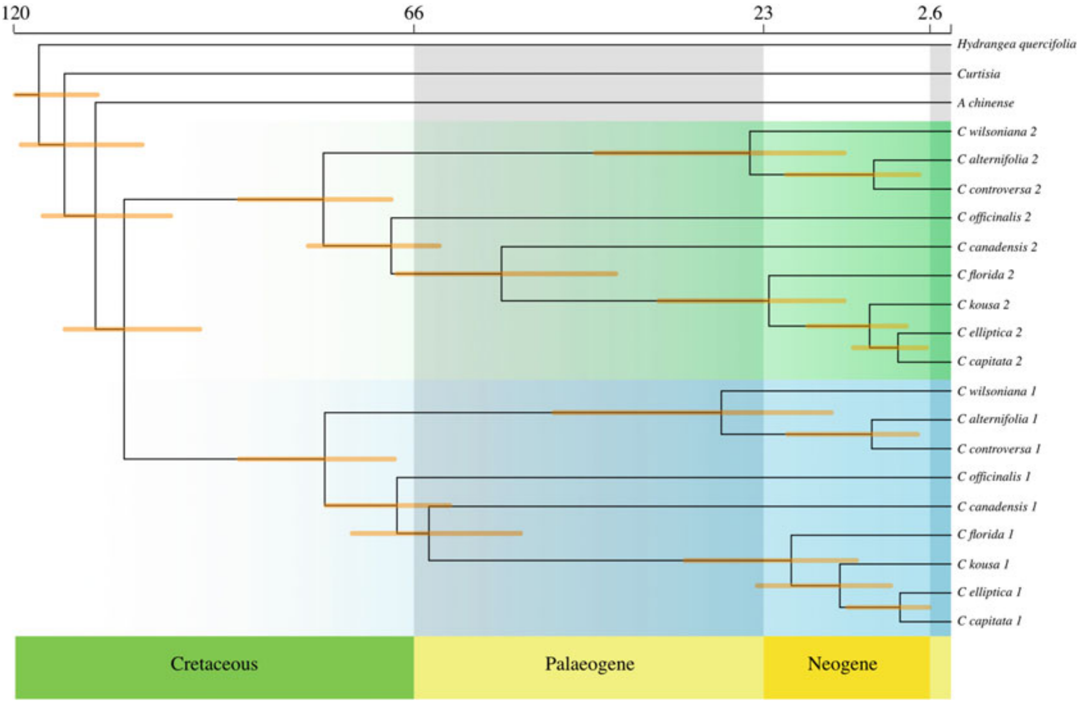

Fig. 5 The posterior estimates of divergence and duplication times derived from the molecular clock analysis. Divergence times are in millions of years before the presence, as indicated by the top bar. Orange bars represent the 95% highest posterior density (HPD) for each node, with each paralogy group colored blue or green

time estimates and to interpret evolutionary history within the context of this uncertainty. In the case of *Cornus*, having run the molecular clock analysis, we can see that the 95% HPD estimates for each node are considerably more precise (Fig. 5), showing the effect of the sequence data on the model. We estimate the duplication event to have occurred within a 109 to 92 Ma (Albian-Cenomanian; middle Cretaceous) interval (Fig. 5).

3 Discussion

We have demonstrated a phylogenetic approach to directly infer the age of WGD events that integrates information from molecular sequence data and a fossil record. Like any analysis, this approach is constrained by the quality of the information provided, relying on a true signal of WGD in the gene families and sufficient information from the fossil record. A major assumption is that gene families containing the signal of a duplication event are derived from WGD rather than local small-scale or chromosomal duplication events. Previous studies have found instances of high-throughput methods identifying WGD events that have more likely been local duplications occurring at a high frequency [36, 37]. Where possible, a high contiguity genome sequence will be able to detect patterns of synteny between duplicate genes and can clarify whether duplications are derived from small- or genome-scale duplication events [37].

Without calibration, relaxed molecular clocks cannot differentiate rate and time since the two are confounded [38]. Thus, systems lacking a reliable fossil record, or those that are particularly difficult to interpret, may suffer a loss of accuracy. The careful consideration of suitable calibrations and the presentation of their provenance is perhaps the most important component of a clock study, and in their absence, results should be treated with caution [15]. Crucially, our approach, through the modeling of fossil calibrations, allows the uncertainty inherent to the fossil record to be incorporated into the analysis.

It is also important to consider exactly what event is being dated given that not all WGD events are the same and their different expectations should be considered in experimental design [10]. During allopolyploidy events, when duplication occurs alongside hybridization, the paralogous genes may have diverged prior to the duplication event itself, in which case the divergence time will reflect the divergence of the two parent species rather than the duplication. Likewise, autopolyploidy events can be followed by a period of tetrasomal inheritance, and paralogs only begin to segregate after a period of fractionation [12]. In this case, the dating will identify the period at which the paralogs diverge, rather than the duplication itself. The extent to which these discrepancies will impact the estimation of the timing of WGD events is unclear. Typically, allopolyploidy occurs between closely related species (though see [39]), and so the difference between the parent divergence and the duplication will be minimal. Similarly, in salmonids, it has been shown that post-WGD, the genome retained a state of tetrasomic inheritance lasting for 17 to 39 million years [40]. The relative contribution of auto- and allopolyploidy in plant evolution is an area of active research [41–43], though the outcomes of the two are predicted to be different [42, 44].

In spite of these caveats, molecular clock methods provide the best means of estimating the time of WGD and in turn facilitate the testing of the most fascinating hypotheses linked to WGD. One of the most prevalent hypotheses surrounding WGD in plants is the apparent clustering of events around the K-Pg boundary [45, 46]. It is proposed that WGD could have provided a selective advantage during the ecological changes in the wake of the mass extinction event, allowing polyploid lineages to survive [47]. The WGD event in *Cornus* is one event that clusters here, originally estimated to have occurred between 85 and 66 Ma and linked to mass extinction event [13]. Instead, we find that the WGD event occurred 109–92 Ma, predating the K-Pg mass extinction event by

43 to 29 million years. We also find that the crown age of *Cornus* also predates the K-Pg mass extinction, as supported by the fossil species *C. piggae* from the Maastrichtian/Campanian [48]. While our results find that the WGD event long predated the K-Pg boundary, this is still compatible with a model wherein polyploid lineages showed greater resilience through the end-Cretaceous mass extinction event.

Another advantage of directly estimating the age of WGD events using the methods outlined in this chapter is that they allow the estimation of both the timing of the WGD event and subsequent speciation events. WGD has been linked to increased rates of diversification in several plant lineages, though often after a "lag" period [5, 49]. This lag has been predicted to reflect the period of diploidization post-WGD, where gene loss, rearrangement, and sub- and neofunctionalization occur [6]. In estimating both WGD and speciation events, this lag can be directly measured. In the case of *Cornus*, we can see that the WGD event predates the divergence of the crown lineage by between 40 and 5 million years. The length of this lag period may be important for hypotheses of causality—estimates of the age of the angiosperm-specific WGD event suggest that it predates the divergence of crown angiosperms by 27 to 65 Ma [11, 50], making a causal relationship between WGD and the success of angiosperms tenuous.

4 Concluding Remarks

We hope to have demonstrated that molecular clock methods provide a means of more accurately and precisely estimating the timing of WGD events. The precision they provide scales with the amount of sequence data available for analysis, and this can be a limiting factor for ancient WGD event, the phylogenetic footprint of which has been eroded through biased gene loss, leading to very few gene families available for dating formative events like the Zeta spermatophyte WGD event [11]. Nevertheless, even a small number of gene families can be sufficient to obtain clade-age estimates with sufficient precision to test macroevolutionary hypotheses. These include, for example, the role of WGD in the origin of angiosperms, the relationship between ploidy and the K-Pg mass extinction, and, indeed, whether the macroevolutionary effect of WGD events lies, if anywhere, with the co-option of redundant duplication genes or their differential loss. Hence, we anticipate that absolute dating methods can support research into the nature, causes, and consequences of WGD events as a generalized phenomenon.

References

1. Clark JW, Donoghue PC (2018) Whole-genome duplication and plant macroevolution. Trends Plant Sci 23(10):933–945
2. Zhang Z, Coenen H, Ruelens P, Hazarika RR, Al HT, Oguis GK, Vandeperre A, van Noort V, Geuten K (2018) Resurrected protein interaction networks reveal the innovation potential of ancient whole-genome duplication. Plant Cell 30(11):2741–2760. https://doi.org/10.1105/tpc.18.00409
3. Barker MS, Li Z, Kidder TI, Reardon CR, Lai Z, Oliveira LO, Scascitelli M, Rieseberg LH (2016) Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. Am J Bot 103(7):1203–1211
4. Vanneste K, Sterck L, Myburg AA, Van de Peer Y, Mizrachi E (2015) Horsetails are ancient polyploids: evidence from Equisetum giganteum. Plant Cell 27(6):1567–1578. https://doi.org/10.1105/tpc.15.00157
5. Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. Am J Bot 105(3): 348–363. https://doi.org/10.1002/ajb2.1060
6. Schranz ME, Mohammadin S, Edger PP (2012) Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr Opin Plant Biol 15(2):147–153
7. Dodsworth S, Chase MW, Leitch AR (2016) Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? Bot J Linn Soc 180(1):1–5
8. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155
9. Vanneste K, Van de Peer Y, Maere S (2013) Inference of genome duplications from age distributions revisited. Mol Biol Evol 30(1): 177–190
10. Doyle JJ, Egan AN (2010) Dating the origins of polyploidy events. New Phytol 186(1): 73–85
11. Clark JW, Donoghue PCJ (1858) 2017 Constraining the timing of whole genome duplication in plant evolutionary history. Proc R Soc B Biol Sci 284:20170912. https://doi.org/10.1098/rspb.2017.0912
12. Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc B Biol Sci 281(1778):20132881. https://doi.org/10.1098/rspb.2013.2881
13. Yu Y, Xiang Q, Manos PS, Soltis DE, Soltis PS, Song B-H, Cheng S, Liu X, Wong G (2017) Whole-genome duplication and molecular evolution in Cornus L. (Cornaceae) – Insights from transcriptome sequences. PLoS One 12(2):e0171361. https://doi.org/10.1371/journal.pone.0171361
14. Donoghue PC, Benton MJ (2007) Rocks and clocks: calibrating the Tree of Life using fossils and molecules. Trends Ecol Evol 22(8): 424–431
15. Parham JF, Donoghue PC, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT (2012) Best practices for justifying fossil calibrations. Syst Biol 61(2):346–359
16. Atkinson BA, Martínez C, Crepet WL (2018) Cretaceous asterid evolution: fruits of Eydeia jerseyensis sp. nov. (Cornales) from the upper Turonian of eastern North America. Ann Bot 123(3):451–460. https://doi.org/10.1093/aob/mcy170
17. Christopher RA (1982) The occurrence of the Complexiopollis-Atlantopollis zone (Palynomorphs) in the Eagle Ford Group (Upper Cretaceous) of Texas. J Paleontol 56:525–541
18. Doyle JA, Robbins EI (1977) Angiosperm pollen zonation of the continental Cretaceous of the Atlantic Coastal Plain and its application to deep wells in the Salisbury Embayment. Palynology 1:41
19. Christopher RA (1979) Normapolles and triporate pollen assemblages from the Raritan and Magothy Formations (Upper Cretaceous) of New Jersey. Palynology 3(1):73–121
20. Gale A, Mutterlose J, Batenburg S, Gradstein F, Agterberg F, Ogg J, Petrizzo M (2020) The Cretaceous period. In: Geologic time scale 2020. Elsevier, pp 1023–1086
21. Barba-Montoya J, dos Reis M, Schneider H, Donoghue PCJ, Yang Z (2018) Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. New Phytol 218(2):819–834. https://doi.org/10.1111/nph.15011
22. Ho SYW, Phillips MJ (2009) Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst Biol 58(3):367–380. https://doi.org/10.1093/sysbio/syp035
23. Benton M, Donoghue P, Asher R (2009) Calibrating and constraining molecular clocks. The Timetree of Life 35:86

24. Warnock RC, Yang Z, Donoghue PC (2012) Exploring uncertainty in the calibration of the molecular clock. Biol Lett 8(1):156–159

25. Puttick MN (2019) MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. Bioinformatics 35(24): 5321–5322

26. Yang Z, Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23(1):212–226

27. Shih PM, Matzke NJ (2013) Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. Proc Natl Acad Sci 201305813. https://doi.org/10.1073/pnas.1305813110

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

29. Warnock RC, Parham JF, Joyce WG, Lyson TR, Donoghue PC (2015) Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proc R Soc B Biol Sci 282(1798): 20141013

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

31. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15(12): 1647–1657

32. Reis MD, Yang Z (2011) Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol 28(7):2161–2172

33. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8(3): 275–282

34. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol 67(5):901

35. Warnock RCM, Yang Z, Donoghue PCJ (1857) 2017 Testing the molecular clock using mechanistic models of fossil preservation and molecular evolution. Proc R Soc B Biol Sci 284:20170227. https://doi.org/10.1098/rspb.2017.0227

36. Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS (2018) Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc Natl Acad Sci 115(18):4713–4718

37. Nakatani Y, Mc Lysaght A (2019) Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. Proc Natl Acad Sci 116(6):1816–1818

38. dos Reis M, Donoghue PCJ, Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet 17(2):71–80. https://doi.org/10.1038/nrg.2015.8

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

40. Gundappa MK, To T-H, Grønvold L, Martin SAM, Lien S, Geist J, Hazlerigg D, Sandve SR, Macqueen DJ (2021) Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. bioRxiv:2021.2006.2005.447185. https://doi.org/10.1101/2021.06.05.447185

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

42. Spoelhof JP, Soltis PS, Soltis DE (2017) Pure polyploidy: closing the gaps in autopolyploid research. J Syst Evol 55(4):340–352

43. Doyle JJ, Sherman-Broyles S (2017) Double trouble: taxonomy and definitions of polyploidy. New Phytol 213(2):487–493

44. Parisod C, Holderegger R, Brochmann C (2010) Evolutionary consequences of autopolyploidy. New Phytol 186(1):5–17

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

46. Vanneste K, Maere S, Van de Peer Y (2014) Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. Philos Trans Royal Soc B Biol Sciences 369(1648): 20130353. https://doi.org/10.1098/rstb.2013.0353

47. Levin DA (2020) Did dysploid waves follow the pulses of whole genome duplications? Plant Syst Evol 306(5):1–4

48. Atkinson BA, Stockey RA, Rothwell GW (2016) Cretaceous origin of dogwoods: an anatomically preserved Cornus (Cornaceae) fruit from the Campanian of Vancouver Island. PeerJ 4:e2808–e2808. https://doi.org/10.7717/peerj.2808

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

50. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473(7345):97–100