PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Clark JW, Donoghue PCJ. 2017 Constraining the timing of whole genome duplication in plant evolutionary history. *Proc. R. Soc. B* **284**: 20170912. http://dx.doi.org/10.1098/rspb.2017.0912

Received: 27 April 2017
Accepted: 1 June 2017

Subject Category:
Palaeobiology

Subject Areas:
evolution, genomics, plant science

Keywords:
genome duplication, plant evolution, polyploidy, molecular clock

Authors for correspondence:
James W. Clark
e-mail: james.clark@bristol.ac.uk
Philip C. J. Donoghue
e-mail: phil.donoghue@bristol.ac.uk

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3807310.

THE ROYAL SOCIETY PUBLISHING

Constraining the timing of whole genome duplication in plant evolutionary history

James W. Clark and Philip C. J. Donoghue

School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

JWC, 0000-0003-2896-1631; PCJD, 0000-0003-3116-7463

Whole genome duplication (WGD) has occurred in many lineages within the tree of life and is invariably invoked as causal to evolutionary innovation, increased diversity, and extinction resistance. Testing such hypotheses is problematic, not least since the timing of WGD events has proven hard to constrain. Here we show that WGD events can be dated through molecular clock analysis of concatenated gene families, calibrated using fossil evidence for the ages of species divergences that bracket WGD events. We apply this approach to dating the two major genome duplication events shared by all seed plants (ζ) and flowering plants (ε), estimating the seed plant WGD event at 399–381 Ma, and the angiosperm WGD event at 319–297 Ma. These events thus took place early in the stem of both lineages, precluding hypotheses of WGD conferring extinction resistance, driving dramatic increases in innovation and diversity, but corroborating and qualifying the more permissive hypothesis of a 'lag-time' in realizing the effects of WGD in plant evolution.

1. Background

The discovery in plant genomes of evidence of recurrent whole genome duplication events (WGD; polyploidy) has reignited debate over its importance in land plant evolution [1,2]. Several causal hypotheses have emerged linking WGD to key innovations [3], increased rates of diversification [4] and extinction resistance that may have facilitated the success of multiple lineages of extant plants [5]. The mechanisms through which genome duplication can result in evolutionary novelty are becoming better understood and the traditional models of neo- and subfunctionalization have now been hybridized with models of dosage balance in attempts to explain how evolutionary innovation can arise post-WGD in the face of extensive gene loss and stabilizing patterns of gene retention [6,7]. Furthermore, there now exist elegant examples of genes and gene families that have taken on new functions (neofunctionalization) following multiple rounds of WGD and then playing a key role in the evolution of plant lineages [8]. The link between polyploidy and diversification remains controversial [9], but there exists some evidence that several of the ancient WGD events in angiosperms correlate with shifts in diversification [4]. Separating the WGD events and the shifts in diversification are a 'lag' of several million years, which has been explained as the period of fractionation post-WGD and, in turn, the feature of WGD that leads to innovation and diversification [10]. However, at the broadest scale, these hypotheses are underpinned by the relative phylogenetic placement and absolute timing of each event. Though the relative phylogenetic timing of plant WGD events is well constrained, their absolute timing is not [9].

Constraining the phylogenetic position of WGD events relies on broad taxonomic sampling of genomic or transcriptomic data. The presence or absence of shared 'age peaks' in Ks plots of synonymous substitution rates between duplicates provides evidence for shared genome duplications [11]. This approach culminated in a survey of 41 plant genomes focusing on angiosperms [5] and more recently several transcriptomes also highlighting the presence of WGD within the evolutionary history of gymnosperms [12] and peat mosses [13]. The number and position of the peaks on the Ks plot also reveals the relative

timing of each event, with multiple peaks representing multiple successive WGDs. The absolute timing of each event can be obtained indirectly by phylogenetically bracketing the event—the event must have occurred along the branch between those lineages that have undergone the WGD and those that have not. However, despite well-sampled exceptions among certain groups of angiosperms [14–16], there are few cases where the sampling of taxa is dense enough to prevent very long branches, and so the ages of genome duplication events must be inferred directly. Direct dates can be obtained by converting the relative timing of peaks on a Ks plot into absolute ages. This has the advantage that it does not require additional taxon sampling and so estimates can be obtained for WGD events isolated on long branches [17]. A major caveat of this approach is that it relies on the assumption of a strict molecular clock that, depending on shifts in the rate of sequence evolution, can lead to inaccurate age estimates. Furthermore, Ks plots are known to saturate beyond a certain age, meaning that they cannot always distinguish more ancient duplications and may lead to artificial peaks in the distribution [18]. More complex relaxed clock methods can be employed in a phylogenetic or phylogenomic approach, whereby the individual gene families containing signal of WGD are reconstructed and individually dated [19]. The distribution of ages obtained can then be plotted to provide a range of estimates for each event. This approach is more powerful and has been used to estimate the ages of multiple WGD events across the angiosperms, where genomic and transcriptomic data are more abundant [19,20]. However, dating individual gene trees does not fully exploit the power of the molecular clock and the power of individual gene trees is likely to diminish over longer periods of evolutionary time. Increasing the amount of sequence data by concatenating multiple gene families into alignments decreases uncertainty in the estimation of relative ages [21], and can be used to date the absolute timing of WGD events [22] yet, to date, studies focusing on WGD in plants have relied on the power of individual gene trees. Directly dating WGD events using concatenated gene trees also provides estimates of the absolute timing of the WGD in relation to subsequent speciation events within the lineage, since gene trees observe species divergences as well as duplication events. Thus, concatenated gene trees have the potential to provide an accurate estimate of the absolute timing of WGD events relative to the diversification events in which they are causally implicated.

The seed plants (Spermatophyta) are the most species rich of extant plant clades, encompassing the gymnosperms and angiosperms (flowering plants). WGD events have been identified at the base of all seed plants (ζ; [12,20]) and at the base of all angiosperms (ε; [20]), and so all extant flowering plants have undergone at least two rounds of genome duplication. Previous attempts to date these events were based on distributions of ages inferred using poorly defined calibrations and penalized likelihood molecular clock methods [20] that have since been found unreliable [23]. The WGD shared by all extant angiosperms has been linked with the 'big bang' diversification of the Mesangiospermae (following a lag period) as well as several major innovations, including the origin of the flower [3,4]. WGD has been thought to be less prevalent within gymnosperms, the sister clade to angiosperms (together comprising Spermatophyta), despite the fact that the ζ WGD is part of their shared evolutionary history. More recent evidence has indicated that WGD has occurred in several gymnosperm lineages and confirmed that the ζ WGD (spermatophyte) was not shared with their sister lineage, the ferns [12].

Conventionally molecular clock dating approaches have sought to minimize the influence of duplication by using only single copy genes. In contrast, we exploit the pattern of paralogy produced by WGD in the evolutionary history of multiple gene families and concatenate them into a partitioned alignment. Combined with broad taxon sampling and multiple fossil calibrations, we demonstrate an approach for dating gene trees to provide well-constrained estimates of the timing of duplication events and attendant speciation events.

2. Material and methods

Gene families containing signal of the ζ (spermatophyte) and ε (angiosperm) WGD events and those that contain the signal of both were catalogued by Jiao *et al*. [4], and from these we expanded orthogroups by obtaining amino acid sequences using Plaza 3.0 (bioinformatics.psb.ugent.be/plaza), and GreenPhyl 4 (www.greenphyl.org). Further sequences were obtained by local BLAST searches of iPlant (www.iplantcollaborative.org). One hundred and twenty-eight species were sampled in total, representing all major lineages of land plants and these are listed in electronic supplementary material, table S1. Four datasets were assembled for all taxa: families containing a clear signal of just the ε WGD event (angiosperm dataset), just the ζ WGD event (spermatophyte dataset), families containing signal of both events ($\zeta + \varepsilon$ dataset), and a combined dataset. To verify a clear signal of the relevant WGD event in each gene family, we built individual gene trees based on multiple amino acid sequence alignments generated using MAFFT while model selection and gene tree reconstructions were performed using IQ-TREE [24]. We opted for a conservative approach, discarding orthogroups that following phylogenetic reconstruction and visual inspection did not clearly reflect the signal of either or both WGDs (e.g. electronic supplementary material, figure S1), had sequence alignments shorter than 100 amino acids, displayed a topology that was incongruent with our current understanding of land plant phylogeny with either the total group seed plants or major lineages within being resolved as non-monophyletic, or were too large with multiple nested duplications, resulting in large numbers of sequences having to be discarded. Of 130 orthogroups surveyed, 12 gene families were found containing a clear signal of the ε WGD. The number of sequences among individual gene families ranged from 87–126 and when concatenated a total of 176 tips. Fourteen further gene families were found for the ζ WGD, representing 189 tips when concatenated and varying from 106 to 149 tips individually. An additional seven gene families were found containing the signal for both, for which 254 tip sequences were assembled when concatenated and individual gene families ranging from 132 to 249 tips. The combined dataset contained 33 gene families, with one node representing ζ, but two representing ε. As 12 gene families contain only one node with the ε duplication, the event was represented only once in the combined analysis, to maximize precision at this node. Similarly, angiosperm gene copies from gene families not containing signal of the ε duplication were randomly assigned to one side of the duplication. Due to differential retention, a copy of each gene paralogue was not present in all families and the number of tips in each gene family is listed in electronic supplementary material, table S3.

Across all analyses, nodes were constrained using 35 fossil calibrations spanning land plant phylogeny defined using best practice [25] (electronic supplementary material, table S2). The

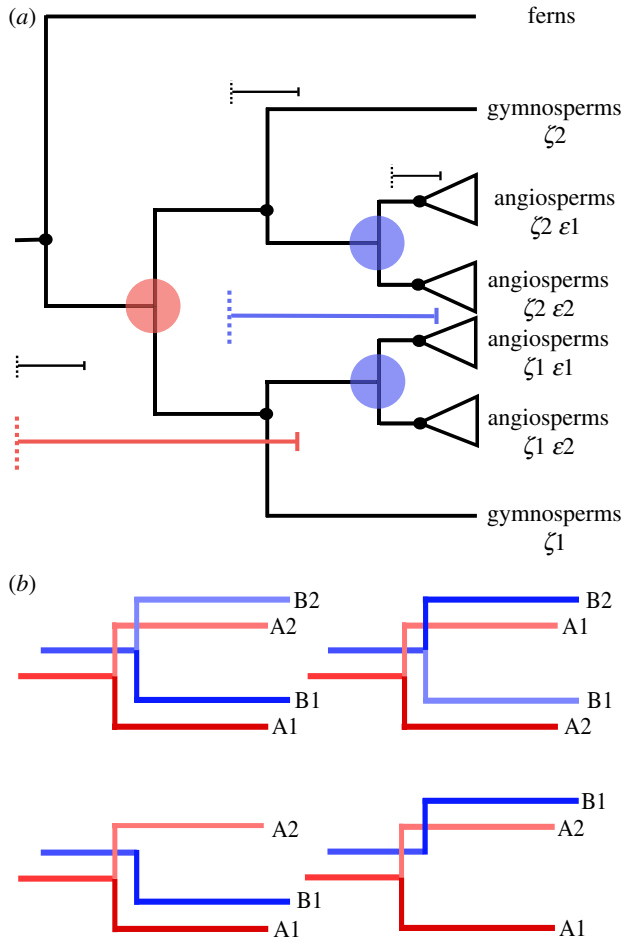

Figure 1. (*a*) An example gene tree showing the seed plant (ζ, red) and angiosperm (ε, blue) duplications. The duplication events are constrained using minima and maxima (coloured brackets) based on fossils used to constrain speciation events (black brackets). (*b*) Gene trees may retain both copies of the duplicate gene (top), or a single copy may be lost (bottom). When concatenating duplicates from different gene families, given that both copies are descended from the same event, their assignment to either side of the duplication is arbitrary. (Online version in colour.)

duplication nodes were constrained temporally to reflect the possibility of the WGD occurring at any point following the divergence of spermatophytes from an ancestral euphyllophyte (ζ WGD event) and for angiosperms from an ancestral spermatophyte (ε WGD event) (figure 1). Calibrations that provided only a minimum age were modelled as a hard minimum bound with a truncated Cauchy distribution ($p = 0.1, c = 0.2$). Calibrations that provided a maximum age were modelled with a soft maximum with a uniform distribution between the minimum and maximum age [26]. Molecular clock analyses were conducted on concatenated alignments using the normal approximation method in MCMCtree under the appropriate model [27]. The normal approximation method provides a fast and efficient way of analysing large datasets using complex models and a relaxed clock and is run under a fixed topology. We ran all analyses on a topology reflecting both WGD events and recent hypotheses of relationships among land plants [28] (electronic supplementary material, figure S2). We also reconstructed the topology based on our own datasets using IQ-TREE and found that it was highly congruent with the constraint tree. Each analysis was run twice independently and regularly checked for convergence and for effective sample sizes greater than 200 using Tracer v. 16 [29].

Assuming autopolyploidy, each WGD event produces two daughter nodes that are created simultaneously and that must have the same age, and so the assignment of each paralogue to either node of the duplication is arbitrary (figure 1). In this way paralogues between the gene families can be concatenated in multiple combinations, so long as they are consistent within each gene family. To explore the impact of different combinations of paralogy groups between gene families, we randomly reassigned groups to either node using the $\zeta + \varepsilon$ dataset containing both duplications.

The extent to which the low number of available gene families impacted on the estimation of dates was explored through infinite sites analyses [30]. The gene families were successively concatenated and the analysis repeated with one more gene family each time. The relationship between the mean age estimates and the widths of the 95% HPDs was used as a measure of the precision of the data versus the uncertainties induced by the fossil calibrations. Higher R^2 values indicate that large HPD widths are due to increasing uncertainty in the fossil record deeper in time. A saturation of the curve suggests that adding further sequence data would not increase the precision of the analysis, since it is limited by the information available in the fossil record.

3. Results

In most Bayesian molecular software, specified node age priors are modified in the construction of the joint time prior to achieve the expectation that only ages compatible with the assumption that ancestral nodes are older than their descendants, are proposed to the MCMC [31,32]. To ensure that these effective priors are biologically reasonable, we estimated them by running the analysis without sequence data. The effective priors are compatible with the original palaeontological and phylogenetic evidence, yielding broad 95% HPDs for the timing of WGDs in all analyses, though both were truncated relative to the specified calibrations. The spans of the 95% HPD for the prior on the ζ and ε WGD events are 81 (434–353 Ma) and 111 (355–244 Ma) million years, respectively (table 1). In the separate analyses of both the ζ and the ε WGD events, the truncation effects on the prior were the same as for the combined analysis, and so the additional nodes in the combined analysis and the $\zeta + \varepsilon$ dataset did not affect the effective prior.

In all instances, the addition of sequence data yielded estimates congruent with, yet more precise than, the joint time prior. Estimates for both WGD events were compared between gene families using the $\zeta + \varepsilon$ dataset, and we found variation in both the width of the 95% HPD and the absolute age estimates, though the overlapping distributions of the HPDs showed that the gene families were congruent. While some gene families produced much more precise estimates, the variation in estimates between all gene families showed a similar level of precision to the joint time prior alone, ranging from 435–346 Ma for the ζ WGD event and 355–244 for the ε WGD event. The $\zeta + \varepsilon$ dataset also allowed us to compare the estimates for the ε duplication, which is represented twice in each gene family, within gene families. We found that the 95% HPD widths for the event varied within gene families, though this is likely due to the absence of paralogues on one side of the duplication. The only family with all paralogues present, CDK, showed estimates consistent in both age and uncertainty across both nodes.

The greatest effect in terms of precision was produced by increasing the amount of sequence data by concatenating the gene families. The effect of missing paralogues across both duplication nodes in the $\zeta + \varepsilon$ dataset was minimized and

Table 1. Ninety-five per cent HPD estimates for the age of both WGD events, summarizing the effective prior, individual gene families (1 to 7), the effects of concatenating gene families, the expanded and combined datasets.

node	effective prior	gene families							concatenated gene families						ε dataset	ζ dataset	combined dataset
		1	2	3	4	5	6	7	1–2	1–3	1–4	1–5	1–6	1–7			
spermatophyte duplication (ζ)	353–434	382–435	346–411	346–411	354–418	354–404	357–415	355–433	390–433	386–430	380–418	380–416	377–408	378–409	—	380–401	381–399
angiosperm duplication (ε) ζ1 ε	244–355	270–339	250–353	248–328	280–354	258–340	249–351	254–356	273–336	268–323	280–323	285–331	282–325	281–323	295–321	—	297–319
angiosperm duplication (ε) ζ2 ε	244–355	267–340	273–344	247–350	245–349	277–362	247–313	245–355	278–333	276–330	276–322	289–338	283–325	276–321	—	—	—

the age estimates for both ε nodes were highly consistent. The ζ + ε concatenation was also considerably more precise than any of the individual gene families (table 1). Multiple concatenations were tested on this dataset, to determine if the assignment of paralogues between duplicates affected the estimates. We did not observe any material differences in age or uncertainty, indicating that the results are robust to the way in which the gene families are concatenated.

The addition of further sequence data for each duplication event in turn produced results of even greater precision. The angiosperm dataset estimated an age of 321–295 Ma for the ε WGD event, almost five times more precise than the joint time prior alone. A similar increase in precision was obtained by the spermatophyte dataset, the ζ duplication estimated to have occurred 400–380 Ma, four times more precise than the joint time prior alone. Based on the largest amount of data, the combined analysis of the combined dataset produced results that were highly congruent with the two individual datasets, if not marginally more precise, estimating 399–381 Ma and 319–297 Ma for the ζ and ε WGD events, respectively (figure 2).

Infinite sites plots suggest that though the R^2 value showed little changed with increased sequence data, the addition of sequence data reduced the uncertainty of estimates (figure 3). With 19 gene families, the amount of error was continuing to decrease, suggesting that additional gene families may increase precision further.

4. Discussion

(a) Inferring the age of whole genome duplication

Our results indicate that the evolutionary history of gene families can be exploited to obtain precise estimates of the age of WGD events. These methods depend on both careful selection of fossil constraints and available gene families containing signal of WGD events, though even with limited sequence data, we greatly improve the precision over the raw calibrations alone.

Both the ε (angiosperm) and ζ (spermatophyte) genome duplication events have been independently reported [12,20], yet we were unable to find large numbers of gene families with clear signal of either or both events. The paucity of available gene families for these WGD events is likely in part a result of our conservative criteria in selecting gene families based on topology. In part, this reflects the limitations of single genes to resolve unequivocal phylogenetic signal for such events over long timescales. However, it also reflects the antiquity of the events, given that retention of genes following a WGD follows a decay pattern and widespread gene loss leads to a gradually decreasing phylogenetic signal over time. It is unsurprising that so few gene families remain with a clear signal of these events and, when considered next to existing evidence for these events [12,20], our findings are entirely compatible with the ε and ζ duplication events. Our results indicate that the evolutionary history of gene families can be exploited to obtain precise estimates of the age of WGD events. Infinite sites plots lead us to expect that the addition of further sequence data will leverage further precision. Similarly, WGD events that are more recent and may contain more genome-wide data, may be dated using the same approach but with greater precision.

Unlike genomic datasets that can be used for gene-tree reconciliation and the construction of Ks plots, the methods

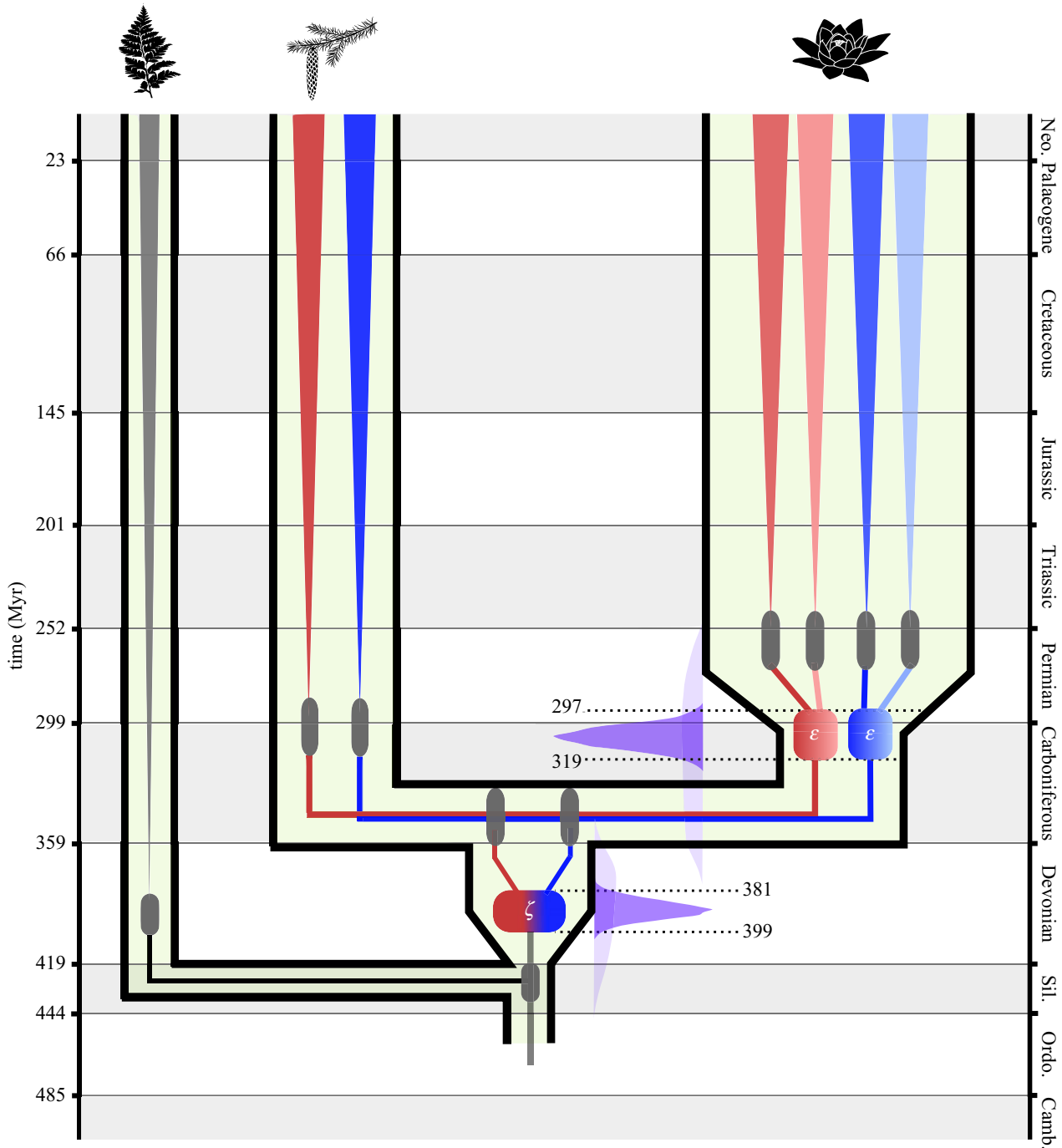

Figure 2. Estimated dates for the occurrence of both the seed plant (ζ) and angiosperm (ε) duplication events based on a molecular clock analysis of 33 concatenated gene families. Age estimates (95% HPD) for the divergences of the major lineages and crown groups represented by grey bars. The age estimates (95% HPD) of two duplication events are represented by coloured boxes, with the subsequent subgenomes represented first by blue and red (ζ), then by lighter and darker shades of each colour (ε). For each duplication event, the effective prior is shown (light blue) next to the posterior distribution (dark blue). (Online version in colour.)

presented here focus solely on the dating of WGD events, rather than their characterization. However, the congruence of age estimates between gene families serves as a test of their coincidence, as anticipated by WGD. The annotation of gene families to either side of the duplication event requires greater care and is a potentially limiting factor on the number of gene families that can be analysed, yet we have demonstrated that even with a relatively small dataset (compared to a genomic dataset), high levels of precision can be achieved. Novel molecular clock approaches such as cross bracing could also be used to increase precision around the duplication nodes, especially as they are so difficult to constrain [33].

An additional caveat is that WGD or polyploidy is often categorized into two distinct classes [34], autopolyploidy and allopolyploidy, traditionally distinguished based on the number of parent species, but also characterized by the patterns of fractionation post-WGD. The mode of duplication may impact our estimates of duplication age [35], as the point at which duplicates coalesce is actually the timing of divergence of the two parental species, or a more ancestral autoploidy event, as opposed to the alloploidy event itself [35]. New methods are emerging to discriminate between auto- and allopolyploidy [36], but these are likely to fail when applied to more ancient genome duplication events. However, allopolyploidy

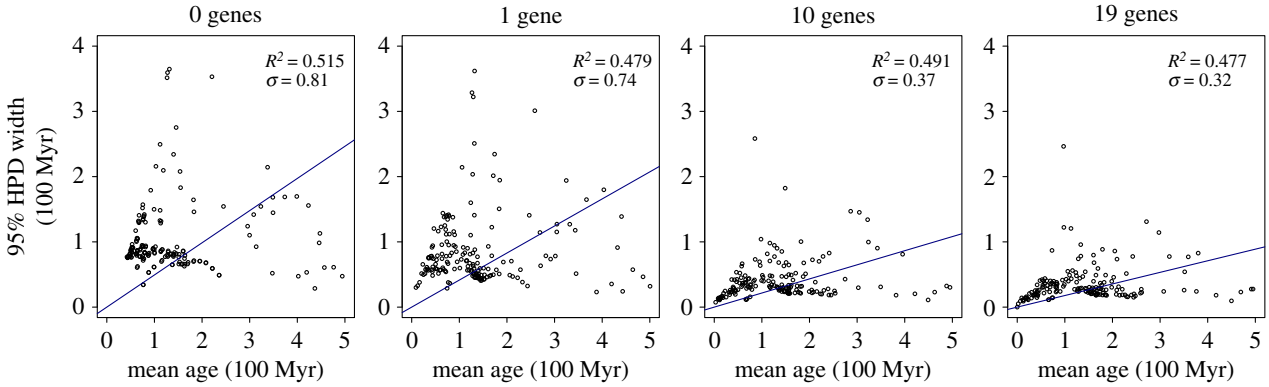

Figure 3. Infinite sites plots for the most complete (angiosperm) dataset, with the regression between the mean age and the 95% HPD shown for 0, 1, 10 and 19 gene datasets. The R^2 and error terms are also shown. (Online version in colour.)

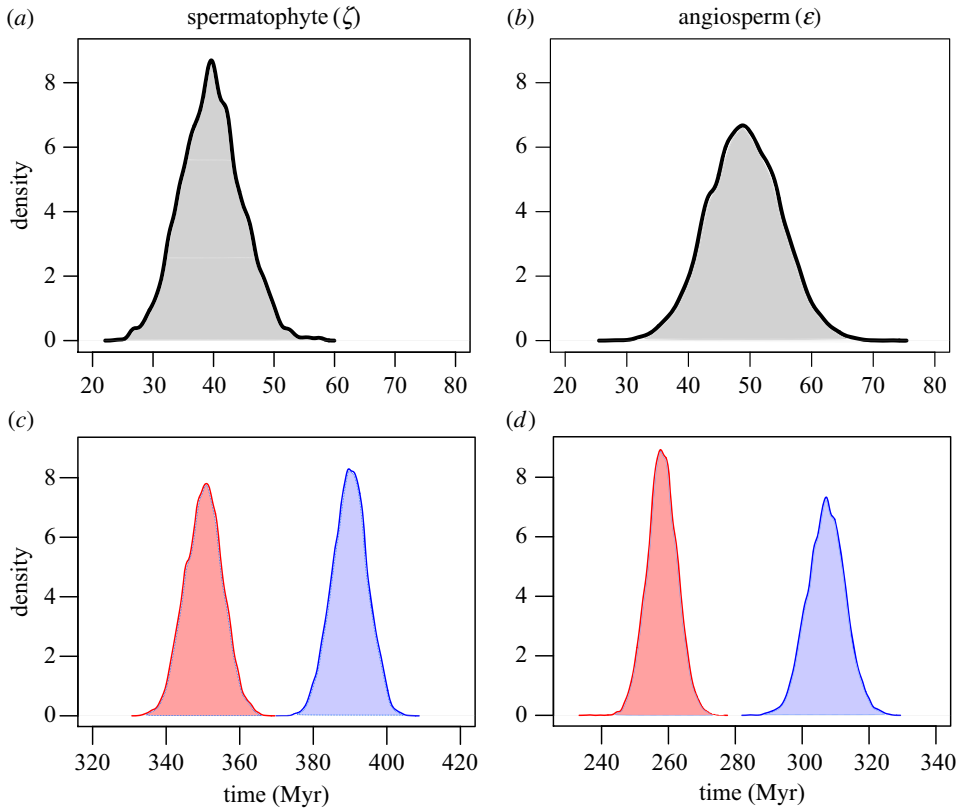

Figure 4. The posterior probabilities of (*a*) the lag between the ζ duplication and the diversification of crown spermatophytes and (*b*) the lag between the ε duplication and the diversification of crown angiosperms. The posterior probabilities of the absolute age of the WGD events (blue) and diversification (red) are also shown for (*c*) ζ and spermatophytes and (*d*) ε and angiosperms. (Online version in colour.)

would only have a large impact on accuracy if hybridization occurred between very distant parent species.

(b) Dating duplication, diversification and innovation

Our most comprehensive analysis of 33 gene families indicated that the genome duplication present in all crown spermatophytes occurred 399–381 Ma, a period spanning the Early to Late Devonian (figure 2). The WGD event present in all crown angiosperms occurred almost 100 Myr later, 319–297 Ma, across the Carboniferous–Permian boundary (figure 2). Gene trees contain both the signal of WGD and species divergence, allow a direct estimation of the age of the WGD event relative to the age of the crown group (figure 4). Both estimates predict that the respective WGD events occurred early in the stem of both lineages, predating the diversification of the crown group by about 50 Myr. These estimates are considerably older than those of Jiao *et al.* [20], yet our estimates for the age of the seed plant (360–340 Ma) and angiosperm (267–247) crown groups are comparable to other molecular clock analyses [37,38], allowing us to reject the notion that the duplications occurred late in the stem lineage. Greater precision in the absolute age of WGD events leveraged by concatenation allows that hypotheses can be more rigorously tested. WGD occurring early in the stem lineage has two implications for current hypotheses regarding the role of WGD in plant evolution.

First is the hypothesis that WGD drives evolutionary success [39–41], or confers extinction resistance [19,42], since the long stem lineages of both groups are, by definition, characterized by extinction. However, many extinct lineages must also share these genome duplications. For example, the ζ

duplication predates the appearance of the earliest seed plants, the pteridosperms and cordaitales, and so WGD cannot have contributed to their diversification or conferred extinction resistance, as has been proposed for the ancient palaeopolyploid *Equisetum* [17]. The long-term evolutionary success of seed plants and especially angiosperms is unquestionable, and there is considerable evidence for the role of gene duplication in the evolution of angiosperms, in particular [3,43], yet our results are more in keeping with the idea of 'rarely successful polyploids' [39]. The challenges faced by polyploids in order to establish and persist may be partially responsible for extinctions in a lineage post-WGD, and it may be the case that extant spermatophytes and angiosperms are the surviving lineages best able to exploit any long-term competitive advantages [42]. Secondly, if their crown clades of seed and flowering plants can be considered to be characterized by evolutionary success, this has been achieved in both lineages after a substantial lag post-WGD. Our results indicate that the lag between the ζ WGD event and the divergence of crown spermatophytes is 22–60 Myr, and 27–65 Myr between the ε WGD event and the divergence of crown angiosperms (figure 4). These are comparable to the results of Tank *et al*. [4], who estimated a 49.2 Myr lag between the ε WGD event and the shift in diversification of angiosperms, though without directly inferring the age of the WGD. Tank *et al*. [4] also estimated that the rate shift in diversification among angiosperms occurred at 213 Ma, following the divergence of Mesangiospermae which, following our age estimates, indicates a lag of 84–106 Myr. Ultimately, these results indicate that more precise age estimates require more precise hypotheses regarding the role of WGD in promoting evolutionary success. Given these long lag periods and that some, though clearly not all, clades that share a history of WGD are diverse or characterized by innovations, it requires more explicit hypotheses regarding which clades are considered successful.

Evidently, we find no direct support for the deterministic role of WGD in driving diversification or innovation. Rather, our data are more compatible with the more permissive model of evolution via genome duplication that emphasizes the importance of the post-WGD period of genome fractionation. During this period, the need to maintain a dosage balance of protein products selects for the maintenance of duplicates, followed by a relaxation of selection allowing sub- and neofunctionalization [7]. An additional consideration is the lineage specific re-diploidization model, which applies when species divergence occurs before the diploidization process in complete [44]. Under this model, the lag is produced by the pattern of tetrasomic inheritance that is characteristic of autopolyploidy, leading to massively delayed functional divergence of duplicate genes. This model also predicts that duplicate genes evolve independently in separate lineages, and that this can explain the divergent evolutionary trajectories of lineages that share the same history of WGD [44]. This more permissive model explains the 'long fuse' or 'lag' found in our results, whereby an early WGD during a lineage's evolution provides a primer for subsequent innovation and diversification, leading to the evolutionary success of both lineages [42]. It also explains the paucity of genes preserving all paralogues anticipated as a phylogenetic footprint of the ζ and ε WGD events, as a consequence of post-duplication dysploidy leading to dosage bias.

The quantification of this lag is clearly relevant to understanding the role of WGD in plant evolution [42]. Our methods are applicable to other WGD events characterized previously within the plant kingdom, including those thought to be associated with increased diversification or the K–Pg boundary [4,5]. Furthermore, these methods could be used to clarify the timing of the proposed WGDs associated with the origins and early evolution of vertebrates [45], which are still undermined by uncertainty around their timing.

5. Conclusion

Accurate and precise estimates of the timing of WGD events are fundamental to our understanding their significance on a macroevolutionary scale and can be achieved by coupling a careful appraisal of the fossil record with molecular clock approaches. We demonstrated that by concatenating multiple gene families with a shared history of WGD into a single alignment, the ages of two ancient WGD events, ε (angiosperm) and ζ (spermatophyte), were estimated to a high degree of precision. Both events were found to occur early in the stem of each lineage, predating the divergence of the crown groups by 50 Myr. These methods can be applied to date any previously characterized WGD event, including those identified in yeasts and vertebrates.

Data accessibility. Electronic supplementary material includes Supplemental Experimental Procedures, three figures and three tables and can be found with this article online. The molecular sequence alignment and trees with fossil calibrations have been deposited in Figshare: https://figshare.com/s/d46377d5ae6999c0cd52.

Authors' contributions. J.W.C. and P.C.J.D. conceived the project and designed the analysis. J.W.C. prepared the datasets and performed the analyses. J.W.C. and P.C.J.D. interpreted the results. J.W.C. wrote the main draft of the manuscript, to which P.C.J.D. contributed.

Competing interests. The authors declare no competing interests.

Funding. This research was funded by Biotechnology and Biosciences Research Council (UK) grant nos. (BB/N000609/1; BB/N000919/1), Natural Environment Research Council grant no. (NE/N002067/1), the Royal Society, and the Wolfson Foundation.

Acknowledgements. We thank members of the Bristol Palaeobiology Research Group for discussion and two anonymous reviewers for their comments and improvements to the manuscript.

References

1. Mayrose I, Zhan SH, Rothfels CJ, Arrigo N, Barker MS, Rieseberg LH, Otto SP. 2015 Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis *et al*. (2014). *New Phytol.* **206**, 27–35. (doi:10.1111/nph.13192)
2. Soltis DE *et al*. 2014 Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose *et al*. (2011). *New Phytol.* **202**, 1105–1117. (doi:10.1111/nph.12756)
3. Soltis PS, Soltis DE. 2016 Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–65. (doi:10.1016/j.pbi.2016.03.015)
4. Tank DC *et al*. 2015 Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* **207**, 454–467. (doi:10.1111/nph.13491)

5. Vanneste K, Baele G, Maere S, Van de Peer Y. 2014 Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347. (doi:10.1101/gr.168997.113)
6. Teufel AI, Liu L, Liberles DA. 2016 Models for gene duplication when dosage balance works as a transition state to subsequent neo- or sub-functionalization. *BMC Evol. Biol.* **16**, 45. (doi:10.1186/s12862-016-0616-1)
7. Conant GC, Birchler JA, Pires JC. 2014 Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* **19**, 91–98. (doi:10.1016/j.pbi.2014.05.008)
8. Edger PP *et al.* 2015 The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl Acad. Sci. USA* **112**, 8362–8366. (doi:10.1073/pnas.1503926112)
9. Kellogg EA. 2016 Has the connection between polyploidy and diversification actually been tested? *Curr. Opin. Plant Biol.* **30**, 25–32. (doi:10.1016/j.pbi.2016.01.002)
10. Dodsworth S, Chase MW, Leitch AR. 2016 Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Bot. J. Linn. Soc.* **180**, 1–5. (doi:10.1111/boj.12357)
11. Lynch M, Conery JS. 2000 The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155. (doi:10.1126/science.290.5494.1151)
12. Li Z *et al.* 2015 Early genome duplications in conifers and other seed plants. *Sci. Adv.* **1**, e1501084. (doi:10.1126/sciadv.1501084)
13. Devos N, Weston DJ, Rothfels CJ, Johnson MG, Shaw AJ. 2016 Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta). *New Phytol.* **211**, 300–318. (doi:10.1111/nph.13887)
14. Estep MC *et al.* 2014 Allopolyploidy, diversification, and the Miocene grassland expansion. *Proc. Natl Acad. Sci. USA* **111**, 15 149–15 154. (doi:10.1073/pnas.1404177111)
15. 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**, 1203–1211. (doi:10.3732/ajb.1600113)
16. Kagale S *et al.* 2014 Polyploid evolution of the Brassicaceae during the Cenozoic era. *Plant Cell* **26**, 2777–2791. (doi:10.1105/tpc.114.126391)
17. Vanneste K, Sterck L, Myburg AA, Van de Peer Y, Mizrachi E. 2015 Horsetails are ancient polyploids: evidence from *Equisetum giganteum*. *Plant Cell* **27**, 1567–1578. (doi:10.1105/tpc.15.00157)
18. Vanneste K, Van de Peer Y, Maere S. 2013 Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30**, 177–190. (doi:10.1093/molbev/mss214)
19. Fawcett JA, Maere S, Van de Peer Y. 2009 Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl Acad. Sci. USA* **106**, 5737–5742. (doi:10.1073/pnas.0900906106)
20. Jiao Y *et al.* 2011 Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100. (doi:10.1038/nature09916)
21. dos Reis M, Donoghue PCJ, Yang Z. 2016 Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**, 71–80. (doi:10.1038/nrg.2015.8)
22. 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* **281**, 20132881. (doi:10.1098/rspb.2013.2881)
23. Thorne JL, Kishino H. 2005 Estimation of divergence times from molecular sequence data. In *Statistical methods in molecular evolution* (ed. R Nielsen), pp. 233–256. New York, NY: Springer.
24. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015 IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274. (doi:10.1093/molbev/msu300)
25. Parham JF *et al.* 2012 Best practices for justifying fossil calibrations. *Syst. Biol.* **61**, 346–359. (doi:10.1093/sysbio/syr107)
26. Warnock RC, Parham JF, Joyce WG, Lyson TR, Donoghue PCJ. 2015 Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. *Proc. R. Soc. B* **282**, 20141013. (doi:10.1098/rspb.2014.1013)
27. Yang Z. 2007 PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)
28. Wickett NJ *et al.* 2014 Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl Acad. Sci. USA* **111**, E4859–E4868. (doi:10.1073/pnas.1323926111)
29. Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014 Tracer v1.6. See http://beast.bio.ed.ac.uk/Tracer.
30. 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**, 212–226. (doi:10.1093/molbev/msj024)
31. Inoue JG, Donoghue PCJ, Yang Z. 2010 The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.* **59**, 74–89. (doi:10.1093/sysbio/syp078)
32. Warnock RCM, Yang Z, Donoghue PCJ. 2012 Exploring uncertainty in the calibration of the molecular clock. *Biol. Lett.* **8**, 156–159. (doi:10.1098/rsbl.2011.0710)
33. 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. USA* **110**, 12 355–12 360. (doi:10.1073/pnas.1305813110)
34. Garsmeur O, Schnable JC, Almeida A, Jourda C, Hont A, Freeling M. 2014 Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.* **31**, 448–454. (doi:10.1093/molbev/mst230)
35. Doyle JJ, Egan AN. 2010 Dating the origins of polyploidy events. *New Phytol.* **186**, 73–85. (doi:10.1111/j.1469-8137.2009.03118.x)
36. Marcet-Houben M, Gabaldón T. 2015 Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol.* **13**, e1002220. (doi:10.1371/journal.pbio.1002220)
37. Murat F, Armero A, Pont C, Klopp C, Salse J. 2017 Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* **49**, 490–496. (doi:10.1038/ng.3813)
38. Foster CSP *et al.* 2017 Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst. Biol.* **66**, 338–351.
39. Arrigo N, Barker MS. 2012 Rarely successful polyploids and their legacy in plant genomes. *Curr. Opin. Plant Biol.* **15**, 140–146. (doi:10.1016/j.pbi.2012.03.010)
40. Madlung A. 2013 Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* **110**, 99–104. (doi:10.1038/hdy.2012.79)
41. Soltis DE, Visger CJ, Soltis PS. 2014 The polyploidy revolution then. . .and now: Stebbins revisited. *Am. J. Bot.* **101**, 1057–1078. (doi:10.3732/ajb.1400178)
42. Fawcett JA, Van de Peer Y. 2010 Angiosperm polyploids and their road to evolutionary success. *Trends Evol. Biol.* **2**, 3. (doi:10.4081/eb.2010.e3)
43. Chanderbali AS 2016 Evolving ideas on the origin and evolution of flowers: new perspectives in the genomic era. *Genetics* **202**, 1255–1265. (doi:10.1534/genetics.115.182964)
44. Robertson FM *et al.* 2017 Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* **18**, 111. (doi:10.1186/s13059-017-1241-z)
45. Donoghue PCJ, Purnell MA. 2005 Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.* **20**, 312–319. (doi:10.1016/j.tree.2005.04.008)