Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/3/7/e1603195/DC1

Supplementary Materials for

Revisiting ancestral polyploidy in plants

Colin Ruprecht, Rolf Lohaus, Kevin Vanneste, Marek Mutwil, Zoran Nikoloski,
Yves Van de Peer, Staffan Persson

Published 5 July 2017, *Sci. Adv.* **3**, e1603195 (2017)
DOI: 10.1126/sciadv.1603195

This PDF file includes:

- Differences in duplication age estimates between (ME)(ME), (ME)(E), and (ME)(M) trees
- Timing and phylogenetic placement of the WGD derived from the *Amborella* synteny data
- table S1. Statistical analyses of modality of duplication age distributions
- table S2. Bayesian information criterion values from the EMMIX analysis to estimate the number of significant components in the duplication age distributions
- fig. S1. Phylogenetic gene trees with (ME)(E) and (ME)(M) topology.
- fig. S2. Distribution of duplication estimates in the (ME)(M) and (ME)(E) trees.
- fig. S3. Correlation of the gene tree topology with the duplication estimate.
- fig. S4. Correlation of the gene tree topology with the duplication estimate.
- fig. S5. Identification of significant components using SiZer.
- fig. S6. Replacing all ME node calibrations does not change the general pattern of the distribution of gene duplication ages estimated by BEAST.
- References (*17–22*)

Supplementary Materials

Differences in duplication age estimates between (ME)(ME), (ME)(E), and (ME)(M) trees

As described in the main text, we found a strong difference in the age estimates of gene duplications between the two main classes of gene tree topologies, i.e., topology (ME)(ME) vs. topologies (ME)(M) and (ME)(E). Interestingly, we also observed a slight difference in the gene duplication ages between the (ME)(M) and (ME)(E) topologies (fig. S1). In the original r8s data (*5*), the age distribution of both topologies showed a strong peak between ~160–220 Mya; however, we found slightly more of the older gene duplication estimates in the (ME)(E) than in the (ME)(M) trees (fig. S2, a and b). Similarly, we found slightly more of the older gene duplication estimates and a wider age distribution in the (ME)(E) trees than in the (ME)(M) trees in our BEAST analysis (fig. S3, d and f). The strong skew of gene duplication ages in the (ME)(ME) trees also found in our BEAST analysis, as described in the main text, was caused by the un-calibrated (MEO)-child clade of the gene duplication nodes (fig. S3, a and b). In the same way, we hypothesized that the slight skew of duplication ages in the (ME)(E) trees compared to the (ME)(M) trees was due to the un-calibrated child clade of the gene duplication nodes in these trees, i.e., the (E)- or (M)-child clade of a gene duplication node.

We therefore extracted the estimated ages for the crown nodes of eudicots (hereafter named E-nodes, see fig. S1a) and crown nodes of monocots (hereafter named M-nodes, see fig. S1b) in each of the child clades of the ME- or gene duplication nodes from our BEAST dataset that used the same calibration constraints as Jiao et al. (*5*) (some gene trees did not contain E- or M-nodes in some (E)- or (M)-clades, if such clades only consisted of one eudicot or monocot gene). For the (ME)(M) trees, we found similar age estimates for the M-nodes in the un-calibrated (M)-child clade of the gene duplication node and those in the (M)-child clade of the calibrated (ME)-node (fig. S3e, dark gray and light gray, respectively). However, for the (ME)(E) trees the two E-nodes in each of the two child clades of a gene duplication node did not have similar age estimates (fig. S3c). We found a peak at ~100–200 Mya for the E-nodes in the un-calibrated (E)-child clade of the gene duplication node (fig. S3c, red) that was about 50 My older than the peak for the E-nodes in the (E)-child clade of the calibrated (ME)-node (fig. S3c, light red); these latter E-nodes are often the calibrated RO-nodes, thus the age estimate of most of the former un-calibrated E-nodes are in disagreement with the RO-calibration in a similar way than the age estimates of the MEO-nodes were in disagreement with the ME-calibration in (ME)(ME) trees. Notably, approximately half of these E-nodes in the un-calibrated (E)-child clade were estimated even older than the ME-

nodes, which were fixed to a maximum of 150 Mya due to the calibration constraints, thus estimating crown divergence of eudicot lineages earlier than 150 Mya (fig. S3c, red vs. blue; very few M-nodes in the un-calibrated (M)-branches of (ME)(M) trees showed a similar pattern, fig. S3e, dark gray vs. blue). This was again surprising, since any crown node of eudicots (or crown node of monocots) should by definition be younger than the (calibrated) split of monocots and dicots. These data suggested that the older gene duplication age estimates in (ME)(E) trees compared to (ME)(M) trees could, at least partially, be caused by older E-nodes in the un-calibrated (E)-clade of (ME)(E) trees (fig. S3, d and f). Albeit less strong, this skew is similar to the effect of the un-calibrated (MEO)-clades that we observed for (ME)(ME) trees (fig. S3, a and b).

Thus, we decided to calibrate also both child clades of a gene duplication node in the (ME)(M) and (ME)(E) trees by using additional calibration constraints on the originally un-calibrated M- and E-child nodes of a gene duplication node (if such child-nodes existed, see Methods). As expected, this resulted in a slight shift to younger duplication age estimates in (ME)(E) trees but almost no such shift in (ME)(M) trees (fig. S4, c–f). Despite these more coherent calibration constraints, a difference in the shape of the age distributions of gene duplications between (ME)(M) and (ME)(E) trees, and a difference of ~50 My in the estimated age of duplications between (ME)(ME) trees and (ME)(M)/(ME)(E) trees remained.

Timing and phylogenetic placement of the WGD derived from the *Amborella* synteny data

The recent study of the *Amborella* genome showed clear synteny-based evidence for one WGD that occurred prior to the diversification of angiosperms (that this WGD was not an independent, *Amborella*-specific polyploidy event was supported by analysis of synteny between *Amborella* and grape (*Vitis vinifera*)) (*14*). The line of evidence used to place this WGD into the common ancestor of angiosperms and not into the common ancestor of seed plants followed three steps of analyses:

(i) As in the Jiao et al. (*5*) study, molecular dating of duplications in phylogenetic trees was performed using exactly the same dating approach (and thus suffering from the same technical issues described herein), but now also including genes from the *Amborella* genome sequence and several additional eudicot and monocot species (Supplementary Material, Section 5 in (*14*)). The same calibration parameters were used with the addition of an age constraint of 125–190 Mya for the split of *Amborella*. They found again a bimodal distribution of gene duplication age estimates

with peaks slightly shifted to an older ~244 Mya and older ~341 Mya, again supporting two WGDs prior to the divergence of *Amborella* (fig. S17 in (*14*)). Likely, the shifts of ~52 and ~22 My, respectively, were mainly due to the additional maximum age constraint of 190 Mya for the divergence of *Amborella*, as this calibration age is very close to the previous younger peak in age estimates for gene duplications at ~192 Mya, attributed to the ancestral angiosperm WGD event. Based on this replicate but potentially unreliable support for two WGDs, it is very plausible to reach the conclusion that the synteny data corresponds to the more recent of the two WGDs, since it is extremely unlikely that traces from an older WGD are kept, while the signs of a recent one are lost.

(ii) The distribution of synonymous substitutions per synonymous site (K_s) of *Amborella* paralogues was analyzed, and the K_s values of gene pairs in the syntenic blocks were correlated with the putative younger WGD in the common ancestor of angiosperms (figs. S18A and S9 in (*14*)). However, the presented data does not seem to be fully consistent with this. First, the overall K_s distribution of *Amborella* (fig. S18A in (*14*)) shows a broad hump at K_s values between ~1.5 and ~3.0. The study uses mixture modelling (as implemented in EMMIX) to look for potential signatures of WGDs in the K_s distribution. EMMIX identified four significant components, two of which overlap within the broad hump and were attributed to the two ancient WGDs; the other two at younger K_s values were ignored. Mixture modelling techniques have been shown to suffer from overfitting (*17*, *2*) and can be misled at higher K_s values by K_s saturation and stochasticity effects (*18*). Thus, the broad hump evident in the *Amborella* K_s distribution could equally well be the signature of only one ancient WGD event, (dispersed by the stochasticity of substitutions over such a long period of time) or of one ancient WGD overlapping with an artificial K_s saturation peak (*18*). Second, and notwithstanding the previous issue, the K_s values of the syntenic gene pairs do in fact not seem to unambiguously correspond to the younger WGD with a peak K_s of around ~1.84 as identified by EMMIX. As far as discernible from the relevant fig. S9 in (*14*), the K_s values of syntenic gene pairs show a peak around a log-transformed K_s of ~0.34–0.40 (fig. S9B, color coded green) and many of individual syntenic gene pairs (dots in fig. S9A in (*14*)) have even higher K_s values (fig. S9B, color coded red/orange/yellow in (*14*)). Log-transformed K_s values of ~0.34–0.40 and higher correspond to K_s values of ~2.19–2.51 and higher, and hence, according to the identified significant components, more likely correspond to the older WGD with a peak K_s of around ~2.56 (fig. S18A, color coded green in (*14*)), placed into the common ancestor of seed plants.

(iii) Gene pairs from the identified *Amborella* syntenic blocks were analyzed for their phylogenetic placement. Of the 466 syntenic *Amborella* gene pairs only 44 could be directly shown to have a duplication on the branch preceding the divergence of extant angiosperms (Supplementary Material, Section 5 and table S10 in (*14*)). In addition, a more focused analysis using phylogenies based on merged and expanded orthogroups was conducted for six out of the 47 identified syntenic blocks. Of the 155 syntenic gene pairs contained on these six blocks 70 could be shown to have a duplication on the branch preceding the divergence of extant angiosperms (bootstrap support ≥ 80%, table S11 in (*14*)). This was presented as phylogenetic evidence confirming that the identified syntenic blocks originated from the putative WGD in the common ancestor of angiosperms. However, it is unclear whether any of these phylogenetic analyses placed (or were even designed to place) these duplications of syntenic pairs into the common ancestor of angiosperms versus the common ancestor of seed plants, or simply just anytime pre-angiosperm divergence. About a third of these phylogenies were unrooted (table S11 in (*14*)), which the authors argue can still be used because the corresponding gene pairs belong to "regions [that] had already been identified as syntenic blocks within the specified K_s window". As discussed in the previous section, there is no equivocal support that this K_s window and the K_s values of individual syntenic gene pairs correspond specifically to a WGD in the common ancestor of angiosperms. Moreover, if our reading of table S11 in (*14*) is correct, more than 60% of the constructed gene trees did not contain gymnosperms in the root, which would be necessary to distinguish whether the gene pairs in the syntenic blocks stem from a WGD that occurred before or after the divergence of gymnosperms.

In summary, the analysis of the *Amborella* K_s distribution (ii) and the phylogenetic analysis of syntenic *Amborella* gene pairs (iii) both independently show that the synteny data very likely belongs to a WGD pre-dating the origin and diversification of extant angiosperms. However, neither analysis is able to establish a more precise timing for this WGD event, i.e., to unambiguously distinguish whether this WGD occurred either before the diversification of angiosperms or already before the diversification of seed plants, and neither analysis is able to provide any substantive evidence for two ancient plant WGDs in this time frame. Only if one assumes, *a priori*, based on the phylogenomic dating in Jiao et al. (*5*) and (i), the existence of two such ancient polyploidy events in plants, could both analyses lend some support to the phylogenetic placement of this WGD into the common ancestor of angiosperms and not into the common ancestor seed plants. Here, we question the underlying phylogenomic dating result and

therefore potentially also the timing and phylogenetic placement, but not the existence, of the WGD derived from the *Amborella* synteny data.

table S1. Statistical analyses of modality of duplication age distributions.

Data	EMMIX BIC[1]	Bimodal coefficient[2]	Dip test[3]	SiZer[4]
Jiao et al. (*5*) original data (Fig. 2a)	3	0.631	< 0.001	2
BEAST (ME-node calibration only, as in Jiao et al. (*5*), Fig. 3c)	2	0.492	0.984	1
BEAST (ME- and MEO, E or M-node calibrations, Fig. 3f)	2	0.450	0.951	1
BEAST (no ME- or MEO-node calibrations, fig. S6a)	2	0.426	0.995	1

[1] number of components of the best mixture model identified by EMMIX using the Bayesian information criterion (BIC, see table S2 for BIC values)

[2] Bimodal coefficient (BC), benchmark value $BC_{crit} \approx 0.555$ is expected for a uniform distribution; values > BC_{crit} indicate bimodality, whereas values < BC_{crit} indicate unimodality (*19*)

[3] p-value of Hartigans' dip test statistic for unimodality as implemented in the R package `diptest`, $p < 0.05$ indicates bimodality (*20*, *21*)

[4] number of components identified with SiZer (*22*) (see fig. S5 for details)

table S2. Bayesian information criterion values from the EMMIX analysis to estimate the number of significant components in the duplication age distributions.

Data	Number of components				
	1	**2**	**3**	**4**	**5**
Jiao et al. original data (Fig. 2a)	8838	8546	**8532**	8544	8561
BEAST (ME-node calibration only, as in Jiao et al. (*5*), Fig. 3c)	8319	**8263**	8275	8291	8311
BEAST (ME- and MEO, E or M-node calibrations, Fig. 3f)	8008	**7976**	7992	8009	8027
BEAST (no ME- or MEO-node calibrations, fig. S6a)	4661	**4656**	4669	4680	4695

The smallest BIC values indicate the most likely number of components and are highlighted in bold. Note that the smallest BIC value for the Jiao et al. (*5*) dataset is corresponding to three components instead of two components, probably due to the fact that we scored only one duplication per tree and therefore slightly less duplications than Jiao et al. (*5*).

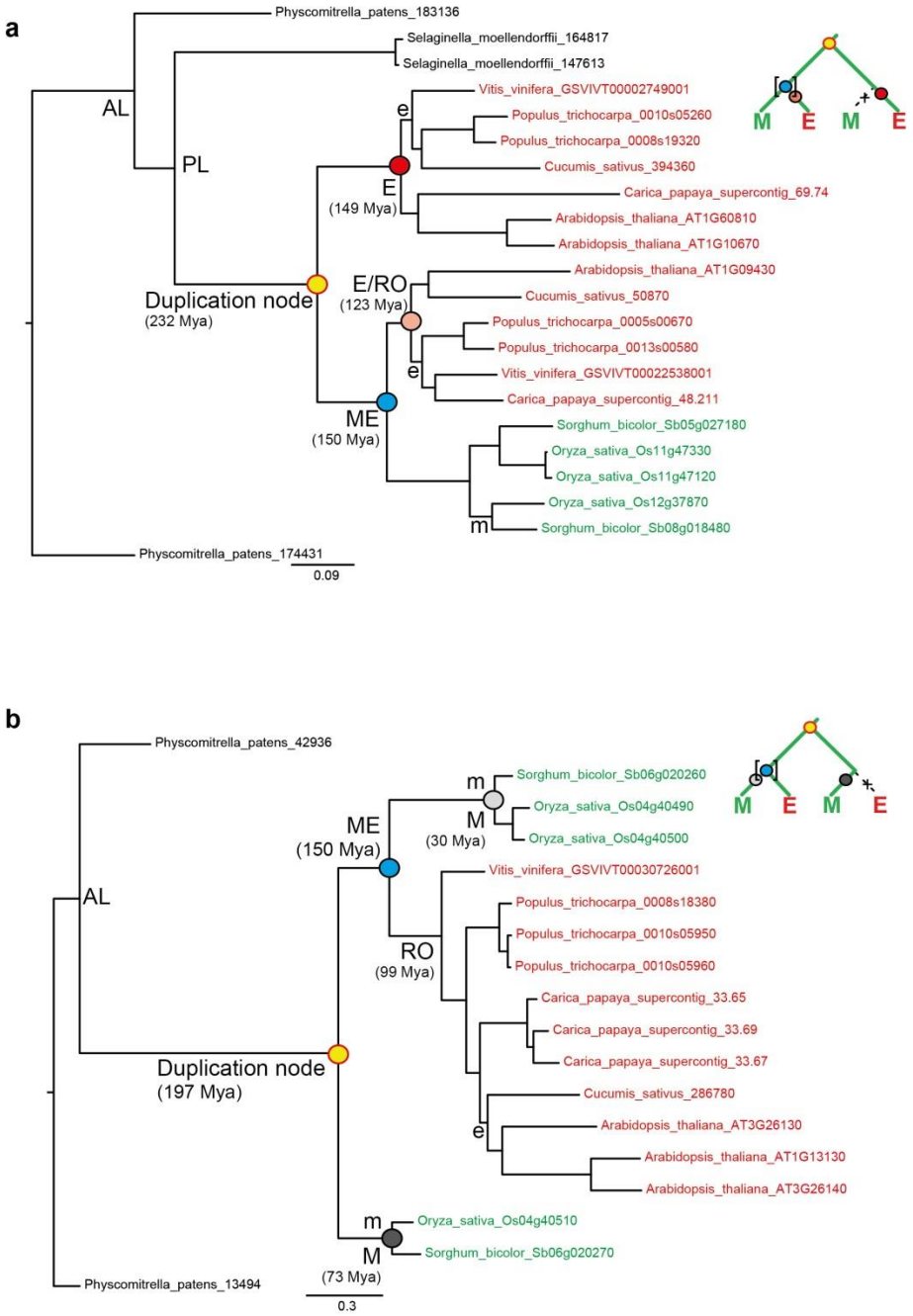

fig. S1. Phylogenetic gene trees with (ME)(E) and (ME)(M) topology. (**a**) Example of gene tree RAxML_1045 from the dataset of Jiao et al. (*5*) with (ME)(E) topology. (**b**) Example of gene tree RAxML_1688 from the dataset of Jiao et al. (*5*) with (ME)(M) topology. Age estimates of nodes extracted from the original r8s output file of Jiao et al. (*5*) are given in parentheses (the calibration of only the ME-node in the Jiao et al. (*5*) r8s analysis is indicated by square brackets in the small schematic trees at the top right). BEAST age estimates were extracted for colored nodes (shown in fig. S3 and S4). m- and e-nodes were calibrated for the set of BEAST analyses shown in fig. S6.

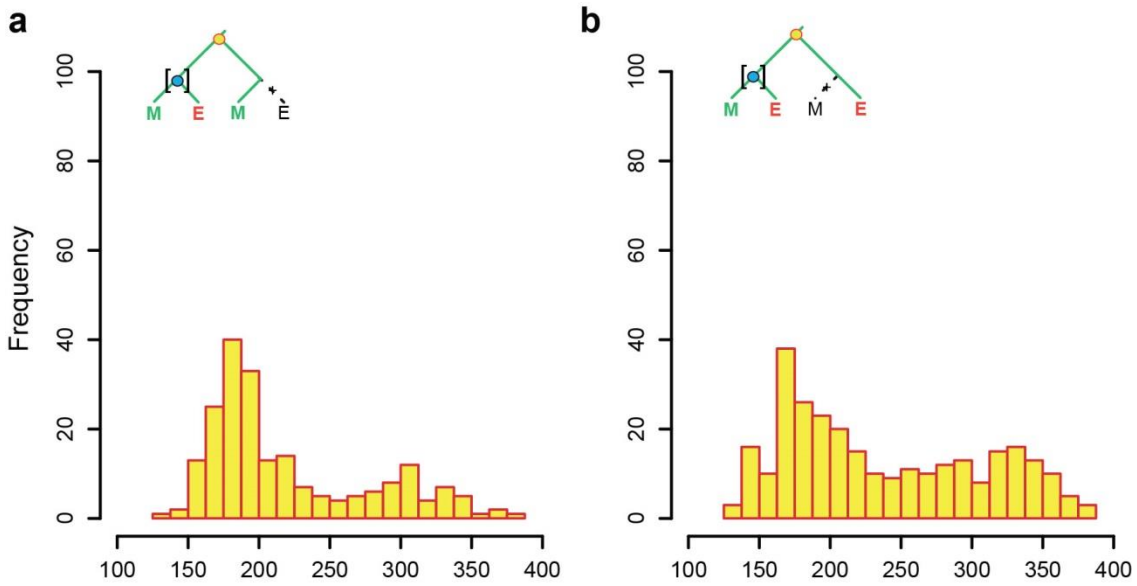

fig. S2. Distribution of duplication estimates in the (ME)(M) and (ME)(E) trees. Age estimates of nodes were extracted from the original r8s output file of Jiao et al. (*5*). (**a**) Age estimates of gene duplication nodes in the (ME)(M) trees (n=208). (**b**) Age estimates of gene duplication nodes in the (ME)(E) trees (n=286). The small schematic trees illustrate the general topology of the corresponding trees (yellow circle indicates the gene duplication node). Square brackets indicate which node/clade had been calibrated.

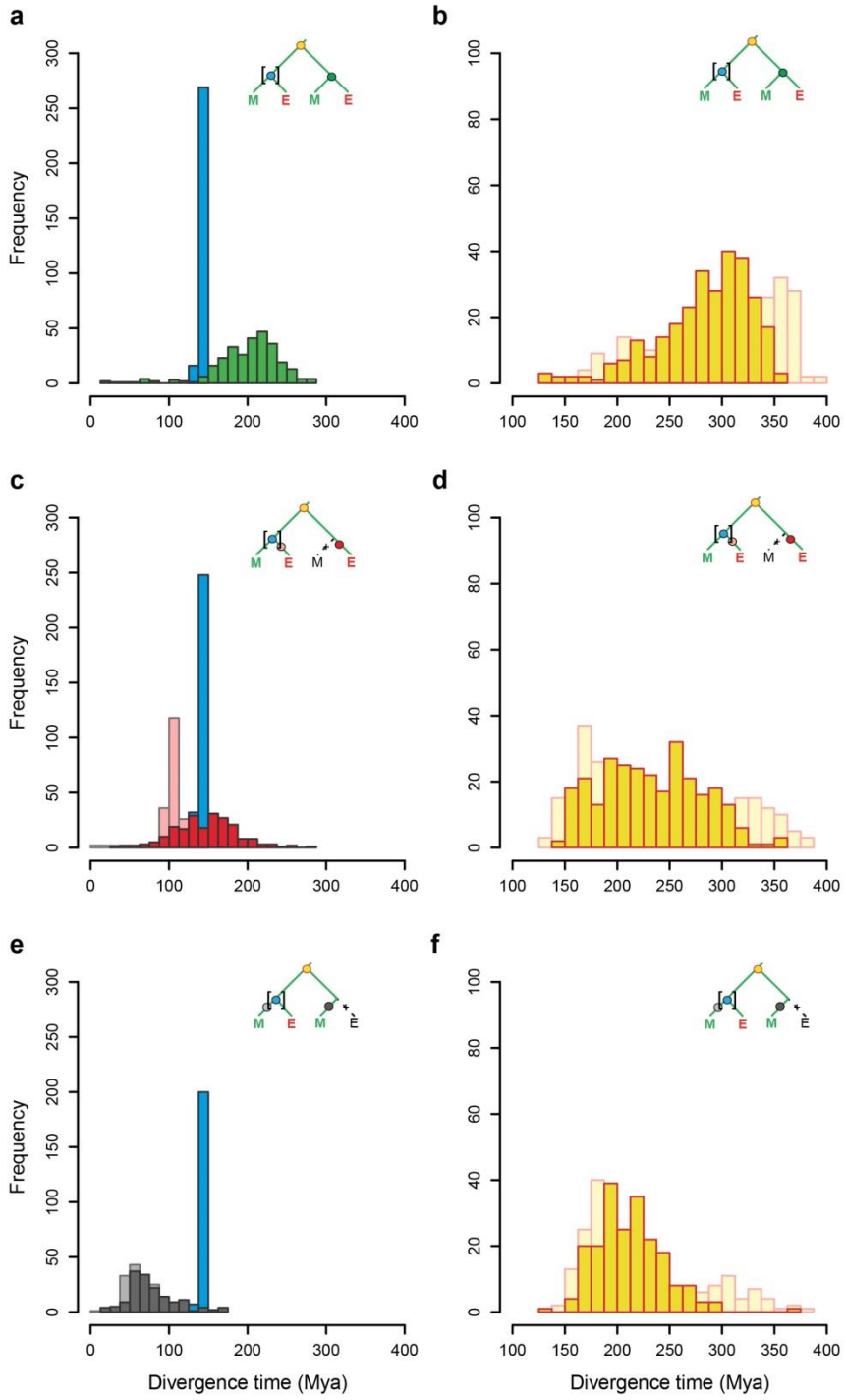

fig. S3. Correlation of the gene tree topology with the duplication estimate. Ages of nodes estimated using BEAST with calibration of only the ME-node in all trees (the same as in Jiao et al. (*5*), illustrated by blue node with square brackets in small schematic trees). (**a**) Age estimates of ME-nodes (blue) and MEO-nodes (green) in (ME)(ME) trees (n=285). (**b**) Age estimates of gene duplication nodes (yellow) in the (ME)(ME) trees (n=285). (**c**) Age estimates in (ME)(E) trees (n=280) of ME-nodes (blue, n=280), E-nodes in the un-calibrated (E)-clade, if such a node exists (red, n=211), and E-child nodes of the calibrated (ME)-node, if such a node exists (light red, n=203). (**d**) Age estimates of gene duplication nodes (yellow) in the (ME)(E) trees (n=280). (**e**), Age estimates in (ME)(M) trees (n=207) of ME-nodes (blue, n=207), M-nodes in the un-calibrated (M)-clade, if such a node exists (dark gray, n=156), and M-child nodes of the calibrated (ME)-node, if such a node exists (light gray, n=159). (**f**) Age estimates of gene duplication nodes (yellow) in the (ME)(M) trees (n=207). In all panels, the small schematic trees illustrate the general topology of the corresponding trees with color of nodes matching color of age estimates displayed in the histograms. Square brackets indicate which node/clade has been calibrated.

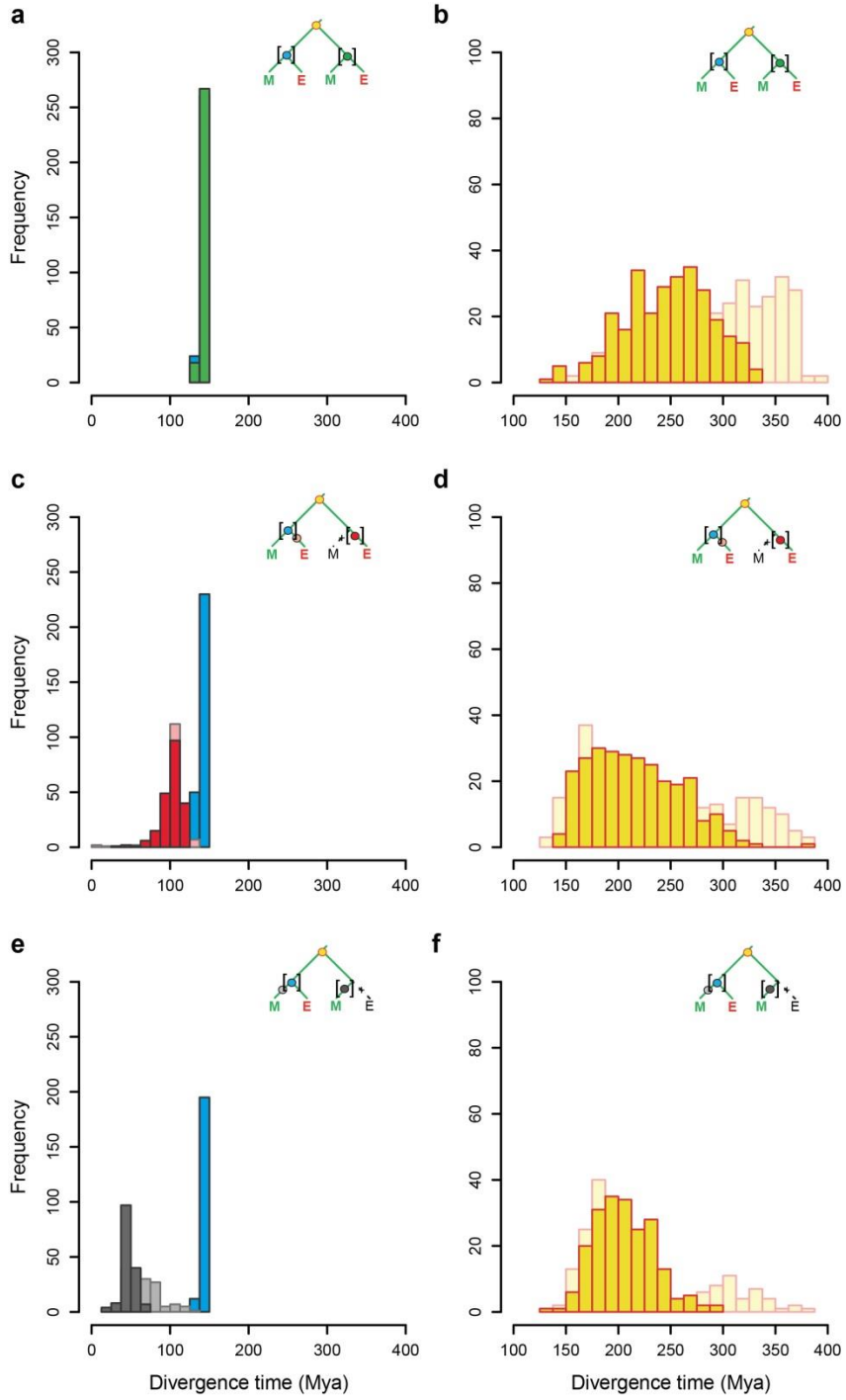

fig. S4. Correlation of the gene tree topology with the duplication estimate. Ages of nodes estimated using BEAST with calibrations of both the ME-node and MEO-, E- or M-node (i.e., both child nodes of a gene duplication node). (**a**) Age estimates of ME-nodes (blue) and MEO-nodes (green) in (ME)(ME) trees (n=285). (**b**) Age estimates of gene duplication nodes (yellow) in the (ME)(ME) trees (n=285). (**c**) Age estimates in (ME)(E) trees (n=280) of ME-nodes (blue, n=280), calibrated E-child nodes of the gene duplication node, if such a node exists (red, n=211), and E-child nodes of the calibrated (ME)-node, if such a node exists (light red, n=203). (**d**) Age estimates of gene duplication nodes (yellow) in the (ME)(E) trees (n=280). (**e**) Age estimates in (ME)(M) trees (n=207) of ME-nodes (blue, n=207), calibrated M-child nodes of the gene duplication node, if such a node exists (dark gray, n=156), and M-child nodes of the calibrated (ME)-node, if such a node exists (light gray, n=159). (**f**) Age estimates of gene duplication nodes (yellow) in the (ME)(M) trees (n=207). In all panels, the small schematic trees illustrate the general topology of the corresponding trees with color of nodes matching color of age estimates displayed in the histograms. Square brackets indicate which node/clade has been calibrated.

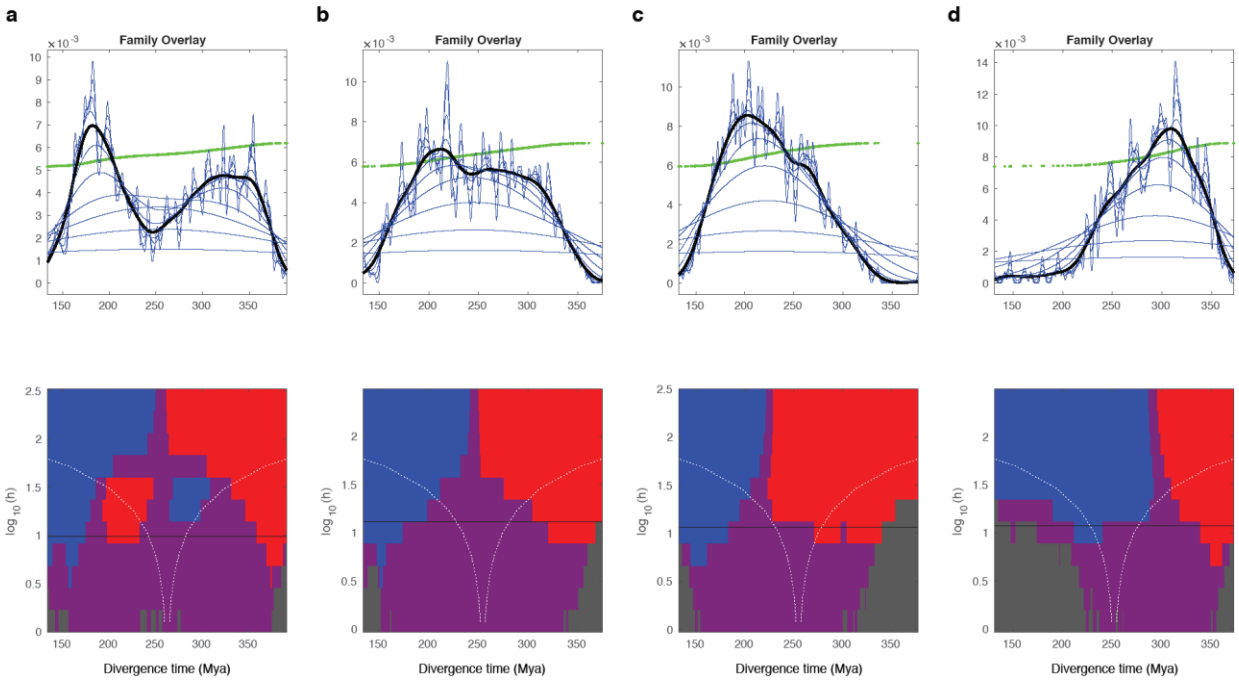

fig. S5. Identification of significant components using SiZer. (**a**) Jiao et al. (*5*) original data (Fig. 2a). (**b**) BEAST data with ME-node calibration only, as in Jiao et al. (*5*) (Fig. 3c). (**c**) BEAST data with ME-node and MEO-, E- or M-node calibrations (Fig. 3f). (**d**) BEAST data without ME- and MEO-node calibrations (fig. S6a). See also table S1.

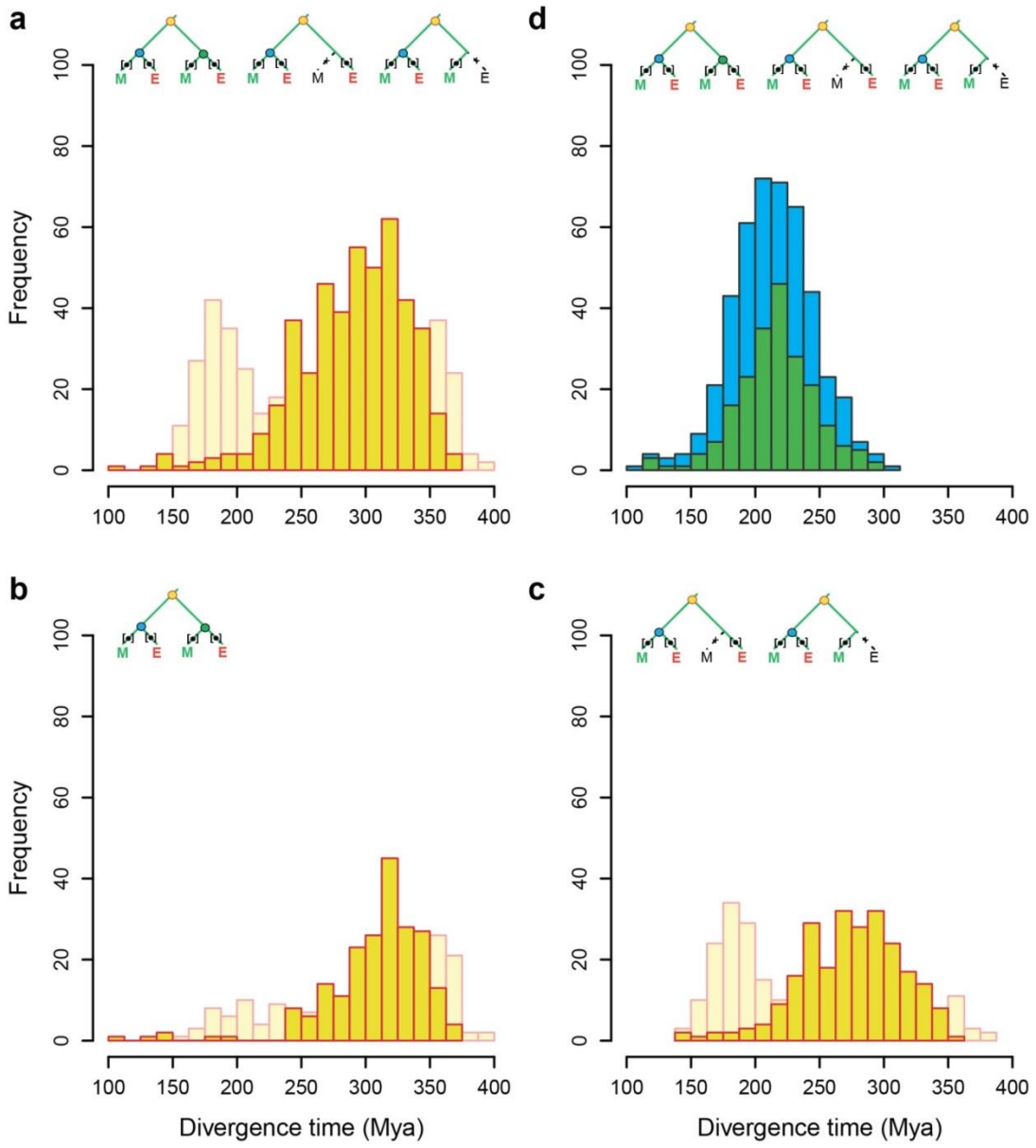

fig. S6. Replacing all ME node calibrations does not change the general pattern of the distribution of gene duplication ages estimated by BEAST. (**a**) Age estimates of gene duplication nodes in all trees (n=455). (**b**) Age estimates of gene duplication nodes in (ME)(ME) trees (n=213). (**c**) Age estimates of gene duplication nodes in (ME)(E) and (ME)(M) trees (n=242). (**d**) Age estimates of ME-nodes (blue, n=455) and MEO-nodes (green, n=213) in all trees. Nodes within all (M)- and (E)-clades were calibrated instead of the ME-nodes in this BEAST analysis, reducing the number of analyzed trees (see Methods). For comparison, the distribution of the corresponding original data from Jiao et al. (*5*) is given in light yellow in the background. In all panels, the small schematic trees illustrate the general topology of the corresponding trees with color of nodes matching color of age estimates displayed in the histograms (yellow circle indicates the gene duplication node; blue and green circles indicate ME- and MEO-nodes, respectively). Square brackets indicate which node/clade has been calibrated.