Supplementary Appendix for

The Origin of the Legumes is a Complex Paleopolyploid Phylogenomic Tangle closely associated with the Cretaceous-Paleogene (K-Pg) Mass Extinction Event

Authors:
Erik J.M. Koenen[1*], Dario I. Ojeda[2,3], Freek T. Bakker[4], Jan J. Wieringa[5], Catherine Kidner[6,7], Olivier J. Hardy[2], R. Toby Pennington[6,8], Patrick S. Herendeen[9], Anne Bruneau[10] and Colin E. Hughes[1]

[1] Department of Systematic and Evolutionary Botany, University of Zurich, Zollikerstrasse 107, CH-8008, Zurich, Switzerland
[2] Service Évolution Biologique et Écologie, Faculté des Sciences, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, 1050, Brussels, Belgium
[3] Norwegian Institute of Bioeconomy Research, Høgskoleveien 8, 1433 Ås, Norway
[4] Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands
[5] Naturalis Biodiversity Center, Leiden, Darwinweg 2, 2333 CR, Leiden, The Netherlands
[6] Royal Botanic Gardens Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, UK
[7] School of Biological Sciences, University of Edinburgh, King's Buildings, Mayfield Rd, Edinburgh, EH9 3JU, UK
[8] Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, EX4 4RJ, UK
[9] Chicago Botanic Garden, 1000 Lake Cook Rd, Glencoe, IL 60022, USA
[10] Institut de Recherche en Biologie Végétale and Département de Sciences Biologiques, Université de Montréal, 4101 Sherbrooke St E, Montreal, QC H1X 2B2, Canada

* Correspondence to be sent to: Zollikerstrasse 107, CH-8008, Zurich, Switzerland; phone: +41 (0)44 634 84 16; email: erik.koenen@systbot.uzh.ch.

Supplementary Appendix S1

In this appendix we describe the fossil calibrations used for the time-scaling analysis, the custom *a priori* fixed local clock models and the alternative fossil calibration priors that were explored, and briefly discuss the effects of alternative fossil calibrations in Cercidoideae and Detarioideae.

Fossil Time-calibration Priors

Non-legume Eudicot Fossils – These were taken from Magallón et al. (2015) and are thoroughly discussed in the supplementary information of that article. The numbers listed in Table 1 match those used in the Supplementary Information Methods 1 of Magallón et al. (2015). We have followed their fossil placements although our more limited taxon sampling means that some minimum ages are placed on deeper nodes. The only exception is the stem node of Fagales (calibration X14), which was here calibrated using the oldest fossil prior used by Xing et al. (2014). All minimum ages were updated to the latest version of the Geologic Time Scale (v. 4.0; Gradstein et al., 2012).

Legume Fossils – The selection of legume fossils used here for calibrating the divergence time estimation analyses (Table 1; Fig. 5) differs from previous legume time tree studies (Lavin et al., 2005; Bruneau et al., 2008; Simon et al., 2009), as well as in the placement of fossils and in the minimum ages that some of these fossils represent. Calibrations Q2 and Z are used for the first time here. Calibrations A, D, F, G, I2, M2 and Q are labelled according to the schemes of Bruneau et al. (2008) and/or Simon et al. (2009), and differences from

previous studies are discussed here. Other fossils used by Bruneau et al. (2008) and/or Simon et al. (2009) are not used here because of our sparser taxon sampling.

First, we did not fix the crown age of the family, which is critical as it is the most important node for which we want to estimate the age. The oldest definitive legume fossil, a fossil wood from the Early Paleocene of Patagonia (Brea et al., 2008), is used to set a minimum age on the stem node of the family at 63.5 Ma (calibration A, same node as in Bruneau et al. (2008) and Simon et al. (2009), but a new fossil and minimum age). This calibration is probably uninformative because of the long stem of the family, but it is included for completeness. Lyson et al. (2019) document newly discovered older legume fossil fruits and leaflets from the Paleocene of Colorado (65.35 Ma). Although these are the earliest definitive legume fossils documented so far, like the Early Paleocene fossil wood from Patagonia these new fossils would have to be placed as minimum age constraints on the legume stem node, and hence are also likely to be uninformative and were not included here. The oldest crown group fossil, bipinnate leaves from the Late Paleocene of Colombia (Wing et al., 2009; Herrera et al., 2019), is placed on the stem node of Caesalpinioideae with a minimum age of 58 Ma (calibration Z), a new calibration that has not been used in previous studies. This calibration renders the calibration of the stem of Papilionoideae (which is sister to Caesalpinioideae), with fossil flowers of *Barnebyanthus buchananensis* from the Paleocene-Eocene boundary at 56 Ma (Crepet and Herendeen, 1992), redundant.

We find the interpretation of some Early and Middle Eocene fossils, that were used in previous studies to calibrate lineages within crown group Cercidoideae and Detarioideae (Bruneau et al., 2008; Simon et al., 2009) to be problematic. Bruneau et al. (2008: Table 3) already pointed out the large discrepancy in age estimates of Detarioideae between calibrated and non-calibrated analyses. Given that this subfamily has a very long stem lineage, placing Early to Middle Eocene fossils within the crown group would require very high

inferred substitution rates along the stem lineage, while at the same time implying a relatively low substitution rate for the Detarioideae crown group lineages (see Results). Cercidoideae are also subtended by a long stem lineage, leading to similar, although less severe, substitution rate discrepancies than in Detarioideae. We investigate and test this with molecular clock analyses with fixed local clocks, as described below. Here, we discuss the interpretation of these fossils as either stem or crown relatives and how we have calibrated lineages from subfamilies Cercidoideae and Detarioideae.

Bauhinia-like bilobed leaves from the Eocene of Tanzania (c. 46 Ma) (Jacobs and Herendeen, 2004) were used by Bruneau et al. (2008) and Simon et al. (2009) to calibrate the stem lineage of *Bauhinia* s.l.. This leaf type is highly characteristic of Cercidoideae and therefore the fossil is certainly representative of the subfamily. However, even though this type of leaf is not found in *Cercis*, which has been found to be sister to the rest of the genera in the subfamily (Bruneau et al., 2008; Wang et al., 2018), it may not provide a strong apomorphy for crown group Cercidoideae. Leaves in *Bauhinia* s.l. are variously bifoliolate, bilobed or entire, implying that entire leaves like those of *Cercis* have evolved multiple times independently, leading to homoplasy. This means that the bilobed leaves may have been present in the most recent common ancestor (MRCA) or stem relatives of Cercidoideae, and evolved to having an entire lamina in *Cercis.* If the Tanzanian fossils are a possible stem-relative of Cercidoideae, we consider the oldest definitive crown group fossil evidence to be the recently described *Cercis* fossil leaves and fruits from the Late Eocene of Oregon (Jia and Manchester, 2014), at c. 36 Ma (calibration C, a slightly older minimum age than used by Bruneau et al. (2008) and Simon et al. (2009)).

Bifoliolate leaves from the same fossil site in Tanzania as the *Bauhinia* fossil were ascribed to *Aphanocalyx* (Detarioideae) (Herendeen and Jacobs, 2000) based on distinctive venation patterns, after comparing the leaves to all extant legume genera with bifoliolate

leaves. The fossil was used to calibrate the stem lineage of that genus by Bruneau et al. (2008) and Simon et al. (2009). The genus is deeply nested within Detarioideae, also meaning that the difference between age estimates from calibrated and uncalibrated analyses is large (46.0 vs 4.4 Ma; Bruneau et al., 2008: Table 3). While venation patterns can be diagnostic in many cases, they are often variable even within modern genera and likely to be homoplasious. Therefore, these fossils might also represent an extinct lineage, possibly a stem relative of Detarioideae, that had evolved similar leaf morphology to extant *Aphanocalyx.* Moreover, the author of the most recent taxonomic account of *Aphanocalyx* (Wieringa, 1999), Jan Wieringa, does not accept this fossil as belonging to the genus. It also does not fit with the morphology-based phylogeny of *Aphanocalyx* which showed that bifoliolate leaves evolved recently and are derived within *Aphanocalyx* (Wieringa, 1999). In general, leaflet numbers are highly variable across Detarioideae, so relatives of fossils should not be sought only among bifoliolate taxa.

Further evidence of Detarioideae from the Eocene is found at two localities within the Claiborne Formation in western Tennessee, USA. Fruits and leaflets from those sites are ascribed to the genus *Crudia* (Herendeen and Dilcher, 1990). As for the *Aphanocalyx* fossil, the affinities of the fossils were carefully evaluated before concluding that they are related to *Crudia.* Bruneau et al. (2008) and Simon et al. (2009) used this fossil to calibrate the stem of *Crudia* at 45 Ma, but as for the *Aphanocalyx* fossil age, an uncalibrated analysis finds a far younger age (6.9 Ma; Bruneau et al., 2008: Table 3), which is likely an underestimate but it illustrates the contrasting signal in the molecular data. Therefore, placement of this fossil within *Crudia* is doubtful and it may rather represent a related but now extinct taxon. While the venation pattern on the fruit valves and leaf epidermal anatomy are important features to link these fossils to *Crudia*, the raised venation on the fruit valves and twisted petiolules that most strongly resemble *Crudia*, are both homoplasious across Detarioideae, further reinforcing the

possibility that the fossils represent a lineage in Detarioideae that is related to *Crudia* and not the genus itself.

Fossil wood, flowers and amber of *Aulacoxylon sparnacense*, which has previously been interpreted as related to the extant genus *Daniellia* (Detarioideae), from the Early Eocene of the Paris basin (De Franceschi and De Ploëg, 2003), provide the most convincing evidence of fossils representing Early to Middle Eocene crown group members of Detarioideae. The fossil wood has vestured pits and resin canals like modern resin-producing Detarioideae, and the amber deposits are chemically similar to the Dominican ambers. Bruneau et al. (2008) considered the wood and flowers similar to *Daniellia*, but suggested they could also belong to a different genus of resin-producing Detarieae. However, it is also possible that resin-production was already present in stem-relatives of Detarioideae. This is quite likely given that this trait is homoplasious across the resin-producing clade, having apparently been independently gained or lost several times, with only about half of the extant genera in the clade producing resin (Fougère-Danezan et al., 2007). If the production of resin evolved in the ancestral lineage of Detarioideae it would not require many more losses to account for the absence of the trait in the other lineages of the subfamily, because the resin-producing clade branches deeply within Detarioideae and the basal relationships of the subfamily are poorly resolved and understood (Bruneau et al., 2008; de la Estrella et al., 2018). Furthermore, the large majority of genera in the subfamily are confined to the large clade of Amherstieae, so perhaps only a single additional loss of the trait in the lineage leading to this clade could have produced this homoplasious pattern. This suggests that the Paris basin fossils could belong to an extinct genus belonging to the stem group of Detarioideae. Therefore, the *Aulacoxylon* fossils can be used either to calibrate the stem node of the resin producing clade (calibration G^g, as done by Bruneau et al, (2008) and Simon et al., (2009) or the stem node of Detarioideae (calibration G), with a minimum age of 53 Ma.

From the rich collections of legume fossils from the Middle to Late Paleocene of Colombia (Wing et al., 2009; Herrera et al., 2019), one fruit morphotype was recently placed in Detarioideae, based on the presence of resin bodies in the fruit wall (Herrera et al., 2019). While this feature may indeed suggest affinity with Detarioideae, resin-producing taxa are also found in Caesalpinioideae, of which other fossils were found at the same fossil site. Therefore, this feature alone is not enough to relate these fossils with certainty to Detarioideae and we maintain the stem group calibration of 53 Ma for Detarioideae (calibration G), based on the *Aulacoxylon* fossils.

For the disputed age of Dominican amber (Iturralde-Vinent and MacPhee, 1996), an intermediate age of 24 Ma was chosen by Bruneau et al. (2008), which was followed by Simon et al. (2009), but it is preferable to not consider an intermediate age as a valid minimum, but rather to use the minimum age that was estimated for Mexican amber that includes flowers of *Hymenaea mexicana*, the extinct species that presumably produced the amber (Poinar and Brown, 2002), and we calibrate the Detarieae s.s. stem node with a minimum age of 22.5 Ma (calibration F, a more inclusive node than in Bruneau et al, (2008) and Simon et al., (2009), and a different fossil age).

The calibration of the stem group of Styphnolobium and Cladrastis (calibration I2) is the same as used in Bruneau et al. (2008) and Simon et al. (2009), but the minimum age was updated to 37.8 Ma according to the latest version of the Geologic Time Scale (v. 4.0; Gradstein et al., 2012), representing the end of the Middle Eocene (end of the Bartonian). Calibration M2 is the same as used in Simon et al., (2009) but since *Robinia* itself is not sampled here, we place it on the stem node of the robinioid clade (represented here by *Lotus japonicus*) and update the minimum age to the Eocene-Oligocene boundary at 33.9 Ma.

Bruneau et al. (2008) and Simon et al. (2009) also set the ages of several fossil calibrations at the midpoint of the Eocene, at 45 Ma. This led to a bias that was observable in

an LTT plot of legumes (Koenen et al., 2013), and here we prefer to use the minimum boundary ages for these fossils. Although most of these calibrations are not used in our analyses due to sparser taxon sampling, we use one of these fossils, *Acacia*-like polyads, to calibrate the minimum stem age of the clade including all *Acacia* s.l. segregates at 33.9 Ma, the Eocene-Oligocene boundary (calibration Q, same node but younger age than Bruneau et al, (2008) and Simon et al., (2009)). Finally, we add calibration Q2, based on Australian Oligocene polyads with pseudocolpi (Miller et al., 2013), which suggest affinity with *Acacia* s.s., and we calibrate the stem node of that genus with a minimum age of 23 Ma, the Oligocene-Miocene boundary.

Custom Fixed Local Clock Models

To specify the different FLC models, we examined root-to-tip length variation across subclades to specify biologically meaningful *a priori* clock partitions (Fig. S5). The 50kb-inversion clade of papilionoid legumes and the asterids (without *Panax ginseng*) have uniformly longer root-to-tip lengths than the remaining taxa across the tree and were each therefore assigned their own local clocks, with a different clock for all remaining taxa in the tree (this model referred to as FLC3, partitioning of taxa is illustrated in Fig. S5a). A more complex model was specified where the rosid rate was decoupled from the background rate and more clock partitions within the legumes were created for the mimosoids together with the *Cassia* clade because of their longer root-to-tip lengths relative to other Caesalpinioideae and most of the rosid clade and also for the combined clade of Cercidoideae and Detarioideae. This more complex model is referred to as FLC6 (Fig. S5b). The most complex model (FLC8; Fig. S5c) was generated by further partitioning the combined clade of Cercidoideae and Detarioideae with a separate local clock for each subfamily, and one on

their combined stem lineages (this most complex partitioning is also indicated with coloured branches in Fig. 6 and S11-S12 and those of the other FLC models in Figs. S9-S10). The Ranunculales that were pruned for the root-to-tip length calculations were included in the background clock for each FLC model.

Alternative Calibration Priors

The separate clock partitions assigned to Cercidoideae and Detarioideae in the FLC8 model are particularly useful for evaluating the controversial placement of Early and Middle Eocene fossils within their crown groups. This was done by running two analyses under the FLC8 model, one with the same priors as the other analyses, and one with similar placements of these calibrations as in Bruneau et al. (2008) and Simon et al. (2009) (Table 1). Calibration C was replaced with a minimum age of 46 Ma on the stem of Bauhinia, based on fossil leaves from Tanzania (Herendeen and Jacobs, 2000; discussed in the previous section). Calibration G^g was applied on the stem of the resin-producing clade (i.e. the crown node of Detarioideae) instead of on the stem node of Detarioideae. Calibration H^g is taken from Bruneau et al. (2008) and Simon et al. (2009), and is added in the alternative analysis to specify a minimum age of 46 Ma on the stem of *Anthonotha*, based on fossil leaves assigned to the closely related genus *Aphanocalyx* from Tanzania (Herendeen and Jacobs, 2000; but see previous section). We refer to this calibration scheme as “alternative prior 1” (Table S2). Since a separate local clock is assigned to the combined stem lineages of Cercidoideae and Detarioideae, substitution rate estimates for stem and crown groups can be compared under both calibration schemes.

Maximum ages of fossil calibrations were set conservatively, and perhaps overly so, which can lead to a poorly formed joint marginal prior on node ages across the tree (Phillips,

2015). Therefore, we also constructed an alternative prior with less conservative maxima as specified in Table S2 (“alternative prior 2”). These maxima represent boundary ages of older epochs from which the crown or stem group is not known, and in line with ages found by Magallón et al. (2015). These analyses serve to test the sensitivity of the UCLN model to the marginal prior.

Effects of Alternative Detariod and Cercidoid Fossil Calibrations

Different interpretations of Eocene fossils of Cercidoideae and Detarioideae (see Supplementary Methods S1) lead to very different crown age estimates for these clades (Fig. S15h-j). As expected, this also leads to very different substitution rates along the stem lineages of these subfamilies, with rates 10-fold higher than within their crown clades when interpreting these fossils as crown group members (Fig. S14). While it cannot be ruled out that the stem lineages of Cercidoideae and Detarioideae experienced such markedly elevated substitution rates, which could potentially be caused by polyploidization at least along the detarioid stem lineage, it is unlikely that rates were five times higher relative to the rest of the eudicots across all 36 nuclear genes analysed, especially as these genes were chosen because of their approximately clock-like evolution, and given that these two clades comprise long-lived woody perennials. The idea that molecular information from extant taxa could inform that particular fossils are too old to belong to a crown clade is controversial. However, the test we have performed here is similar to the cross-validation method proposed by Near et al. (2005), which also uses molecular data to discover fossil calibration points that do not fit well with a larger set of fossils. Favouring those calibrations that do not lead to extreme substitution rate shifts is more parsimonious, and we believe that additional evidence is necessary to justify the inference of such a strong shift in substitution rates as that observed

in the FLC8 analysis with alternative prior 1 (Fig. S12). While there seems little doubt that the Early Eocene fossils from the Mahenge in Tanzania and the Paris Basin in France do represent Cercidoideae and Detarioideae, the extreme substitution rate heterogeneity implied by their treatment as crown group members suggest that they may better be reinterpreted as stem-relatives of these subfamilies (see additional discussion about the affinities of these fossils above).

References

Brea M., Zamuner A.B., Matheos S.D., Iglesias A., Zucol A.F. 2008. Fossil wood of the Mimosoideae from the early Paleocene of Patagonia, Argentina. Alcheringa. 32:427–441.

Bruneau A., Mercure M., Lewis G.P., Herendeen P.S. 2008. Phylogenetic patterns and diversification in the caesalpinioid legumes. Botany. 86:697–718.

Crepet W.L., Herendeen PS. 1992. Papilionoid flowers from the early Eocene of southeastern North America. In: Herendeen P.S., Dilcher D.L., editors, Advances in legume systematics part 4: The fossil record. Richmond, UK: Royal Botanic Gardens, Kew. p. 43–55.

De Franceschi D., De Ploëg G. 2003. Origine de l’ambre des faciès sparnaciens (Éocène inférieur) du Bassin de Paris: le bois de l’ambre producteur. Geodiversitas. 25:633–647.

de la Estrella M., Forest F., Klitgård B., Lewis G.P., Mackinder B.A., de Queiroz L.P., Bruneau A. 2018. A new phylogeny-based tribal classification of subfamily Detarioideae, an early branching clade of florally diverse tropical arborescent legumes. Sci. Rep. 8(1):6884.

Fougère-Danezan M., Maumont S., Bruneau A. 2007. Relationships among resin-producing Detarieae sl (Leguminosae) as inferred by molecular data. Syst. Bot. 32(4):748–761.

Gradstein F.M., Ogg J.G., Schmitz M.D., Ogg G.M. 2012. The Geologic Time Scale 2012. Boston, USA: Elsevier.

Herendeen P.S., Dilcher D.L. 1990. Reproductive and vegetative evidence for the occurrence of *Crudia* (Leguminosae, Caesalpinioideae) in the Eocene of southeastern North America. Bot. Gaz. 151:402–413.

Herendeen P.S., Jacobs B.F. 2000. Fossil legumes from the Middle Eocene (46.0 Ma) Mahenge Flora of Singida, Tanzania. Am. J. Bot. 87:1358–1366.

Herrera F., Carvalho M.R., Wing S.L., Jaramillo C., Herendeen P.S. 2019. Middle to Late Paleocene Leguminosae fruits and leaves from Colombia. Austr. Syst. Bot. 32:385-408.

Iturralde-Vinent M.A., MacPhee R.D.E. 1996. Age and paleogeographical origin of Dominican amber. Science. 273:1850–1852.

Jacobs B.F., Herendeen P.S. 2004. Eocene dry climate and woodland vegetation in tropical Africa reconstructed from fossil leaves from northern Tanzania. Palaeogeogr. Palaeocl. 213:115–123.

Jia H., Manchester S.R. 2014. Fossil Leaves and Fruits of *Cercis* L. (Leguminosae) from the Eocene of Western North America. Int. J. Plant Sci. 175:601–612.

Koenen E.J.M., De Vos J.M., Atchison G.W., Simon M.F., Schrire B.D., De Souza E.R., de Queiroz L.P., Hughes C.E. 2013. Exploring the tempo of species diversification in legumes. S. Afr. J. Bot. 89:19–30.

Lavin M., Herendeen P.S., Wojciechowski M.F. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. Syst. Biol. 54:575–594.

Lyson, T.R., Miller, I.M., Bercovici, A.D., Weissenburger, K., Fuentes, A.J., Clyde, W.C., Hagadorn, J.W., Butrim, M.J., Johnson, K.R., Fleming, R.F., Barclay, R.S. 2019. Exceptional continental record of biotic recovery after the Cretaceous–Paleogene mass extinction. Science p.eaay2268.

Magallón S., Gómez-Acevedo S., Sánchez-Reyes L.L., Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. 207:437–453.

Miller J.T., Murphy D.J., Ho S.Y.W., Cantrill D.J., Seigler D. 2013. Comparative dating of *Acacia*: combining fossils and multiple phylogenies to infer ages of clades with poor fossil records. Aust. J. Bot. 61:436–445.

Near T.J., Meylan P.A., Shaffer H.B., Meyer A.E.A. 2005. Assessing concordance of fossil calibration points in molecular clock studies: An example using turtles. Am. Nat. 165(2):137–146.

Phillips M.J. 2015. Geomolecular dating and the origin of placental mammals. Syst. Biol. 65(3):546–557.

Poinar Jr G.O., Brown A.E. 2002. *Hymenaea mexicana* sp. nov. (Leguminosae: Caesalpinioideae) from Mexican amber indicates Old World connections. Bot. J. Linn. Soc. 139:125–132.

Simon M.F., Grether R., de Queiroz L.P., Skema C., Pennington R.T., Hughes C.E. 2009. Recent assembly of the Cerrado, a Neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. Proc. Natl. Acad. Sci. USA. 106:20359–20364.

Wang Y.-H., Wicke S., Wang H., Jin J.-J., Chen S.-Y., Zhang S.-D., Li D.-Z., Yi T.-S. 2018. Plastid genome evolution in the early-diverging legume subfamily Cercidoideae (Fabaceae). Front. Plant Sci. 9:138.

Wieringa J.J. 1999. *Monopetalanthus* exit. A systematic study of *Aphanocalyx*, *Bikinia*, *Icuria*, *Michelsonia* and *Tetraberlinia* (Leguminosae, Caesalpinioideae). Agric. Univ. Wagening. Pap. 99(4):1–320.

Wing S.L., Herrera F., Jaramillo C.A., Gómez-Navarro C., Wilf P., Labandeira C.C. 2009. Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest record of Neotropical rainforest. Proc. Natl. Acad. Sci. USA. 106:18627–18632.

Xing Y.X., Onstein R.E., Carter R.J., Stadler T., Linder H.P. 2014. Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity and turnover rates are disconnected. Evolution. 68:2821–2832.

Supplementary Appendix S2

We downloaded the gene trees used for the MAPS analysis for the Caesalpinioideae WGD by the One Thousand Plant Transcriptomes Initiative (2019) from https://bitbucket.org/barkerlab/1kp/src/master/ (data set E24). Visual inspection of these gene trees clearly shows that many of the included terminals likely represent transcriptomic diversity rather than paralogs, because for each species mono- and paraphyletic groups of highly similar sequences that likely represent different alleles and/or isoforms were present. Because these are not indicative of an ancient WGD we reduced this transcriptomic diversity using the mask_tips_by_taxonID_transcripts.py script of Yang & Smith (2014). The majority of these gene trees also had multiple representatives of all six taxa, suggesting that many of these gene trees represent larger gene families derived from more ancient gene duplications. Therefore, we only included gene trees in which the most distant outgroup taxon, *Quillaja saponaria*, was only present once, such that this taxon could be used for rooting the gene trees. Using this approach, 863 cleaned and filtered gene trees (Data S8) remained for analysis with Phyparts (Smith et al., 2015) and Notung (Stolzer et al., 2012). Phyparts was run in duplication mode with default settings, in Notung we performed a phylogenomics analysis with cost of duplication and loss set to 1.5 and 0.1, respectively, and horizontal transfers not allowed.

Both the Phyparts and Notung analyses find the highest numbers of gene duplications mapping to the crown node of Leguminosae (Appendix S2 Figure), in line with the results found in the analyses using our data set (Fig. 1 in the main text). Higher numbers of duplications shared with *Gompholobium polymorphum* (Papilionoideae) compared to those

restricted to Caesalpinioideae are also apparent in this re-analysis of the One Thousand Plant Transcriptomes Initiative (2019) gene tree data (Figure).

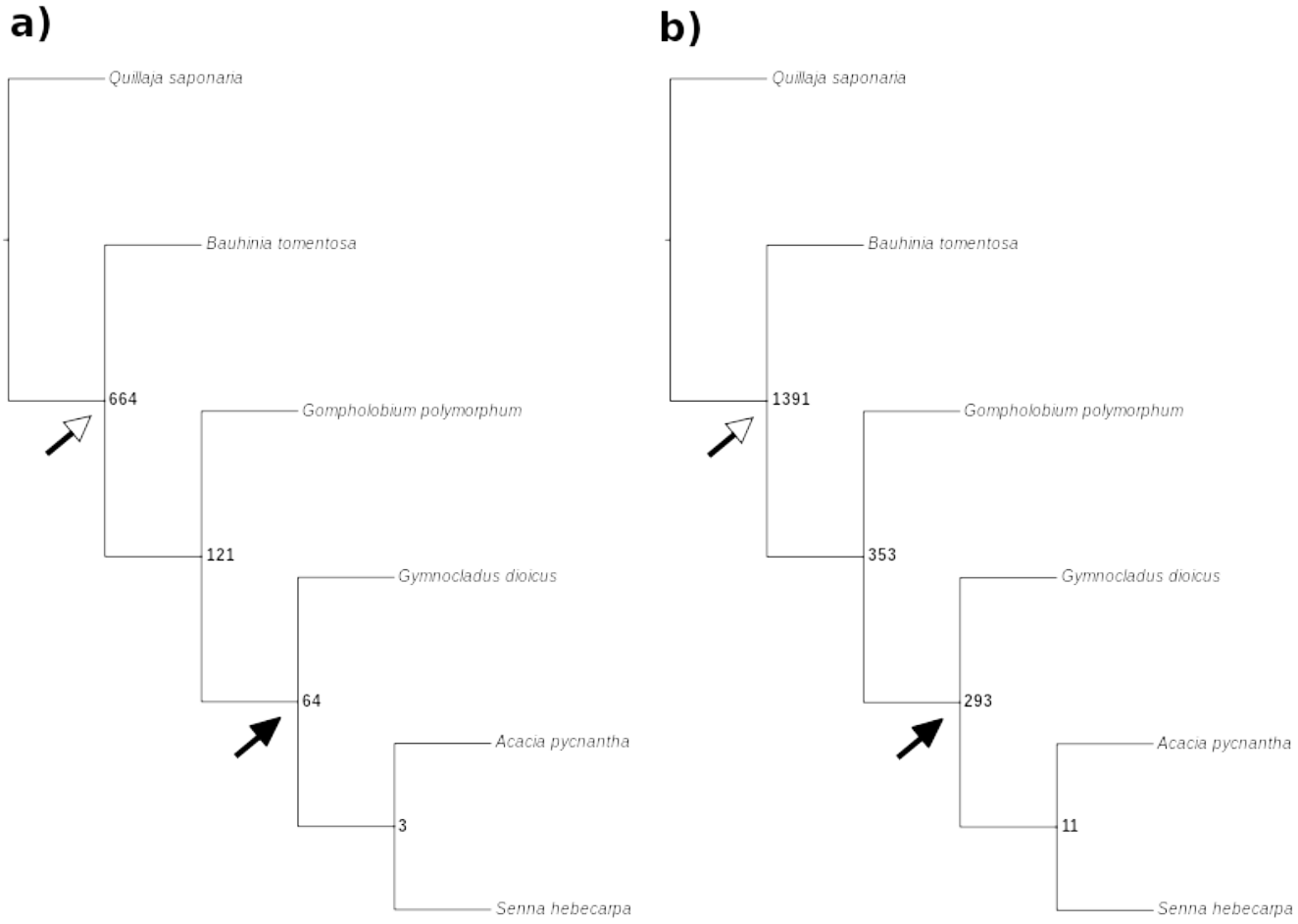

Figure. Numbers of gene duplications in gene trees obtained after filtering data from the One Thousand Plant Transcriptomes Initiative (2019) used for testing the placement of a WGD in Caesalpinioideae, mapped across the species tree using a) Phyparts and b) Notung. The crown nodes of Leguminosae and Caesalpinioideae are indicated with open and filled arrows, respectively.

These results question the evidence presented by the One Thousand Plant Transcriptomes Initiative (2019) based on their MAPS (Li et al., 2015) analysis for a WGD specific to Caesalpinioideae. Perhaps the small and biased sample of taxa, including just a

single representative of Papilionoideae, the sister clade of Caesalpinioideae with which a WGD may likely be shared (see main text), could have led to an inflated number of gene duplications mapping on the Caesalpinioideae crown node in the case of missing data in the larger, unfiltered data set. Taken together, these results suggest that high-throughput analysis of WGD placements, without careful selection of representative taxa and rigorous data filtering may lead to erroneous inferences.

References

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H. and Barker, M.S., 2015. Early genome duplications in conifers and other seed plants. *Science advances*, *1*(10), p.e1501084.

One Thousand Plant Transcriptomes Initiative, 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:150.

Stolzer M., Lai H., Xu M., Sathaye D., Vernot B. and Durand D. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28(18):i409-i415.

Yang Y., Smith S.A. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-loverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Table S1. Taxon occupancy per analysis (number of gene trees | number of sequences)

Species	Phyparts (n=8,038)	Notung (n=8,324)	WGDgc (n=8,324)	GRAMPA (n=4,731)	species tree dating supermatrix (n=36)	legume WGD dating (n=863)	Caesalpinioideae WGD dating (n=246)	Detarioideae WGD dating (n=250)	Papilionoideae WGD dating (n=272)
Acacia koa	6917 | 8314	7256 | 8061			35	783 | 1000			
Acrocarpus fraxinifolius	5002 | 5765	5141 | 5536			27	582 | 678			
Afzelia bella	5165 | 6074	5371 | 5929	5371 | 5929		30	565 | 699		220 | 288	
Albizia julibrissin	6905 | 8980	7183 | 8544	7183 | 8544	3914 | 4548	30	793 | 1142	237 | 400		
Alnus serrulata	5253 | 5768	3633 | 3667	3633 | 3667	1389 | 1390	28	572 | 577	126 | 126	145 | 145	145 | 145
Amaranthus hypochondriacus	3582 | 3725				29				
Anthonotha fragrans	6413 | 8046	6740 | 7793	6740 | 7793	3727 | 3862	35	704 | 932		250 | 391	
Apios americana	7278 | 9576	7711 | 9249	7711 | 9249		35	809 | 1138			
Aquilegia coerulea					34				
Arabidopsis thaliana	4278 | 4537	1706 | 1724				249 | 249			
Arachis ipaensis	6679 | 8933	6748 | 8204	6748 | 8204		35	724 | 1002			272 | 365
Astragalus membranaceus	6104 | 7190	6521 | 7152			31				
Bauhinia tomentosa	6428 | 7153	6745 | 6907	6745 | 6907	3696 | 3727	35	712 | 731			
Bituminaria bituminosa	5810 | 6986	6142 | 6861			30				
Cajanus cajan	7393 | 11163	7654 | 10264	7654 | 10264		33				
Cannabis sativa	6606 | 7331	4424 | 4462	4424 | 4462	1580 | 1589	36	663 | 665	138 | 139	133 | 139	157 | 157
Carica papaya	5368 | 5900	2882 | 2897			26	476 | 478			
Castanea mollissima	6600 | 7797	4683 | 4783	4683 | 4783	1765 | 1777	33	685 | 695	149 | 151	153 | 155	171 | 172
Cercis canadensis	4449 | 4925	4579 | 4655	4579 | 4655		24	498 | 510			
Chamaecrista fasciculata	2206 | 2374	2100 | 2165			11				
Cicer arietinum	7237 | 9901	7508 | 9294	7508 | 9294		36	806 | 1190			
Citrus sinensis	6474 | 7320	3675 | 3704	3675 | 3704		34	598 | 600			
Cladrastis lutea	3779 | 4244	3867 | 4115			19				
Codariocalyx motorius	6008 | 7305	6379 | 7240			33				
Copaifera officinalis	5750 | 6983	6075 | 6849	6075 | 6849		29	658 | 842		240 | 365	
Cucumis sativus	5749 | 6290	3236 | 3280	3236 | 3280	932 | 933	36	484 | 486	86 | 88	79 | 81	118 | 121
Desmanthus illinoensis	1714 | 1919	1614 | 1718			8				
Elaeocarpus photiniifolia	1067 | 1144	435 | 440			4	92 | 92			
Entada abyssinica	6799 | 8535	7112 | 8193	7112 | 8193	3919 | 4435	33	790 | 1109	246 | 384		
Eucalyptus grandis	6224 | 7061	2946 | 3012			36	468 | 476			
Fragaria vesca	5752 | 6532	3949 | 4032	3949 | 4032	1473 | 1485	28	605 | 610	131 | 132	130 | 131	135 | 136
Gleditsia triacanthos	4090 | 4809	4234 | 4627			21	418 | 502			

Glycine max	7871 | 12902	8211 | 11828	8211 | 11828		36	863 | 1551			272 | 525
Glycyrrhiza lepidota	6287 | 7531	6657 | 7399			32				
Gossypium raimondii	7198 | 8632	4309 | 4445	4309 | 4445		36	684 | 703			
Gymnocladus dioicus	1764 | 2002	1827 | 1956			4				
Inga spectabilis	7029 | 8767	7405 | 8534	7405 | 8534	4092 | 4650	35	796 | 1115	239 | 394		
Juglans regia	6882 | 8080	4953 | 5078	4953 | 5078	1915 | 1928	36	736 | 749	171 | 173	180 | 185	188 | 190
Lactuca sativa	4808 | 5134				33				
Lathyrus sativus	5628 | 6723	5923 | 6628			27				
Lens culinaris	2061 | 2339	2001 | 2170			12				
Linum usitatissimum	5136 | 5472	2330 | 2337				360 | 361			
Lotus japonicus	6656 | 8793	6864 | 8210	6864 | 8210		34	766 | 1039			
Lupinus angustifolius	4102 | 4612	4232 | 4503			19				
Lupinus polyphyllus	1962 | 2194	1919 | 2041			6				
Manihot esculenta	7119 | 8412	4219 | 4312	4219 | 4312		35	675 | 682			
Medicago truncatula	7634 | 11190	7952 | 10426	7952 | 10426	4240 | 4545	36	841 | 1311			269 | 464
Microlobius foetidus	7074 | 9088	7305 | 8596	7305 | 8596	3966 | 4583	33	799 | 1140	235 | 396		
Mimulus guttatus	5367 | 5767				34				
Morus notabilis	6386 | 7223	4292 | 4340	4292 | 4340	1507 | 1516	36	665 | 667	140 | 141	131 | 137	150 | 150
Nelumbo nucifera	4957 | 5050				35				
Paeonia lactiflora	2888 | 3096				14				
Panax ginseng	2198 | 2333								
Papaver somniferum					32				
Phaseolus vulgaris	7746 | 11727	8089 | 10882	8089 | 10882		36	857 | 1411			271 | 485
Pisum sativum	6251 | 7822	6554 | 7556			30				
Polygala lutea	4072 | 4349	3074 | 3108	3074 | 3108	1335 | 1339		366 | 366	68 | 69	76 | 78	90 | 92
Populus trichocarpa	7134 | 11983	4190 | 6196	4190 | 6196		36	664 | 1062			
Primula veris	3765 | 3997				29				
Prioria balsamifera	6287 | 7625	6646 | 7485	6646 | 7485		35	686 | 886		244 | 361	
Prosopis alba	2799 | 3274	2790 | 3069			11				
Prunus persica	7260 | 8487	5045 | 5156	5045 | 5156	1823 | 1837	36	764 | 772	162 | 164	150 | 153	168 | 169
Punica granatum	3825 | 4205	1649 | 1684			21	283 | 289			
Quillaja saponaria	4680 | 5024	4214 | 4239	4214 | 4239	2114 | 2117	24	515 | 516	145 | 145	164 | 164	145 | 145
Salix purpurea	7058 | 11780	4128 | 6094	4128 | 6094		35	655 | 1050			
Senna hebecarpa	1643 | 1835	1573 | 1658			6				
Solanum tuberosum	5070 | 5529				28				
Styphnolobium japonicum	6964 | 7781	7233 | 7421			34	757 | 780			

Theobroma cacao	7156 | 8588	4304 | 4428	4304 | 4428		36	683 | 701
Trifolium pratense	4936 | 5893	5126 | 5729			19	
Tripterygium wilfordii	4551 | 4847	2446 | 2458			25	380 | 382
Vicia faba	4277 | 5054	4350 | 4833			25	
Vigna radiata	6526 | 9119	6726 | 8500	6726 | 8500		33	
Vitis vinifera	5768 | 6576				32	
Xanthocercis zambesiaca	6496 | 7518	6934 | 7444			34	716 | 813
Zenia insignis	5442 | 6012	5569 | 5684	5569 | 5684	2824 | 2868	25	610 | 616

Table S2. Age intervals specified for the fossil calibration priors under different alternative priors.

Calibration	Definition	MRCA taxon 1	MRCA taxon 2	Prior	Alternative prior 1	Alternative prior 2
eudicots						
26	CG eudicots	*Aquilegia coerulea*	*Medicago truncatula*	normal (mean 126.0, stdev 1.0)	normal (mean 126.0, stdev 1.0)	normal (mean 126.0, stdev 1.0)
27	CG Ranunculales	*Aquilegia coerulea*	*Papaver somniferum*	uniform (min 113.0, max 126.0)	uniform (min 113.0, max 126.0)	uniform (min 113.0, max 126.0)
38	CG Pentapetalae	*Nelumbo nucifera*	*Medicago truncatula*	uniform (min 100.0, max 126.0)	uniform (min 100.0, max 126.0)	uniform (min 100.0, max 126.0)
48	SG Ericales	*Primula veris*	*Solanum tuberosum*	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 113.0)
94	SG Myrtaceae	*Eucalyptus grandis*	*Punica granatum*	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 100.0)
105	SG Brassicales	*Carica papaya*	*Theobroma cacao*	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 126.0)	uniform (min 89.8, max 100.0)
112	CG Rosaceae	*Fragaria vesca*	*Prunus persica*	uniform (min 49.4, max 126.0)	uniform (min 49.4, max 126.0)	uniform (min 49.4, max 66.0)
116	SG Cannabaceae	*Cannabis sativa*	*Morus notabilis*	uniform (min 66.0, max 126.0)	uniform (min 66.0, max 126.0)	uniform (min 66.0, max 83.6)
122	SG Juglandaceae	*Alnus serrulata*	*Juglans regia*	uniform (min 64.4, max 126.0)	uniform (min 64.4, max 126.0)	uniform (min 64.4, max 83.6)
133	SG Populus	*Populus trichocarpa*	*Salix purpurea*	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 56.0)
X14	SG Fagales	*Alnus serrulata*	*Medicago truncatula*	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 126.0)	uniform (min 83.6, max 126.0)
legumes						
A	SG Leguminosae	*Medicago truncatula*	*Quillaja saponaria*	uniform	uniform	uniform

				(min 63.5, max 126.0)	(min 63.5, max 126.0)	(min 63.5, max 100.0)
C	SG Cercis	*Cercis canadensis*	*Bauhinia tomentosa*	uniform (min 36.0, max 126.0)		uniform (min 36.0, max 83.6)
C&	SG Bauhinia	*Bauhinia tomentosa*	*Cercis canadensis*		uniform (min 46.0, max 126.0)	
F	CG Resin-producing clade	*Copaifera officinalis*	*Prioria balsamifera*	uniform (min 22.5, max 126.0)	uniform (min 22.5, max 126.0)	uniform (min 22.5, max 66.0)
G	SG Detarioideae	*Copaifera officinalis*	*Bauhinia tomentosa*	uniform (min 53.0, max 126.0)		uniform (min 53.0, max 83.6)
G&	SG Resin-producing clade	*Copaifera officinalis*	*Anthonotha fragrans*		uniform (min 53.0, max 126.0)	
H&	CG Amherstieae	*Afzelia bella*	*Anthonotha fragrans*		uniform (min 46.0, max 126.0)	
I2	SG Styphnolobium/Cladrastis	*Styphnolobium japonicum*	*Medicago truncatula*	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 126.0)	uniform (min 37.8, max 66.0)
M2	SG Robinioid clade	*Lotus japonicus*	*Medicago truncatula*	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 66.0)
Q	SG Acacieae/Ingeae	*Albizia julibrissin*	*Prosopis alba*	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 126.0)	uniform (min 33.9, max 66.0)
Q2	SG Acacia s.s.	*Acacia koa*	*Albizia julibrissin*	uniform (min 23.0, max 126.0)	uniform (min 23.0, max 126.0)	uniform (min 23.0, max 56.0)
Z	SG Caesalpinioideae	*Albizia julibrissin*	*Medicago truncatula*	uniform (min 58.0, max 126.0)	uniform (min 58.0, max 126.0)	uniform (min 58.0, max 83.6)

Table S3. Crown node age estimates and priors (95% HPD intervals) for selected nodes in the different analyses.

Clade	Eurosids	Fabales	Leguminosae	Cercidoideae	Detarioideae	Caesalpinioideae	Papilionoideae
Standard prior							
Marginal prior	108.24 - 123.51	84.93 - 112.50	80.03 - 109.70	36.00 - 80.60	28.82 - 87.20	56.37 - 95.91	59.24 - 97.17
UCLN	105.49 - 113.40	78.45 - 95.39	65.47 - 86.45	36.00 - 53.97	25.47 - 42.98	54.11 - 74.49	55.19 - 73.58
RLC	113.45 - 117.94	96.60 - 103.85	73.46 - 81.18	39.34 - 46.74	31.52 - 36.43	55.76 - 63.75	49.05 - 54.38
strict clock	110.02 - 112.48	79.68 - 83.77	66.94 - 69.55	36.00 - 36.66	26.25 - 28.71	56.01 - 59.12	56.89 - 59.47
FLC 3 clocks	111.78 - 114.08	80.83 - 85.21	65.99 - 68.85	36.00 - 36.85	27.69 - 30.59	56.09 - 59.04	47.39 - 50.03
FLC 6 clocks	111.17 - 113.52	80.31 - 84.71	65.74 - 68.81	36.00 - 36.86	27.53 - 30.94	56.10 - 59.20	47.24 - 49.86
FLC 8 clocks	110.97 - 113.37	79.56 - 84.07	64.63 - 67.64	36.00 - 52.41	27.00 - 49.18	55.53 - 58.51	46.98 - 49.60
Alternative prior 1 (Bruneau et al. 2008 for Cercidoideae and Detarioideae)							
Marginal prior	105.79 - 122.12	86.70 - 113.29	81.58 - 110.21	46.00 - 85.81	53.02 - 90.87	57.41 - 97.44	60.58 - 98.35
FLC 8 clocks	111.07 - 113.42	79.59 - 84.08	64.81 - 67.96	57.35 - 63.89	55.38 - 63.49	55.73 - 58.76	46.96 - 49.59
Alternative prior 2 (tighter maxima)							
Marginal prior	101.13 - 119.59	80.02 - 100.00	73.26 - 96.55	36.00 - 71.05	28.69 - 75.26	53.24 - 80.94	54.14 - 79.20
UCLN	105.15 - 112.81	79.62 - 93.31	69.98 - 84.50	36.00 - 52.87	25.22 - 45.43	57.36 - 72.64	57.08 - 70.59

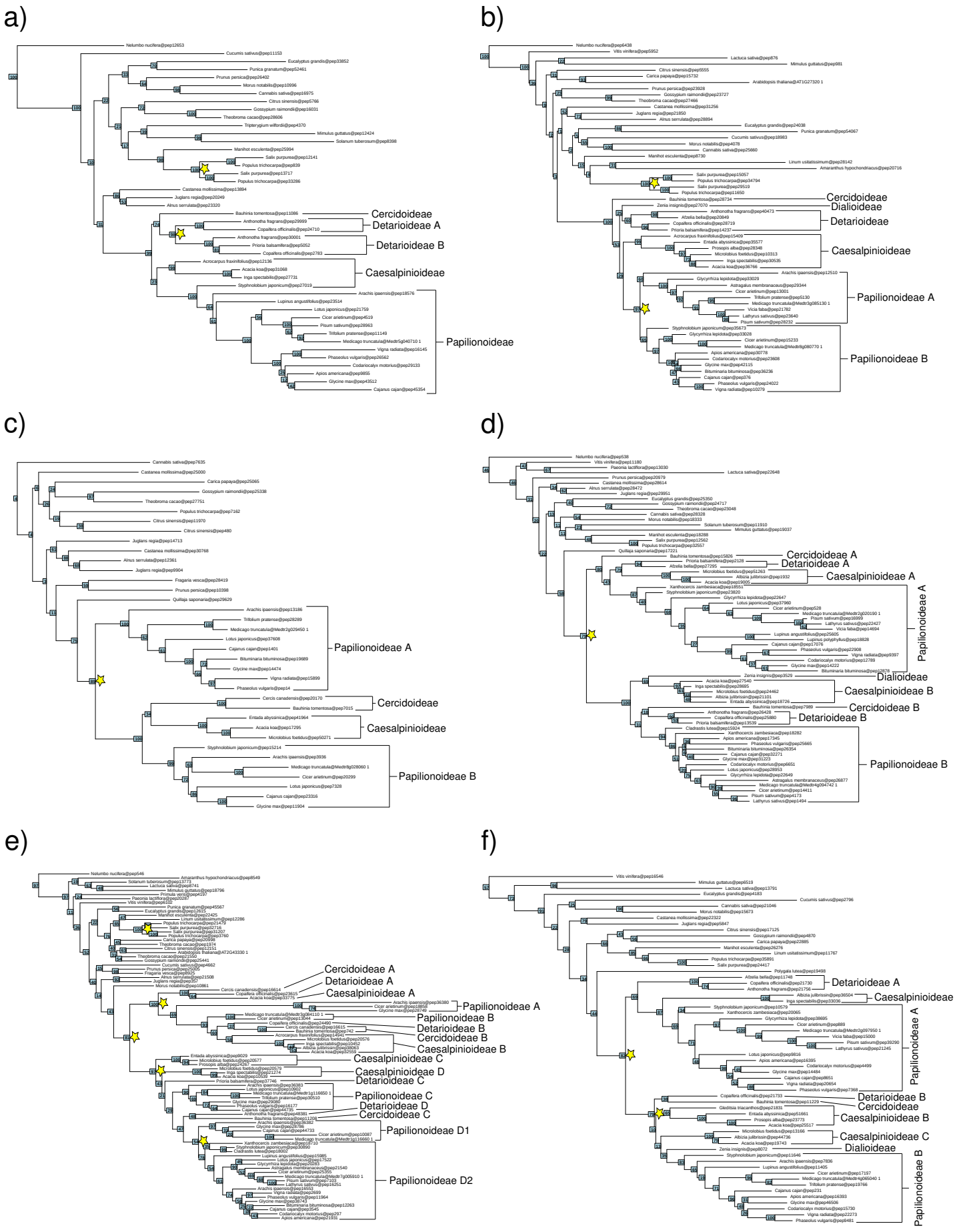

Figure S1. Examples of homolog clusters with gene duplications in legumes that passed the bootstrap filter. Yellow stars behind nodes indicate locations of gene duplications, numbers on nodes indicate bootstrap support. The plotted gene trees are extracted from (A) cluster3675_1rr_1rr, showing a duplication subtending Detarioideae, (B) cluster1032_1rr_1rr, showing a duplication subtending Papilionoideae, (C) cluster1248_1rr_1rr and (D) cluster2941_1rr_1rr, both with a duplication subtending the legume family. Trees for (E) cluster51_7rr_1rr and (F) cluster544_1rr_1rr show evidence of more than one duplication, including one specific to Papilionoideae in the former.

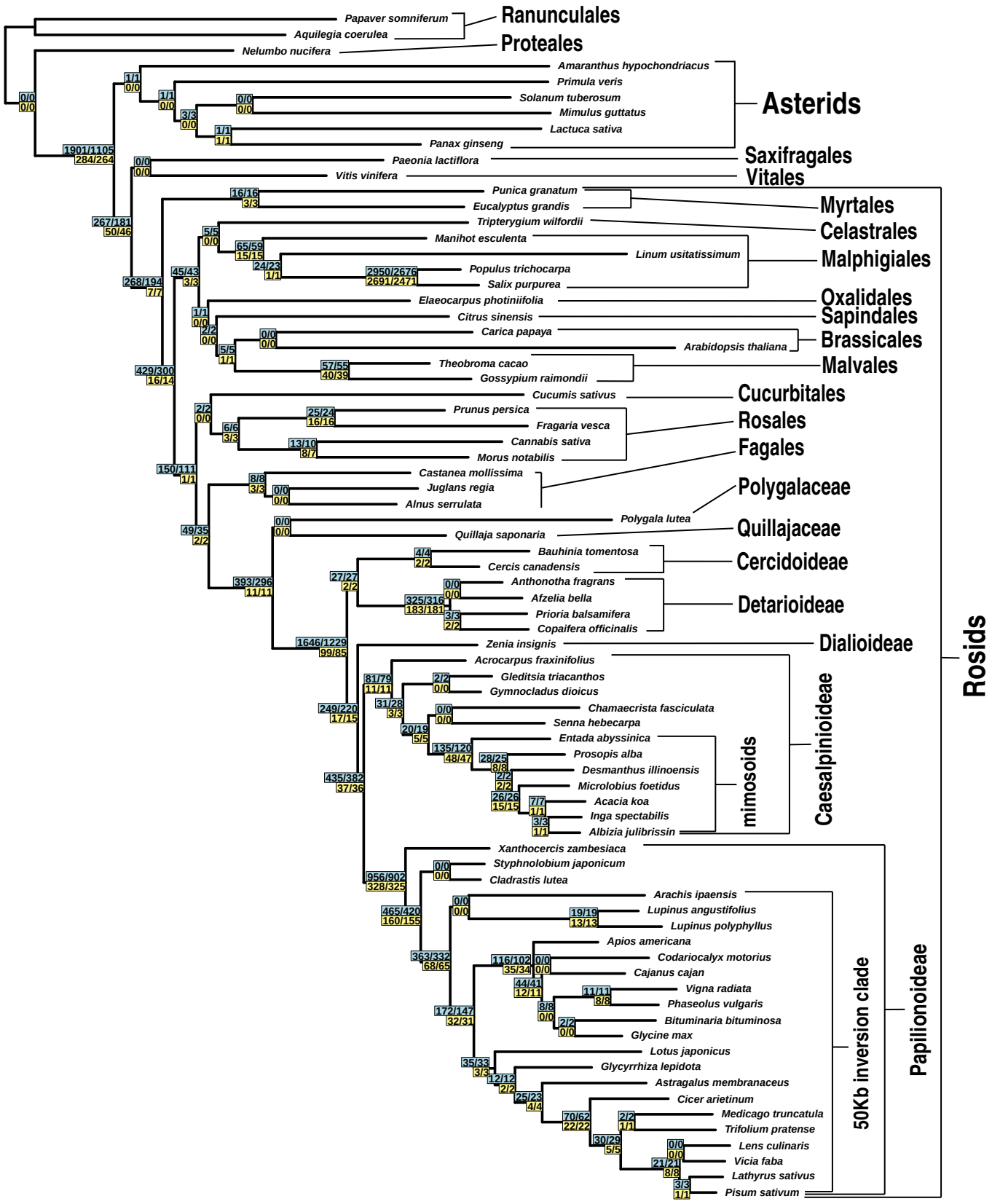

Figure S2. Numbers of gene duplications mapped across the phylogeny with Phyparts. The topology used is the ML topology of the nuclear concatenated alignment of 1,103 genes, duplications were counted from 8,038 homolog clusters. Numbers above branches (with blue background) and below branches (with yellow background) represent numbers of duplications and numbers of homolog trees with duplications, without or with a bootstrap filter of 50%, respectively.

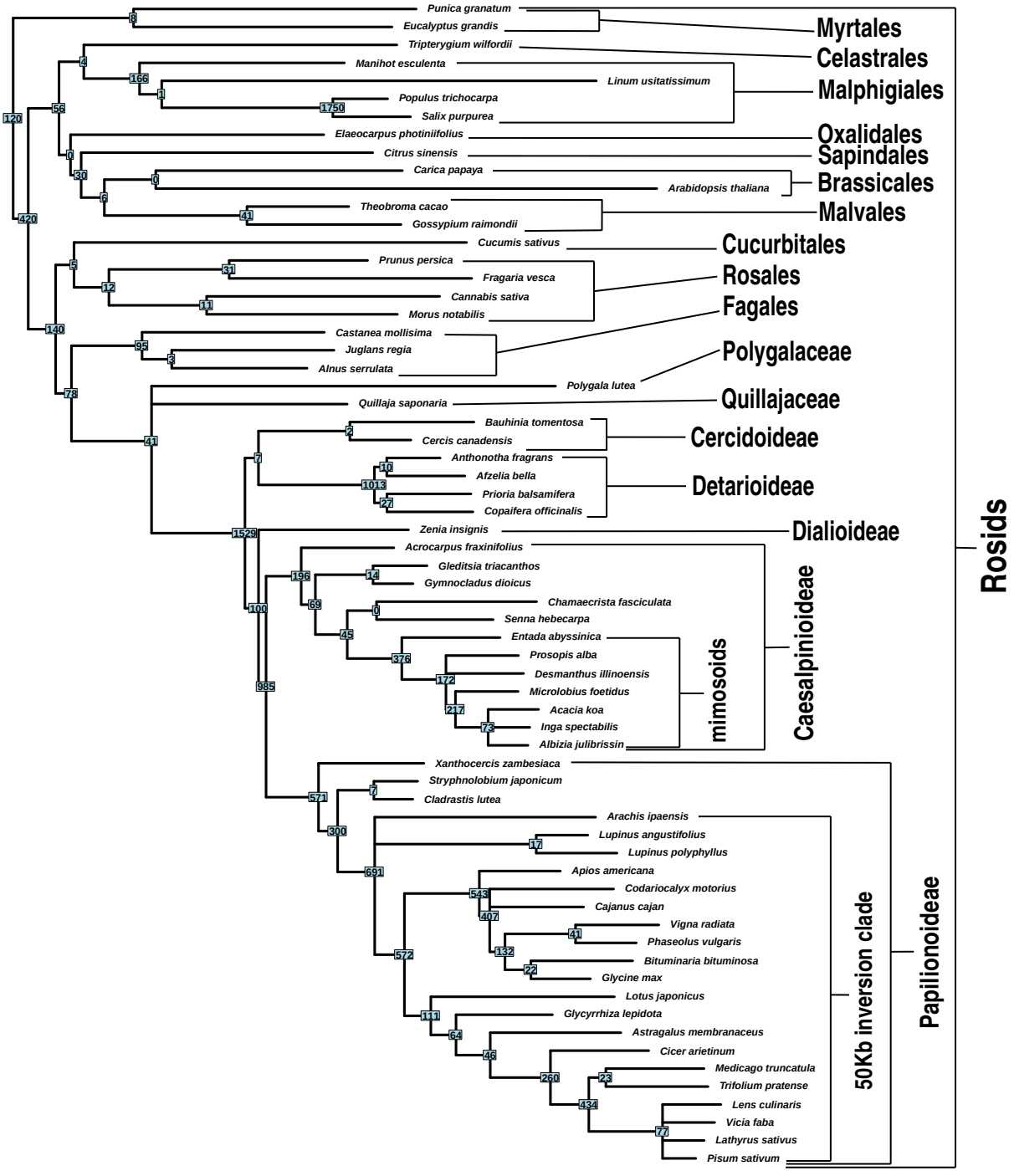

Figure S3. Numbers of gene duplications as estimated by Notung, mapped across the species tree with six polytomies that were introduced manually to account for incomplete lineage sorting. The topology used is the rosid portion of the ML topology of the nuclear concatenated alignment of 1,103 genes of Koenen et al. (2019), duplications were counted from 8,324 homolog clusters.

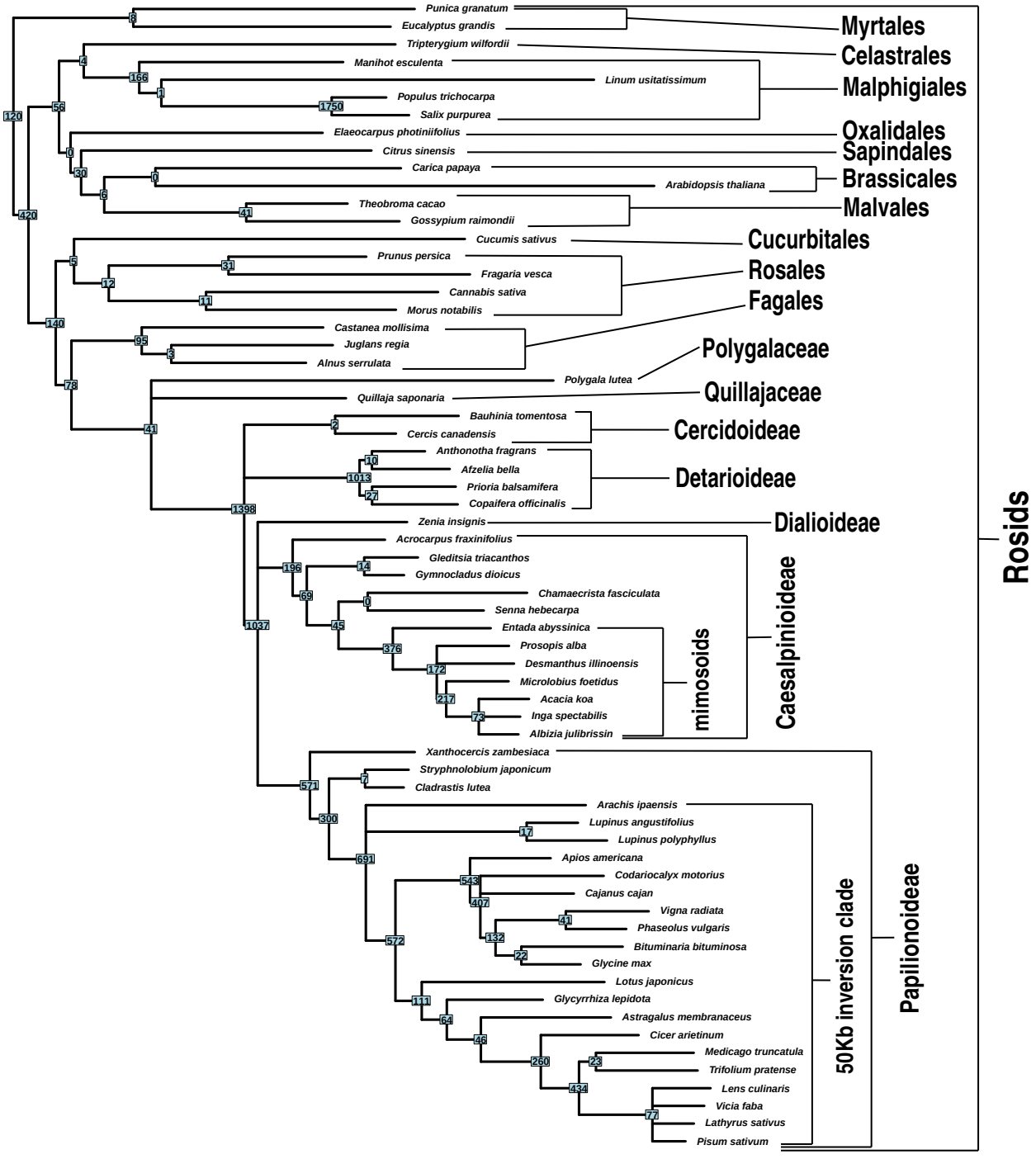

Figure S4. Numbers of gene duplications as estimated by Notung, mapped across the species tree with eight polytomies, including two along the legume backbone, that were introduced manually to account for incomplete lineage sorting. The topology used is the rosid portion of the ML topology of the nuclear concatenated alignment of 1,103 genes of Koenen et al. (2019), duplications were counted from 8,324 homolog clusters.

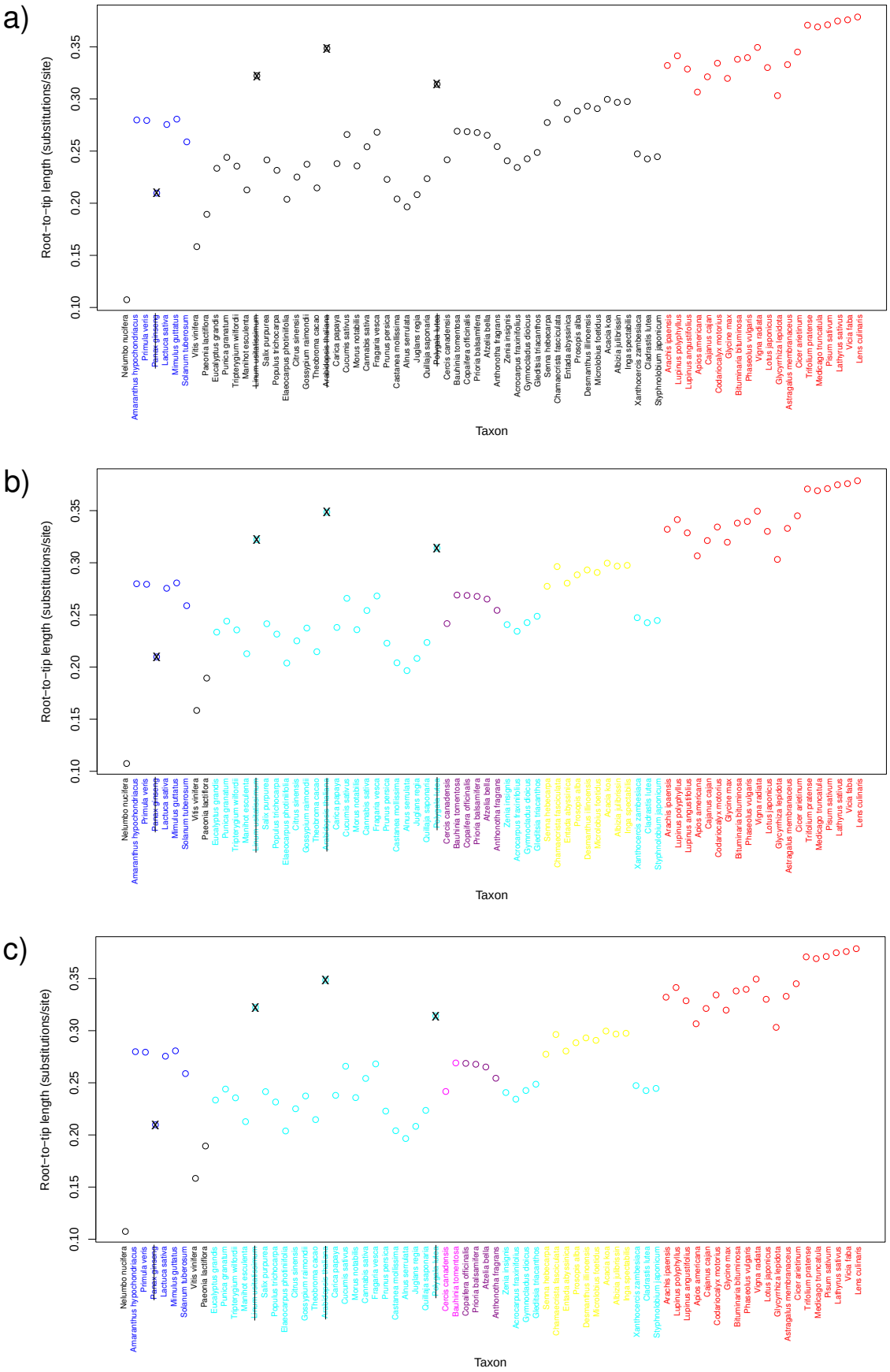

Figure S5. Root-to-tip lengths per taxon with partitions of fixed local clocks indicated. Pruned taxa with outlier root-to-tip lengths are indicated with an X, partitions are indicated with colours. (A) FLC3, (B) FLC6, (C) FLC8.

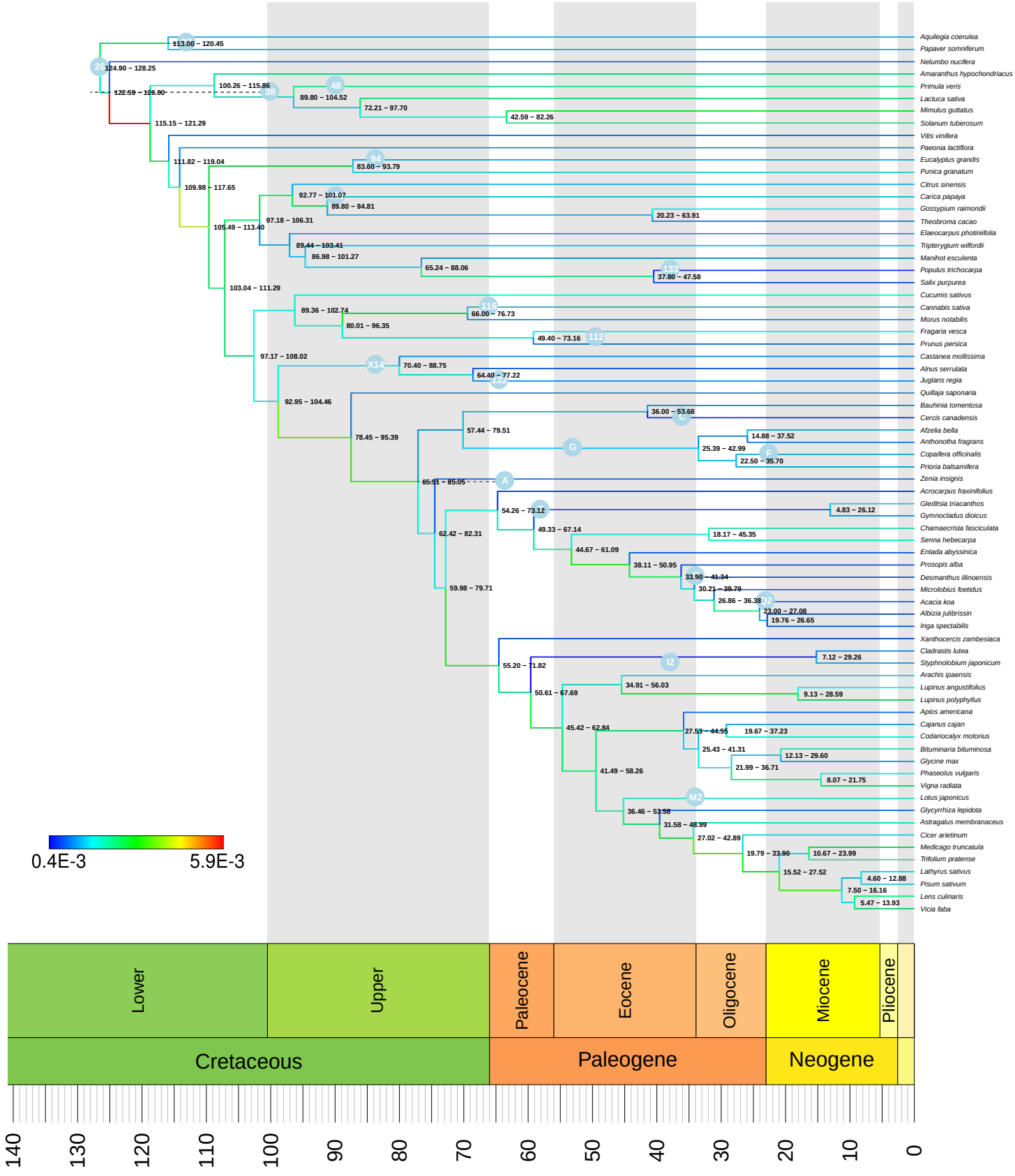

Figure S6. Chronogram estimated under the UCLN clock model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by coloured branches, as indicated by the colour legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

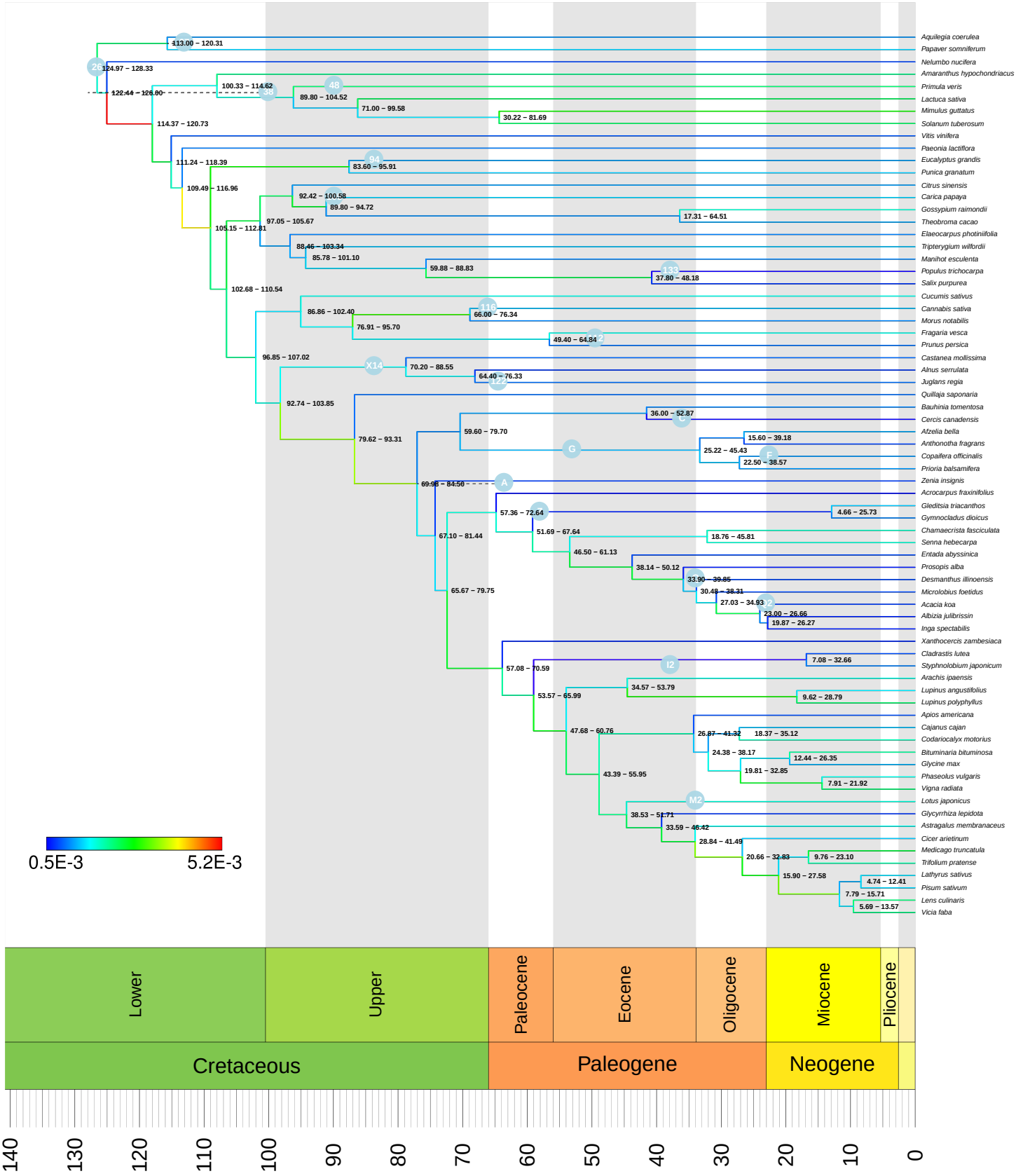

Figure S7. Chronogram estimated under the UCLN clock model, with alternative prior 2. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by coloured branches, as indicated by the colour legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

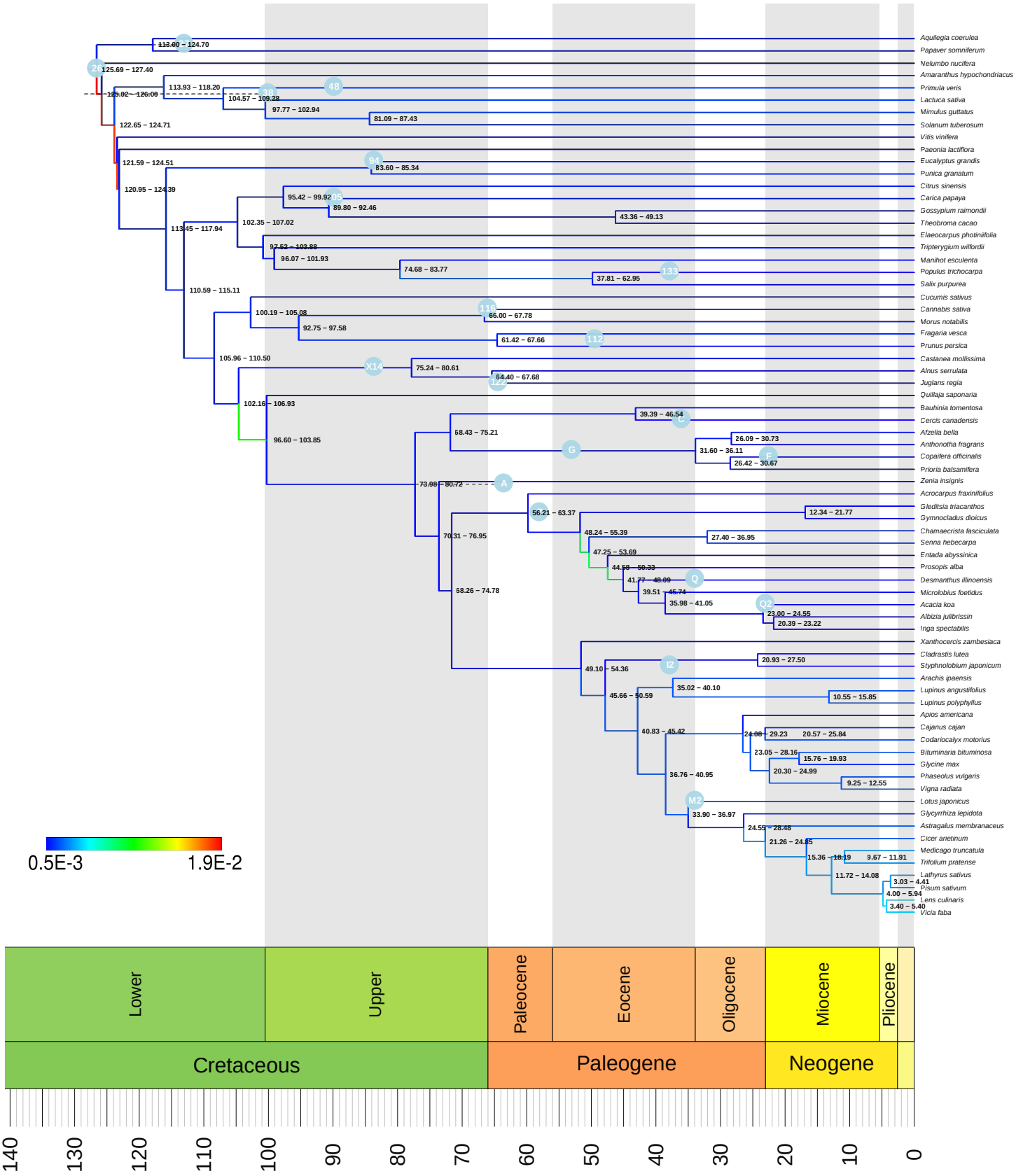

Figure S8. Chronogram estimated under the RLC model. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by coloured branches, as indicated by the colour legend, in substitutions per site per million years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

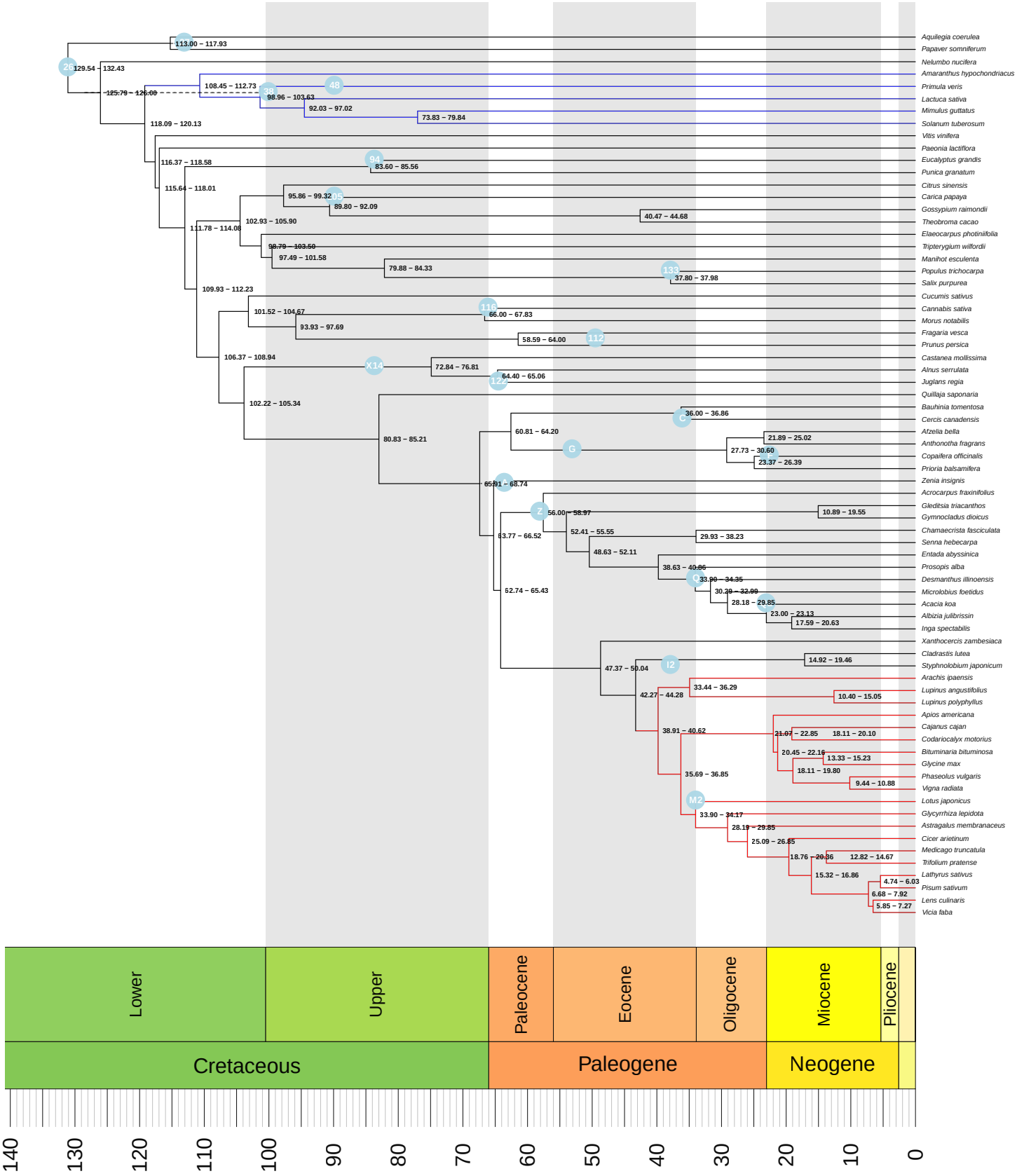

Figure S9. Chronogram estimated under the FLC3 model. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by coloured branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

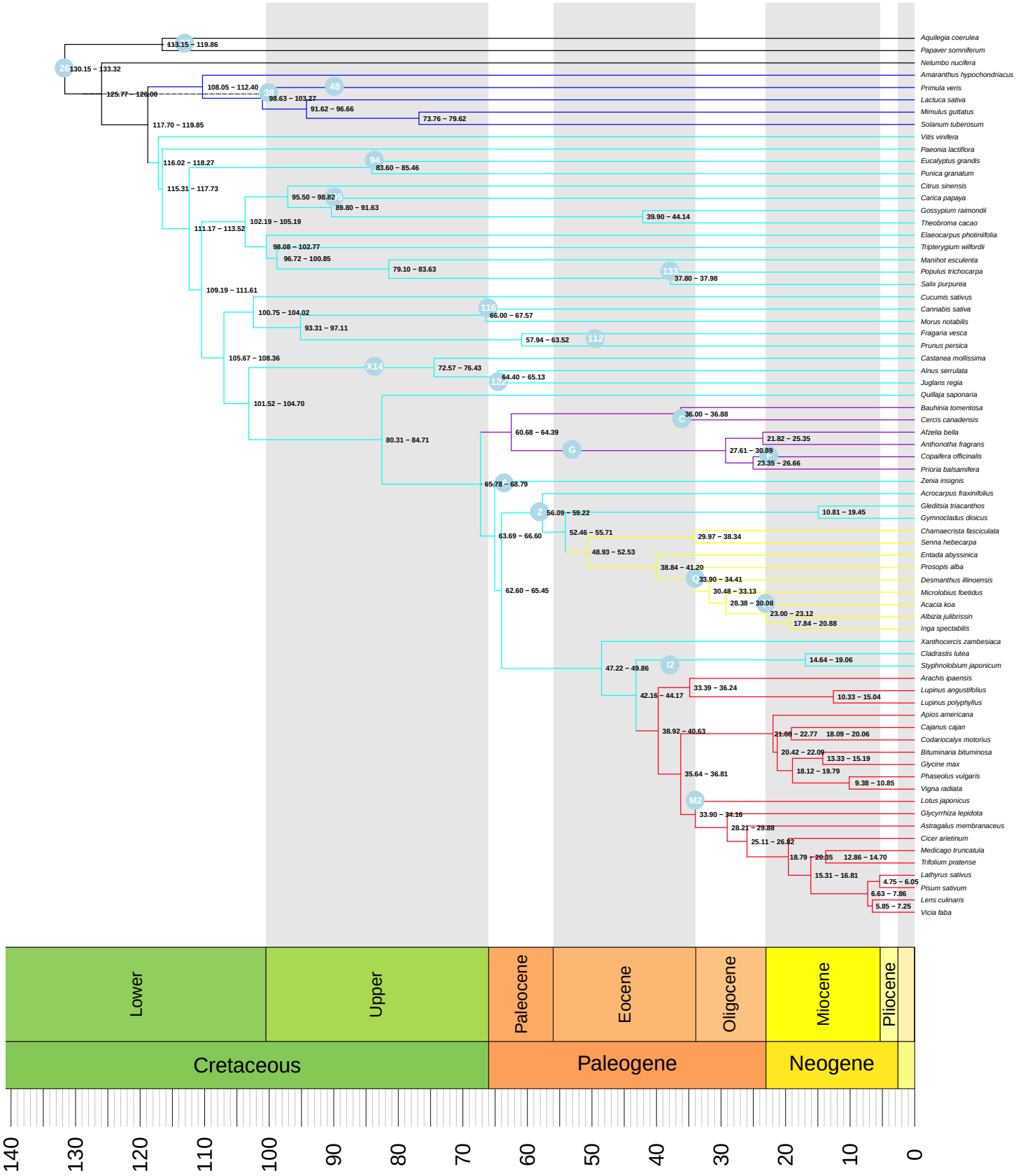

Figure S10. Chronogram estimated under the FLC6 model. Numbers behind nodes indicate 95% HPD intervals. Cock partitions are indicated by coloured branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

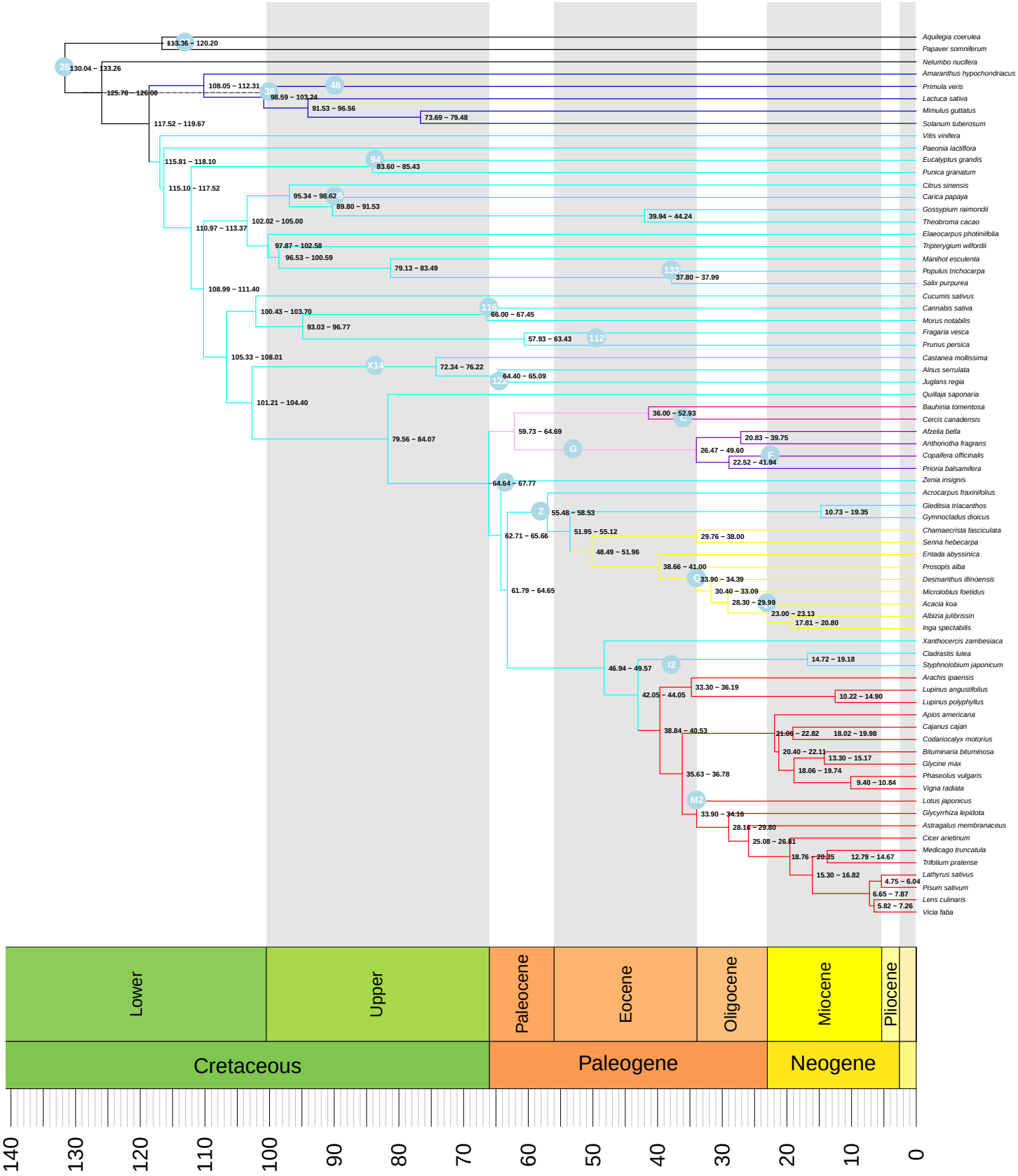

Figure S11. Chronogram estimated under the FLC8 model. Numbers behind nodes indicate 95% HPD intervals. Cock partitions are indicated by coloured branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

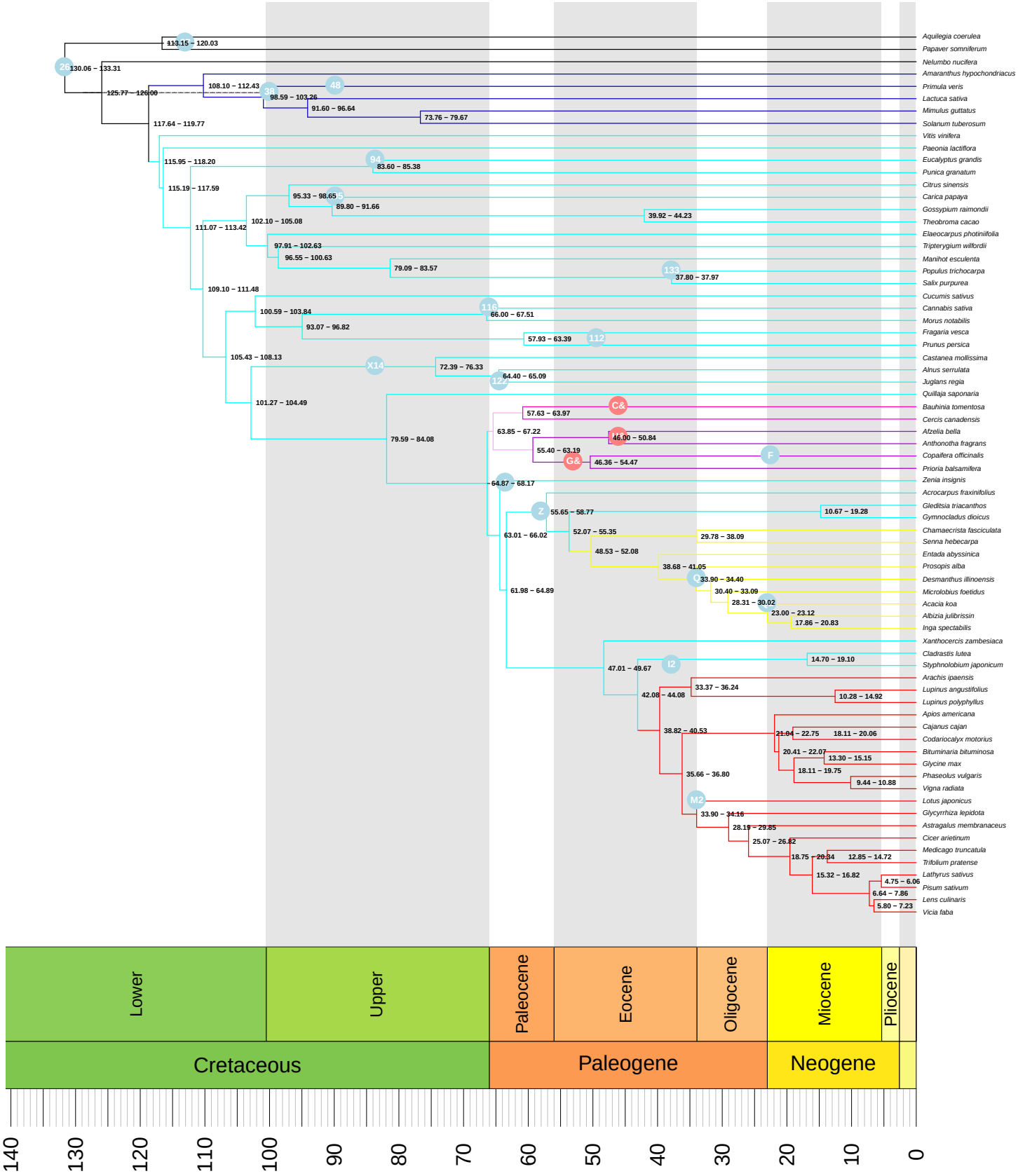

Figure S12. Chronogram estimated under the FLC8 model, with alternative prior 1. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by coloured branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles, with alternative calibrations as red circles.

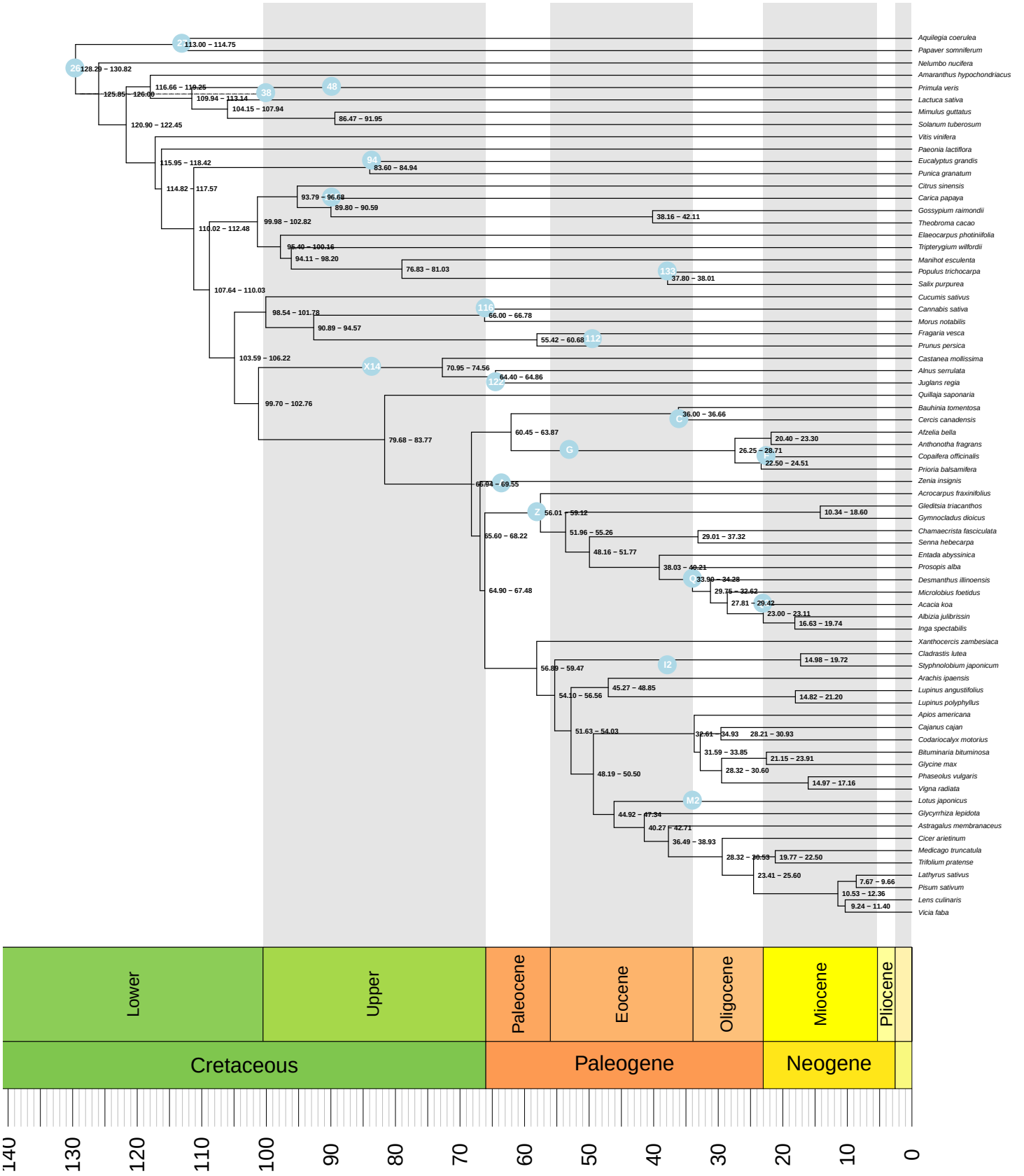

Figure S13. Chronogram estimated under the STRC model. Numbers behind nodes indicate 95% HPD intervals. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

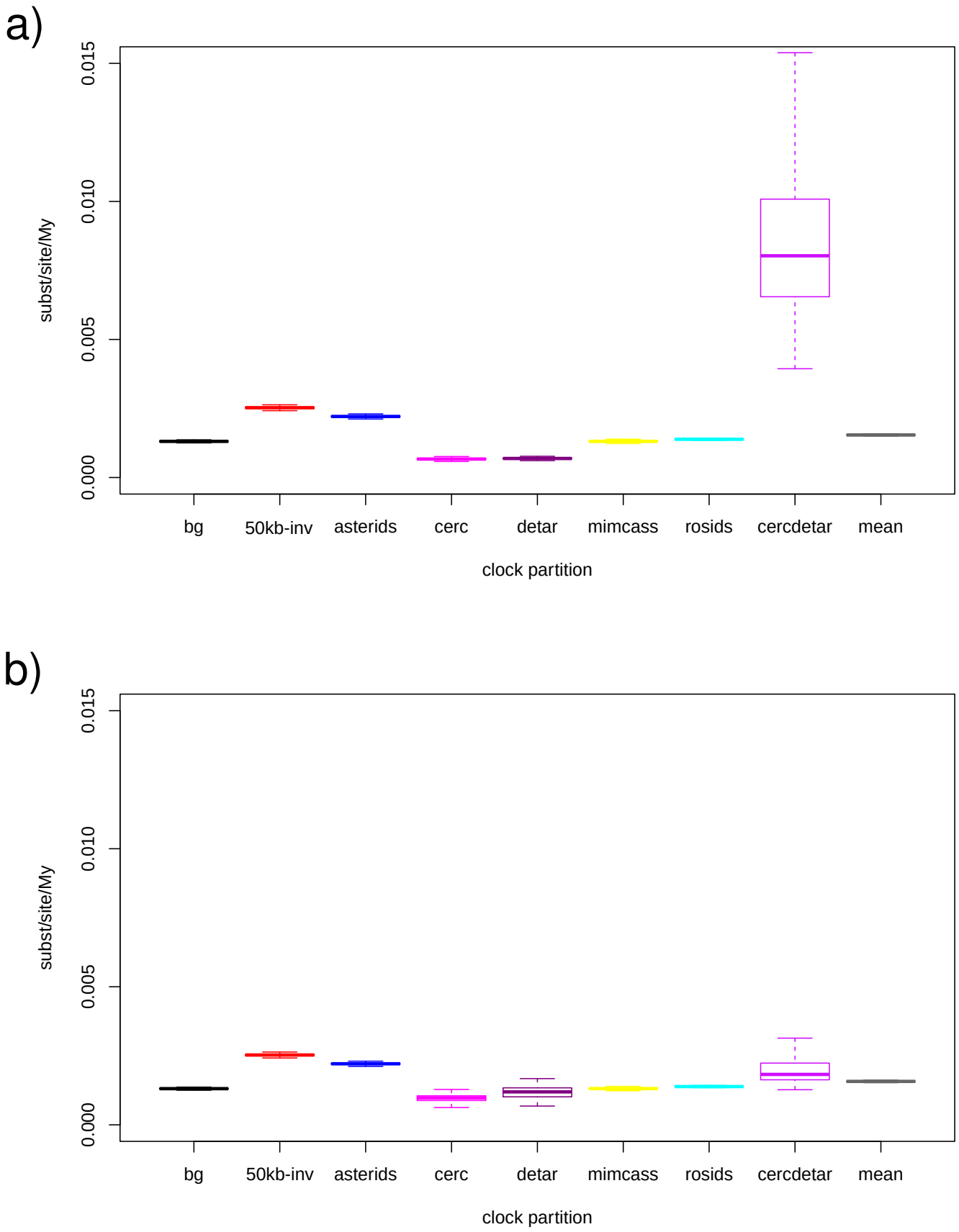

Figure S14. Substitution rates as estimated in FLC8 analyses for the different clock partitions. Boxplots for each partition for (A) alternative prior 1 and (B) the “normal” prior setting. Colours correspond to the partitions as shown in Figs 5, S14, S15 and S18.

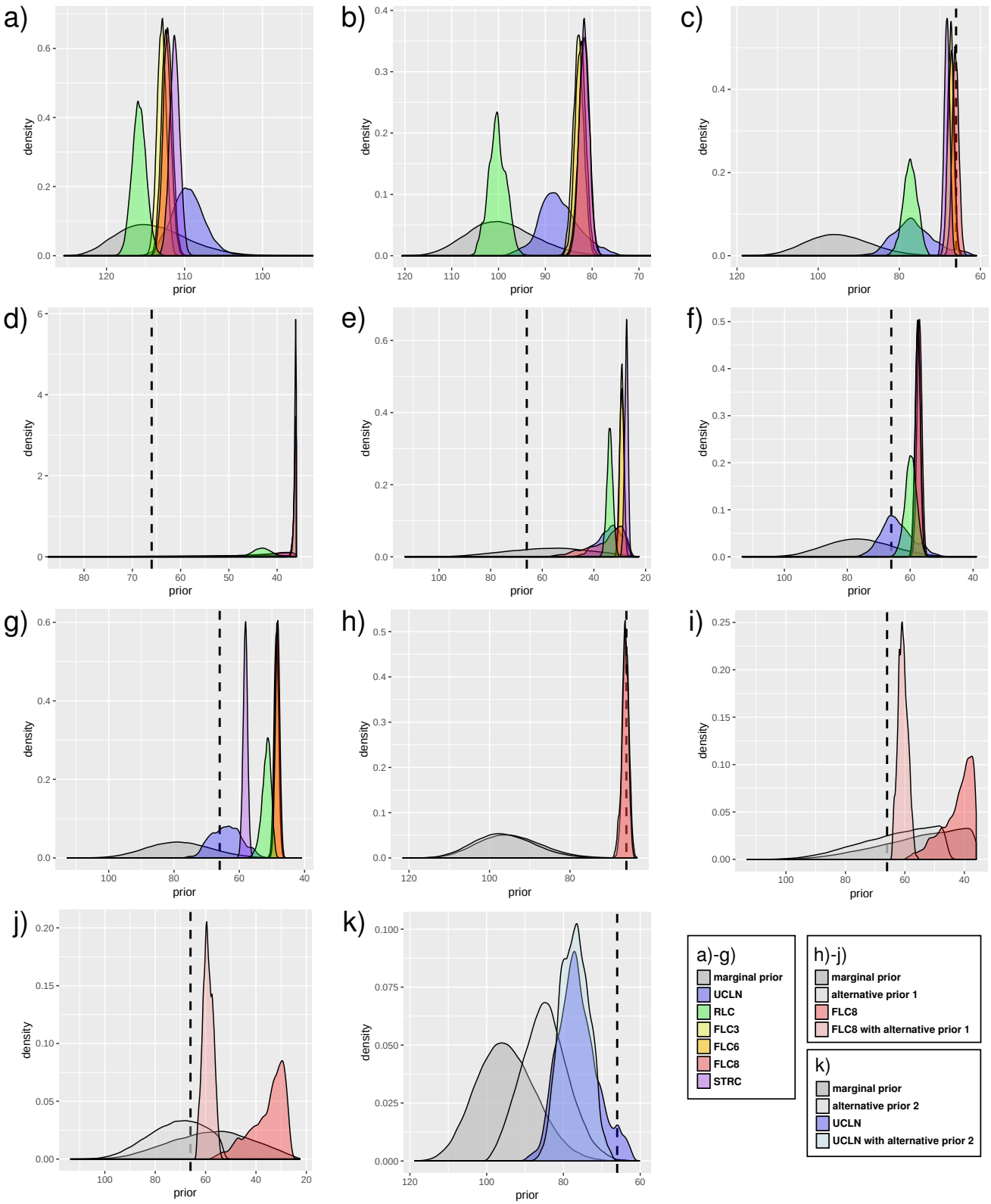

Figure S15. Prior and posterior densities of age estimates of selected nodes under different clock models (a - g) and alternative priors (h - k). Density plots are drawn for crown groups of (a) Eurosids, (b) Fabales, (c) Leguminosae, (d) Cercidoideae, (e) Detarioideae, (f) Caesalpinioideae, (g) Papilionoideae; and (h) Leguminosae, (i) Cercidoideae and (j) Detarioideae under standard and alternative prior 1; and (k) the legume crown node under standard and alternative prior 2. Colours used to indicate clock models or priors are as per the legend in the lower right corner. The vertical dashed line in (c - k) indicate the Cretaceous-Paleogene (K-Pg) boundary.

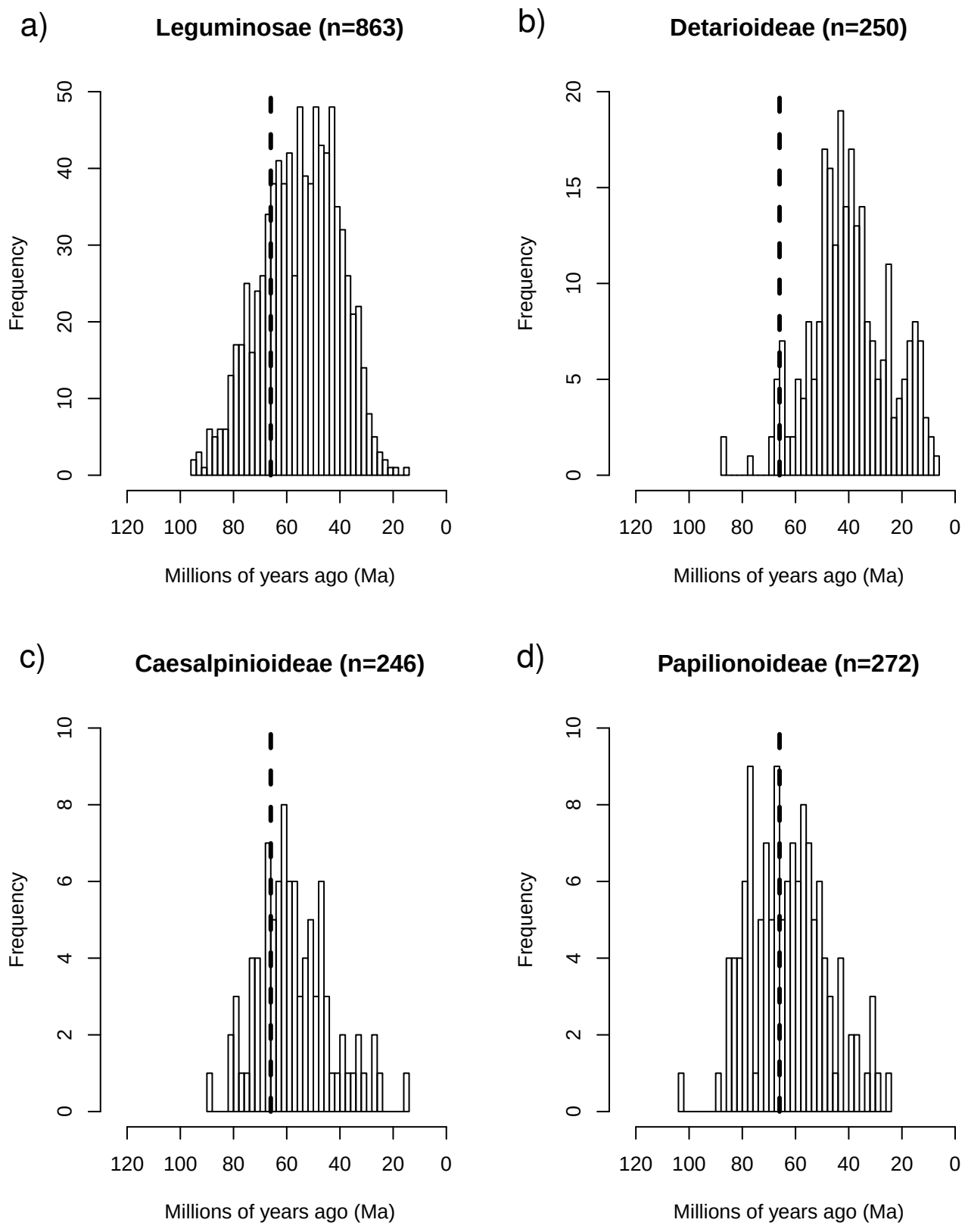

Figure S16. Age estimates of duplication nodes. Histograms of age estimates for (a) the duplications mapped to the legume crown node in the Notung analysis and for duplication nodes in gene trees with only (b) Detarioideae, (c) Caesalpinioideae and (d) Papilionoideae included.