The Origin and Early Evolution of the Legumes are a Complex Paleopolyploid Phylogenomic Tangle closely associated with the Cretaceous-Paleogene (K-Pg) Boundary

Running head:
Phylogenomic complexity and polyploidy in legumes

Authors:
Erik J.M. Koenen[1*], Dario I. Ojeda[2,3], Royce Steeves[4,5], Jérémy Migliore[2], Freek T. Bakker[6], Jan J. Wieringa[7], Catherine Kidner[8,9], Olivier Hardy[2], R. Toby Pennington[8,10], Patrick S. Herendeen[11], Anne Bruneau[4] 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] Institut de Recherche en Biologie Végétale and Département de Sciences Biologiques, Université de Montréal, 4101 Sherbrooke St E, Montreal, QC H1X 2B2, Canada
[5] Fisheries & Oceans Canada, Gulf Fisheries Center, 343 Université Ave, Moncton, NB E1C 5K4, Canada
[6] Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands
[7] Naturalis Biodiversity Center, Leiden, Darwinweg 2, 2333 CR, Leiden, The Netherlands
[8] Royal Botanic Gardens, 20a Inverleith Row, Edinburgh EH3 5LR, U.K.
[9] School of Biological Sciences, University of Edinburgh, King's Buildings, Mayfield Rd, Edinburgh, UK
[10] Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, EX4 4RJ, U.K.
[11] Chicago Botanic Garden, 1000 Lake Cook Rd, Glencoe, IL 60022, U.S.A.

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

Abstract – The consequences of the Cretaceous-Paleogene (K-Pg) boundary (KPB) mass extinction for the evolution of plant diversity are poorly understood, even although evolutionary turnover of plant lineages at the KPB is central to understanding the assembly of the Cenozoic biota. One aspect that has received considerable attention is the apparent concentration of whole genome duplication (WGD) events around the KPB, which may have played a role in survival and subsequent diversification of plant lineages. In order to gain new insights into the origins of Cenozoic biodiversity, we examine the origin and early evolution of the legume family, one of the most important angiosperm clades that rose to prominence after the KPB and for which multiple WGD events are found to have occurred early in its evolution. The legume family (Leguminosae or Fabaceae), with c. 20.000 species, is the third largest family of Angiospermae, and is globally widespread and second only to the grasses (Poaceae) in economic importance. Accordingly, it has been intensively studied in botanical, systematic and agronomic research, but a robust phylogenetic framework and timescale for legume evolution based on large-scale genomic sequence data is lacking, and key questions about the origin and early evolution of the family remain unresolved. We extend previous phylogenetic knowledge to gain insights into the early evolution of the family, analysing an alignment of 72 protein-coding chloroplast genes and a large set of nuclear genomic sequence data, sampling thousands of genes. We use a concatenation approach with heterogeneous models of sequence evolution to minimize inference artefacts, and evaluate support and conflict among individual nuclear gene

trees with internode certainty calculations, a multi-species coalescent method, and phylogenetic supernetwork reconstruction. Using a set of 20 fossil calibrations we estimate a revised timeline of legume evolution based on a selection of genes that are both informative and evolving in an approximately clock-like fashion. We find that the root of the family is particularly difficult to resolve, with strong conflict among gene trees suggesting incomplete lineage sorting and/or reticulation. Mapping of duplications in gene family trees suggest that a WGD event occurred along the stem of the family and is shared by all legumes, with additional nested WGDs subtending subfamilies Papilionoideae and Detarioideae. We propose that the difficulty of resolving the root of the family is caused by a combination of ancient polyploidy and an alternation of long and very short internodes, shaped respectively by extinction and rapid divergence. Our results show that the crown age of the legumes dates back to the Maastrichtian or Paleocene and suggests that it is most likely close to the KPB. We conclude that the origin and early evolution of the legumes followed a complex history, in which multiple nested polyploidy events coupled with rapid diversification are associated with the mass extinction event at the KPB, ultimately underpinning the evolutionary success of the Leguminosae in the Cenozoic.

Keywords: Cretaceous-Paleogene (K-Pg) boundary, Leguminosae, Fabaceae, Incomplete Lineage Sorting, Whole Genome Duplication events, paleopolyploidy, phylogenomics

The Cretaceous-Paleogene (K-Pg) boundary (KPB), 66 Million years ago (Ma), is defined by the mass extinction event that famously killed the non-avian dinosaurs and led to major turnover in the earth's biota. The Chicxulub meteorite impact is generally thought to have been the cause of the mass extinction, but Deccan trap flood basalt volcanism likely contributed or may have been the primary cause, in line with previous global mass extinctions that are all related to volcanism (Keller, 2014). The KPB event determined in significant part the composition of the Earth's modern biota, because many lineages that were successful in repopulating the planet and diversifying in the wake of the KPB have remained abundant and diverse throughout the Cenozoic until the present. Probably the best-known examples of successful post-KPB lineages are the mammals and birds, both inconspicuous elements of the Cretaceous fauna, while their core clades Placentalia and Neoaves became ubiquitous throughout Cenozoic fossil faunas. Plants were also severely affected by the KPB, with a clear shift in floristic composition and a drop in macrofossil species richness of up to 78% reported across boundary-spanning fossil sites in North-America (Wilf & Johnson, 2004; McElwain & Punyasena, 2007; Vajda & Bercovici, 2014). In addition, a global fungal spike followed by a global fern spike in the palynological record (Vajda et al., 2001; Barreda et al., 2012) are consistent with sudden ecosystem collapse and a recovery period characterized by low diversity vegetation dominated by ferns. Although the KPB is not considered a major extinction event for plants as no plant family appears to have been

lost at the KPB (McElwain & Punyasena, 2007; Cascales-Miñana & Cleal, 2014), a sudden increase in net diversification rate in the Paleocene has been inferred from a large paleobotanical data set (Silvestro et al., 2015), suggesting increased origination following the KPB. Arguably, analyses of global plant fossil data suffer from the poor rock record in the Maastrichtian just prior to the KPB (Nicholls & Johnson, 2008) and are limited to inferences at family or genus level due to the nature of palaeontological data, thereby potentially underestimating global extinction rates at the species level.

For individual plant lineages, macro-evolutionary dynamics relative to the KPB extinction event have received less attention than prominent vertebrate clades. However, given that plants are the main primary producers and structural components of terrestrial ecosystems, the shaping and diversification of Cenozoic biota cannot be fully understood without understanding the consequences of the KPB for evolutionary turnover of plant diversity. The legume family (Leguminosae or Fabaceae), perhaps more than any other plant clade, appears to parallel Placentalia and Neoaves. No fossils are known that pre-date the KPB and are clearly identifiable to the legume family (Herendeen & Dilcher, 1992), but the family was already abundant and diverse in one of the earliest examples of modern type rainforests in the Paleocene (Wing et al., 2009; Herrera et al., submitted). The oldest known fossils that are already referable to (stem groups of) subfamilies are from close to the Paleocene-Eocene Thermal Maximum (PETM) (morphotype # CJ76 of c. 58 Ma (Wing et al., 2009) can be referred to Caesalpinioideae and *Barnebyanthus buchananensis* of c. 56 Ma to Papilionoideae

(Crepet & Herendeen, 1992)) and legumes are a ubiquitous element of many Eocene, Oligocene and Neogene floras (Herendeen & Dilcher, 1992). Today, it is the third most species-rich angiosperm family, and arguably the most spectacular evolutionary and ecological radiation of any angiosperm family (McKey, 1994). Leguminosae is subdivided into six subfamilies (Fig. 1A-F; LPWG, 2017), which share the defining feature of the family, the fruit (referred to as the “legume” or “pod”) (Fig. 1G). It is the second most cultivated plant family after the Poaceae, and its species serve many purposes for humans, including timber, ornamentals, fodder crops and perhaps most notably, a large set of globally important pulse crops (Fig. 1I). A key trait of many legumes is the ability to fix atmospheric nitrogen via symbiosis with “rhizobia”-bacteria in root nodules (Fig. 1H), which leads to enriched soil, high nitrogen content in the leaves, and protein-rich seeds. The fact that legume species are diverse, omnipresent and often abundant in nearly all vegetation types across the planet, ranging in habit from large rainforest trees to small temperate herbs (Fig. 1J-L), means that legumes are an excellent study system to understand plant evolution in the Cenozoic.

The rapid appearance of legume diversity shortly after the first occurrence in the fossil record has been likened to the 'abominable mystery' of the sudden appearance of the angiosperms (Sanderson, 2015). The legume phylogeny also suggests rapid early evolution of legume diversity with very short internodes subtending the six major lineages following the origin of the family (Lavin et al., 2005; LPWG, 2017) as well as at the base of subfamilies Detarioideae, Caesalpinioideae (Bruneau et al., 2008; LPWG,

2017) and Papilionoideae (Cardoso et al., 2012, 2013; LPWG, 2017). Just as for Placentalia (Teeling & Hedges, 2013) and Neoaves (Suh et al., 2015; Suh, 2016), this apparently rapid early diversification of legumes poses problems for phylogeny inference. In particular, the first few dichotomies in the phylogeny of the family have been difficult to resolve, as have deep divergences in Detarioideae, Caesalpinioideae and Papilionoideae (LPWG, 2013 & 2017). In this study, we attempt to resolve the deep-branching relationships in the legume family by using much larger molecular sequence data sets than those previously used in legume phylogenetics. Moreover, previous legume phylogenies have been mainly inferred from chloroplast markers (Wojciechowski et al., 2004; Lavin et al., 2005; Bruneau et al., 2008; Simon et al., 2009; Cardoso et al., 2012, 2013; LPWG, 2017). In addition to analysing nearly all protein-coding genes from the chloroplast genome, here we also analyse thousands of gene alignments from the nuclear genome.

Unlike birds and mammals, whole genome duplication (WGD) events are common in angiosperms, and such events have been suggested to be significantly concentrated around the KPB (Fawcett et al., 2009; Vanneste et al., 2014; Lohaus & Van de Peer, 2016). This is explained by the idea that polyploid lineages could have had enhanced survival and establishment across the KPB (Lohaus & Van de Peer, 2016) as well as greater potential to diversify rapidly thereafter relative to diploids (Levin & Soltis, 2018). WGDs have also been found to have occurred multiple times during the early evolution of the legumes (Cannon et al., 2015) and could have contributed to the initial

rapid diversification of the family, as well to the difficulties of resolving relationships among the six subfamilies. There is considerable uncertainty about how many WGDs were involved in the early evolution of legumes and in the placements of possible WGDs on the legume phylogeny. From whole genome sequencing studies, it has been known for some time that several papilionoids share a WGD event (Cannon et al., 2006; Mudge et al., 2005), but recently it has been suggested that several other legume lineages have also undergone independent WGDs (Cannon et al., 2015). Indeed, Cannon et al. (2015) showed that the papilionoid WGD is shared by all members of that subfamily using phylogenetic methods, and used age estimates from K_s plots to infer additional independent WGDs early in the evolution of subfamilies Caesalpinioideae, Cercidoideae and Detarioideae. However, in the absence of data for several critical legume lineages, the phylogenetic positions of these additional putative WGDs remain uncertain. A more recent study (Wong et al., 2017) suggested instead that all legumes share the same WGD, based on rate-corrected K_s plots and a genetic linkage map of *Acacia* that suggested mimosoids (Caesalpinioideae) and Papilionoideae retained an orthologous duplicated chromosomal segment. From homolog gene family trees generated prior to separating orthologs from paralogs (Yang & Smith, 2014; Smith et al., 2015; Yang et al., 2015), we map the number of gene duplications over the legume phylogeny to evaluate how many early legume WGDs occurred and where they are located on the phylogeny.

While the legumes are not known with certainty from any Cretaceous fossil site, the family has a long stem lineage dating back to c. 80 – 100 Ma (Wang et al., 2009; Magallón et al., 2015). This long ghost lineage means that the timing of the initial radiation of the family, as well as of legume WGDs, and notably whether they pre- or post-date the KPB, are uncertain. In Placentalia and Neoaves, divergence time estimation has led to much debate, with some studies using molecular sequence data for divergence time estimation suggesting that both clades originated and diversified well before the KPB, implying that many lineages of both clades survived the end-Cretaceous event (Cooper & Penny, 1997; Jetz et al., 2012; Meredith et al., 2011). However, like the legumes, both groups first appear in the Paleocene fossil record. A phylogenetic study of mammals combining both molecular sequence data and morphological characters to enable inclusion of fossil taxa, found only a single placental ancestor crossing the KPB (O'Leary et al., 2013; but see Springer et al., 2013; dos Reis et al., 2014). Alternatively, it has been argued that diversification of Placentalia followed a "soft explosive" model, with a few lineages crossing the KPB followed by rapid ordinal level radiation during the Paleocene (Phillips, 2015; Phillips & Fruciano, 2018). Recent time-calibrated phylogenies for birds showed the age of Neoaves to also be close to the KPB (Jarvis et al., 2014; Claramunt & Cracraft, 2015; Prum et al., 2015), with initial rapid post-KPB divergence represented by a hard polytomy (Suh, 2016). For legumes, it is similarly unlikely that modern subfamilies of legumes have Cretaceous crown ages. These clades, in particular Papilionoideae, Caesalpinioideae and Detarioideae, appear

to have rapidly diversified following their origins, which would imply mass survival of very large numbers of legume lineages across the KPB. Diversification into the six main lineages of legumes appears to have occurred rapidly (Lavin et al., 2005), with long stem branches leading to each of the modern subfamilies. Therefore, two hypotheses seem plausible: (1) the legumes have a Cretaceous crown age and diversified into the six subfamilies prior to the KPB, while crown radiations of the subfamilies occurred in the wake of the mass extinction event, corresponding to a “soft explosive” model, or (2) a single legume ancestor crossed the KPB and rapidly diversified into six main lineages in the wake of the mass extinction event, corresponding to a “hard-explosive” model, with the subsequent subfamily radiations related to the Paleocene-Eocene Thermal Maximum (PETM) and/or Eocene climatic optimum. Currently available molecular crown age estimates for the family range from c. 59 to 64 Ma (Lavin et al., 2005; Bruneau et al., 2008; Simon et al., 2009). These studies, however, lacked extensive sampling of outgroup taxa and relied instead on fixing the stem age of the legumes, thereby compromising the ability to estimate the crown age of the family. Furthermore, these earlier studies relied exclusively on chloroplast sequences, for which evolutionary rates are known to vary strongly across legumes (Lavin et al., 2005), such that nuclear gene data are likely to be better suited for estimating divergence times (Christin et al., 2014).

In this study, using large genomic-scale data sets, we aim to resolve the deep divergences in the legume family, find the phylogenetic locations of WGDs and estimate the timing of these. We analyse these new datasets with Maximum Likelihood (ML)

analysis, Bayesian inference, a multi-species coalescent summary method and filtered supernetwork reconstruction to resolve the deep-branching relationships in the family. In particular, we focus on the relationships among the six major lineages recently recognized as subfamilies (LPWG, 2017). Sister-group relationships between subfamilies Papilionoideae and Caesalpinioideae (sensu LPWG, 2017), and of the clade combining the two with the newly recognized Dialioideae, were previously known (Lavin et al., 2005; Bruneau et al., 2008; LPWG, 2017). However, the relationships between the clade comprising those three subfamilies and the other three subfamilies Cercidoideae, Detarioideae and Duparquetioideae remained difficult to resolve (cf. Bruneau et al., 2008; LPWG, 2017). Having inferred the most likely species-tree topology, we evaluate numbers of supporting and conflicting bipartitions for critical nodes across gene trees. To infer likely locations of WGDs, we count the number of gene duplications present in nuclear homolog clusters and map these across the species tree. Finally, we perform molecular clock dating on a selection of informative and clock-like nuclear genes with 20 fossil calibration points, to infer whether the origin of the legumes and WGDs in the early evolution of the family are related to the K-Pg mass extinction event.

MATERIAL & METHODS

DNA/RNA Extraction and Sequencing

244

245 For the newly generated chloroplast gene data, DNA was extracted from fresh
246 leaves, leaf tissue preserved in silica-gel or herbarium specimens, using the Qiagen
247 DNeasy Plant Mini Kit. Sequencing libraries were prepared using the NEBNext Ultra
248 DNA Library Prep Kit for Illumina. They were then sequenced on the Illumina HiSeq
249 2000 sequencing platform, at low coverage ('genome-skimming') or as part of hybrid
250 capture experiments for a separate study on mimosoid legumes (Koenen et al.,
251 unpublished data). RNA was extracted from fresh leaves using the Qiagen RNeasy
252 Plant Mini Kit. RNA sequencing libraries were prepared using the Illumina TruSeq RNA
253 Library Prep Kit and sequenced on the Illumina HiSeq 2000 sequencing platform. All lab
254 procedures were performed according to the specifications and protocols provided by
255 the manufacturers of the kits.

256

257 *Sequence Assembly*

258

259 Raw reads for the chloroplast DNA data were cleaned and filtered using the
260 following steps: (1) Illumina adapter sequence artifacts were trimmed using
261 Trimmomatic v. 0.32 (Bolger et al., 2014), (2) overlapping read pairs were merged with
262 PEAR v. 0.9.8 (Zhang et al., 2014) and (3) low quality reads were discarded and low
263 quality bases at the end of the reads were trimmed with Trimmomatic v. 0.32 (using
264 settings MAXINFO:40:0.1 LEADING:20 TRAILING:20). The quality-filtered reads were

then assembled into contigs using the SPAdes assembler v. 3.6.2 (Bankevich et al., 2012). For RNA data, raw reads were quality-filtered using the FASTX-toolkit v. 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to remove low quality reads (less than 80% of bases with a quality score of 20 or higher), TagDust v. 1.12 (Lassmann et al., 2009) to remove adapter sequences and PRINSEQ-lite v. 0.20.4 (Schmieder & Edwards, 2011) to trim low quality bases off the ends of reads. Transcriptome assembly was then performed on the quality-filtered reads using Trinity (Grabherr et al., 2011; Release 2012-06-08), with default settings.

Chloroplast Proteome Alignment

DNA sequences of protein-coding chloroplast genes were newly generated as described above, or extracted from several different data sources, as specified for each accession in Table S1. Sequence data were extracted directly from annotated plastomes in Genbank, by blast searches from *de novo* assembled contigs and from transcriptomes using custom Python scripts. Sequences for some outgroup taxa (data from Moore et al., 2010) were downloaded separately per gene from Genbank. For each gene, a codon alignment was inferred using MACSE v. 1.01b (Ranwez et al., 2011). Phylogenetic trees were then inferred for each gene separately to screen for erroneously aligned sequences with RAxML v. 8.2 (Stamatakis, 2014). For some species, individual gene sequences that led to anomalously long terminal branches

were then removed. The genes *accD* and *clpP* were removed completely. The gene alignments were concatenated and the full alignment was visually checked and obvious misalignments were resolved. Furthermore, sequence errors (single A/T indels) that caused frameshift mutations were corrected and the accuracy of the alignment at codon level was assessed and corrected if necessary. For a few genes where the ends of coding sequences had varying lengths, all sites between the first and last stop codon in the alignment were excluded, since they were poorly aligned. Finally, using BMGE v. 1.12 (Block Mapping and Gathering with Entropy; Criscuolo & Gribaldo, 2010) the codon alignment was translated to amino acid sequences.

Nuclear Gene Data and Matrix Assembly

Whole genome and transcriptome data were downloaded from various sources and augmented with newly generated transcriptome sequence data for six Caesalpinioideae and Detarioideae taxa (see Table S2). Peptide sequences were downloaded from annotated genomes, or were extracted from transcriptome assemblies using TransDecoder (http://transdecoder.github.io/). To assemble the nuclear peptide sequence data into aligned gene matrices, we used the pipeline of Yang & Smith (2014). We performed mcl clustering as described in Yang & Smith (2014), with a hit fraction cut-off of 0.75, inflation value of 2 and a minimum log-transformed e-value of 30. These settings lead to clusters with good overlap between sequences and good

alignability (omitting genes that are too variable), although we may have lost a few short gene clusters. Next, the homolog gene clusters were subjected to two rounds of alignment with MAFFT v. 7.187 (Katoh & Standley, 2013), gene tree inference inference with RaxML v. 8.2 (Stamatakis, 2014), and pruning and masking of tips and cutting deep paralogs as described in Yang & Smith (2014). In the first round we used 0.3 and 1.0 as relative and absolute cut-offs for trimming tips, respectively, and 0.5 as the minimum cut-off for cutting deep paralogs, and keeping all clusters with a minimum of 25 taxa for the second round. In the second round we used more stringent cut-off values (0.2 and 0.5 for trimming tips and 0.4 for cutting deep paralogs). See Yang & Smith (2014) for more information on these parameter settings. One-to-one orthologs and rooted ingroup (RT) homologs were then extracted from the homolog cluster trees, with a minimum aligned length of 100 amino acids for each homolog. One-to-one orthologs are those homolog gene clusters in which each taxon is represented only by a single gene copy. RT homologs are extracted by orienting homolog cluster trees by rooting them on the outgroup (in our case *Aquilegia coerulea* and *Papaver somniferum*), and then detecting gene duplications and pruning the paralog copies with fewer taxa present until each taxon is represented by a single copy. The outgroup is pruned as well, and clusters without outgroup in which each taxon is only present once are also included, meaning that all 1-to-1 orthologs are also in the RT homolog set. See Yang & Smith (2014) for a more detailed description of how these homologs are extracted. Sequences with more than 50% gaps and all sites with more than 5% missing data were removed from the

328 homolog alignments using BMGE. For the 1-to-1 orthologs that were used for species
329 tree inference, alignments with fewer than 50 taxa were discarded, for the larger set of
330 RT homologs that were used for counting of supporting and conflicting bipartitions,
331 alignments with fewer than 25 taxa were discarded.

332

333 *Phylogenetic Inferences*

334

335 Maximum likelihood (ML) and Bayesian analyses were run in RaxML v. 8.2
336 (Stamatakis, 2014) and Phylobayes-MPI 1.7 (Lartillot et al., 2013), respectively. For the
337 ML analysis using nucleotide sequences of the chloroplast alignment, we used
338 PartitionFinder 2 (Lanfear et al., 2017) to estimate partitions, with a minimum length per
339 partition set to 500 nucleotides, and allowing different codon positions per gene to be in
340 different partitions. The resulting 16 partitions were run with the GTR + GAMMA model,
341 and 1000 rapid bootstrap replicates were carried out. For the amino acid sequences,
342 the ML analyses of both the chloroplast alignment and the concatenated alignment of
343 nuclear 1-to-1 orthologs were analyzed with the LG4X model, without partitioning, as
344 the model accounts for substitution rate heterogeneity across the alignment by
345 estimating 4 different LG substitution matrices (Le et al., 2012). For the chloroplast
346 alignment, 1000 rapid bootstrap replicates were additionally carried out. Gene trees of
347 1-to-1 orthologs and RT homologs were estimated with RAxML using the WAG + G
348 model, with 100 rapid bootstrap replicates. We then calculated 80% majority-rule

consensus trees for each ortholog or homolog and used these to calculate Internode Certainty All (ICA) values using RAxML, to include only nodes that received 80% or greater bootstrap support in the individual gene trees. Bayesian analyses were performed with the CATGTR model, with invariant sites deleted and default settings for other options in Phylobayes. Analyses were run until the chain reached convergence (usually after 10-20k cycles), with at least two independent chains run for each data set. To perform Bayesian analyses on the complete nuclear gene data set in a computationally tractable manner, we ran 25 gene jack-knifing replicates without replacement, dividing the total number of genes over 5 subsets with 5 replicates. These subsampled replicates were run in Phylobayes-MPI, with a starting tree derived from the analysis sampling the 100 genes with the longest gene tree length, using the CATGTR model with constant sites deleted, for 1000 cycles each. We found that all 25 chains had converged after a few hundred cycles, and discarded the first 500 cycles of each as burn-in. A majority-rule consensus tree was constructed using sumtrees.py (from the Dendropy library (Sukumaran et al., 2010)) from 12500 total posterior trees, representing the MCMC cycles 501-1000 of each replicate. For both the ML and Bayesian analyses, concatenated alignments were not partitioned. Instead we rely on the LG4X and CATGTR models to take rate heterogeneity into account, since these models describe heterogeneity across alignments more accurately than partitioning by gene and/or codon since the substitution process also varies across gene sequences and codon positions. For the multi-species coalescent analysis, we used ASTRAL

370 (Mirabab et al., 2014) on the 1,103 gene trees estimated with RAxML, using local
371 posterior probability and quartet support to evaluate the inferred topology (Sayyari &
372 Mirabab, 2016).

373

374 *D_n/D_s Ratio Analyses for cpDNA*

375

376 The codon alignments for each chloroplast gene were analyzed individually using
377 the branch model test in PAML v. 4.9 (Yang, 2007), to test if higher substitution rates in
378 the 50-kb inversion and vicioid clades of Papilionoideae were related to differing
379 selective pressures. These clades were partitioned separately to allow for the estimation
380 of independent rates of synonymous and non-synonymous substitution rates for each of
381 these clades relative to the rest of the tree. Since the vicioid clade is nested in the 50-kb
382 inversion clade, the rates reported for the latter clade are estimated without the vicioid
383 clade taxa. While this test does not evaluate selective pressures for specific sites, it
384 does give an indication whether genes evolve neutrally or are under purifying or positive
385 selection.

386

387 *Counting Supporting Bipartitions for Key Nodes across Gene Trees*

388

389 Using a custom python script, numbers of matching and alternative bipartitions
390 across gene trees were counted for particular nodes labeled A-H in Figure 3A in the

legume phylogeny. For this purpose, we assessed monophyly of each of the subfamilies and combinations (clades) of subfamilies, against the outgroup, across all gene trees. For each gene tree, we first assessed whether all 6 groups (5 subfamilies plus the outgroup) are present and gene trees with missing groups were not taken into account. Next, we evaluated whether the gene tree includes a matching bipartition for the family, each subfamily and for all possible combinations of subfamilies. A matching bipartition means that all taxa of a subfamily or combination of subfamilies are separated from all other taxa in the gene tree, thus constituting support for that clade to be monophyletic. For combinations of subfamilies, the subfamilies themselves do not necessarily need to be monophyletic, but all taxa within those subfamilies should be separated from all other taxa to constitute a matching bipartition, and thus to be a supported clade in the gene tree. For well supported clades, we expect matching bipartitions for a majority of gene trees. For poorly supported clades, we expect most gene trees to be uninformative due to low phylogenetic signal, hence a low number of matching bipartitions, and possibly relatively high numbers of conflicting bipartitions. All counts were done for ML gene trees of RT homologs, and with 50 and 80% bootstrap cutoffs. The recently published DiscoVista software package (Sayyari et al., 2018) allows similar evaluations of conflicting and supporting bipartitions to those described here to be made and visualized.

Phylogenetic Supernetwork Analysis

412

413 We used SplitsTree4 to draw a filtered supernetwork (Whitfield et al., 2008) of the
414 1,103 1-to-1 orthologs, using the 80% majority-rule consensus trees to only include
415 well-supported bipartitions to infer the network. All gene trees were pruned for simplified
416 visualization, focusing on the deep divergences within the legume family. All taxa
417 outside the nitrogen-fixing clade comprising Cucurbitales, Fabales, Fagales, Rosales,
418 as well as a subset of taxa in the relatively densely sampled Papilionoideae and
419 Caesalpinioideae were pruned, preferentially keeping taxa that were sampled in as
420 many gene trees as possible. The mintrees parameter was set to 552 (at least 50% of
421 the number of orthologs) and the maximum distortion parameter was set to 0.

422

423 *Gene Duplication Mapping*

424

425 We used the homolog clusters generated from the Yang & Smith (2014) pipeline
426 prior to extracting 1-to-1 and RT orthologs to map duplications onto the species tree.
427 First, all sites with more than 5% missing data were removed with BMGE, to reduce the
428 amount of missing data. Also all sequences with more than 75% gaps were removed, to
429 avoid having fragmented paralog sequences present, which could inflate the number of
430 gene duplications. These data removal steps also led to the elimination of some clusters
431 with large amounts of missing data. Tree estimation was then repeated on these
432 clusters, with RAxML using the WAG + G model and 100 rapid bootstrap replicates.

433 Next, rooted ingroup clades were extracted from the resulting homolog trees with the
434 extract_clades.py script that is included with the Yang & Smith (2014) pipeline. To
435 extract the clades, we only considered *Aquilegia* and *Papaver* as outgroup taxa,
436 because the outgroup is not included in the extracted clades, and this way we could
437 maximize the number of taxa per extracted clade. However, because of uncertain
438 relationships along the backbone of Pentapetalae, we observed that the clusters were
439 often not correctly rooted. This does not have much effect for the number of duplications
440 that are observed near the tips, but it does lead to erroneous mapping near the base of
441 the tree. Therefore, we rooted the extracted clades with the Phyx package (Brown et al.,
442 2017), using a list of the non-legume taxa ordered by their phylogenetic relationships,
443 rooting the trees on the taxon that is most distantly related to legumes. Clusters that
444 included only legume species, without any outgroup taxa present, were excluded. From
445 the resulting multi-labeled trees (i.e. each taxon can be present multiple times,
446 representing different paralogs), duplications were mapped onto the species tree, with
447 and without a 50% bootstrap cut-off, using phyparts (Smith et al., 2015).
448
449 *Divergence time analyses*
450
451 Fossils used to calibrate molecular clock analyses are listed in Table 1 and are
452 discussed in Methods S1.

453 Using SortaDate (Smith et al., 2018b), we analyzed all gene trees to estimate the
454 total tree length (a proxy for sequence variation or informativeness), root-to-tip variance
455 (a proxy for clock-likeness) and compatibility of bipartitions with the ML tree that was
456 inferred using the full data set (the RAxML tree inferred with the LG4X model, shown in
457 Figure 3A). We then selected the best genes for dating based on cutoff values that were
458 arbitrarily chosen from the estimated values across gene trees: (1) total tree length
459 greater than 5, (2) root-to-tip variance less than 0.005 and (3) at least 10% of the
460 bipartitions in common with the ML tree. This yielded 36 genes, which were
461 concatenated to have a total aligned length of 14462 amino acid sites. We also used the
462 'pxlstr' program of the Phyx package (Brown et al., 2017) to calculate taxon-specific
463 root-to-tip lengths from the ML tree, after pruning the Ranunculales, on which the tree
464 was rooted. The values obtained were then used to define local clocks as described
465 below. *Arabidopsis thaliana, Linum usitatissimum* and *Polygala lutea* were removed
466 because of much higher root-to-tip lengths relative to their closest relatives. *Panax*
467 *ginseng* was also removed because of a low root-to-tip length relative to the other
468 sampled asterids, leaving a total of 72 taxa.

469 We used BEAST v.1.8.4 (Drummond et al., 2012) with various clock models to
470 estimate divergence time estimates across the phylogeny based on the alignment of the
471 selected 36 genes and the fossil calibrations described above. All analyses were run
472 with the LG + G model of amino acid substitution and the birth-death tree prior, and
473 using the ML tree to fix the topology. Fossil calibration priors were set as uniform priors

between the minimum age as specified in Table 1 and a maximum age of 126 Ma (oldest fossil evidence of eudicots) as listed in Table S4, with the exception of the root node, for which we used a normal prior at 126 Ma with a standard deviation of 1.0 and truncated to minimum and maximum ages of 113 (the Aptian-Albian boundary) and 136 Ma (the oldest crown angiosperm fossil, see Magallón et al. (2015)). With these settings, we ran analyses under the uncorrelated lognormal (UCLN), strict (STRC), random (RLC) and 3 different fixed local (FLC) clock models. To specify the different FLC models, we looked at root-to-tip length variation across subclades to specify biologically meaningful *a priori* clock partitions (Fig. S19). The 50kb-inversion clade of papilionoid legumes and the asterids (without *Panax ginseng*) have uniformly longer root-to-tip lengths than the other taxa across the tree and were therefore assigned their own local clock, with a different clock for the remaining taxa in the tree (this model referred to as FLC3, partitioning of taxa is illustrated in Supplementary Figure S19A). 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 for the combined clade of Cercidoideae and Detarioideae as well. This more complex model is referred to as FLC6 (Fig. S19B). The most complex model (FLC8; Fig. S19C) 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 colored branches in Figures 6 and S16-17 and those of the other FLC models in Figures S14-15). The Ranunculales that were pruned for the root-to-tip length calculations were included in the background clock for each FLC model.

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 (see Methods S1). This was done by running two analyses under the FLC8 model, one with the same priors as the other analyses, and one where calibrations C and G were changed and another calibration (H^g) was added to use similar placements of these calibrations as in Bruneau et al. (2008) and Simon et al. (2009) (Table 1 & Methods S1). We refer to this calibration scheme as “alternative prior 1” (Table S4). 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 S4 (“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.

Analyses sampling from the prior (without data) were run for 100 million generations, the strict clock and FLC3 and FLC6 analyses were run for 25 million generations and all other clock analyses were run for 50 million generations, and convergence was confirmed with Tracer v1.7.1 (Rambaut et al., 2018). For the non-prior analyses, the first 10% of the total number of generations was discarded as burn-in before summarizing median branch lengths and substitution rates with TreeAnnotator from the BEAST package.

RESULTS

The chloroplast alignment includes 72 protein-coding genes, for 157 taxa (including 111 legume species; Table S1), with a total aligned length of 75,282 bp or 25,094 amino acid residues. From transcriptomes and fully sequenced genomes, we gathered 9,282 homologous nuclear encoded gene clusters for 76 taxa including 42 legume species (Table S2). From these clusters, we extracted protein alignments of 1,103 1-to-1 orthologs for species tree inference with a total aligned length of 325,134 amino acids when concatenated, and 7,621 Rooted Ingroup (RT) homologs for additional gene tree inference. We also extracted 8,038 rooted clades from the homolog clusters to map the locations of gene duplications. The alignments, gene trees and species trees are available in TreeBASE (accession number XXXX) and on Dryad (doi: XXXX).

537

538 *Inferring the Species Tree*

539

540 Our analyses reveal that both the chloroplast and nuclear data sets resolve all
541 subfamilies as monophyletic with full support and most relationships among the
542 subfamilies are also robustly resolved (Figs 2, 3A-C & S1-7), with the notable exception
543 of the root node. The clade consisting of Papilionoideae, Caesalpinioideae and
544 Dialioideae is recovered in all analyses, with *Duparquetia* as the sister-group to this
545 clade as inferred from chloroplast data. *Duparquetia* is not sampled for nuclear data,
546 therefore transcriptome or genome sequencing is necessary for this taxon to confirm the
547 relationship found here. The root node of the legume family is more difficult to resolve,
548 and the chloroplast and nuclear data sets lead to conflicting topologies. The chloroplast
549 alignment supports Cercidoideae as sister to the rest of the family when analysing
550 protein sequences with ML under the LG4X model (58% bootstrap support; Fig. S1) and
551 Bayesian inference under the CATGTR model (0.98 posterior probability (pp); Figs 2 &
552 S2). When analysing chloroplast nucleotide sequences, we recovered the same
553 relationship in a partitioned ML analysis under the GTR model (recovered in 66% of the
554 bootstrap replicates; Fig. S3), but a Bayesian analysis under the CATGTR model does
555 not resolve the root, the majority-rule consensus tree showing a polytomy of
556 Cercidoideae, Detarioideae and a robustly supported clade formed by the other four
557 subfamilies (Fig. S4). To resolve deep divergences, amino acid sequences are more

suitable because they are less saturated with substitutions (silent substitutions are absent), and less prone to long branch attraction (LBA). Additionally, the LG4X and CAT models better account for heterogeneous substitution rates across sites in the alignment (Lartillot & Philippe, 2004; Le et al., 2012). Taken together, this suggests that the sister-group relationship of Cercidoideae with the rest of the family is the most likely rooting as inferred from chloroplast data, but given the low bootstrap support values, phylogenetic signal with regards to the root node appears to be limited. A notable observation is that the chloroplast genome evolves markedly faster in the 50kb-inversion clade of Papilionoideae than in other legumes (with even higher rates apparent in the vicioid clade), as is evident from both the nucleotide and amino acid alignments (Figs 2C & S1-4), suggesting that this pattern is not driven solely by synonymous substitutions. However, branch model D_n/D_s ratio tests do not find any evidence of differential selection acting on chloroplast genes across the different clades (Fig. 2D), and suggest that the majority of chloroplast genes across legumes are under purifying selection.

In contrast to the results obtained with chloroplast data, in all analyses of the 1,103 nuclear 1-to-1 orthologs, we recover a sister-group relationship between Cercidoideae and Detarioideae, with this clade sister to the clade comprising of Dialioideae, Caesalpinioideae and Papilionoideae (note that Duparquetioideae is not sampled) (Figs 3A-C & S5-7). We inferred an ML tree of the concatenated alignment with the LG4X model, and calculated Internode Certainty All (ICA) values from bootstrapped gene trees on this topology (Fig. 3A & S5). Only bipartitions that received

579 >80% bootstrap support were considered. The internode certainty metric was
580 introduced to assess phylogenetic conflict among loci and identify internodes with high
581 certainty, to be used in particular in phylogenomic studies where bootstrap values are
582 often inflated (Salichos & Rokas, 2013). The sister-group relationship between
583 Cercidoideae and Detarioideae is well-supported, receiving an ICA value of 0.85. A
584 Bayesian jackknifing analysis with the CATGTR model infers a nearly identical topology
585 to the ML topology (Fig. 3B & S6), with posterior probability of 0.91 in support of this
586 same relationship. The multi-species coalescent species-tree inferred with ASTRAL
587 (Mirabab et al., 2014), which accounts for incomplete lineage sorting (ILS), is also
588 consistent with that relationship (Fig. 3C & S7), with the Cercidoideae/Detarioideae
589 clade supported by a local posterior probability of 0.95 (Sayyari & Mirabab, 2016). In
590 summary, all analyses of nuclear protein alignments lend strong support for a sister-
591 group relationship between Cercidoideae and Detarioideae.
592
593 *Evaluation of Gene Tree Support and Conflict*
594
595 While the chloroplast and nuclear phylogenies show a different topology with
596 regards to the first two dichotomies within the legumes, the different types of analyses
597 performed on the nuclear data set all yield the same topology at the base of the family
598 (Figs 3A-C). Because the nuclear data set consists of 1,103 unlinked loci sampled from
599 across the nuclear genome compared to the single locus that the chloroplast genome

constitutes, this topology should be considered to be more likely. However, when evaluating gene tree conflict, it appears that a large number of conflicting bipartitions exist, with the most prevalent being nearly as frequent across gene trees as compatible bipartitions (pie charts in Figure 3A). The quartet support as calculated by ASTRAL is also low (37%, with alternative quartet supports 33% and 30%; pie charts in Figure 3C). The relationships among the remaining three sampled subfamilies are also supported by significantly fewer bipartitions and lower quartet support than for example the legume crown node (pie charts in Figures 3A & C). Furthermore, the filtered supernetwork shows a complex tangle of gene tree relationships at the base of the legumes (Fig. 4).

Rather than relying solely on ICA and quartet support values, we sought to evaluate in a more intuitive way how much support and conflict there is among gene trees for the deepest divergences in the legume family. For nodes labeled A-H in Figure 3A, we counted how often a bipartition that is equivalent to that node in the species tree is encountered across gene trees, and how often those bipartitions received at least 50 or 80% bootstrap support. We did this on all RT homologs (n=7,621) in which all subfamilies and the outgroup were represented by at least one taxon each, leading to 3,473 gene trees being considered. This shows that the legume family as a whole, and the four subfamilies for which more than one taxon was sampled (nodes C, D, G and H), are all found to be monophyletic across the majority of gene trees (Fig. 5A & Table 2), and those bipartitions mostly receive at least 50 or 80% bootstrap support. Nodes B, E and F, that is, the relationships among the subfamilies, are recovered in many fewer

621 gene trees, especially when only considering bipartitions with at least 50 or 80%
622 bootstrap support. For these nodes, we then checked how often the most important
623 conflicting bipartitions were present (Figs 5B-D & Table 2). These conflicting bipartitions
624 are each less prevalent than those found by the concatenated ML and Bayesian
625 analyses as well as by ASTRAL, confirming that the recovered topology represents the
626 relationships among legume subfamilies that is supported by the largest fraction of the
627 genomic data used here. But it also shows that there is significant and well-supported
628 gene tree conflict, in line with the complicated tangle and short edges observed in the
629 filtered supernetwork at the base of the legumes (Fig. 4).
630
631 *Inferring Phylogenetic Locations of WGDs*
632
633 To map gene duplications over the species tree, we first removed fragmentary
634 sequences and gappy sites from the 9,282 homolog clusters, after which 640 clusters
635 with large amounts of missing data were eliminated. From trees that were inferred from
636 the remaining 8,642 homologs, we extracted 8,038 rooted clades. Exemplar homolog
637 trees with gene duplications are shown in Figure S8. We find significantly elevated
638 numbers of gene duplications at several nodes where WGDs are hypothesized to have
639 occurred, including the previously documented *Salix/Populus* clade (Tuskan et al.,
640 2006) and one subtending Pentapetalae, consistent with the known *gamma*
641 hexaploidization associated with that clade (Jiao et al., 2012) (Figs 3D & S9). For the

642 Pentapetalae clade, many homologs show more than one gene duplication at that node,
643 given that the number of duplications (1,901) is nearly twice as high than the number of
644 homologs with duplications (1,105), as expected for two consecutive rounds of WGD.
645 Part of these duplications may also stem from older events, since missing data for the
646 three non-Pentapetalae taxa in our dataset could mean that we do not find duplicates of
647 older events in these taxa. In the legumes, high numbers of gene duplications at
648 particular nodes suggest that there were three early WGD events, one at the base of
649 the family, and one each subtending subfamilies Papilionoideae and Detarioideae (Figs
650 3D & S9). When applying a bootstrap filter to the homolog trees (≥50% bootstrap
651 support), numbers of gene duplications are considerably lower, but the pattern is the
652 same (Figs 3D & S9). At the root of the family, the number of gene duplications drops
653 from 1,646 to 99 when applying this bootstrap filter, in line with the difficulty of resolving
654 the deepest dichotomies of the legume phylogeny. Notably, for the legume crown node
655 we also find evidence for a significant part of homologs having had more than one gene
656 duplication, because 1,646 duplications from only 1,229 homologs map on that node.
657 This would suggest multiple rounds of WGD (e.g. Figs S8E & F), although some of
658 these can be attributed to duplications in both paralog copies of genes duplicated at the
659 *gamma* event, while for many others support values across the tree are low. For other
660 hypothesized WGDs, the numbers of homologs with more than one duplication for those
661 nodes are much lower, suggesting they involved a single round of WGD.
662

663 *Divergence Time Estimation*

664

665 To establish whether the origin of legumes and the early WGD events are closely
666 associated with the KPB, we performed clock dating in a Bayesian framework. Because
667 the chloroplast phylogeny shows large root-to-tip length variation (Fig. 2), we refrained
668 from using the chloroplast data to infer divergence time estimates, and instead rely on
669 the better suited nuclear data for this purpose as suggested by Christin et al. (2014).
670 We selected 36 relatively highly informative and clock-like nuclear genes and 20 fossil
671 calibrations (Table 1 and Methods S1). The oldest definitive fossil evidence of crown
672 group legumes is from the Late Paleocene, consisting of bipinnate leaves from c. 58 Ma
673 (Wing et al., 2009; Herrera et al., submitted) and papilionoid-like flowers from c. 56 Ma
674 (Crepet & Herendeen, 1992), representing Caesalpinioideae and Papilionoideae
675 respectively. The older fossil woods with vestured pits, from the Early Paleocene of
676 Patagonia (Brea et al., 2008) and the Middle Paleocene of Mali (Crawley, 1988), could
677 represent stem relatives of the family (vestured pits are found in Papilionoideae,
678 Caesalpinioideae and Detarioideae, so this is likely an ancestral legume trait). Based on
679 this fossil evidence, c. 58 Ma can be considered the minimum age of the legume crown
680 node. Molecular age estimates (95% HPD intervals) for the crown node range from
681 65.47-86.45 Ma and 73.46-81.18 Ma under the uncorrelated log-normal relaxed clock
682 (UCLN) and the random local clock (RLC) models, respectively, to minima and maxima
683 between 64.63 and 68.85 Ma under various fixed local clock (FLC) models (Table S3),

684 the latter suggesting a close association of the origin of the legumes with the KPB (Fig.
685 6). Maximum clade credibility (MCC) trees for all clock analyses, with 95% HPD
686 intervals indicated, are included in Supplementary Figures S10-17, and 95% HPD
687 intervals for nodes A-H are listed in Table S3.

688 Placement of Eocene fossils of Detarioideae and Cercidoideae within the crown
689 groups of those clades (Bruneau et al., 2008; Simon et al., 2009; de la Estrella et al.,
690 2017), yields older crown group estimates for these clades. However, with these
691 calibrations (alternative prior 1 in Table S4), a more than 10-fold higher substitution rate
692 along the stem lineages of these two subfamilies relative to the rates within both crown
693 clades is inferred (c. 8.82×10^{-3} vs 0.69×10^{-3} substitutions per site per million years,
694 with identical rates estimated independently for Cercidoideae and Detarioideae; Fig.
695 S18A). This rate is also nearly five times higher than the mean rate across the tree as a
696 whole (1.54×10^{-3} substitutions per site per million years), while the crown clades are
697 estimated to have rates about half as high as the mean. Analyses with the same clock
698 partitioning but calibrated with Late Eocene *Cercis* fossils and Mexican amber
699 (*Hymenaea*) as the oldest crown group evidence for Cercidoideae and Detarioideae,
700 respectively, do not infer such strong substitution rate shifts, with all clock partitions
701 across the phylogeny estimated to have a substitution rate ranging from 0.96×10^{-3} to
702 2.53×10^{-3} substitutions per site per million years (Fig. S18B). Either way, different
703 placements of these fossils have little influence on the crown age estimates for the
704 family in the FLC analyses (Figs S15 & S16, Table 3).

705

706 DISCUSSION

707

708 In this study, we present significant advances in our understanding of the origin
709 and early evolution of the legume family. All the different species tree analyses of the
710 nuclear genomic data yielded the same most likely topology with regards to
711 relationships among subfamilies and the root of the legumes. Detailed evaluation of
712 supporting and conflicting bipartitions across gene trees show that these relationships
713 are the most prevalent, but we also found many conflicting bipartitions, and the
714 chloroplast phylogeny also shows a different rooting of the family. Furthermore, we find
715 evidence for three WGD events early in the evolution of the family, which further
716 complicate the phylogenomic tangle at the base of the family. Time-calibration of the
717 species tree suggests a close association of this complex origin of the legumes with the
718 KPB. We discuss these findings and their relevance to understanding the evolution of
719 the third largest angiosperm family, the likely complications caused by WGDs on
720 phylogenetic inferences in deep time and the consequences of the KPB mass extinction
721 event on plant evolution in the Cenozoic.

722

723 *Substitution Rate Variation in Legume Chloroplast Genomes*

724

The chloroplast data set has the advantage of denser taxon sampling (including subfamily Duparquetioideae) compared to the nuclear genomic data. However, the chloroplast data are less useful for phylogenomic analysis, being effectively a single locus in the absence of recombination in plastid genomes. Furthermore, chloroplast genes have highly heterogeneous substitution rates across legumes, leading to a well-resolved topology in core Papilionoideae but poor resolution in other lineages, particularly Caesalpinioideae (Figs 2C & S1-4). It has long been known that there is significant variation in rates of chloroplast sequence evolution among plant lineages (Bousquet et al., 1992) and previous analyses of single chloroplast genes (Lavin et al., 2005) and legume chloroplast genomes (Dugas et al., 2015; Schwarz et al., 2017; Wang et al., 2018) have suggested substantial variation in rates of molecular evolution among legume lineages. Because branch model D_n/D_s ratio tests do not provide evidence for different selective forces on photosynthesis genes across legumes, this pattern may rather be related to life-history strategies in the 50Kb-inversion clade, which includes many herbaceous plants of short stature (Lanfear et al., 2013) and shorter generation times (Smith & Donoghue, 2008), especially in the vicioid clade where the highest rates are found (Fig. 2C).

Resolving the Deep-branching Relationships in the Leguminosae

745 The difficulty of obtaining resolution for the deep divergences in the legume
746 family is in part caused by lack of phylogenetic signal in a large fraction of the sampled
747 genes (pie charts in Figure 3A), with too few substitutions having accumulated along the
748 deepest short internodes due to rapid early divergence of the six principal legume
749 lineages. Lack of phylogenetic signal could potentially be explained by rapid
750 diversification which, especially in combination with extinction of stem-relatives, causes
751 alternations of long and short internodes, leading to “bushy” phylogenies that are
752 extremely difficult to resolve (Rokas & Carrol, 2006). However, for a significant
753 proportion of those genes that do have sufficient phylogenetic signal, we find strongly
754 supported conflicting evolutionary histories. Putting aside methodological issues such
755 as poor orthology inference for a number of genes, this conflict is likely to be caused by
756 incomplete lineage sorting (ILS) (Pamilo & Nei, 1988; Maddison, 1997). Together with
757 the complexity depicted in the supernetwork (Fig. 4), the strongly supported conflicting
758 gene trees suggest that a fully bifurcating tree is an oversimplified representation of the
759 initial radiation of the legumes. As we show here, genes have many different
760 evolutionary histories across the early divergences of legumes (Table 2), while the
761 species tree merely represents the dominant evolutionary history. In the case of
762 complete lack of phylogenetic signal, or equally prevalent conflicting evolutionary
763 histories without a single dominant one, this would constitute a hard polytomy, meaning
764 (nearly) instantaneous speciation of three or more lineages, as demonstrated for
765 Neoaves (Suh, 2016). Alternatively, a phylogenetic network can provide a better

766 representation of evolutionary relationships when there is significant gene tree conflict.
767 In the legumes, there does appear to be one dominant evolutionary history in the
768 relationships among subfamilies, suggesting that the root of the family is strictly
769 speaking not a hard polytomy. Nevertheless, the short internodes leading to lack of
770 phylogenetic signal and significant conflict among gene trees at the base of the legumes
771 suggest that the first few divergences in the family occurred within a short time span,
772 leading to ILS. Indeed, strong gene tree conflict caused by ILS has been shown to be
773 relatively common when internodes are short due to rapid speciation and this provides
774 an explanation as to why many relationships are contentious (e.g. Pollard et al., 2006;
775 Suh et al., 2015; Moore et al., 2017). In such cases, it is essential that phylogenomic
776 studies explicitly evaluate conflicting phylogenetic signals across the genome. By taking
777 into account alternative topologies that are supported by significant numbers of gene
778 trees (Fig. 5) and inferring a phylogenetic network (Fig. 4), the phylogenomic complexity
779 of the initial radiation of the legumes is revealed.
780
781 *Locating WGD Events on the Phylogeny*
782
783 Numbers of gene duplications mapped onto the species tree provide evidence for
784 three WGD events early in the evolution of the legume family, one shared by the whole
785 family, plus independent nested WGDs subtending subfamilies Detarioideae and
786 Papilionoideae. We note that several nodes that immediately follow the most likely

787 locations of hypothesized WGD events also show elevated numbers of gene
788 duplications (Figs 3D & S9). This is most likely caused by missing data for some taxa.
789 For example, sequences for *Xanthocercis zambesiaca*, *Cladrastis lutea* and
790 *Styphnolobium japonicum* are derived from transcriptomes, while in the core
791 Papilionoideae, several accessions are represented by fully sequenced genomes and
792 therefore have higher gene sampling. Alternatively, paralog copies for a subset of genes
793 could have been lost in lineages outside the core Papilionoideae. These gene sampling
794 issues mean that a considerable number of gene duplications are likely to be mapped
795 onto the second and third divergences in the subfamily, even though they probably stem
796 from the same WGD event shared by the subfamily as a whole. Similar patterns are
797 apparent at the bases of the legumes and of Pentapetalae (Figs 3D & S9). At the bases
798 of subfamily Caesalpinioideae and the Mimosoid clade, we also find modestly elevated
799 numbers of gene duplications, but fewer than for the three main duplication events (Figs
800 3D & S9). This could indicate a partial genome duplication shared by all
801 Caesalpinioideae and another one shared by all mimosoids. Alternatively, it could reflect
802 higher gene coverage in the mimosoid transcriptomes relative to the other
803 Caesalpinioideae, in which case many of the gene duplications currently depicted as
804 subtending the Mimosoid clade should potentially map at the base of the
805 Caesalpinioideae. It is therefore possible that another WGD has occurred at the base of
806 the Caesalpinioideae, as suggested by Cannon et al. (2015), but the rather low
807 numbers of gene duplications inferred from our data cannot be considered as strong

evidence for that. Cannon et al. (2015) also hypothesized another WGD early in the evolution of subfamily Cercidoideae, shared by Cercis and Bauhinia. Our results do not support this, and furthermore, in Bauhinia s.l. (Cercidoideae) the most common haploid chromosome number is n=14, while Cercis has n = 7. This suggests that an early WGD in Cercidoideae was not shared by Cercis. This is further supported by a densely sampled phylogenetic analysis of the LegCyc gene in Cercidoideae, which is duplicated in all Cercidoideae except Cercis, the sister group to the rest of the subfamily (Carole Sinou, unpublished data). Cannon et al. (2015) further suggested that the ancestral legume most likely had a haploid chromosome number of n = 6 or 7 and had independently doubled in most lineages to arrive at n = 14, the haploid chromosome number that is most commonly found across legume subfamilies except Detarioideae (n = 12) and core Papilionoideae (Cannon et al., 2015: Fig. 1; chromosome counts for Duparquetia are not available). This would imply that Cercis, with n = 7, would have retained the ancestral haploid chromosome number. Indeed, given our results it is likely that the mrca of Cercidoideae and Detarioideae would have had a haploid chromosome number of 6 or 7, followed by independent WGDs in Bauhinia s.l. and Detarioideae to arrive at n = 14 and n = 12, respectively. However, the mrca of Dialioideae, Caesalpinioideae and Papilionoideae most likely had a haploid chromosome number of n = 14, followed by reductions in chromosome number in Chamaecrista and Papilionoideae (Cannon et al., 2015: Fig. 1). Even after an additional WGD in Papilionoideae, extant members of the subfamily still have chromosome numbers <14

(Cannon et al., 2015: Fig. 1), suggesting extensive genomic rearrangement. That leaves the chromosome number of the mrca of all legumes uncertain, being either n = 6 or 7, or n = 14, suggesting either chromosome number reduction in some lineages, or potentially inheritance of different ploidy levels in different lineages from an ancestral polyploid complex. In conclusion, we find evidence that supports many of the findings of Cannon et al. (2015), but our results suggest an additional WGD event that is shared by all legumes, in line with the findings of Wong et al. (2017). Our study expands the taxon sampling of Cannon et al. (2015), but has the same limitation in that a large number of accessions are based on transcriptome data and are thus not sampling complete exomes. Denser sampling of completely sequenced legume genomes will be needed to resolve the number and placement of WGD events with higher confidence, precision and accuracy.

Estimating the Timeline of Legume Evolution

Our divergence time analyses update previous analyses of Lavin et al. (2005), Bruneau et al. (2008) and Simon et al. (2009), and provide, to our knowledge, the first divergence time estimates for legumes based on nuclear genomic data as well as the first molecular clock dating estimate for the crown age of the legumes. The age estimates under the FLC models and the strict clock model are mostly rather similar, but the RLC and UCLN models, that relax the clock assumption more, lead to older

divergence time estimates. By allowing independent substitution rates on all branches, these models are potentially overfitting the data, to attempt to satisfy the marginal prior on node ages (Brown & Smith, 2017). As inferred from analyses run without data, the marginal prior that is constructed across all nodes of the tree, can be considered as "pseudo-data" (Brown & Smith, 2017), derived from the node calibration priors (based on fossil ages) and the branching process prior (constant birth-death model in our case), and should therefore not be overly informative on node ages. FLC and strict clock models lend greater weight to the molecular data and can overrule the marginal prior distributions on divergence times whilst still respecting hard maximum and minimum bounds of the fossil constraints on calibrated nodes, as suggested by our results. It is also clear from running analyses without data, that the marginal age prior on the (uncalibrated) crown node of the legumes is rather poorly informed, with the 95% HPD interval between 79.37-109.20 Ma (Fig. 6B), the minimum being much older than the oldest legume fossils, presumably caused by overly conservative maximum bounds on calibrated nodes (Phillips, 2015). UCLN and RLC analyses also inferred relatively high substitution rates for a few deep branches in the outgroup during the Lower Cretaceous, relative to the more derived and terminal branches of the tree (Figs S10 & S12), presumably to satisfy the poorly informed marginal priors. Phillips (2015) suggested that setting less conservative maxima on priors could remedy this problem, but our analysis with such prior settings shows little effect (Fig. S11), with some of the deepest branches still having much higher estimated substitution rates. Since there is

871 no evidence, nor any reason to assume, that substitution rates along those branches
872 should be elevated relative to terminal branches, we conclude that this is indeed caused
873 by overfitting of rate heterogeneity across branches under the influence of the marginal
874 prior. Furthermore, the RLC analyses fitted c. 45 local clocks across the phylogeny, a
875 rather high number relative to the total of 142 branches in the tree (implying a separate
876 clock for every 3 branches), which is also indicative of overfitting. At the same time, this
877 could be seen as evidence that the data are not the product of clock-like evolution, but it
878 becomes difficult to estimate how much the clock deviates if the marginal prior on node
879 ages is too influential. A more pragmatic approach is to use FLC analyses, by defining
880 local clocks based on root-to-tip length distributions across clades and pruning outlier
881 taxa (see Methods and Fig. S19). This approach accounts in large part for the violation
882 of the molecular clock but it does not relax the clock to the extent that the marginal prior
883 on node ages is given excessive weight relative to the molecular signal. Furthermore,
884 because the genes we selected for divergence time estimation are reasonably clock-like
885 and highly informative, it is desirable that these data inform the node ages with sufficient
886 weight. One drawback of using this approach is that the relatively large amount of
887 sequence data in combination with the FLC model results in estimates that appear
888 unrealistically precise, and the discovery of new fossils may well prove the legumes to
889 be slightly older. Nevertheless, the evidence presented here suggests that the legume
890 crown age dates back to the Maastrichtian or Early Paleocene, likely within one or two

million years before or after the KPB, although such high precision is not warranted due to the idiosyncrasies of the molecular clock.

Polyploidy (Senchina, et al., 2003) as well as the KPB itself (Berv & Field, 2018), have been implicated as potentially causing transient substitution rate increases, raising the possibility that substitution rates during the early evolution of the legumes could have deviated temporarily but markedly from the "background" rate of Cretaceous rosids. This would render the ages inferred for the first few dichotomies as well as those of the subfamilies less certain. The age estimates inferred for these nodes rely in large part on the assumption that the substitution rate did not vary significantly within the different clock partitions, and most importantly within the rosid partition which includes most of the branches along the backbone of the family and the stem lineage subtending it. The WGD events along the stem lineages of the family, and subfamilies Papilionoideae and Detarioideae could have affected substitution rates along those branches. By selecting for smaller stature and shorter generation times and reducing population sizes (Berv & Field, 2018), the KPB could additionally have resulted in increased rates along some or all of the stem lineages of the subfamilies, and, in the case of "hard" explosive diversification after the KPB, perhaps also along the legume stem lineage. A third factor that could influence node age estimates along the backbone of the family, is the strong gene tree incongruence observed for nodes B, E and F (Fig. 5), which is also observed among the 36 genes that were used for time-scaling. The divergence time analyses need to accommodate this incongruence within a single

912 topology, meaning that additional substitutions need to be inferred for conflicting gene
913 trees, which can inflate the branch lengths between rapid speciation events (Mendes &
914 Hahn, 2016). Taken together, these three factors could mean that the timeframe for the
915 early evolution of the legumes appears inflated in our results, with (some of the)
916 subfamily ages likely being slightly older than estimated here, as well as divergence of
917 the subfamilies happening nearly instantaneously (hence the gene tree incongruence
918 and lack of phylogenetic signal), rather than spanning the c. 3 -5 million years inferred
919 here (Figs 6 & S10-17). Potentially, even the legume crown age could be slightly older
920 due to the effects of polyploidy, but not due to the KPB, because if the crown is older,
921 the stem lineage would not have crossed the KPB.

922 Different interpretations of Eocene fossils of Cercidoideae and Detarioideae (see
923 Methods S1) lead to very different crown age estimates for these clades. As expected,
924 this also leads to very different substitution rates along the stem lineages of these
925 subfamilies, whit rates increasing 10-fold when interpreting these fossils as crown group
926 members. While it cannot be ruled out that the stem lineages of Cercidoideae and
927 Detarioideae experienced such markedly elevated substitution rates, it is unlikely that
928 rates were five times higher relative to the rest of the eudicots across all 36 nuclear
929 genes analysed, especially as these genes were chosen because of their approximately
930 clock-like evolution, and given that these two clades comprise long-lived woody
931 perennials. The idea that molecular information from extant taxa could inform that
932 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. S16). 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 in Supplementary Methods S1).

The Impact of the KPB on Plant Diversification

The impacts of the KPB mass extinction event on plant diversity are the focus of debate, with several studies claiming that extinction was less severe for plants than across marine and terrestrial faunas (Nicholls & Johnson, 2008; Cascales-Miñana & Cleal, 2014; Silvestro et al., 2015). However, our results suggest that the massive KPB turnover event likely played a critical role in the evolution of plant taxa. Our analyses indicate that the origin of crown group legumes is closely associated with the KPB. The

954 analyses employing FLCs even suggest that potentially only a single legume ancestor
955 crossed the KPB to give rise to the six main lineages during the early Paleocene,
956 conforming to a “hard explosive” model. However, across the different analyses, part of
957 the posterior density of the crown age estimate falls in the late Maastrichtian,
958 suggesting a “soft explosive” model, with the six main lineages diverging in the Late
959 Cretaceous and crossing the KPB, giving rise to the crown groups of the modern
960 subfamilies in the Cenozoic. These different explosive models have been used to
961 describe the origin and early diversification of the placental mammals, although other
962 studies have lent support to “short fuse” or “long fuse” models (summarized in Phillips,
963 2015: Fig. 1). For birds, the timing of diversification relative to the KPB has also been
964 controversial (Ksepka & Phillips, 2015), but it now appears likely that the Neoaves
965 underwent explosive radiation from a single ancestor that crossed the KPB (Suh, 2016).
966 Apart from Placentalia and Neoaves, recent studies on frogs (Feng et al., 2017) and
967 fishes (Alfaro et al., 2018) have also demonstrated rapid diversification following the
968 KPB, suggesting this is a common pattern across many terrestrial and marine animal
969 groups. We present here, to our knowledge, the first example of a major plant family
970 whose origin and initial diversification appears to be closely linked to the KPB. This is
971 notable because a recent family-level paleobotanical study suggested that the KPB did
972 not constitute a mass extinction event for plants (Cascales-Miñana & Cleal, 2014).
973 Phylogenetic studies in some plant families originating in the Cretaceous also lack any
974 evidence of a significant effect of the KPB on diversification (e.g. Annonaceae

(Couvreur et al., 2011a) and Arecaceae (Couvreur et al., 2011b)), except for the smaller plant family Menispermaceae (Wang et al., 2012), which shows increased diversification following the KPB. In contrast, fern diversification appears to have been strongly affected, with some groups of ferns showing much reduced diversity in the Cenozoic compared to earlier times (Lehtonen et al., 2017), and especially epiphytic groups of ferns showing increased diversification rates since the KPB (Schuettpelz & Pryer, 2009). Furthermore, the generic-level study of Silvestro et al. (2015) showed high extinction rates for non-flowering plant groups during the late Cretaceous, and elevated origination rates for angiosperms during the Paleocene, in line with the pattern we observe for the origins of legume diversity. Thus, even if extinction was less severe for plants than for animals at the KPB, the Paleocene was nevertheless a time of major origination of lineages across biota, and we expect further examples of KPB-related accelerated plant diversification to be discovered when inferring larger angiosperm timetrees.

Implications for our Understanding of the Evolution of Legume Diversity and Traits

Rapid divergence of the six main lineages of legumes is clearly relevant to our understanding of the evolution of legume diversity and the appearance of key traits. Over the last few decades, the prevailing characterization of legume evolution has been that of mimosoids and papilionoids as “derived” clades that evolved from a paraphyletic “grade” of caesalpinioid legumes (e.g. LPWG, 2013). However, we show that all six

996 subfamilies diverged across a short time span after the origin of the legume crown
997 group, with long stem lineages subtending each subfamily, suggesting that none of the
998 modern subfamilies should be seen as diverging earlier or later than any other. The
999 complex phylogenomic paleopolyploid tangle documented here means that it will be
1000 extremely difficult to reconstruct trajectories of trait evolution across the first few
1001 divergences within the family. For example, it is not clear how to understand the
1002 evolution of floral diversity across the family and what the ancestral legume flower
1003 would look like. That makes it questionable, for example, to what extent the specialized
1004 and strongly canalized zygomorphic papilionoid flowers are derived within the family.
1005 Fossil papilionoid flowers from the Paleocene (Crepet & Herendeen, 1992) are among
1006 the oldest evidence of the family in the fossil record. The higher morphological diversity
1007 of flowers in other subfamilies may well have evolved in parallel or even later than the
1008 papilionoid flower, given the crown age estimates that we find in Bayesian clock
1009 analyses (Figs 6, S10-17 and Table S3).

1010 While we are not able at this point to confidently distinguish between a “hard” or
1011 “soft explosive” model of early diversification of the family, it is clear that the early
1012 radiations of the legume subfamilies all occurred in the Cenozoic. While stem age
1013 estimates of each subfamily are remarkably close to each other, crown age estimates
1014 are strikingly different (but see the discussion above on potential effects of polyploidy
1015 and the KPB on substitution rates and ages of subfamilies). Caesalpinioideae are found
1016 to have the oldest crown age (late Paleocene), followed by Papilionoideae with a crown

age in the Early Eocene. Both of these subfamilies therefore likely diversified considerably during the PETM and Eocene climatic optimum, when tropical forests extended far into the Northern Hemisphere. This is in line with the numerous legume fossil taxa known from the Eocene of North America, often of uncertain affinities, but with a majority ascribed to Caesalpinioideae and Papilionoideae (Herendeen, 1992). There is also fossil evidence of Early and Middle Eocene stem-relatives of Cercidoideae and Detarioideae (as discussed above and in Methods), but their crown group divergences are most likely placed in the Late Eocene or Oligocene. Our results suggest extinction of stem-relatives of these two subfamilies, most likely related to Late Eocene and Oligocene cooling, and subsequent diversification of the crown groups during the Oligocene and Miocene, when both groups become diverse at several fossil sites (e.g. Wang et al., 2014; Lin et al., 2015; Poinar, 1991; Poinar and Brown, 2002). Although it remains uncertain whether the crown group divergence of Detarioideae occurred in the (Late) Eocene or the Oligocene, the younger age of the subfamily inferred here contrasts with previous views of the evolutionary trajectories of this subfamily dating back into the Paleocene, comprising relatively slowly evolving lineages (de la Estrella et al., 2017), and with Amazonian subclades within Detarioideae conforming to the museum model of tropical rainforest diversification (Schley et al., 2018). This has important implications for our understanding of the origins of tropical African plant diversity, since Detarioideae dominate the canopy of many equatorial African rainforests, as well as being an important group in African savannas (de la

1038 Estrella et al., 2017). Our results for Detarioideae suggest that the extant diversity in
1039 tropical Africa, in particular the large diversity in tribe Amherstieae, is of relatively recent
1040 origin following a major turnover event at the Eocene-Oligocene boundary, which also
1041 affected other plant groups such as palms (Pan et al., 2006). This more recent
1042 diversification of detarioids is also more in line with the widely proposed recent
1043 assembly of the savanna biome (Cerling et al., 1997; Bouchenak-Khelladi et al., 2009;
1044 Maurin et al., 2014).
1045
1046 *The Added Complications of Paleopolyploidy on Evolutionary Inferences in Deep Time*
1047
1048 The recent proliferation of genomic data is revealing just how prevalent repeated
1049 WGDs have been in the history of the angiosperms (e.g. Wendel, 2015; Soltis et al.,
1050 2016; Yang et al., 2018) and how many large angiosperm clades are characterized by
1051 genome triplications (e.g. Pentapetalae, Brassicaceae, Asteraceae, Solanaceae). Here
1052 we show that there were also multiple WGDs during the early history of the legumes,
1053 including a WGD subtending the family as a whole. It has been suggested that
1054 angiosperm WGDs are non-randomly distributed through time and significantly
1055 clustered around the KPB (Fawcett et al., 2009; Vanneste et al., 2014; Lohaus & Van de
1056 Peer, 2016). The WGD that we identify that is shared by all legumes is also temporally
1057 close to the KPB (Fig. 6), lending further support to the idea that polyploid survival and
1058 establishment were enhanced at or soon after the KPB with its associated rapid

turnover of lineages (Lohaus & Van de Peer, 2016; Levin & Soltis, 2018). WGDs have also been hypothesized to trigger accelerated rates of lineage diversification at least in some lineages, albeit potentially after a time lag (Schranz et al., 2012; Tank et al., 2015; Landis et al., 2018; Smith et al., 2018a). The three legume WGDs we detected are each followed by rapid divergence of lineages as indicated by short internodes (Figs 2, 3 & 6). Polyploidy could have helped ancestral legumes and other plant lineages to both survive the mass extinction event and rapidly diversify owing to differential gene loss and other processes of diploidization (Adams & Wendel, 2005; Dodsworth et al., 2016). Increased polyploid speciation and reduced diploid speciation in the wake of the KPB (Levin & Soltis, 2018) would then lead to over-representation of these WGD-derived lineages in the extant flora and clustering of WGDs around the KPB. On the other hand, many paleopolyploidy events that significantly pre- and post-date the KPB are known (e.g. Angiospermae (Jiao et al., 2011), Pentapetalae (Jiao et al., 2012), Salicaceae (Tuskan et al., 2006), Caryophyllales (Yang et al., 2018), *Gossypium* (Wendel, 2015)), including in legumes (e.g. *Glycine*, Genisteae, the *Leucaena* group, *Vachellia)*, and more extensive sampling of recently diversified groups may well reveal a weaker pattern of clustering around the KPB.

The WGD events subtending all legumes and subfamilies Detarioideae and Papilionoideae are likely to have contributed to the difficulties of obtaining phylogenetic resolution for the deep nodes in these clades (Cardoso et al., 2012 & 2013; de la Estrella 2018). WGDs may have promoted increased lineage diversification rates

resulting in short internodes and ILS. If the polyploidy event happened some time before the first divergences in the legume family, or in the case of allopolyploidy, this could have led to divergent gene copies prior to lineage splitting which should make orthology detection easier. However, if the polyploidy event happened shortly before rapid cladogenesis, potentially a large fraction of paralogous gene copies would not have diverged at this point, making orthology detection challenging. In both cases, paralogous or homoeologous gene copies will have subsequently been differentially lost, pseudogenized or sub- or neo-functionalized, further complicating correct orthology detection. Together with ILS, this could explain the large fraction of gene trees supporting alternative topologies at the root of the legumes. It is notable that several other large plant clades, such as Pentapetalae (Zeng et al., 2017), Asteraceae (Barker et al., 2016; Huang et al., 2016) and Brassicaceae (Couvreur et al., 2010; Huang et al., 2015), also appear to show similar lack of resolution in clades subtended by WGDs to that revealed here for the legume family and subfamilies Papilionoideae and Detarioideae. This suggests that the association of polyploidy with rapid divergence, which leads to a lack of phylogenetic signal and gene tree conflict, is potentially a common feature in the evolution of angiosperms and the origination of major plant clades.

A large number of homolog clusters do not show gene duplications at the base of the legumes or any of the subfamilies, suggesting that loss of paralog copies is widespread, as observed for ancient WGDs more generally (Adams & Wendel, 2005;

Dehal & Boore, 2005; Brunet et al., 2006; Scannel et al., 2007). If many of those losses occurred along the stem lineages of the six subfamilies after their divergence, this could lead to different paralog copies being retained in different lineages, adding to conflict among gene trees. Loss of paralog copies along stem lineages of subfamilies will also make it difficult to distinguish whether a gene duplication corresponds to the WGD shared by all legumes, or whether it represents a nested WGD such as those subtending Detarioideae and Papilionoideae. Lack of support in those homolog trees showing gene duplications further complicates this issue, making it potentially extremely challenging to accurately reconstruct the history of WGDs. Given these difficulties, sampling a wider range of complete genomes will be important, since with transcriptome data it is unknown whether duplicate gene copies are lost or simply not expressed in the tissue from which the RNA was extracted. Furthermore, increased taxon sampling will help to counteract negative impacts of missing data, since particular duplicate gene copies may have been lost in all species sampled here, but not necessarily across the whole clade or subfamily which those species represent. Despite all these complications, a clear pattern of either high or low numbers of gene duplications is observed when mapping duplications from 8,038 extracted clades from homolog trees across the species tree (Figs 3D & S8). This suggests that when summarizing gene duplications over a sufficiently large data set, it is still possible to make sense of the confusing topological differences that are observed and hence to accurately map WGD events. This leads us to propose the hypothesis presented in Figure 7 to reconcile the

1122 complicated topological patterns observed across gene trees. In this hypothesis, the six
1123 major legume lineages (i.e. subfamilies) diverged rapidly one after another from a
1124 polyploid ancestor. The different gene copies would still be nearly identical at the
1125 moment of cladogenesis and would diverge into paralog copies in each lineage
1126 independently, making it impossible to infer relationships between paralog copies from
1127 different subfamilies, consistent with lack of phylogenetic signal in most clusters.
1128 Coupled with differential loss of paralog copies, the diversity of topologies and the lack
1129 of support that we observe in homolog trees is exactly what would be expected from the
1130 sort of evolutionary history depicted in Figure 7.

1131 A polyploid ancestor reconciles the complex patterns of gene duplications
1132 observed in the homolog clusters, suggesting we have six legume lineages derived from
1133 a recently polyploidized ancestor. This raises a number of important questions: Did the
1134 polyploidization event involve hybridization, leading to allopolyploidy? Was the ancestor
1135 tetraploid or did it have a higher ploidy level? Did all six lineages inherit the same ploidy
1136 level? Alternatively, given that polyploidization results in immediate reproductive
1137 barriers, perhaps divergence of these six lineages was even facilitated by differing
1138 ploidy levels, with all modern legume taxa derived from an ancestral polyploid complex?

1139 These questions are difficult to answer for an event that occurred 66 Ma and for
1140 which much of the evidence has been obscured by subsequent genome reorganization
1141 and loss of the large majority of duplicate gene copies. Over such timescales, it appears
1142 nearly impossible to distinguish between autopolyploidization or allopolyploidization

between two species that had only recently diverged, or multiple recurrent WGDs in a polyploid complex, or even to disentangle the impacts of possible reticulation from the effects of ILS. On the one hand, a hybridization event involving WGD could explain the strong gene tree conflict that we observe. However, equally this conflict could be explained by ILS alone. The first few divergences within the family occurred within less than 5 Myr (Fig. 6 & S10-17), and this is probably an overestimate due to gene tree incongruence (Mendes & Hahn, 2016). With a sufficiently large effective population size, the majority of loci would not yet have reached reciprocal monophyly over such a short time. Furthermore, the lack of resolution among the different gene copies in the majority of homolog trees suggests that genes did not diverge significantly prior to the WGD, therefore ruling out the possibility of allopolyploidization of two divergent lineages.

We hypothesize a polyploid ancestor of all legumes, but the ploidy level of this ancestor remains uncertain. Some of the gene trees suggest that multiple rounds of WGD occurred at the base of the legumes, prior to further WGDs that occurred independently in subfamilies Detarioideae and Papilionoideae (Fig. S8 E&F). Indeed, of the 794 homolog trees in which pan-legume duplications occurred, for 166 trees more than one duplication was mapped to the legume crown node. Some of these homolog clusters have low support values, so not all of them lend strong support to multiple rounds of WGD. Nevertheless, many of them clearly show more than two well supported duplicated clusters per subfamily, including for subfamilies other than Detarioideae and Papilionoideae. Therefore, the possibility of a hexaploid or octoploid

1164 legume ancestor, akin to events in Angiospermae (Jiao et al., 2011), Pentapetalae (Jiao
1165 et al., 2012), Asteraceae (Huang et al., 2016) and cotton (*Gossypium*) (Paterson et al.,
1166 2012), should also be considered given the evidence presented here. To further
1167 enhance knowledge on legume molecular biology and genome evolution, an obvious
1168 next step will be to sequence multiple complete genomes for all six legume subfamilies
1169 and the other Fabales families, something that will be forthcoming as part of the 10KP
1170 initiative (Cheng et al., 2018). This would potentially make it possible to disentangle the
1171 early genome evolution of legumes by comparing conserved synteny blocks, detecting
1172 genomic rearrangements and reconstructing chromosome evolution and the ancestral
1173 legume karyotype, as has recently been done for vertebrates (Sacerdot et al., 2018)
1174 and birds (Damas et al., 2018), as well as providing ample other opportunities to further
1175 enhance our understanding of legume evolution and diversification.

1176 Ancient polyploidy not only provides a possible explanation for the difficulties in
1177 resolving the root of the legumes, it could also explain the sudden appearance of
1178 diverse legume fossil taxa in the Paleogene. A polyploid ancestor of all legumes would
1179 have provided a much expanded genomic substrate for rapid evolution and
1180 diversification of legume traits, with further rounds of genome duplication leading to an
1181 even further expanded genomic evolutionary substrate independently in Papilionoideae
1182 and Detarioideae and potentially several other legume lineages. In this sense the
1183 parallels to the sudden rise of the angiosperms (Sanderson, 2015) are even more

compelling given that angiosperms are also subtended by one or two ancient WGD events (Jiao et al., 2011; Ruprecht et al., 2017).

Concluding Remarks

It is becoming increasingly clear that the origin and early evolution of the legumes followed a complex scenario with multiple nested polyploidy events, and rapid divergence of the six main lineages against the background of a mass extinction event that led to major turnover in the Earth's biota and biomes. WGD likely contributed to the survival and evolutionary diversification of the legumes in the wake of the KPB mass extinction event, and contributed to the rise to ecological dominance of legumes in early Cenozoic tropical forests. At the same time, these events make it more difficult to reconstruct aspects of the early evolutionary history of the clade, including evolutionary relationships, divergence time estimates and the phylogenetic location of the WGD events themselves. The similarities between legumes and other major Cenozoic clades such as mammals and birds are striking. All three of these prominent Cenozoic clades show recalcitrant basal polytomies and parallel trajectories of rapid early divergence closely associated with the KPB, further emphasizing the importance of the KPB mass extinction event and the earth system succession that followed in its aftermath (Hull, 2015) in shaping the modern biota.

FUNDING

This work was supported by the Swiss National Science Foundation (Grant 31003A_135522 to C.E.H.); the Department of Systematic & Evolutionary Botany, University of Zurich; the Natural Sciences and Engineering Research Council of Canada (Grant to A.B.), the U.K. National Environment Research Council (Grant NE/I027797/1 to R.T.P.), and the Fonds de la Recherche Scientifique of Belgium (Grant J.0292.17 to O.H.).

ACKNOWLEDGEMENTS

We thank the S3IT of the University of Zurich for the use of the ScienceCloud computational infrastructure and the Functional Genomics Center Zurich (FGCZ) for library preparation and sequencing.

REFERENCES

Adams K.L., Wendel J.F. 2005. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8(2):135–141.

Alfaro M.E., Faircloth B.C., Harrington R.C., Sorenson L., Friedman M., Thacker C.E., Oliveros C.H., Černý D., Near T.J. 2018. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. *Nat. Ecol. Evol.* 2:688–696.

1226 Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M.,
1227 Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N.,
1228 Tesler G., Alekseyev M.A., Pevzner PA. 2012. SPAdes: A New Genome
1229 Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput.*
1230 *Biol.* 19(5):455–477.
1231 Barker M.S., Li Z., Kidder T.I., Reardon C.R., Lai Z., Oliveira L.O., Scascitelli M.,
1232 Rieseberg L.H. 2016. Most Compositae (Asteraceae) are descendants of a
1233 paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae.
1234 *Am. J. Bot.* 103:1203–1211.
1235 Barreda V.D., Cúneo N.R., Wilf P., Currano E.D., Scasso R.A., Brinkhuis H. 2012.
1236 Cretaceous/Paleogene Floral Turnover in Patagonia: Drop in Diversity, Low
1237 Extinction, and a Classopollis Spike. PLoS ONE 7(12):e52455.
1238 Berv J.S., Field D.J. 2017. Genomic signature of an avian Lilliput Effect across the K-Pg
1239 extinction. Syst. Biol. 67(1):1–13.
1240 Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
1241 sequence data. *Bioinformatics.* 30(15):2114–2120.
1242 Bouchenak-Khelladi Y., Anthony Verboom G., Hodkinson T.R., Salamin N., Francois O.,
1243 Chonghaile G.N., Savolainen V. 2009. The origins and diversification of C4
1244 grasses and savanna-adapted ungulates. *Glob. Change Biol.* 15(10):2397–2417.

1245 Bousquet J., Strauss S.H., Doerksen A.H., Price R.A. 1992. Extensive variation in
1246 evolutionary rate of rbcL gene sequences among seed plants. *Proc. Natl. Acad.*
1247 *Sci. USA.* 89:7844–7848.
1248 Brea M., Zamuner A.B., Matheos S.D., Iglesias A., Zucol A.F. 2008. Fossil wood of the
1249 Mimosoideae from the early Paleocene of Patagonia, Argentina. *Alcheringa.*
1250 32:427–441.
1251 Brown J.W., Smith S.A. 2017. The past sure is tense: on interpreting phylogenetic
1252 divergence time estimates. *Syst. Biol.* 67:340–353.
1253 Brown J.W., Walker J.F., Smith S.A. 2017. Phyx: phylogenetic tools for unix.
1254 *Bioinformatics.* 33:1886–1888.
1255 Bruneau A., Mercure M., Lewis G.P., Herendeen P.S. 2008. Phylogenetic patterns and
1256 diversification in the caesalpinioid legumes. *Botany*. 86:697–718.
1257 Brunet F.G., Crollius H.R., Paris M., Aury J.M., Gibert P., Jaillon O., Laudet V.,
1258 Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-
1259 genome duplication in teleost fishes. *Mol. Biol. Evol.* 23(9):1808–1816.
1260 Cannon S.B., Sterc L., Rombauts S., Sato S., Cheung F., Gouzy J., Wang X., Mudge J.,
1261 Vasdewani J., Schiex T., Spannagl M. 2006. Legume genome evolution viewed
1262 through the *Medicago truncatula* and *Lotus japonicus* genomes *Proc. Natl. Acad.*
1263 *Sci. USA.* 103:14959–14964.
1264 Cannon S.B., McKain M.R., Harkess A., Nelson M.N., Dash S., Deyholos M.K., Peng Y.,
1265 Joyce B., Stewart Jr C.N., Rolf M., Kutchan T. 2015. Multiple polyploidy events in

1266 the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.*
1267 32(1):193–210.
1268 Cardoso D., de Queiroz L.P., Pennington R.T., de Lima H.C., Fonty E., Wojciechowski
1269 M.F., Lavin M. 2012. Revisiting the phylogeny of papilionoid legumes: New
1270 insights from comprehensively sampled early-branching lineages. *Am. J. Bot.*
1271 99:1991–2013.
1272 Cardoso D., Pennington R.T., de Queiroz L.P., Boatwright J.S., Van Wyk B.-E.,
1273 Wojciechowski M.F., Lavin M. 2013. Reconstructing the deep-branching
1274 relationships of the papilionoid legumes. *S. Afr. J. Bot.* 89:58–75.
1275 Cascales-Miñana B., Cleal C.J. 2014. The plant fossil record reflects just two great
1276 extinction events. *Terra Nova.* 26:195–200.
1277 Cerling T.E., Harris J.M., Macfadden BJ.., Leakey M.G., Quade J., Eisenmann V.,
1278 Ehleringer J.R. 1997. Global vegetation change through the Miocene/Pliocene
1279 boundary. *Nature.* 389:153–158.
1280 Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.-M., Li F.-
1281 W., Melkonian B., Mavrodiev E.V., Sun W., Fu Y., Yang H., Soltis D.E., Graham
1282 S.W., Soltis P.S., Liu X., Xu X., Wong G.K.-S. 2018. 10KP: a phylodiverse
1283 genome sequencing plan. *GigaScience*. 7:1–9.
1284 Christin P.-A., Spriggs E., Osborne C.P., Strömberg C.A.E., Salamin N., Edwards E.J.
1285 2014. Molecular dating, evolutionary rates, and the age of the grasses. *Syst.*
1286 *Biol.* 63:153–165.

1287 Claramunt S., Cracraft J. 2015. A new time tree reveals Earth history's imprint on the
1288 evolution of modern birds. *Sci. Adv.* 1(11):e1501005.
1289 Cooper A., Penny D. 1997. Mass survival of birds across the Cretaceous-Tertiary
1290 Boundary: molecular evidence. *Science.* 275:1109–1113.
1291 Couvreur T.L.P., Franzke A., Al-Shehbaz I.A., Bakker F.T., Koch M.A., Mummenhoff K.
1292 2010. Molecular phylogenetics, temporal diversification, and principles of
1293 evolution in the mustard family (Brassicaceae). *Mol. Biol. Evol.* 27:55–71.
1294 Couvreur T.L.P., Pirie M.D., Chatrou L.W., Saunders R.M.K., Su Y.C.F., Richardson J.E.,
1295 Erkens R.H.J. 2011a. Early evolutionary history of the flowering plant family
1296 Annonaceae: steady diversification and boreotropical geodispersal. *J. Biogeogr.*
1297 38:664–680.
1298 Couvreur T.L.P., Forest F., Baker W.J. 2011b. Origin and global diversification patterns
1299 of tropical rain forests: inferences from a complete genus-level phylogeny of
1300 palms. *BMC Biol.* 9:44.
1301 Crawley M. 1988. Palaeocene wood from the Republic of Mali. *Bull. Br. Mus. (Nat. Hist.)*
1302 *Geol.* 44:3–14.
1303 Crepet W.L., Herendeen PS. 1992. Papilionoid flowers from the early Eocene of
1304 southeastern North America. In: Herendeen P.S., Dilcher D.L., editors, *Advances*
1305 *in legume systematics part 4: The fossil record*. Richmond, UK: Royal Botanic
1306 Gardens, Kew. p. 43–55.

1307 Criscuolo A., Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a
1308 new software for selection of phylogenetic informative regions from multiple
1309 sequence alignments. *BMC Evol. Biol.* 10:210.
1310 Damas J., Kim J., Farré M., Griffin D.K., Larkin D.M. 2018. Reconstruction of avian
1311 ancestral karyotypes reveals differences in the evolutionary history of macro- and
1312 microchromosomes. *Genome Biol.* 19(1):155.
1313 De Franceschi D., De Ploëg G. 2003. Origine de l'ambre des faciès sparnaciens
1314 (Éocène inférieur) du Bassin de Paris: le bois de l'ambre producteur.
1315 *Geodiversitas.* 25:633–647.
1316 Dehal P., Boore J.L. 2005. Two rounds of whole genome duplication in the ancestral
1317 vertebrate. *PLoS Biol.* 3(10):e314.
1318 de la Estrella M., Forest F., Wieringa J.J., Fougère-Danezan M., Bruneau A. 2017.
1319 Insights on the evolutionary origin of Detarioideae, a clade of ecologically
1320 dominant tropical African trees. *New Phytol.* 214(4):1722–1735.
1321 de la Estrella M., Forest F., Klitgård B., Lewis G.P., Mackinder B.A., de Queiroz L.P.,
1322 Bruneau A. 2018. A new phylogeny-based tribal classification of subfamily
1323 Detarioideae, an early branching clade of florally diverse tropical arborescent
1324 legumes. *Sci. Rep.* 8(1):6884.
1325 Dodsworth S, Chase M.W., Leitch A.R. 2016. Is post-polyploidization diploidization the
1326 key to the evolutionary success of angiosperms? *Bot. J. Linn. Soc.* 180(1):1–5.

1327 dos Reis M., Donoghue P.C.J., Yang Z. 2014. Neither phylogenomic nor
1328 palaeontological data support a Palaeogene origin of placental mammals. *Biol.*
1329 *Lett.* 10:20131003.
1330 Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with
1331 BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29:1969–1973.
1332 Dugas D.V., Hernandez D., Koenen E.J., Schwarz E., Straub S., Hughes C.E., Jansen
1333 R.K., Nageswara-Rao M., Staats M., Trujillo J.T., Hajrah N.H. 2015. Mimosoid
1334 legume plastome evolution: IR expansion, tandem repeat expansions, and
1335 accelerated rate of evolution in clpP. *Sci. Rep.* 5:16958.
1336 Fawcett J.A., Maere S., Van de Peer Y. 2009. Plants with double genomes might have
1337 had a better chance to survive the Cretaceous – Tertiary extinction event. *Proc.*
1338 *Natl. Acad. Sci. USA.* 106:5737–5742.
1339 Feng Y.-J., Blackburn D.C., Liang D., Hillis D.M., Wake D.B., Cannatella D.C., Zhang P.
1340 2017. Phylogenomics reveals rapid, simultaneous diversification of three major
1341 clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. *Proc. Natl.*
1342 *Acad. Sci. USA.* 114(29):E5864–E5870.
1343 Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X.,
1344 Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A.,
1345 Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N.,
1346 Regev A. 2011. Trinity: reconstructing a full-length transcriptome without a
1347 genome from RNA-Seq data. *Nat. Biotechnol.* 29(7):644–652.

1348 Herendeen P.S., Dilcher D.L. 1992. *Advances in legume systematics part 4. The fossil*
1349 *record*. Richmond, UK: Royal Botanic Gardens, Kew..
1350 Herendeen P.S. 1992. The fossil history of the Leguminosae from the Eocene of
1351 southeastern North America. In: Herendeen P.S., Dilcher D.L., editors, *Advances*
1352 *in legume systematics part 4. The fossil record*. Richmond, UK: Royal Botanic
1353 Gardens, Kew. pp. 85–160.
1354 Herendeen P.S., Jacobs B.F. 2000. Fossil legumes from the Middle Eocene (46.0 Ma)
1355 Mahenge Flora of Singida, Tanzania. *Am. J. Bot.* 87:1358–1366.
1356 Huang C.-H., Sun R., Hu Y., Zeng L., Zhang N., Cai L., Zhang Q., Koch M.A., Al-
1357 Shehbaz I., Edger P.P., Pires J.C., Tan D.-Y., Zhong Y., Ma H. 2015. Resolution of
1358 Brassicaceae phylogeny using nuclear genes uncovers nested radiations and
1359 supports convergent morphological evolution. *Mol. Biol. Evol.* 33:394–412.
1360 Huang C.-H., Zang C., Liu M., Hu Y., Gao T., Qi J., Ma H. 2016. Multiple polyploidization
1361 events across Asteraceae with two nested events in the early history revealed by
1362 nuclear phylogenomics. *Mol. Biol. Evol.* 33:2820–2835.
1363 Hull P. 2015. Life in the aftermath of mass extinctions. Curr. Biol. 25:R941–R952.
1364 Jacobs B.F., Herendeen P.S. 2004. Eocene dry climate and woodland vegetation in
1365 tropical Africa reconstructed from fossil leaves from northern Tanzania.
1366 *Palaeogeogr. Palaeocl.* 213:115–123.

1367 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., Faircloth B.C.,
1368 Nabholz B., Howard J.T., Suh A. 2014. Whole genome analyses resolve the early
1369 branches in the tree of life of modern birds. *Science.* 346:1320–1331.
1370 Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of
1371 birds in space and time. *Nature.* 491(7424):444–448.
1372 Jia H., Manchester S.R. 2014. Fossil Leaves and Fruits of *Cercis* L. (Leguminosae)
1373 from the Eocene of Western North America. *Int. J. Plant Sci.* 175:601–612.
1374 Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E.,
1375 Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum
1376 S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral
1377 polyploidy in seed plants and angiosperms. *Nature.* 473:97–100.
1378 Jiao Y., Leebens-Mack J., Ayyampalayam S., Bowers J.E., McKain M.R., McNeal J.,
1379 Rolf M., Ruzicka D.R., Wafula E., Wickett N.J., Wu X., Zhang Y., Wang J., Zhang
1380 Y., Carpenter E.J., Deyholos M.K., Kutchan T.M., Chanderbali A.S., Soltis P.S.,
1381 Stevenson D.W., McCombie R., Pires J.C., Wong G.K.-S., Soltis D.E.,
1382 DePamphilis C.W. 2012. A genome triplication associated with early
1383 diversification of the core eudicots. *Genome Biol.* 13(1):R3.
1384 Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7:
1385 improvements in performance and usability. *Mol. Biol. Evol.* 30(4):772–780.
1386 Keller G. 2014. Deccan volcanism, the Chicxulub impact, and the end-Cretaceous mass
1387 extinction: Coincidence? Cause and effect?, in Keller G., and Kerr A.C., eds.,

1388 Volcanism, Impacts, and Mass Extinctions: Causes and Effects. *Geol. S. Am. S.*
1389 505:57–89.
1390 Ksepka D.T., Phillips M.J. 2015. Avian diversification patterns across the K-Pg
1391 boundary: influence of calibrations, datasets, and model misspecification. *Ann.*
1392 *Mo. Bot. Gard.* 100(4):300–328.
1393 Landis J.B., Soltis D.E., Li Z., Marx H.E., Barker M.S., Tank D.C., Soltis P.S. 2018.
1394 Impact of whole-genome duplication events on diversification rates in
1395 angiosperms. *Am. J. Bot.* 105(3):348–363.
1396 Lanfear R., Ho S.Y.W., Davies T.J., Moles A.T., Aarssen L., Swenson N.G., Warman L.,
1397 Zanne A.E., Allen A.P. 2013. Taller plants have lower rates of molecular evolution.
1398 *Nat. Commun.* 4:1879.
1399 Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2017. PartitionFinder 2:
1400 new methods for selecting partitioned models of evolution for molecular and
1401 morphological phylogenetic analyses. *Mol. Biol. Evol.* 34(3):772–773.
1402 Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities
1403 in the amino-acid replacement process. *Mol. Biol. Evol.* 21:1095–1109.
1404 Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI: phylogenetic
1405 reconstruction with infinite mixtures of profiles in a parallel environment. *Syst.*
1406 *Biol.* 62:611–615.

1407 Lassmann T., Hayashizaki Y., Daub C.O. 2009. TagDust—a program to eliminate
1408 artifacts from next generation sequencing data. *Bioinformatics.* 25(21):2839–
1409 2840.
1410 Lavin M., Wojciechowski M.F., Gasson P., Hughes C., Wheeler E. 2003. Phylogeny of
1411 robinioid legumes (Fabaceae) revisited: *Coursetia* and *Gliricidia* recircumscribed,
1412 and a biogeographical appraisal of the Caribbean endemics. *Syst. Bot.* 28:387–
1413 409.
1414 Lavin M., Herendeen P.S., Wojciechowski M.F. 2005. Evolutionary rates analysis of
1415 Leguminosae implicates a rapid diversification of lineages during the Tertiary.
1416 *Syst. Biol.* 54:575–594.
1417 Le Q., Dang C., Gascuel O. 2012. Modeling protein evolution with several amino acid
1418 replacement matrices depending on site rates. *Mol. Biol. Evol.* 29:2921–2936.
1419 Lehtonen S., Silvestro D., Karger D.N., Scotese C., Tuomisto H., Kessler M., Peña C.,
1420 Wahlberg N., Antonelli A. 2017. Environmentally driven extinction and
1421 opportunistic origination explain fern diversification patterns. *Sci. Rep.* 7(1):4831.
1422 Levin D.A., Soltis D.E. 2018. Factors promoting polyploid persistence and diversification
1423 and limiting diploid speciation during the K–Pg interlude. *Curr. Opin. Plant Biol.*
1424 42:1–7.
1425 Lin Y., Wong W.O., Shi G., Shen S., Li Z. 2015. Bilobate leaves of *Bauhinia*
1426 (Leguminosae, Caesalpinioideae, Cercideae) from the middle Miocene of Fujian

1427 Province, southeastern China and their biogeographic implications. *BMC Evol.*
1428 *Biol.* 15:252.
1429 Lohaus R., Van de Peer Y. 2016. Of dups and dinos: evolution at the K/Pg boundary.
1430 *Curr. Opin. Plant Biol.* 30:62–69.
1431 LPWG (Legume Phylogeny Working Group). 2013. Legume phylogeny and
1432 classification in the 21st century: progress, prospects and lessons for other
1433 species-rich clades. *Taxon.* 62:217–248.
1434 LPWG (Legume Phylogeny Working Group). 2017. A new subfamily classification of the
1435 Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon.*
1436 66:44–77.
1437 Maddison W.P. 1997. Gene trees in species trees. *Syst. Biol.* 46(3):523–536.
1438 Magallón S., Gómez-Acevedo S., Sánchez-Reyes L.L., Hernández-Hernández T. 2015.
1439 A metacalibrated time-tree documents the early rise of flowering plant
1440 phylogenetic diversity. *New Phytol.* 207:437–453.
1441 Maurin O., Davies T.J., Burrows J.E., Daru B.H., Yessoufou K., Muasya A.M., Bank M.,
1442 Bond W.J. 2014. Savanna fire and the origins of the ‘underground forests’ of
1443 Africa. *New Phytol.* 204(1):201–214.
1444 McElwain J.C., Punyasena S.W. 2007. Mass extinction events and the plant fossil
1445 record. *Trends Ecol. Evol.* 22:548–557.
1446 McKey D. 1994. Legumes and nitrogen: The evolutionary ecology of a nitrogen-
1447 demanding lifestyle. In: Sprent J.I., McKey D., editors. *Advances in legume*

1448 *systematics part 5. The nitrogen factor.* Richmond, UK: Royal Botanic Gardens,
1449 Kew. p. 211–228.

1450 Mendes F.K., Hahn M.W. 2016. GVajda V., Bercovici A. 2014. The global vegetation
1451 pattern across the Cretaceous–Paleogene mass extinction interval: A template
1452 for other extinction events. Global and Planet. Change. 122:29-49.ene tree
1453 discordance causes apparent substitution rate variation. Syst. Biol. 65(4):711–
1454 721.

1455 Meredith R.W., Janecka J.E., Gatesy J., Ryder O.A., Fisher C.A., Teeling E.C., Goodbla
1456 A., Eizirik E., Simão TL., Stadler T., Rabosky D.L. 2011. Impacts of the
1457 Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification.
1458 *Science.* 334(6055):521–524.

1459 Miller J.T., Murphy D.J., Ho S.Y.W., Cantrill D.J., Seigler D. 2013. Comparative dating of
1460 *Acacia*: combining fossils and multiple phylogenies to infer ages of clades with
1461 poor fossil records. *Aust. J. Bot.* 61:436–445.

1462 Mirarab S., Reaz R., Bayzid M.dS., Zimmermann T., Swenson M.S., Warnow T. 2014.
1463 ASTRAL: genome-scale coalescent-based species tree estimation.
1464 *Bioinformatics.* 30:i541–i548.

1465 Moore M.J., Soltis P.S., Bell C.D., Burleigh J.G., Soltis D.E. 2010. Phylogenetic analysis
1466 of 83 plastid genes further resolves the early diversification of eudicots. *Proc.*
1467 *Natl. Acad. Sci. USA.* 107:4623–4628.

1468 Moore A.J., De Vos J.M., Hancock L.P., Goolsby E., Edwards E.J. 2017. Targeted
1469 enrichment of large gene families for phylogenetic inference: phylogeny and
1470 molecular evolution of photosynthesis genes in the portullugo clade
1471 (Caryophyllales). *Syst. Biol.* 67:367–383.
1472 Mudge J., Cannon S.B., Kalo P., Oldroyd G.E.D., Roe B.A., Town C.D and Young N.D.
1473 2005. Highly syntenic regions in the genomes of soybean, *Medicago truncatula*,
1474 and *Arabidopsis thaliana*. *BMC Plant Biol.* 5:15.
1475 Near T.J., Meylan P.A., Shaffer H.B., Meyer A.E.A. 2005. Assessing concordance of
1476 fossil calibration points in molecular clock studies: An Example Using Turtles.
1477 *Am. Nat.* 165(2):137–146.
1478 Nicholls D.J., Johnson K.R. 2008. *Plants and the K-T boundary*. Cambridge, UK:
1479 Cambridge University Press.
1480 O'leary M.A., Bloch J.I., Flynn J.J., Gaudin T.J., Giallombardo A., Giannini N.P.,
1481 Goldberg S.L., Kraatz B.P., Luo Z.X., Meng J., Ni X. 2013. The placental mammal
1482 ancestor and the post–K-Pg radiation of placentals. *Science* 339(6120):662–667.
1483 Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol.*
1484 *Evol.* 5(5):568–583.
1485 Pan A.D., Jacobs B.F., Dransfield J., Baker WJ. 2006. The fossil history of palms
1486 (Arecaceae) in Africa and new records from the Late Oligocene (28–27 Mya) of
1487 north-western Ethiopia. *Bot. J. Linn. Soc.* 151(1):69–81.

1488 Paterson AH.., Wendel J.F., Gundlach H., Guo H., Jenkins J., Jin D., Llewellyn D.,
1489 Showmaker K.C., Shu S., Udall J., Yoo .MJ. 2012. Repeated polyploidization of
1490 *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature.*
1491 492(7429):423–427.
1492 Phillips M.J. 2015. Geomolecular dating and the origin of placental mammals. *Syst.*
1493 *Biol.* 65(3):546–557.
1494 Phillips M.J., Fruciano C. 2018. The soft explosive model of placental mammal
1495 evolution. *BMC Evol. Biol.* 18:104.
1496 Poinar Jr G.O. 1991. *Hymenaea protera* sp. n. (Leguminoseae, Caesalpininoideae)
1497 from Dominican amber has African affinities. *Experientia* 47:1075–1082.
1498 Poinar Jr G.O., Brown A.E. 2002. *Hymenaea mexicana* sp. nov. (Leguminosae:
1499 Caesalpinioideae) from Mexican amber indicates Old World connections. *Bot. J.*
1500 *Linn. Soc.* 139:125–132.
1501 Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene
1502 trees with species tree in *Drosophila*: evidence for incomplete lineage sorting.
1503 *PLoS Genet.* 2(10):e173.
1504 Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon
1505 A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-
1506 generation DNA sequencing. *Nature.* 526:569–573.

1507 Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior
1508 summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67(5):901–
1509 904.
1510 Ranwez V., Harispe S., Delsuc F., Douzery E.J.P. 2011. MACSE: Multiple Alignment of
1511 Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS ONE*
1512 6(9): e22594.
1513 Rokas A., Carroll S.B. 2006. Bushes in the Tree of Life. *PLoS Biol.* 4:e352.
1514 Ruprecht C., Lohaus R., Vanneste K., Mutwil M., Nikoloski Z., Van de Peer Y., Persson
1515 S. 2017. Revisiting ancestral polyploidy in plants. *Sci. Adv.* 3(7):e1603195.
1516 Sacerdot C., Louis A., Bon C., Berthelot C., Crollius H.R. 2018. Chromosome evolution
1517 at the origin of the ancestral vertebrate genome. *Genome Biol.* 19(1):166.
1518 Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong
1519 phylogenetic signals. *Nature.* 497(7449):327–331.
1520 Sanderson M.J. 2015. Back to the past: a new take on the timing of flowering plant
1521 diversification. *New Phytol.* 207(2):257–259.
1522 Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support
1523 from quartet frequencies. *Mol. Biol. Evol.* 33:1654–1668.
1524 Sayyari E., Whitfield J. B., Mirarab S. 2018. DiscoVista: Interpretable visualizations of
1525 gene tree discordance. *Mol. Phylogenet. Evol.* 122:110–115.
1526 Scannell D.R., Frank A.C., Conant G.C., Byrne K.P., Woolfit M., Wolfe K.H. 2007.
1527 Independent sorting-out of thousands of duplicated gene pairs in two yeast

1528 species descended from a whole-genome duplication. *Proc. Natl. Acad. Sci.*
1529 *USA.* 104(20):8397–8402.
1530 Schley R.J., de la Estrella M., Pérez-Escobar O.A., Bruneau A., Barraclough T., Forest
1531 F., Klitgård B. 2018. Is Amazonia a 'museum' for Neotropical trees? The evolution
1532 of the Brownea clade (Detarioideae, Leguminosae). *Mol. Phylogenet. Evol.*
1533 126:279–292.
1534 Schmieder R., Edwards R. 2011. Quality control and preprocessing of metagenomic
1535 datasets. *Bioinformatics.* 27(6):863–864.
1536 Schranz M.E., Mohammadin S., Edger P.P. 2012. Ancient whole genome duplications,
1537 novelty and diversification: the WGD Radiation Lag-Time Model. *Curr. Opin.*
1538 *Plant Biol.* 15:147–153.
1539 Schwarz E.N., Ruhlman T.A., Weng M.-L., Khiyami M.A., Sabir J.S.M., Hajarah N.H.,
1540 Alharbi N.S., Rabah S.O., Jansen R.K. 2017. Plastome-wide nucleotide
1541 substitution rates reveal accelerated rates in Papilionoideae and correclations
1542 with genome features across legume subfamilies. *J. Mol. Evol.* 84:187–203.
1543 Schuettpelz E., Pryer K.M. 2009. Evidence for a Cenozoic radiation of ferns in an
1544 angiosperm-dominated canopy. *Proc. Natl. Acad. Sci. USA.* 106(27):11200–
1545 11205.
1546 Senchina D.S., Alvarez I., Cronn R.C., Liu B., Rong J., Noyes R.D., Paterson A.H., Wing
1547 R.A., Wilkins T.A., Wendel J.F. 2003. Rate variation among nuclear genes and
1548 the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* 20(4):633–643.

1549 Silvestro D., Cascales-Miñana B., Bacon C.D., Antonelli A. 2015. Revisiting the origin
1550 and diversification of vascular plants through a comprehensive Bayesian analysis
1551 of the fossil record. *New Phytol.* 207(2):425–436.
1552 Simon M.F., Grether R., de Queiroz L.P., Skema C., Pennington R.T., Hughes C.E.
1553 2009. Recent assembly of the Cerrado, a Neotropical plant diversity hotspot, by
1554 in situ evolution of adaptations to fire. *Proc. Natl. Acad. Sci. USA.* 106:20359–
1555 20364.
1556 Smith S.A., Donoghue M.J. 2008. Rates of molecular evolution are linked to life history
1557 in flowering plants. *Science.* 322:86–89.
1558 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets
1559 reveals conflict, concordance, and gene duplications with examples from animals
1560 and plants. *BMC Evol. Biol.* 15:150.
1561 Smith S.A., Brown J.W., Yang Y., Bruenn R., Drummond C.P., Brockington S.F., Walker
1562 J.F., Last N., Douglas N.A., Moore M.J. 2018a. Disparity, diversity, and
1563 duplications in the Caryophyllales. *New Phytol.* 217(2):836–854.
1564 Smith S.A., Brown J.W., Walker J.F. 2018b. So many genes, so little time: a practical
1565 approach to divergence-time estimation in the genomic era. *PLoS One.*
1566 13(5):e0197433.
1567 Soltis D.E., Visger C.J., Marchant D.B., Soltis P.S. 2016. Polyploidy: pitfalls and paths to
1568 a paradigm. *Am. J. Bot.* 103:1146–1166.

1569 Springer M.S., Meredith R.W., Teeling E.C., Murphy W.J. 2013. Technical comment on
1570 "The placental mammal ancestor and the post–K-Pg radiation of placentals".
1571 *Science.* 341:613.
1572 Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-
1573 analysis of large phylogenies. *Bioinformatics.* 30:1312–1313.
1574 Suh A., Smeds L., Ellegren H. 2015. The dynamics of incomplete lineage sorting across
1575 the ancient adaptive radiation of neoavian birds. *PLoS Biol*. 13(8):e1002224.
1576 Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root
1577 of Neoaves. *Zool. Scr.* 45:50–62.
1578 Sukumaran J., Holder M.T. 2010. DendroPy: A Python library for phylogenetic
1579 computing. *Bioinformatics.* 26:1569–1571.
1580 Tank D.C., Eastman J.M., Pennell M.W., Soltis P.S., Soltis D.E., Hinchliff C.E., Brown
1581 J.W., Sessa E.B., Harmon L.J. 2015. Nested radiations and the pulse of
1582 angiosperm diversification: increased diversification rates often follow whole
1583 genome duplications. *New Phytol.* 207:454–467.
1584 Teeling E.C., Hedges S.B. 2013. Making the impossible possible: rooting the tree of
1585 placental mammals. *Mol. Biol. Evol.* 30:1999–2000.
1586 Tuskan G.A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., Putnam N.,
1587 Ralph S., Rombauts S., Salamov A., Schein J. 2006. The genome of black
1588 cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science.* 313(5793):1596–1604.

1589 Vajda V., Raine J.I., Hollis C.J. 2001. Indication of global deforestation at the
1590 Creataceous-Tertiary boundary by New Zealand fern spike. *Science.* 294:1700–
1591 1702.
1592 Vajda V., Bercovici A. 2014. The global vegetation pattern across the Cretaceous–
1593 Paleogene mass extinction interval: A template for other extinction events. *Global*
1594 *and Planet. Change.* 122:29–49.
1595 Vanneste K., Baele G., Maere S., Van de Peer Y. 2014. Analysis of 41 plant genomes
1596 supports a wave of successful genome duplications in association with the
1597 Cretaceous–Paleogene boundary. *Genome Res.* 24(8):1334–1347.
1598 Wang H., Moore M.J., Soltis P.S., Bell C.D., Brockington S.F., Alexandre R., Davis C.C.,
1599 Latvis M., Manchester S.R., Soltis D.E. 2009. Rosid radiation and the rapid rise
1600 of angiosperm-dominated forests. *Proc. Natl. Acad. Sci. USA.* 106(10):3853–
1601 3858.
1602 Wang W., Ortiz R.D.C., Jacques F.M.B., Xiang X.-G., Li H.-L., Lin L., Li R.-Q., Liu Y.,
1603 Soltis P.S., Soltis D.E., Chen Z.-D. 2012. Menispermaceae and the diversification
1604 of tropical rainforests near the Cretaceous–Paleogene boundary. *New Phytol.*
1605 195:470–478.
1606 Wang Q., Song Z., Chen Y., Shen S., Li Z. 2014. Leaves and fruits of *Bauhinia*
1607 (Leguminosae, Caesalpinioideae, Cercideae) from the Oligocene Ningming
1608 Formation of Guangxi, South China and their biogeographic implications. *BMC*
1609 *Evol. Biol.* 14:88.

1610 Wang Y.-H., Wicke S., Wang H., Jin J.-J., Chen S.-Y., Zhang S.-D., Li D.-Z., Yi T.-S.
1611 2018. Plastid genome evolution in the early-diverging legume subfamily
1612 Cercidoideae (Fabaceae). *Front. Plant Sci.* 9:138.
1613 Wendel J.F. 2015. The wondrous cycles of polyploidy in plants. *Am. J. Bot.* 102:1753–
1614 1756.
1615 Whitfield J., Cameron S.A., Huson D., Steel M. 2008. Filtered Z-closure supernetworks
1616 for extracting and visualizing recurrent signal from incongruent gene trees. *Syst.*
1617 *Biol.* 57:939–947.
1618 Wilf P., Johnson K.R. 2004. Land plant extinction at the end of the Cretaceous: a
1619 quantitative analysis of the North Dakota megafloral record. *Paleobiology*.
1620 30:347–368.
1621 Wing S.L., Herrera F., Jaramillo C.A., Gómez-Navarro C., Wilf P., Labandeira C.C. 2009.
1622 Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest
1623 record of Neotropical rainforest. *Proc. Natl. Acad. Sci. USA.* 106:18627–18632.
1624 Wojciechowski M.F., Lavin M., Sanderson M.J. 2004. A phylogeny of legumes
1625 (Leguminosae) based on analysis of the plastid matK gene resolves many well-
1626 supported subclades within the family. *Am. J. Bot.* 91(11):1846–1862.
1627 Wong M.L.L., Vaillancourt R.E., Freeman J.S., Hudson C.J., Bakker FT.., Cannon C.H.,
1628 Ratnam W. 2017. Novel insights into karyotype evolution and whole genome
1629 duplications in legumes. *BioRxiv.* 099044.

1630 Xing Y.X., Onstein R.E., Carter R.J., Stadler T., Linder H.P. 2014. Fossils and a large
1631 molecular phylogeny show that the evolution of species richness, generic
1632 diversity and turnover rates are disconnected. *Evolution.* 68:2821–2832.
1633 Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum
1634 likelihood. *Mol. Biol. Evol.* 24:1586–1591.
1635 Yang Y., Smith S.A. 2014. Orthology inference in nonmodel nrganisms using
1636 transcriptomes and low-loverage genomes: improving accuracy and matrix
1637 occupancy for phylogenomics. *Mol. Biol. Evol.* 31:3081–3092.
1638 Yang Y., Moore M.J., Brockington S.F., Soltis D.E., Wong G.K., Carpenter E.J., Zhang
1639 Y., Chen L., Yan Z., Xie Y., Sage R.F. 2015. Dissecting molecular evolution in the
1640 highly diverse plant clade Caryophyllales using transcriptome sequencing. *Mol.*
1641 *Biol. Evol.* 32(8):2001–2014.
1642 Yang Y., Moore M.J., Brockington S.F., Mikenas J., Olivieri J., Walker J.F., Smith S.A.
1643 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent
1644 polyploidy events within Caryophyllales, including two allopolyploidy events. *New*
1645 *Phytol.* 217:855–870.
1646 Zhang J., Kobert K., Flouri T., Stamatakis A. 2014. PEAR: a fast and accurate Illumina
1647 Paired-End reAd mergeR. *Bioinformatics.* 30(5):614–620.
1648 Zeng L., Zhang N., Zhang Q., Endress P.K., Huang J and Ma H. 2017. Resolution of
1649 deep eudicot phylogeny and their temporal diversification using nuclear genes
1650 from transcriptomic and genomic datasets. *New Phytol.* 214:1338–1354.

Table 1. Fossil calibrations used in the divergence time analyses.

Calibration[a]	Definition	Fossil	Age (Ma)
eudicots			
26	CG eudicots	Tricolpate pollen; England and Gabon[b]	126[c]
27	CG Ranunculales	*Teixeiraea lusitanica* – flower; Portugal[b]	113
38	CG Pentapetalae	Pentamerous flower with distinct calyx and corolla; USA[b]	100
48	SG Ericales	*Pentapetalum trifasciculandricus* – flowers; USA[b]	89.8
94	SG Myrtaceae	"Flower number 3" from the Table Nunatak Formation, Antarctica[b]	83.6
105	SG Brassicales	*Dressiantha bicarpelata* – flowers; USA[b]	89.8
112	CG Rosaceae	*Prunus wutuensis* – fruits; China[b]	49.4
116	SG Cannabaceae	*Aphananthe cretacea* and *Gironniera gonnensis* – fruits; Germany[b]	66
122	SG Juglandaceae	*Polyptera manningi* – fruits; USA[b]	64.4
133	SG *Populus*	*Populus wilmattae* – leaves, infructescences and fruits; USA[b]	37.8
X14	SG Fagales	*Protofagacea allonensis* – flowers; USA[d]	83.6
legumes			
A	SG Leguminosae	*Paracacioxylon frenguellii* – wood with vestured pits; Argentina[e]	63.5
C	SG *Cercis*	*Cercis parvifolia* – leaves and C. herbmeyeri – fruits; USA[f]	36
C[g]	SG *Bauhinia*	cf. *Bauhinia* – simple leaf with bilobed lamina; Tanzania[h]	46
F	SG Resin-producing clade	*Hymenaea mexicana* – vegetative and floral remains in amber; Mexico[i]	22.5
G	SG Detarioideae	*Aulacoxylon sparnacense* – wood and	53

		amber; France[j]	
G[g]	SG Resin-producing clade	same as G	53
H[g]	CG Amherstieae	*Aphanocalyx singidaensis* – bifoliolate leaves; Tanzania[k]	46
I2	SG *Styphnolobium*/*Cladrastis*	*Styphnolobium* and *Cladrastis* – leaves and fruits; USA[l]	37.8
M2	SG Robinioid clade	*Robinia zirkelii* – wood; USA[m]	33.9
Q	SG Acacieae/Ingeae	Flattened polyads with 16 pollen grains; Brazil, Colombia, Cameroon and Egypt[n]	33.9
Q2	SG *Acacia* s.s.	Polyads with pseudocolpi; Australia[o]	23
Z	SG Caesalpinioideae	Bipinnate leaves; Colombia[p]	58

CG = Crown group; SG = Stem group; Ma = Million years ago.

[a] numbers 26, 27, 38, 48, 94, 105, 112, 116, 122 and 133 refer to calibrations from Magallón et al. (2015) as listed in their Supplementary Information Methods S1; letters A, D, F, G, I2, M2 and Q refer to calibrations from Bruneau et al. (2008) and/or Simon et al. (2009)

[b] Magallón et al. (2015) and references therein

[c] prior set as normal with standard deviation of 1.0, and truncated between minimum and maximum bounds of 113 and 136 Ma, respectively

[d] Xing et al. (2014) and reference therein

[e] Brea et al. (2008)

[f] Jia & Manchester (2014)

[g] alternative prior 1 as used in FLC analysis with 8 local clocks

[h] Jacobs & Herendeen (2004)

[i] Poinar & Brown (2002)

[j] De Franceschi & De Ploëg (2003)

[k] Herendeen & Jacobs (2000)

[l] Herendeen (1992)

[m] Lavin et al. (2003) and references therein

[n] Simon et al. (2009): Supplementary Information and references therein

[o] Miller et al. (2013)

[p] Wing et al. (2009)

1672 **Figure captions**

1673 FIGURE 1. Diversity, ecology and economic importance of legumes. The family is
1674 subdivided into subfamilies (A) Cercidoideae (*Bauhinia madagascariensis*), (B)
1675 Detarioideae (*Macrolobium* sp.), (C) Duparquetioideae (*Duparquetia orchidacea*), (D)
1676 Dialioideae (*Baudouinia* sp.), (E) Caesalpinioideae (*Mimosa pectinatipinna*) and (F)
1677 Papilionoideae (*Medicago marina*). While the family has a very diverse floral
1678 morphology, the fruit (G), which comes in many shapes and is most often referred to as
1679 'pod' or 'legume', is the defining feature of the family (fruit shown is of *Brodriguesia*
1680 *santosii*). A large fraction of legume species is known to fix atmospheric nitrogen
1681 symbiotically with 'rhizobia', bacteria that are incorporated in root nodules, for example
1682 in *Lupinus nubigenus* (H). Economically, the family is the second most important of
1683 flowering plants after the grasses, with a wide array of uses, including timber,
1684 ornamentals, fodder crops, and notably, pulse crops such as peanuts (Arachis), beans
1685 (Phaseolus), chickpeas (Cicer) and lentils (Lens) (I). Also ecologically, legumes are
1686 extremely diverse and important, occurring and often dominating globally across
1687 disparate ecosystems, including wet tropical forest, for example *Albizia grandibracteata*
1688 in the East African Albertine Rift (J), savannas, seasonally dry tropical forests, and semi-
1689 arid thorn-scrub, for example *Mimosa delicatula* in Madagascar (K) and temperate
1690 woodlands and grasslands, for example *Vicia sylvatica* in the European Alps (L). --
1691 Photos A, B, D, F, J, K, L by Erik Koenen, C by Jan Wieringa and E, G, H, I by Colin
1692 Hughes.

1693 FIGURE 2. Phylogeny of legumes based on Bayesian analyses of 72 protein coding
1694 chloroplast genes under the CATGTR model in Phylobayes. (A) majority-rule consensus
1695 tree of the amino acid alignment, showing only the Fabales portion of the tree, outgroup
1696 taxa pruned, (B) complete tree including outgroup taxa, (C) Root-to-tip lengths
1697 measured from the legume crown node in amino acid substitutions per site and (D) D_n/
1698 D_s ratios for background, the 50 Kb inversion clade (excluding the vicioid clade) and
1699 vicioid clade tree partitions. Majority-rule consensus trees for both the amino acid and
1700 nucleotide alignments with tip labels for all taxa and support values indicated are
1701 included in supporting information (Figs S1-2).

1702 FIGURE 3. Congruent relationships among subfamilies when using different types of
1703 phylogenetic analysis, and phylogenetic locations of WGDs, as inferred from nuclear
1704 gene data. Support is indicated with coloured symbols on nodes for simplicity of
1705 presentation, as indicated in the legends; figures annotated with actual support values
1706 are included as Figures S5-7. (A) ML phylogeny estimated with RAxML under the LG4X
1707 model from a concatenated alignment of 1,103 nuclear orthologs. Support indicated
1708 represents Internode Certainty All (ICA) values, estimated with RAxML from 80%
1709 bootstrap threshold consensus gene trees of the same 1,103 orthologs. For the first four
1710 divergences in the legume family, pie charts indicate the proportions of gene trees
1711 supporting the relationship shown (blue), supporting the most prevalent conflicting
1712 bipartition (yellow), supporting other conflicting bipartitions (red) and genes without
1713 phylogenetic signal, i.e. no bootstrap support (gray). Numbers of bipartitions for the pie
1714 charts are derived from phyparts analyses with a 50% bootstrap support filter. Labelled
1715 nodes A-H are analysed in more detail in Figure 5. (B) Bayesian gene jackknifing
1716 majority-rule consensus tree of concatenated alignments of c. 220 genes per replicate,
1717 support indicated represents posterior probability averaged over 25 replicates for 500
1718 posterior trees each (in total 12,500 posterior trees). (C) Phylogeny estimated under the
1719 multi-species coalescent with ASTRAL from gene trees, support indicated represents
1720 local posterior probability. Pie charts show relative quartet support for the first (blue) and
1721 the two (yellow and red) alternative quartets. (D) Gene duplications in 8,038 homolog
1722 clusters mapped onto the ML species tree topology. The size of the circles on nodes is

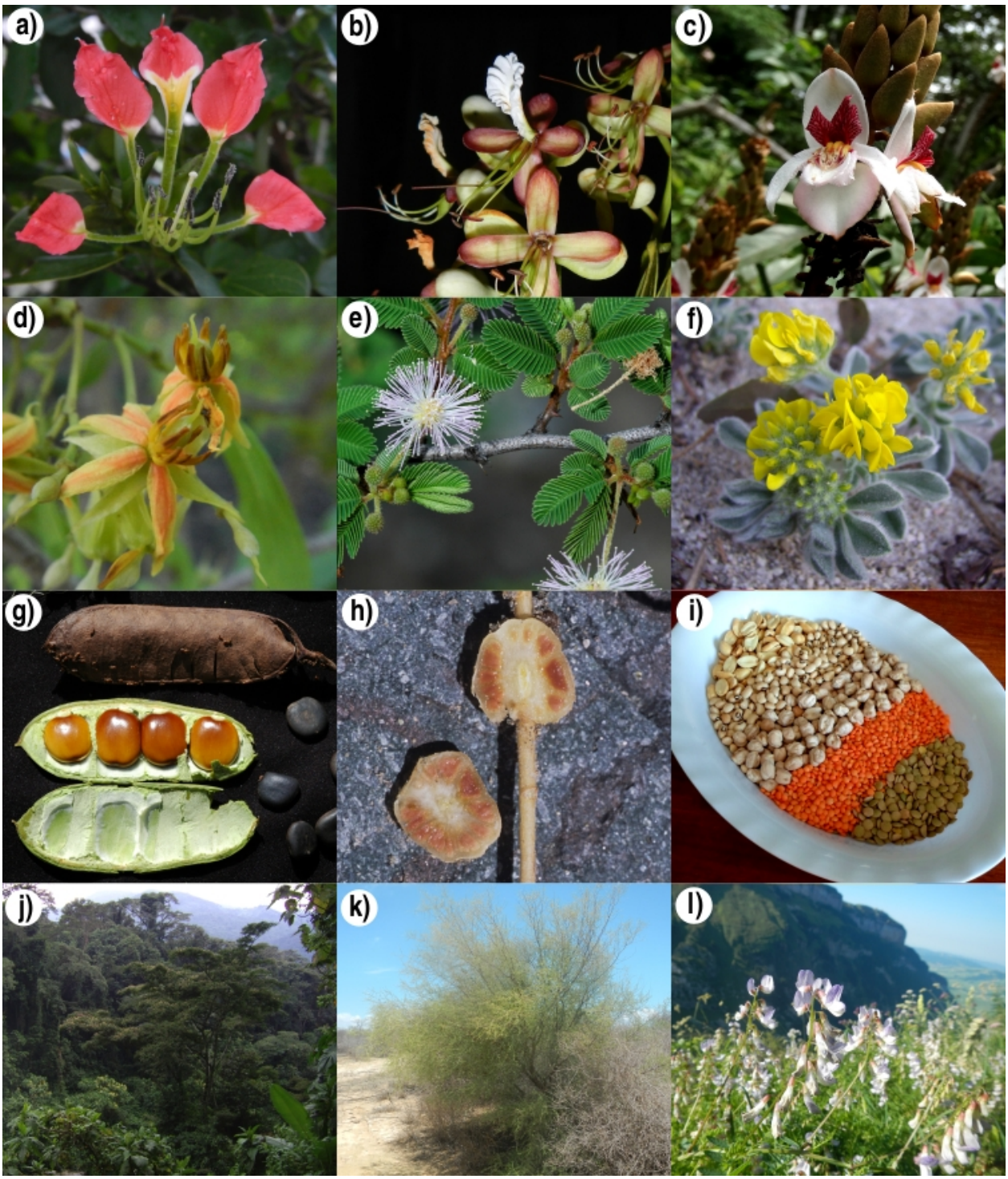
proportional to the number of gene duplications inferred. For hypothesized WGD events, the number of gene duplications without/with bootstrap filter is indicated. See Figure S9 for the number of gene duplications for all nodes in the phylogeny.

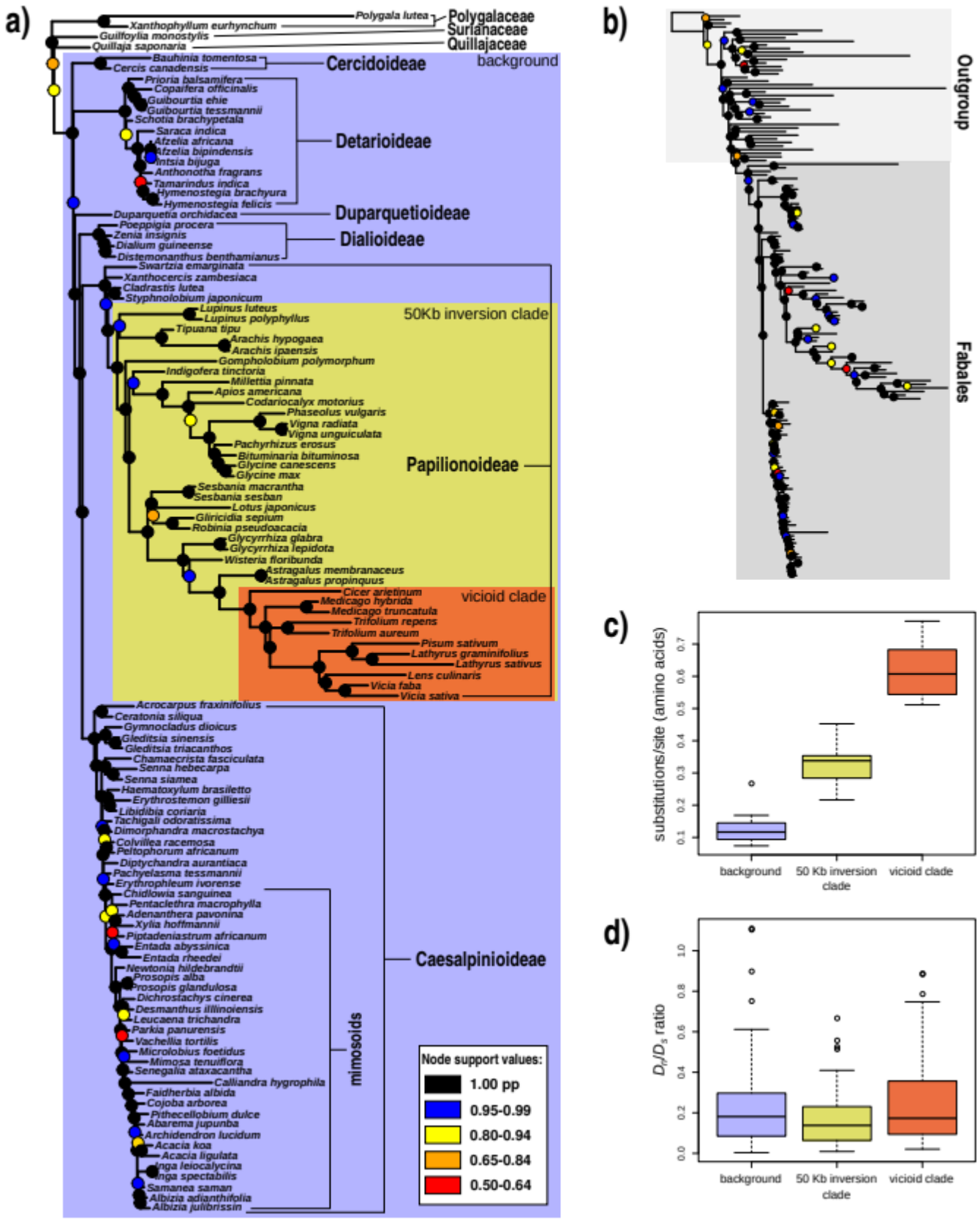
1726 FIGURE 4. A filtered supernetwork shows tangles of gene tree relationships at the bases
1727 of the legumes, and subfamilies Detarioideae and Papilionoideae, that correspond to
1728 WGDs. The filtered supernetwork was inferred from the 1,103 1-to-1 ortholog gene tree
1729 set, only bipartitions that received more than 80% bootstrap support in gene tree
1730 analyses were included. Edge lengths and colours are by their weight, a measure of
1731 prevalence of the bipartition that the edge represents among the gene trees. Ellipses
1732 with dashed outline indicate increased complexity at putative locations of WGDs.

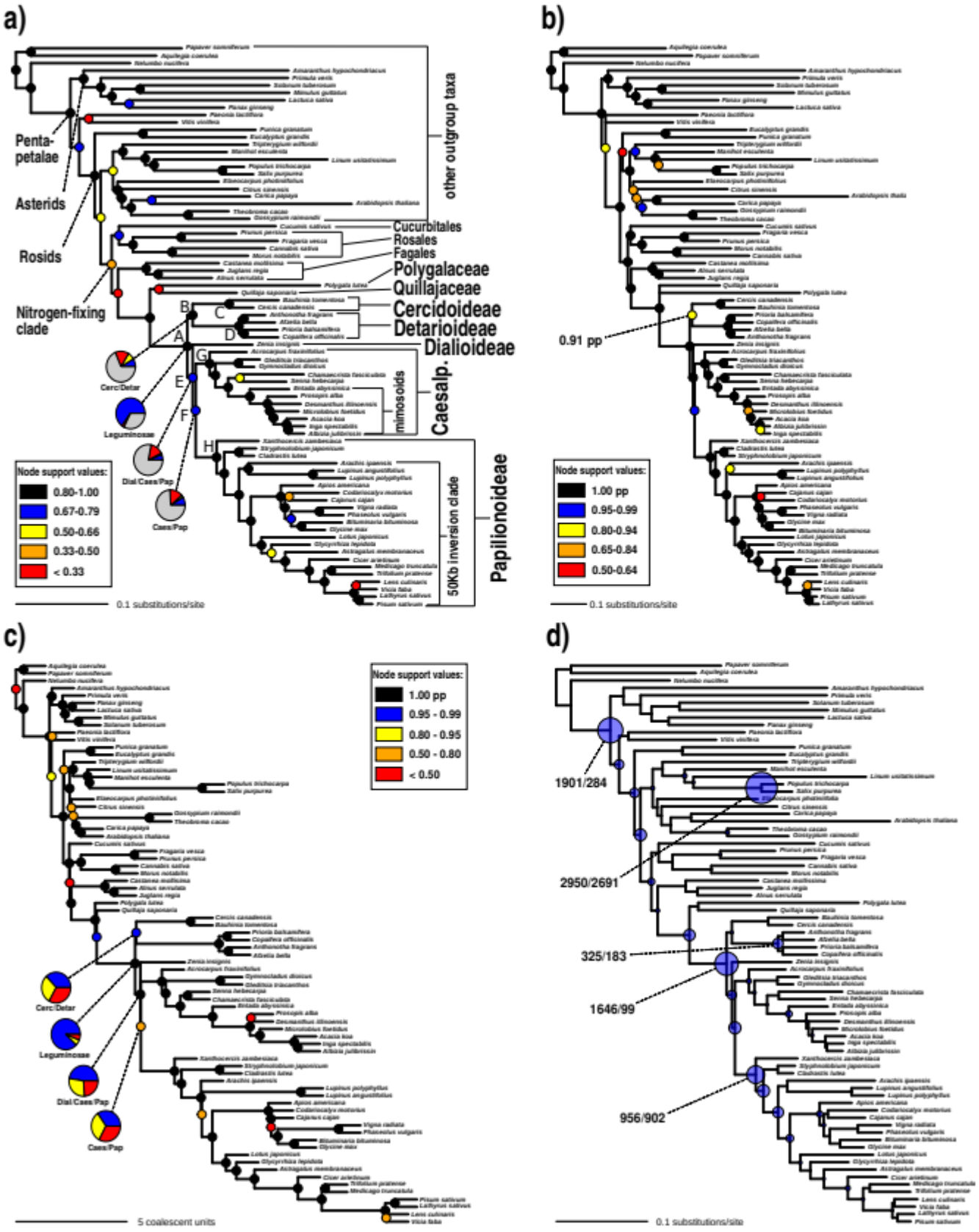
1733 FIGURE 5. Leguminosae and its subfamilies are each supported by a large fraction of
1734 gene trees, in contrast to relationships among the subfamilies. (A) Prevalence of
1735 bipartitions that are equivalent to nodes A-H (see Fig. 3A), among the 3,473 gene trees
1736 inferred from the RT homolog clusters (including 1-to-1 orthologs) in which all five
1737 subfamilies and the outgroup were included. Numbers of bipartitions are shown as
1738 counted from the best-scoring ML gene trees as well as taking only bipartitions with
1739 more than 50 and 80% bootstrap support into account, as indicated in the legend. (B-D)
1740 Prevalence of bipartitions for nodes B, E and F plotted next to the most common
1741 alternative bipartitions. The locations of the stars in the illustrations indicate the
1742 internodes of the phylogeny that are equivalent to the bipartitions for which counts are
1743 plotted below, as counted from the ML estimates and for bipartitions with at least 50 or
1744 80% bootstrap support. Colors of the stars correspond to the colors of the bars in the
1745 barplots.

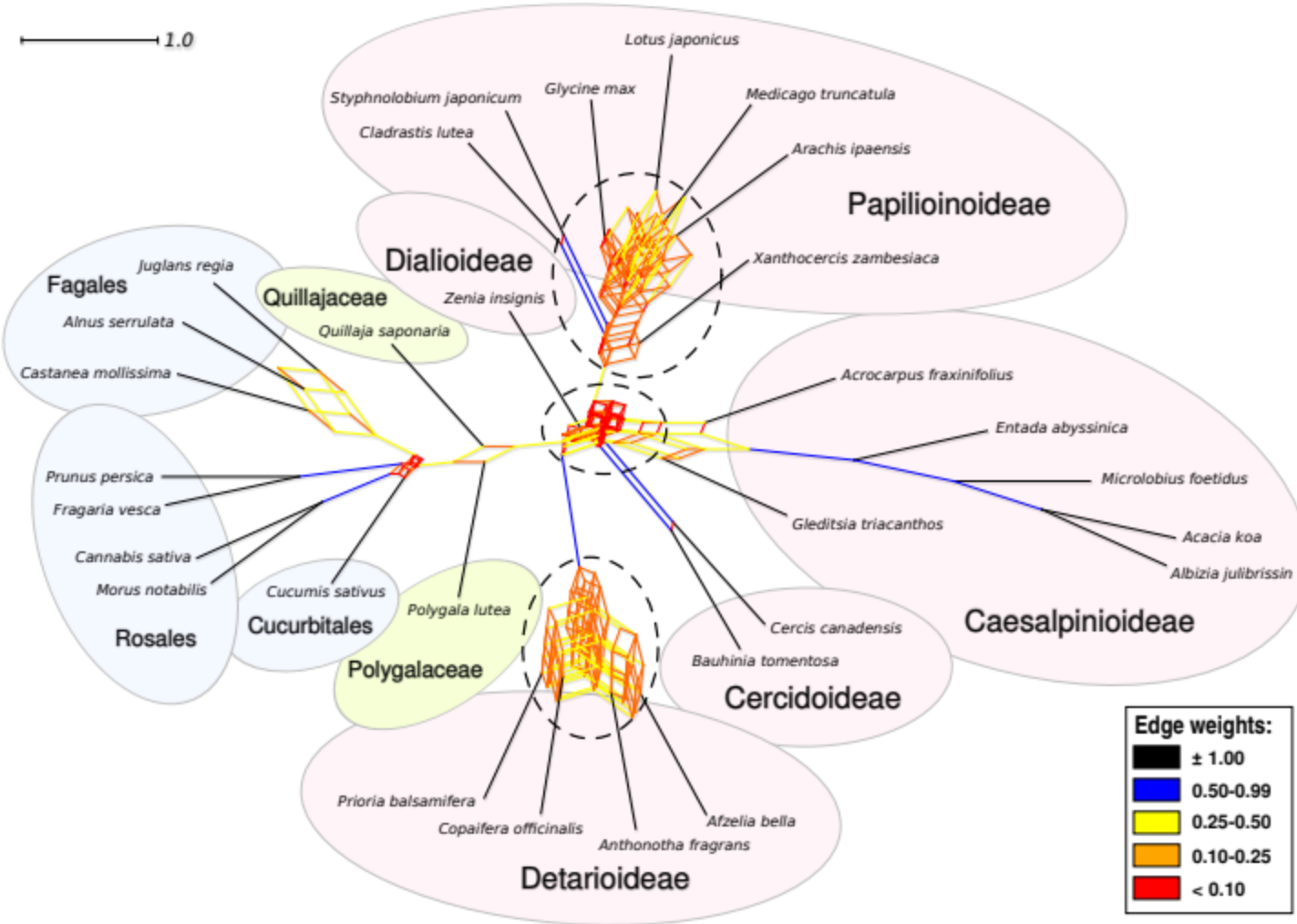
1746 FIGURE 6. The origin of the legumes is closely associated with the KPB. (A) Chronogram
1747 estimated with 8 fixed local clocks (FLC8 model) in BEAST, with the clock partitions
1748 indicated by colored branches, from an alignment of 36 genes selected as both clock
1749 like and highly informative and hence well-suited for clock analyses. Blue shading
1750 represents 500 post-burnin trees ('densitree' plot) to indicate posterior distributions of
1751 node ages. Yellow stars indicate putative legume WGD events. Labelled circles plotted
1752 across the phylogeny indicate placement and age of fossil calibrations listed in Table 1.
1753 (B) Prior and posterior distributions for the age of legumes under different clock models.
1754 The marginal prior distribution is plotted in grey, UCLN in blue, RLC in green, STRC in
1755 purple and FLC3 in yellow, FLC6 in orange and FLC8 in red.

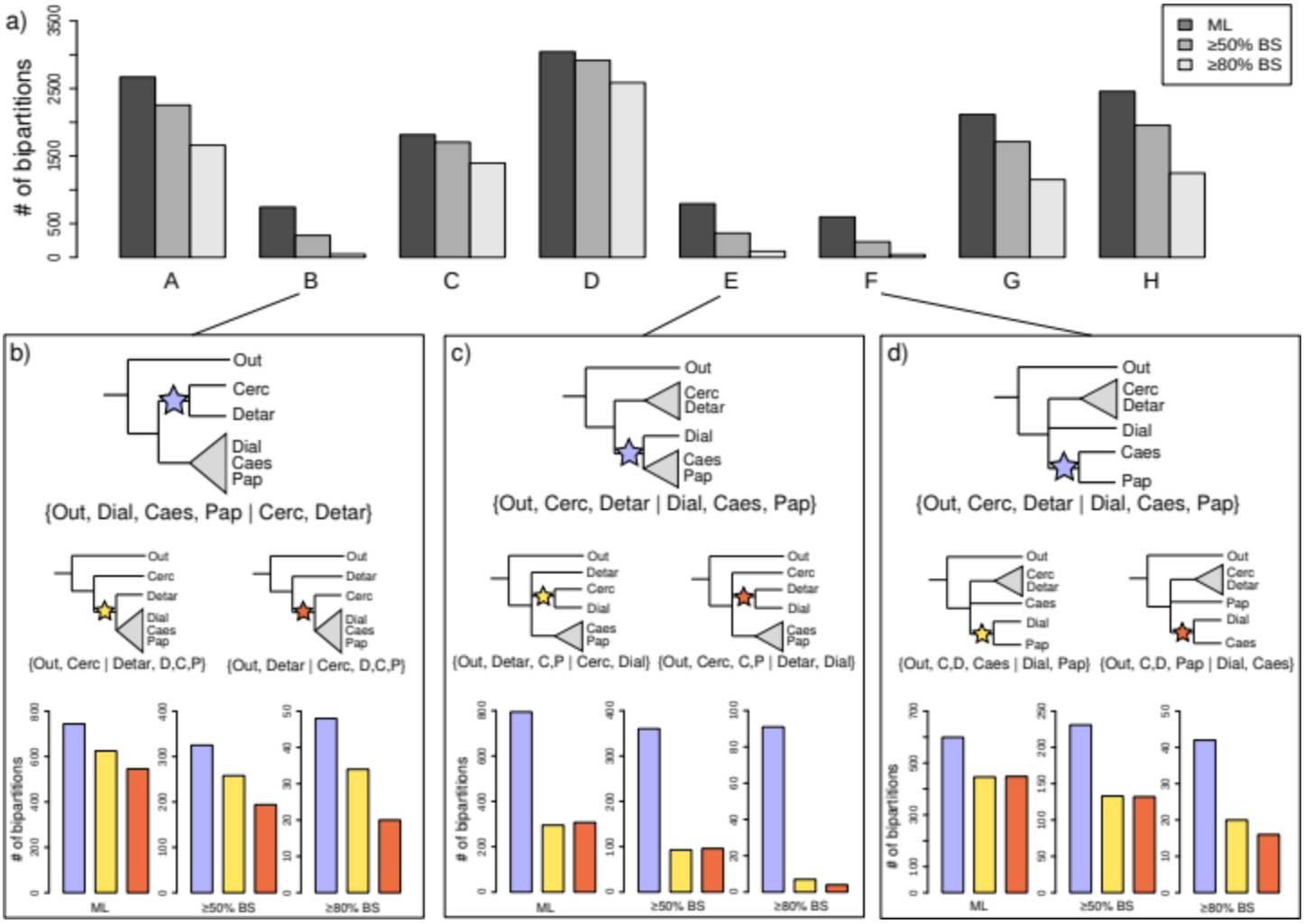
FIGURE 7. Differential loss of paralog copies combined with ILS leads to complex patterns of gene tree evolution. Gene trees 1, 2 and 3 (red, blue and yellow, respectively) are examples of increasingly complex hypothetical evolutionary gene histories, as reconciled with the species tree. Gene 1 loses one paralog copy prior to speciation, and the remaining copy yields the species tree topology in the absence of ILS. Gene 2 is modelled on the homolog cluster2941_1rr_1rr (Fig. S8D), where both duplicated copies are lost or not sampled in a few lineages and there is also ILS. Gene 3 is modelled on the homolog cluster544_1rr_1rr (Fig. S8F) and shows a hypothetical evolutionary history where two rounds of pan-legume WGD occurred in quick succession, with different paralog copies lost either early or late in some lineages and there is also ILS. Blue ovals indicate WGD events, triangles indicate gene duplications and circles indicate coalescences. † = gene loss, o = sampled, x = not sampled, Cerc = Cercidoideae, Detar = Detarioideae, Dupar = Duparquetioideae, Dial = Dialioideae, Caes = Caesalpinioideae and Pap = Papilionoideae.

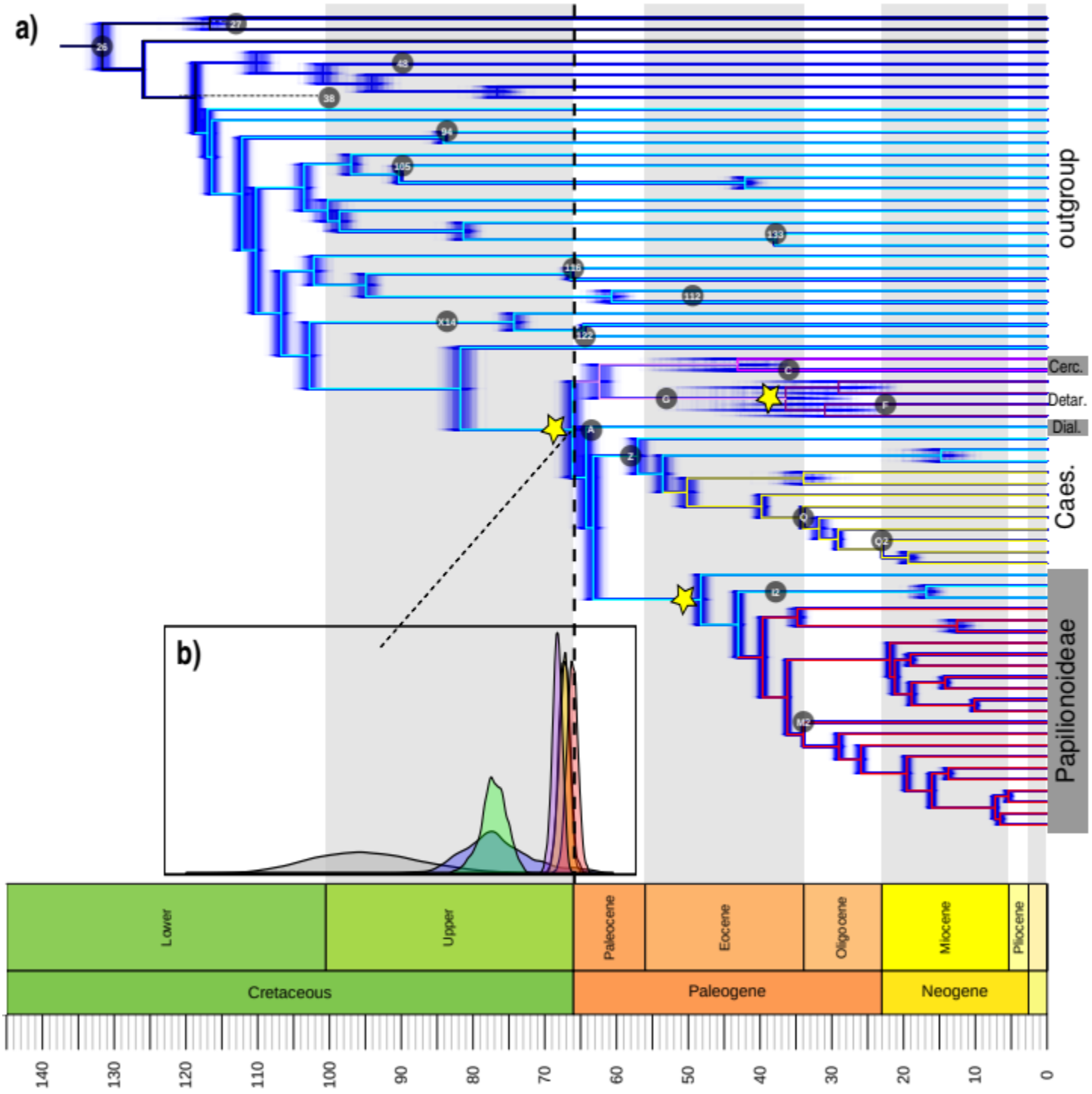


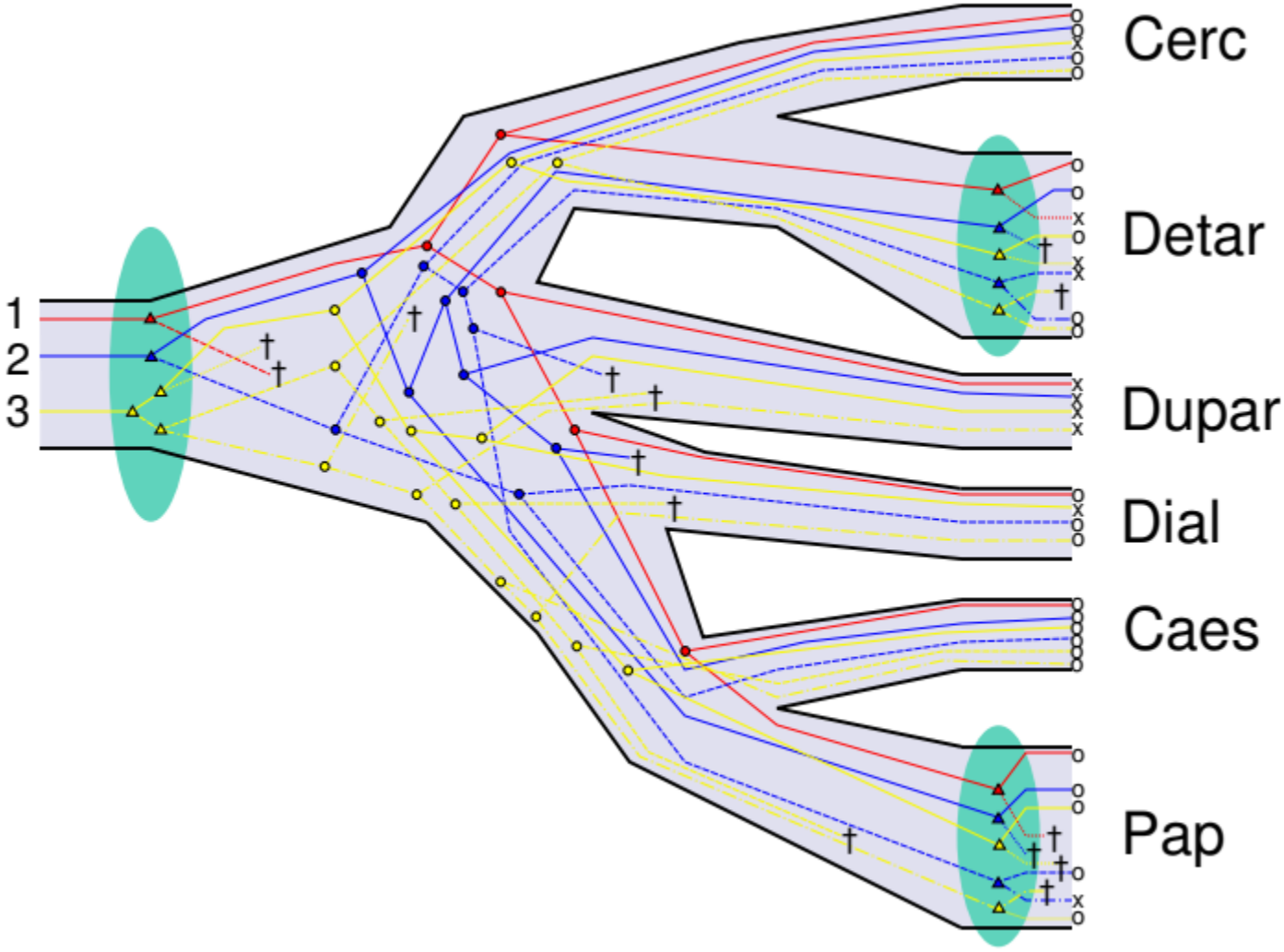












Online Appendices

Methods S1. Discussion on fossils used for calibrating divergence time analyses.

Table S1. Accession information for the taxa included in the chloroplast alignment.

Table S2. Accession information for the taxa included in the nuclear genomic and transcriptomic data set.

Table S3. Counts of bipartitions representing nodes A-H and conflicting bipartitions representing other subfamily relationships among 3,473 gene trees.

Table S4. Age intervals specified for the fossil calibration priors under different alternative priors.

Table S5. Node age estimates and priors (95% HPD intervals) of nodes A-H in the different analyses.

Figure S1. ML topology as inferred by RAxML from amino acid alignment of chloroplast genes under the LG4X model. Numbers on nodes indicate bootstrap percentages estimated from 1000 replicates.

1798

1799 **Figure S2.** Bayesian majority-rule consensus tree inferred with Phylobayes from amino
1800 acid alignment of chloroplast genes under the CATGTR model. Numbers on nodes
1801 indicate posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

1802

1803 **Figure S3.** ML topology as inferred by RAxML from nucleotide alignment of chloroplast
1804 genes under the GTR + G model. Numbers on nodes indicate bootstrap percentages
1805 estimated from 1000 replicates.

1806

1807 **Figure S4.** Bayesian majority-rule consensus tree inferred with Phylobayes from
1808 nucleotide alignment of chloroplast genes under the CATGTR model. Numbers on
1809 nodes indicate the posterior probabilities (pp) from 9000 post-burn-in MCMC cycles.

1810

1811 **Figure S5.** ML topology as inferred by RAxML from a concatenated alignment of 1,103
1812 nuclear genes, under the LG4X model. Numbers on nodes indicate Internode Certainty
1813 All (ICA) values, as estimated from gene trees of the same 1,103 genes.

1814

1815 **Figure S6.** Bayesian gene jackknifing majority-rule consensus tree inferred with
1816 Phylobayes from a concatenated alignment of 1,103 nuclear genes. Numbers on nodes
1817 indicate posterior probabilities (pp), averaged over 500 posterior trees each, for 25
1818 replicates (12,500 posterior trees in total).

Figure S7. Phylogeny estimated under the multi-species coalescent with ASTRAL. Support values indicated represent local posterior probability (blue rectangles) and quartet support (yellow rectangles).

Figure S8. 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 S9. Numbers of gene duplications mapped across the phylogeny. 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.

1840 **Figure S10.** Chronogram estimated under the UCLN clock model. Numbers behind
1841 nodes indicate 95% HPD intervals. Substitution rate is indicated by colored branches,
1842 as indicated by the color legend, in substitutions per site per million years. Fossil
1843 calibrations as listed in Table 1 are indicated by blue labeled circles.

1844

1845 **Figure S11.** Chronogram estimated under the UCLN clock model, with alternative prior
1846 2. Numbers behind nodes indicate 95% HPD intervals. Substitution rate is indicated by
1847 colored branches, as indicated by the color legend, in substitutions per site per million
1848 years. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

1849

1850 **Figure S12.** Chronogram estimated under the RLC model. Numbers behind nodes
1851 indicate 95% HPD intervals. Substitution rate is indicated by colored branches, as
1852 indicated by the color legend, in substitutions per site per million years. Fossil
1853 calibrations as listed in Table 1 are indicated by blue labeled circles.

1854

1855 **Figure S13.** Chronogram estimated under the FLC3 model. Numbers behind nodes
1856 indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil
1857 calibrations as listed in Table 1 are indicated by blue labeled circles.

1858

Figure S14. Chronogram estimated under the FLC6 model. Numbers behind nodes indicate 95% HPD intervals. Cock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S15. Chronogram estimated under the FLC8 model. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles.

Figure S16. Chronogram estimated under the FLC8 model, with alternative prior 1. Numbers behind nodes indicate 95% HPD intervals. Clock partitions are indicated by colored branches. Fossil calibrations as listed in Table 1 are indicated by blue labeled circles, with alternative calibrations as red circles.

Figure S17. 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 S18. 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. Colors correspond to the partitions as shown in Figures 5, S14, S15 and S18.

1880 **Figure S19.** Root-to-tip lengths per taxon with partitions of fixed local clocks indicated.

1881 Pruned taxa with outlier root-to-tip lengths are indicated with an X, partitions are

1882 indicated with colors. (A) FLC3, (B) FLC6, (C) FLC8.