Figure S1. Phylogenetic relationships among 50 selected taxa of angiosperms used as a constraint tree for analyses of rate variation of *clpP* gene. (a) Cladogram of Geraniaceae taken from '(b)' tree. *C*. = *California*, *F*. = *Francoa*, *G*. = *Geranium*, *H*. = *Hypseocharis*, *M*. = *Melianthus*, *P*. = *Pelargonium*, *V*. = *Viviania*. Branches of Geraniaceae are numbered for use in likelihood ratio tests. (b) Maximum likelihood phylogenetic tree inferred from five plastid genes used as a constraint tree for rate variation of *clpP* gene. Bootstrap support values >50% are shown on the branches. Scale bar represents the number of substitution per site.

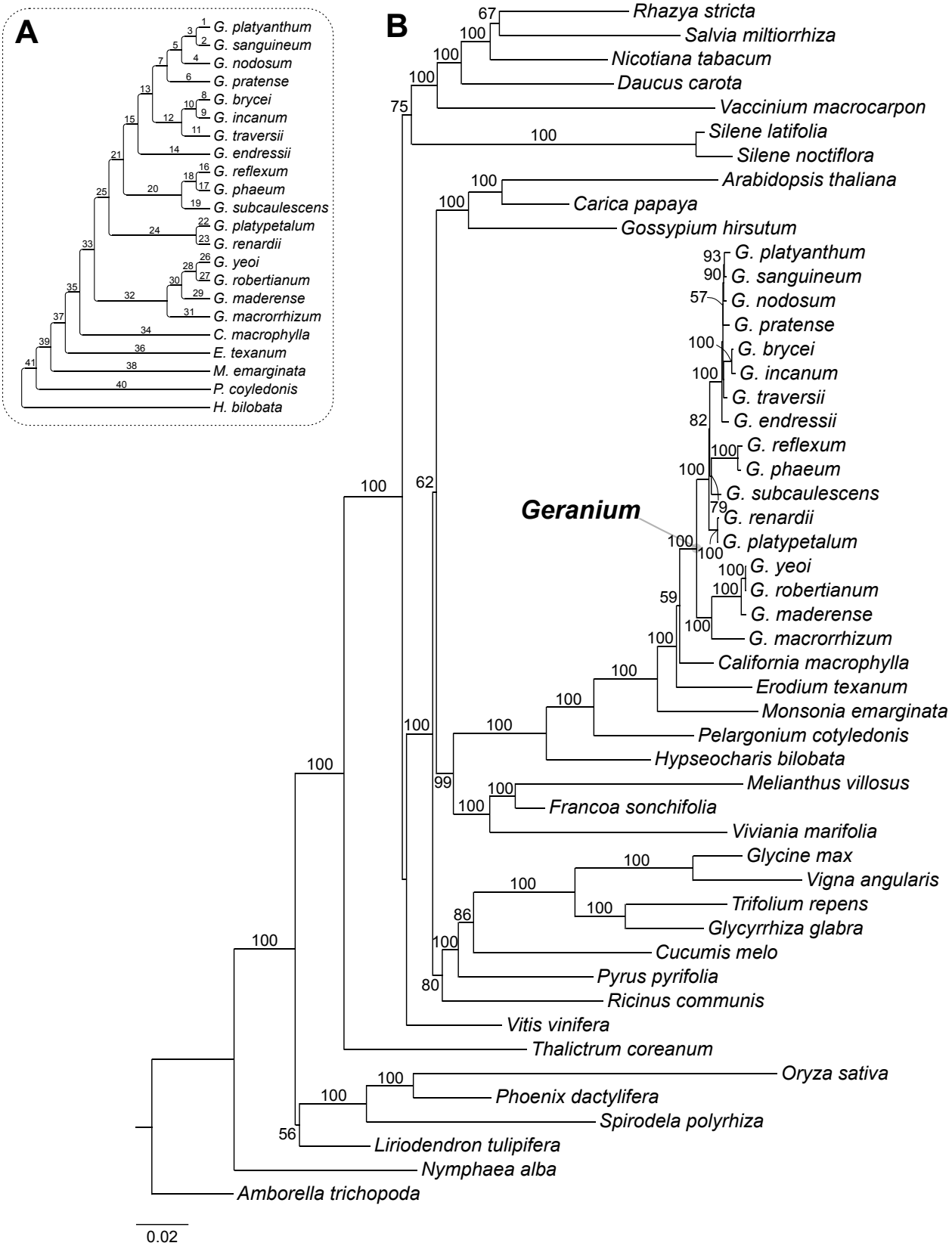

Figure S2. Maximum likelihood phylogenetic tree inferred from five plastid genes for 44 taxa of angiosperms used as a constraint tree for analyses of rate variation of *accD* gene. Bootstrap support values >50% are shown on the branches. Scale bar represents the number of substitution per site.

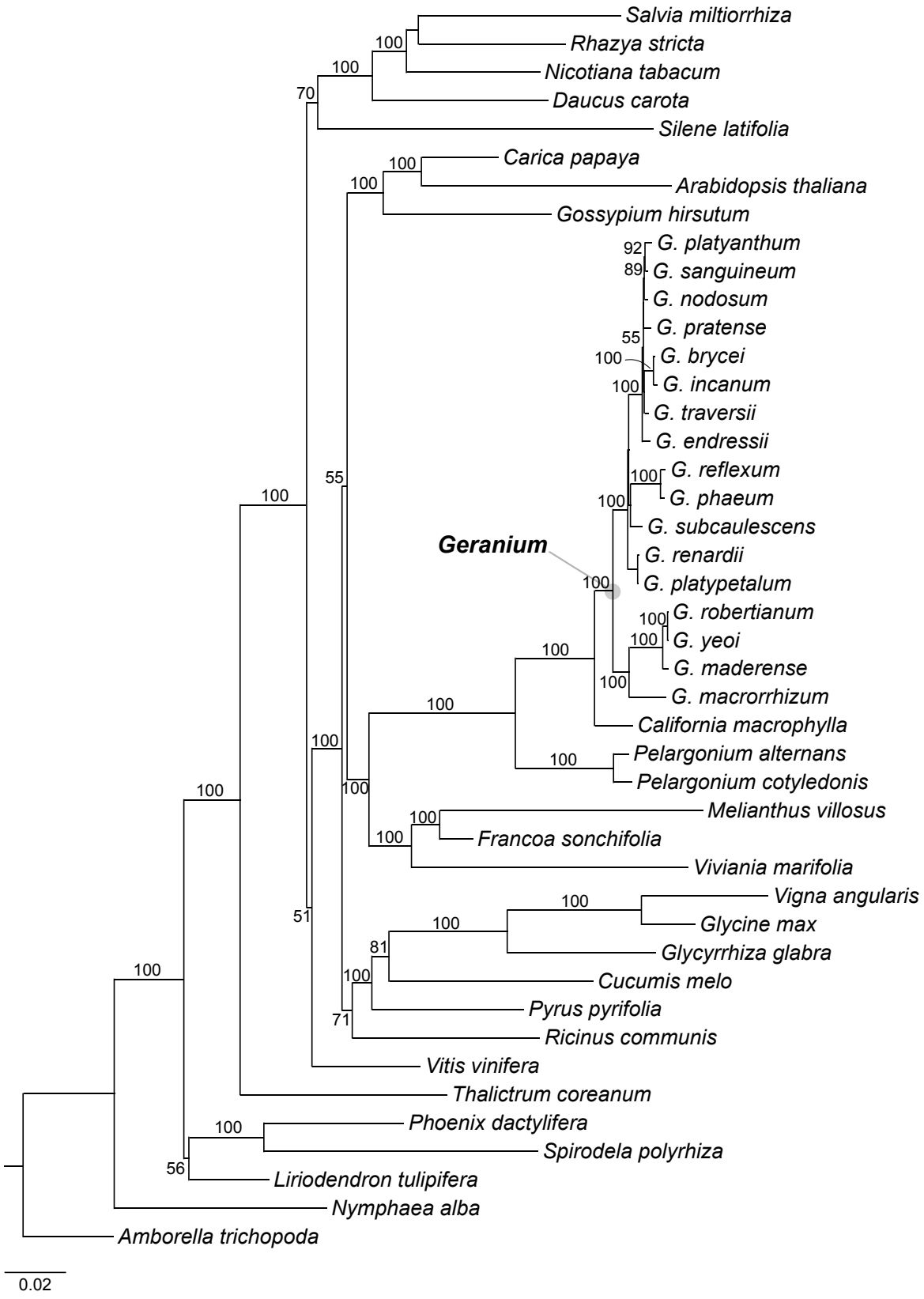

Figure S3. Phylogenetic relationships among 17 species of *Geranium* and three related species of Geraniaceae. (a) Maximum likelihood phylogram inferred from 98 organelle genes from 17 *Geranium* and three outgroups used as a constraint tree for rate analyses. Bootstrap support >50% is species shown on the branches. Subgeneric classification follows Aedo *et al*. (1998). (b) The tree in (a) was modified manually by adding two branches for the evolutionary rate variation of duplicated genes and branch numbers are labeled. The numbers '1' and '2' after each species in the cladogram represent paralogs in species that have experienced gene duplications (see Fig. 3). This cladogram was used as the species tree in subsequent analyses. Each branch and species in *Geranium* is given a number that is used for likelihood ratio tests reported in table S9. Scale bar indicates the number of substitutions per site.

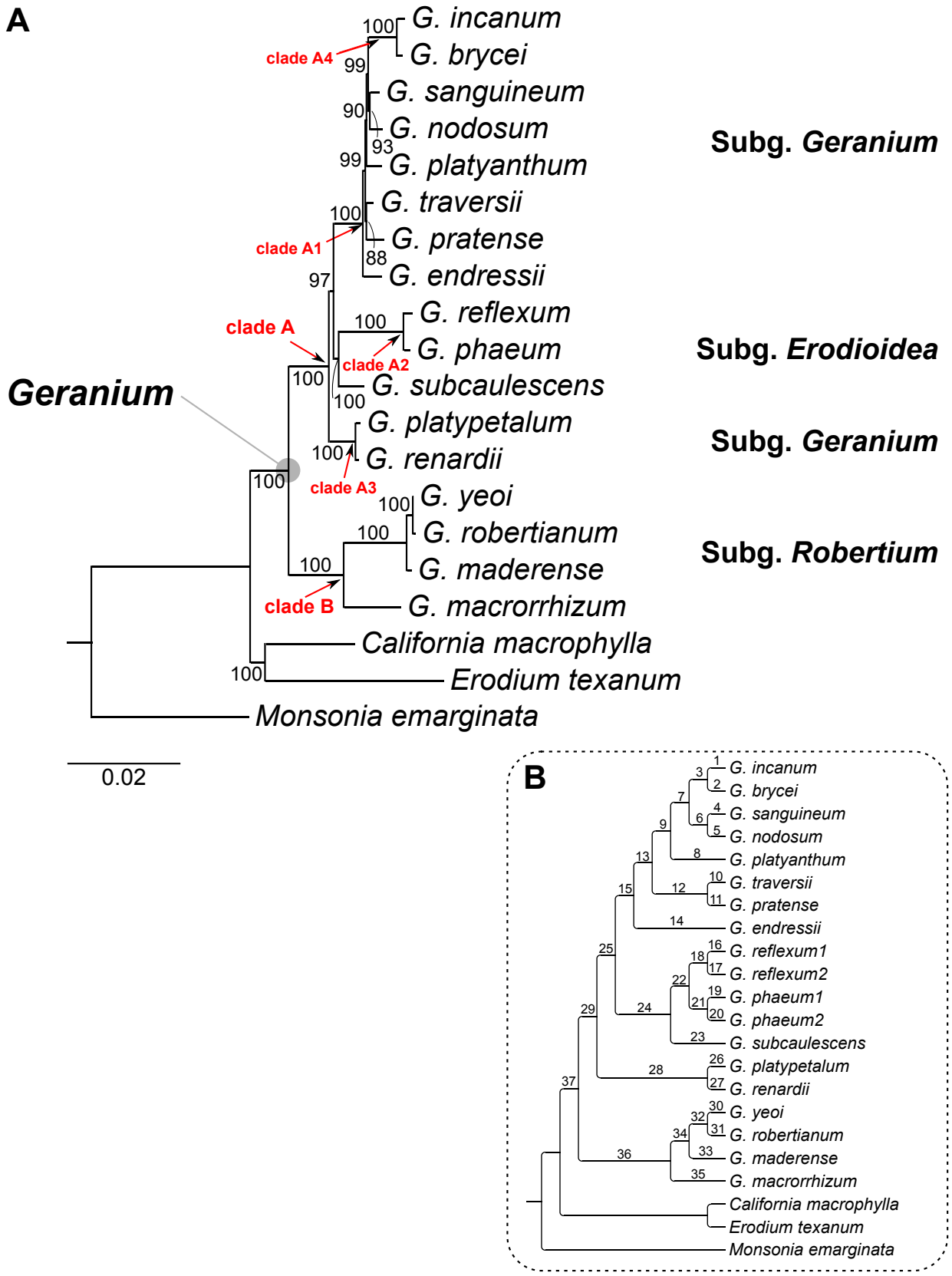

Figure S4. Nonsynonymous (d_N) and synonymous (d_S) substitutions in *Geranium* plastid (red) and mitochondrial (blue) individual genes or groups of genes. The box represents values between quartiles, solid lines extend to minimum and maximum values, outliers are shown as circles and horizontal lines in boxes show median values.

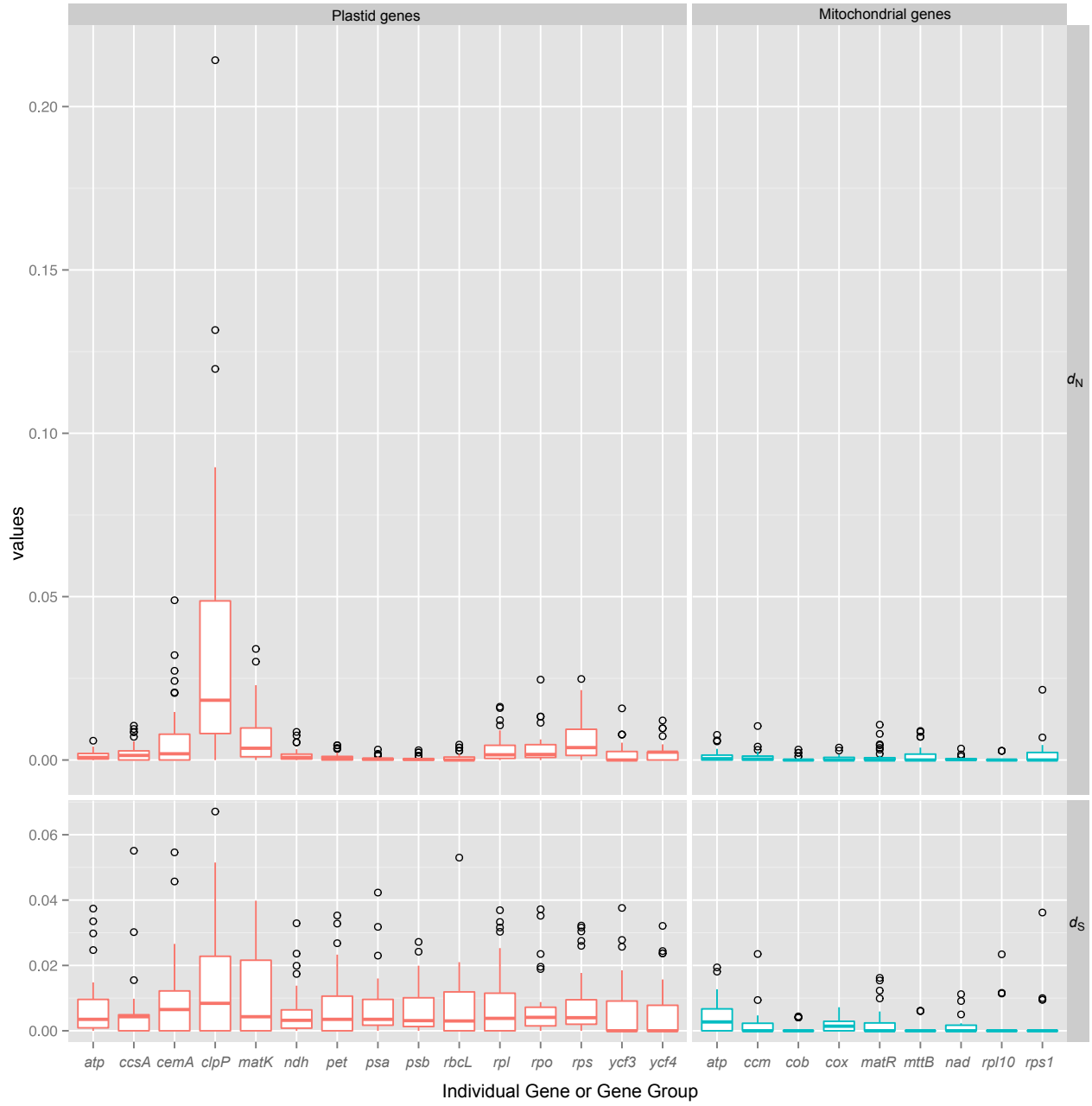

Figure S5. Nonsynonymous (*dN*) and synonymous (*dS*) sequence divergence among *Geranium* mitochondrial genes. Colored bars indicate *dN* (red) and *dS* (blue) values. Absent bars in some species indicate that the values of dN or dS are zero.

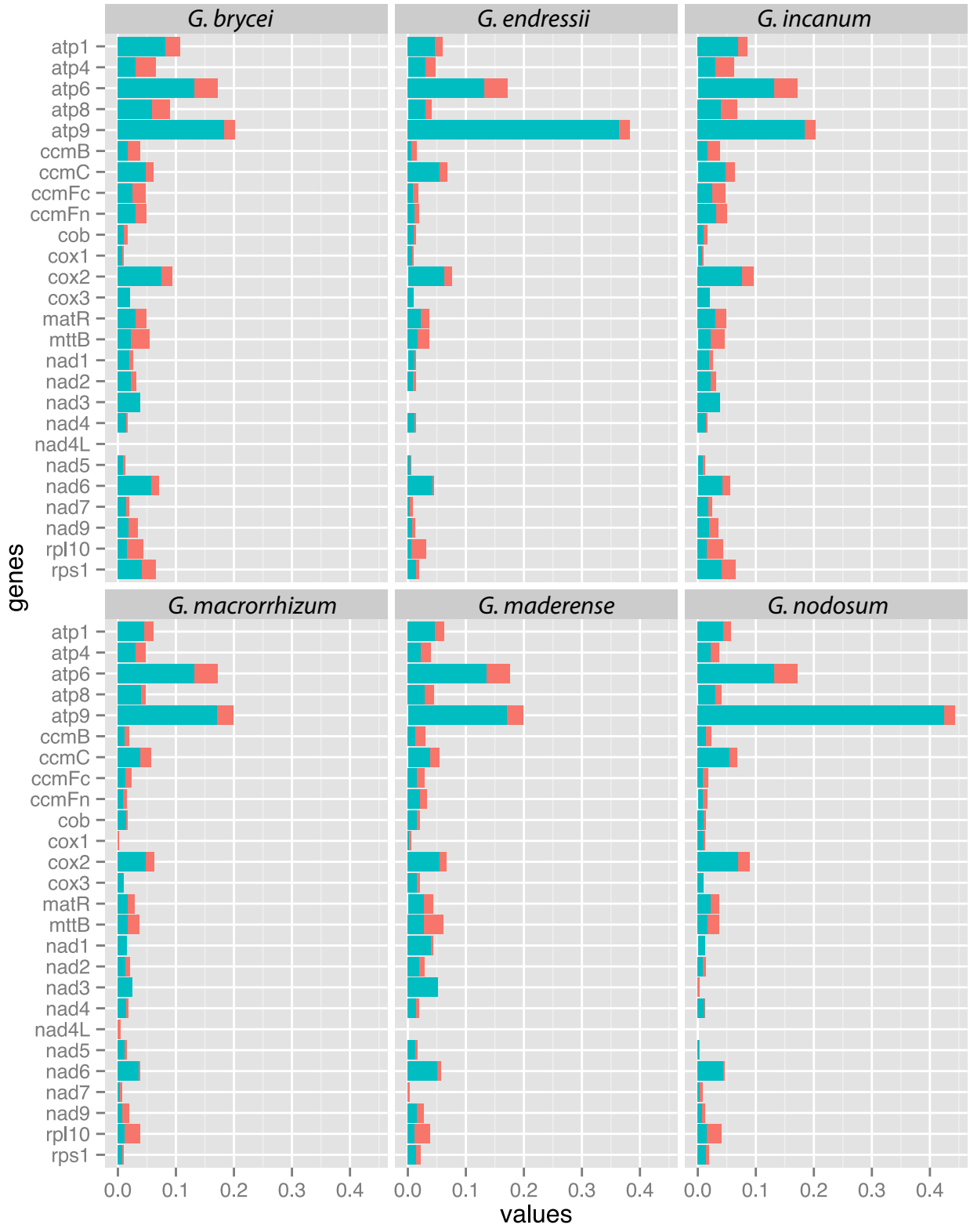

Figure S5. (continued)

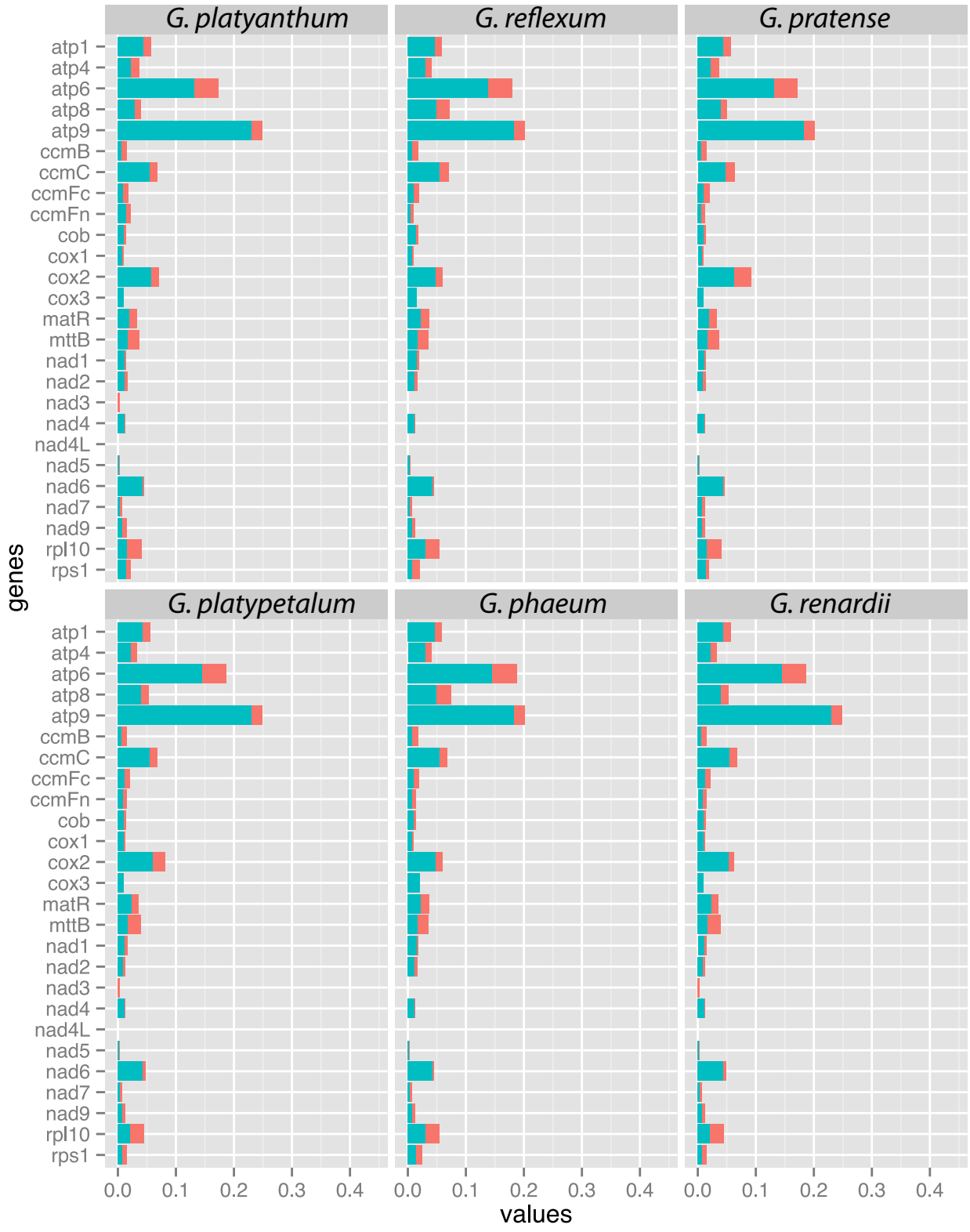

Figure S5. (continued)

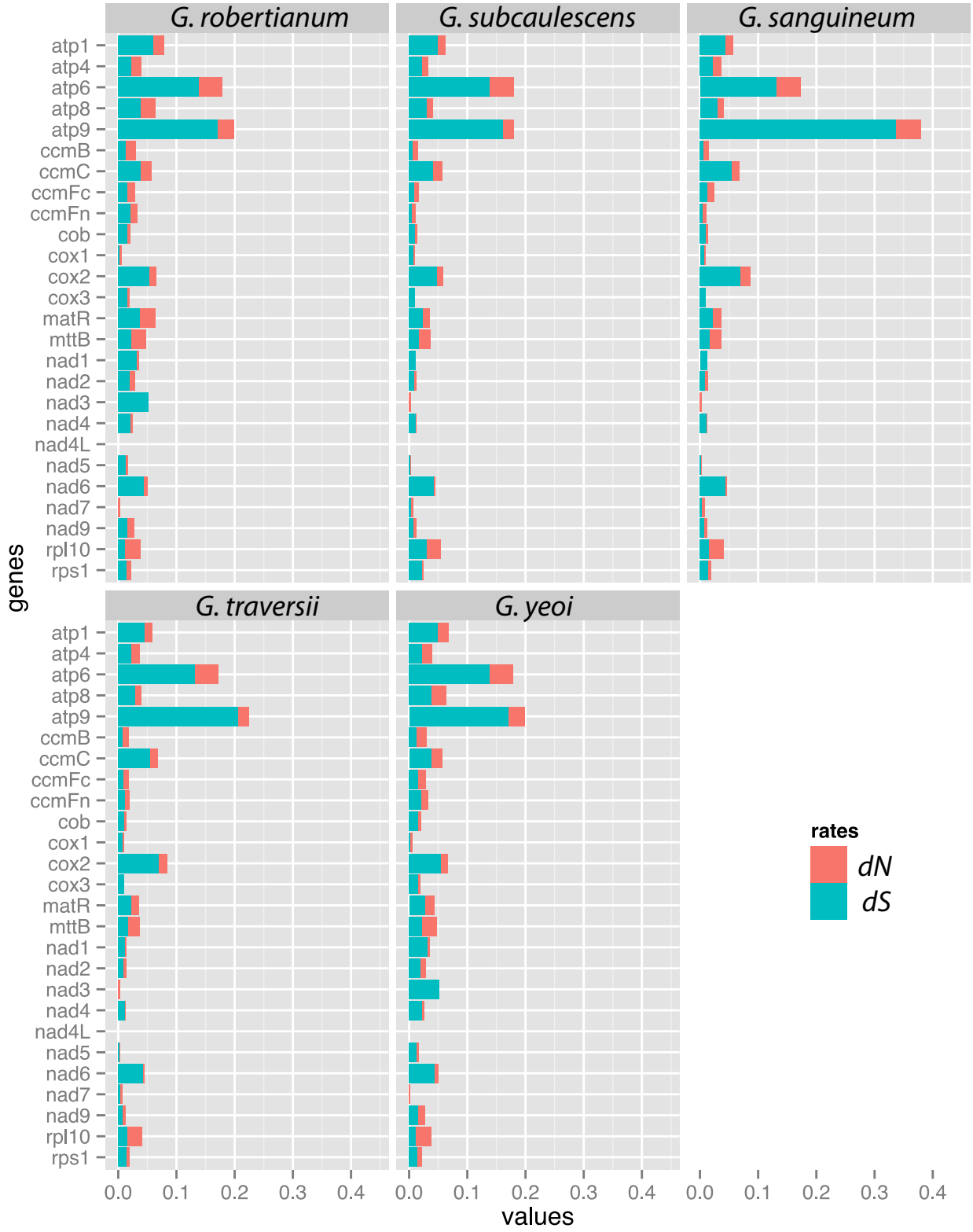

Figure S6. Nonsynonymous sequence divergence in plastid-encoded genes and gene groups. Clades A2 and B showed accelerated rates of sequence evolution and are highlighted in shaded gray. Branch lengths in all trees are drawn to the same scale based on the number of nonsynonymous substitution per site. Major clades and subclades (see supplementary fig. S1A) are labeled on *clpP* tree. *C.* = *California*, *E.* = *Erodium*, *G.* = *Geranium*, *M.* = *Monsonia*.

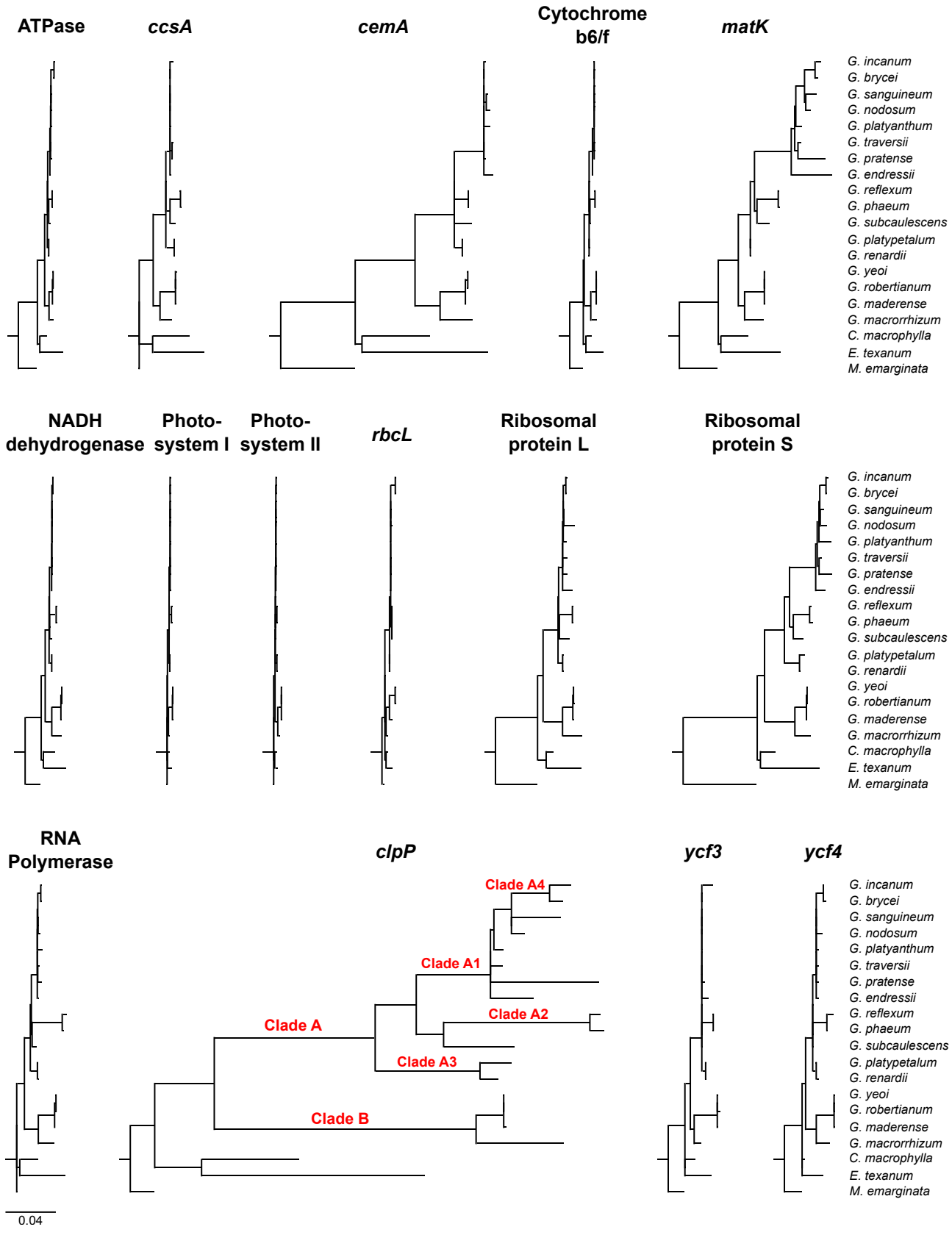

Figure S7. Synonymous sequence divergence in mitochondrial-encoded genes and gene groups. Clades A2 and B have accelerated rates of sequences and are highlighted in shaded gray. Branch lengths in all trees are drawn to the same scale based on the number of synonymous substitution per site. Major clades and subclades (see Fig. S3A) are labeled on ATPase tree. *C.* = *California*, *E.* = *Erodium*, *G.* = *Geranium*, *M.* = *Monsonia*.

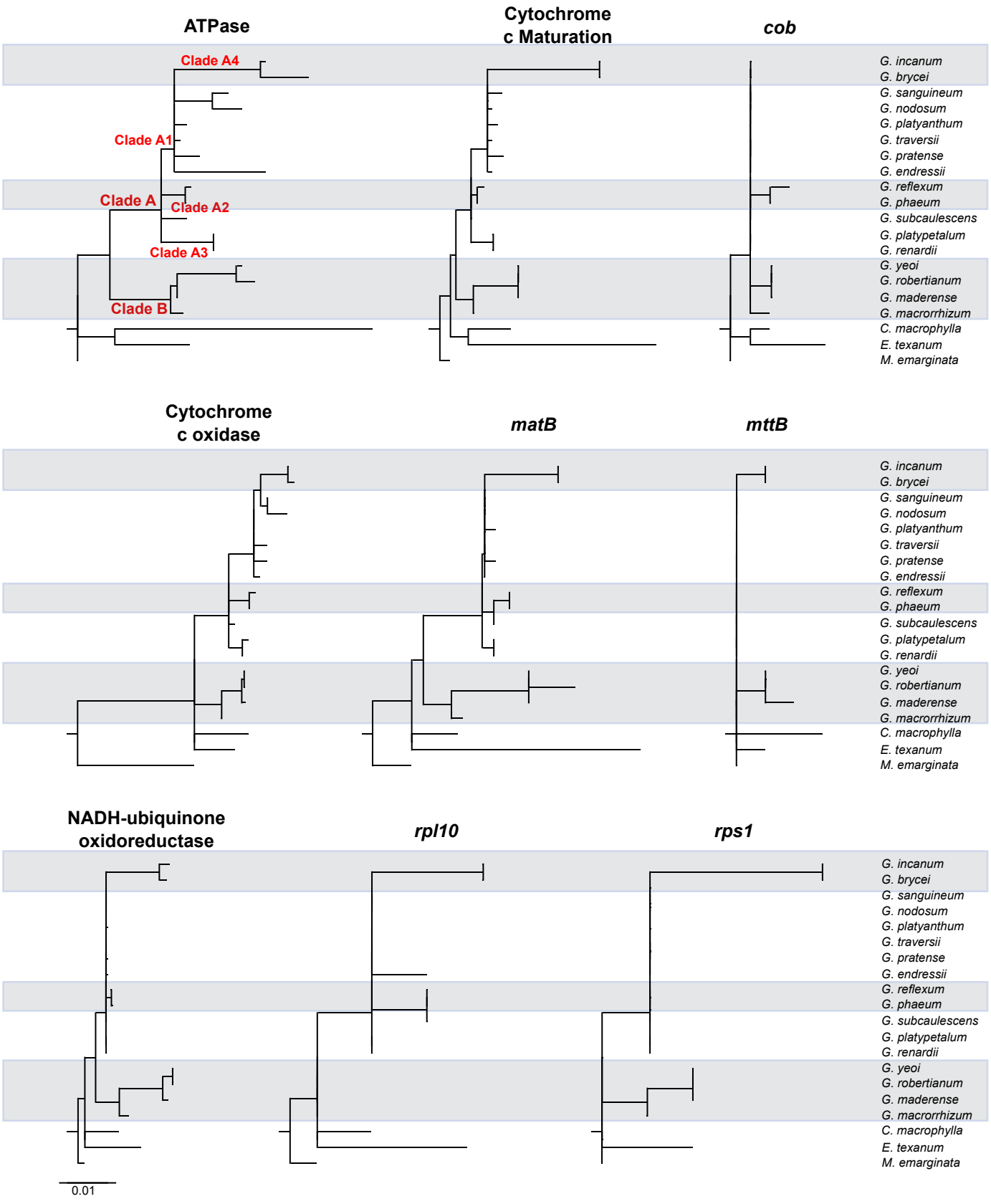

Figure S8. The effective number of codons (EN_C) plotted against GC content at third synonymous site positions (GC_{3S}) for plastid (a) and mitochondrial (b) genes in *Geranium*. Line shows the expected EN_C as a function of GC_{3S} given random codon usage (Wright 1990).

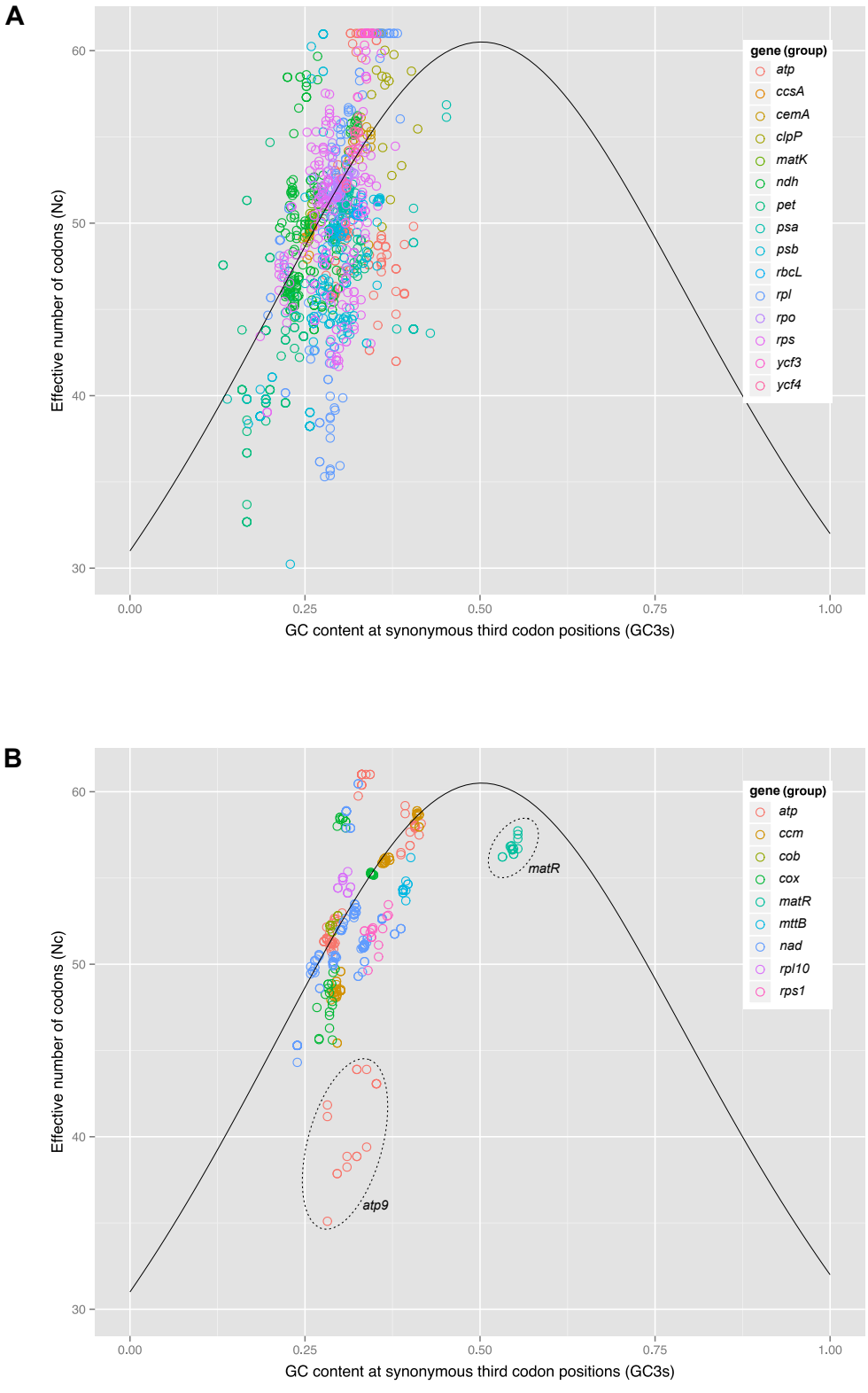

Figure S9. Correlation between synonymous substitution rates and (a) GC content in third synonymous site positions (GC_{3S}), and (b) the effective number of codons (EN_C). Linear regression analyses using gene or gene groups (*atp9*, blue; *atp* gene group, green; other genes, red). Significance of fit was evaluated by Pearson's correlation in the R package.

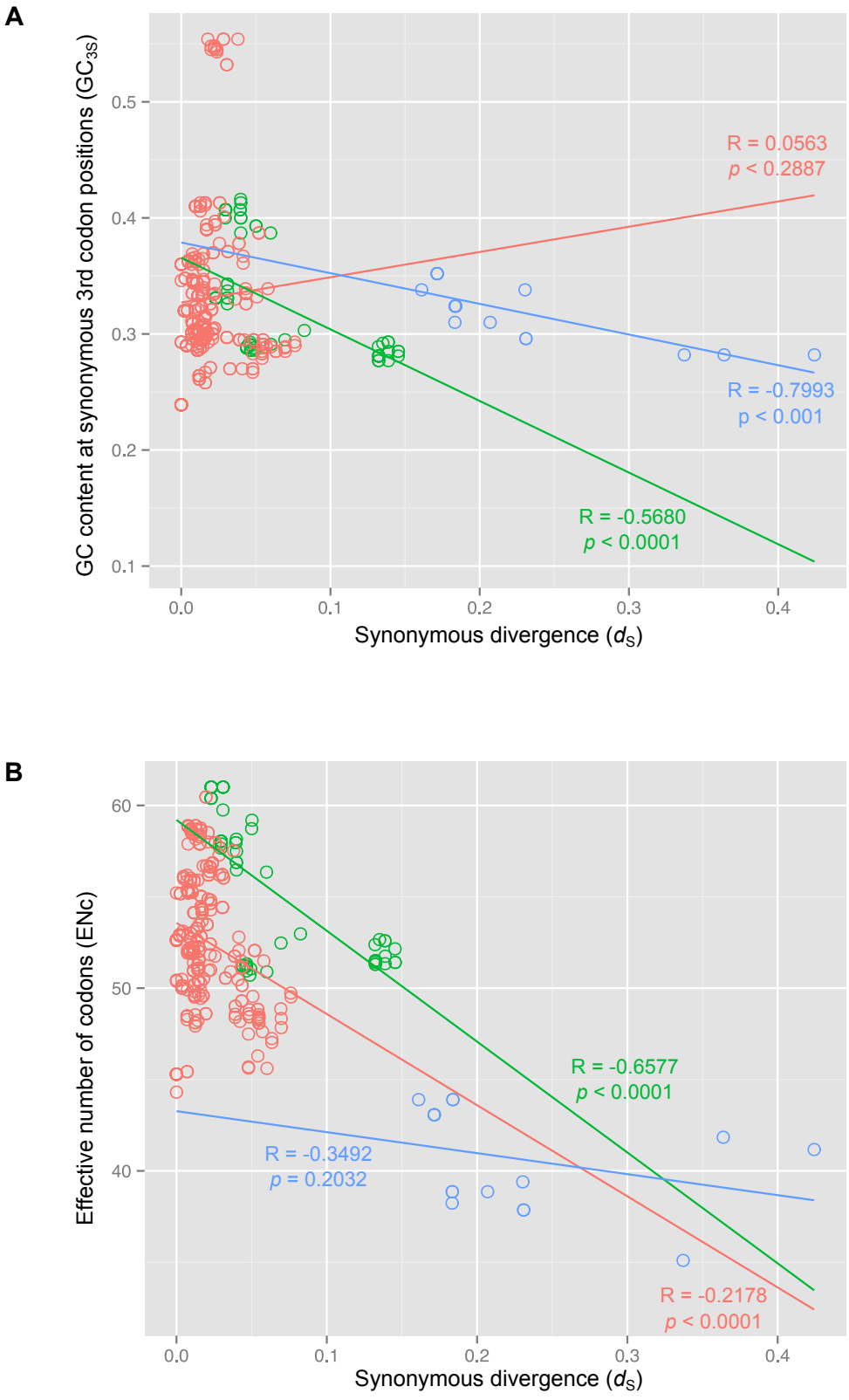

Figure S10. Nonsynonymous (d_N) and synonymous (d_S) values for four mitochondrial genes. Color bars indicate the values of d_N (red) and d_S (blue). Significance of fit was evaluated by Wilcoxon rank sum tests in the R package. Numbers at nodes on tree are bootstrap values. *G.* = *Geranium*. Scale bar indicates the number of substitutions per site.

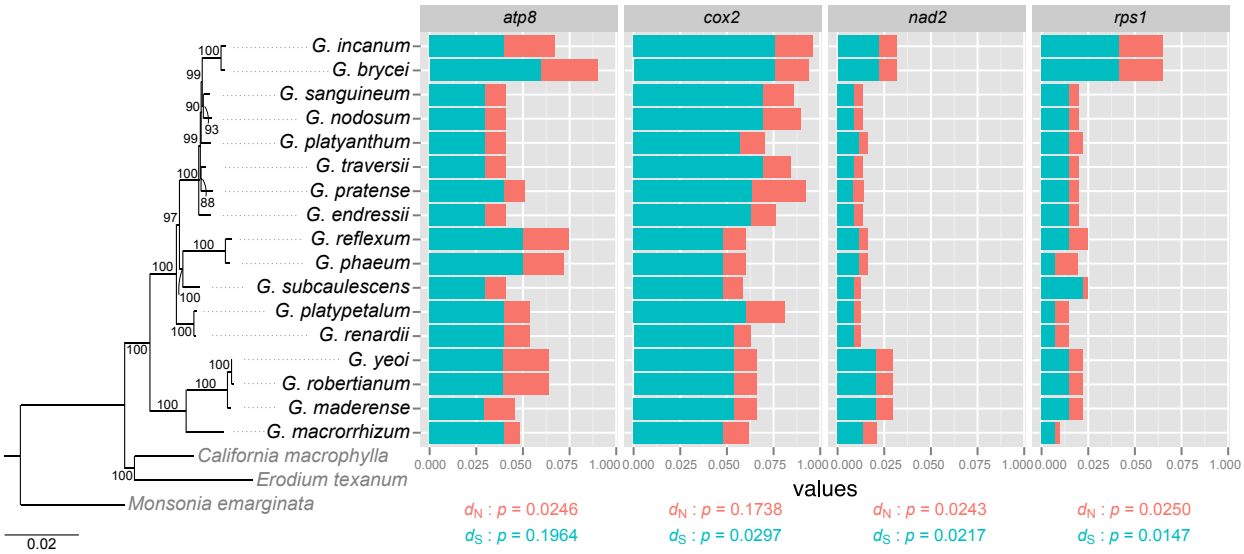

Figure S11. Phylograms of plastid-encoded polymerase (PEP) genes *rpoA*, *rpoB*, *rpoC1* and *rpoC2* showing nonsynonymous (d_N) and synonymous (d_S) substitution rates. *Geranium reflexum* and *G. phaeum* are highlighted in shaded gray. All trees are drawn to the same scale. *C.* = *California*, *E.* = *Erodium*, *G.* = *Geranium*, *M.* = *Monsonia*.copy Scale bar indicates the number of substitutions per site. For *rpoA*, copy 1 of this gene was used for *G. reflexum* and *G. phaeum*.

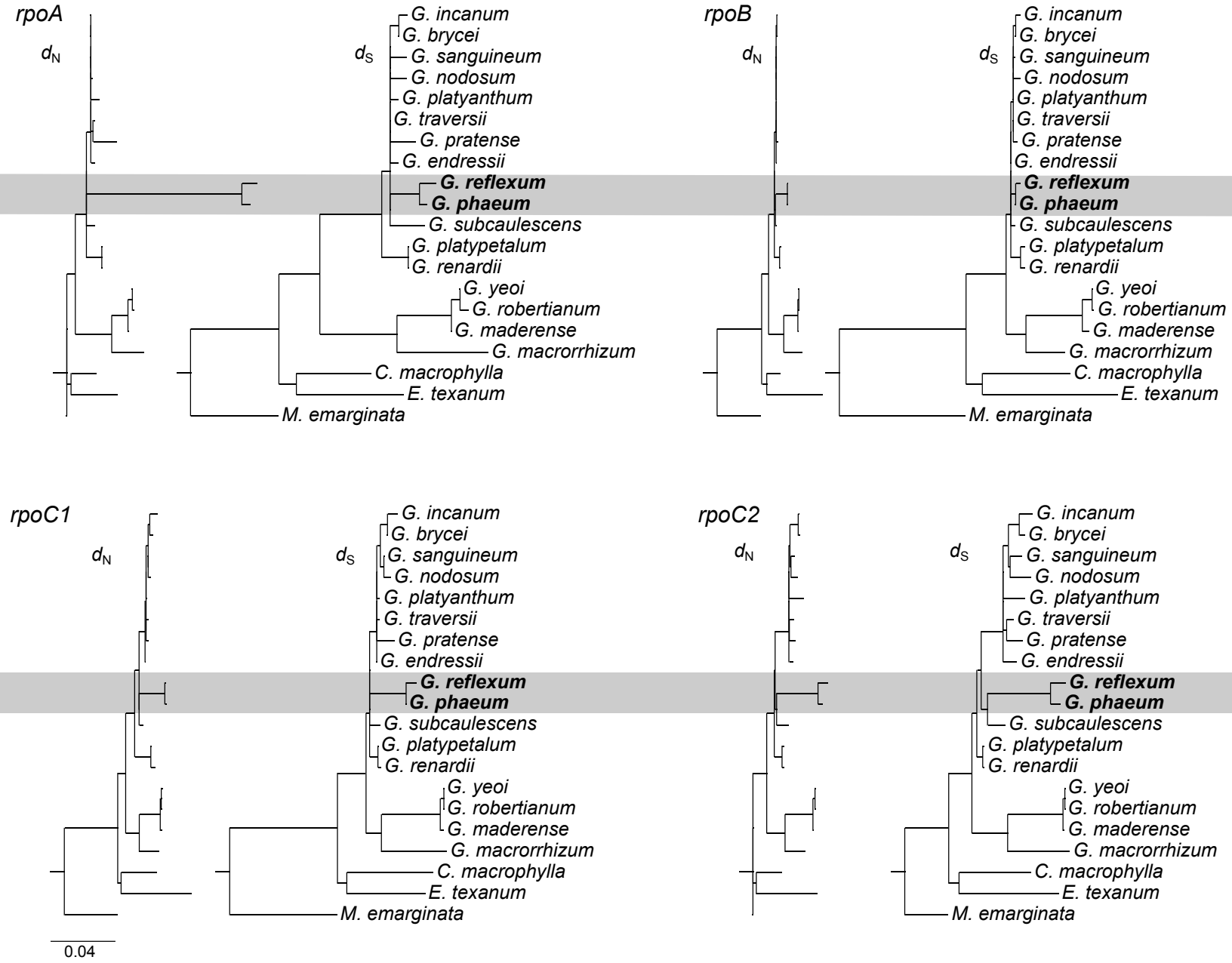

Figure S12. Rapid structural evolution of the plastid-encoded *clpP* gene. (a) Schematic diagram of the variable structure of *clpP* among selected angiosperms. Arrowheads indicate the positions of the first (black) and second (gray) intron. Pink boxes indicate the conserved domain of caseinolytic protease. *Geranium* and *Monsonia* highlighted in red have a variable length at N-terminus. (b) Phylograms showing nonsynonymous (d_N) and synonymous (d_S) substitution rates for the *clpP* genes among representative angiosperms, including Geraniaceae. Branch lengths are drawn to the same scale based on d_N (left) and d_S (right) substitutions per site. Branches with significantly higher d_N/d_S ratios detected by likelihood ratio test (LRT) are marked with asterisks (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.0001$ after Bonferroni correction). Colored asterisks indicate different models (i.e. black, 'branch leading to' and red, 'branch within'). Clades A, B, A1, A2, and A3 corresponding to different lineages of *Geranium* (supplementary fig. S1A) are highlighted in gray. Closed black circles on branches indicate lineages that have divergent *clpP* genes. *C.* = *California*, *F.* = *Francoa*, *G.* = *Geranium*, *H.* = *Hypseocharis*, *M.* = *Melianthus*, *P.* = *Pelargonium*, *V.* = *Viviania*. Gray rectangles at tips of d_S tree indicates presence or absence of introns 1 and 2 in *clpP* gene. Scale bar indicates the number of substitutions per site.

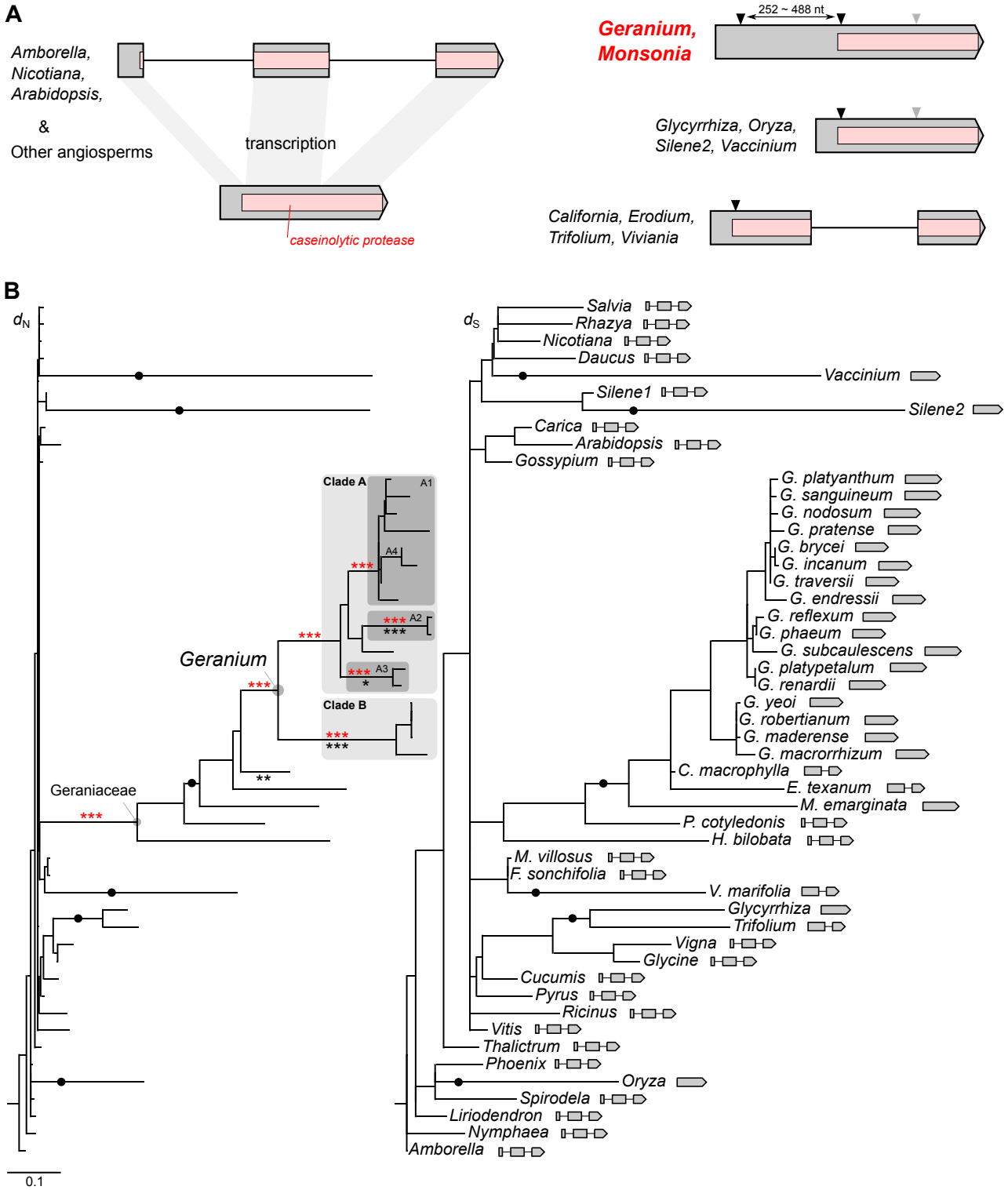

Figure S13. Amino acid alignment of plastid-encoded *clpP* gene for *Nicotiana tabacum*, *Geranium* (G.) and *Monsonia emarginata*. Red box indicates caseinolytic protease conserved domain.

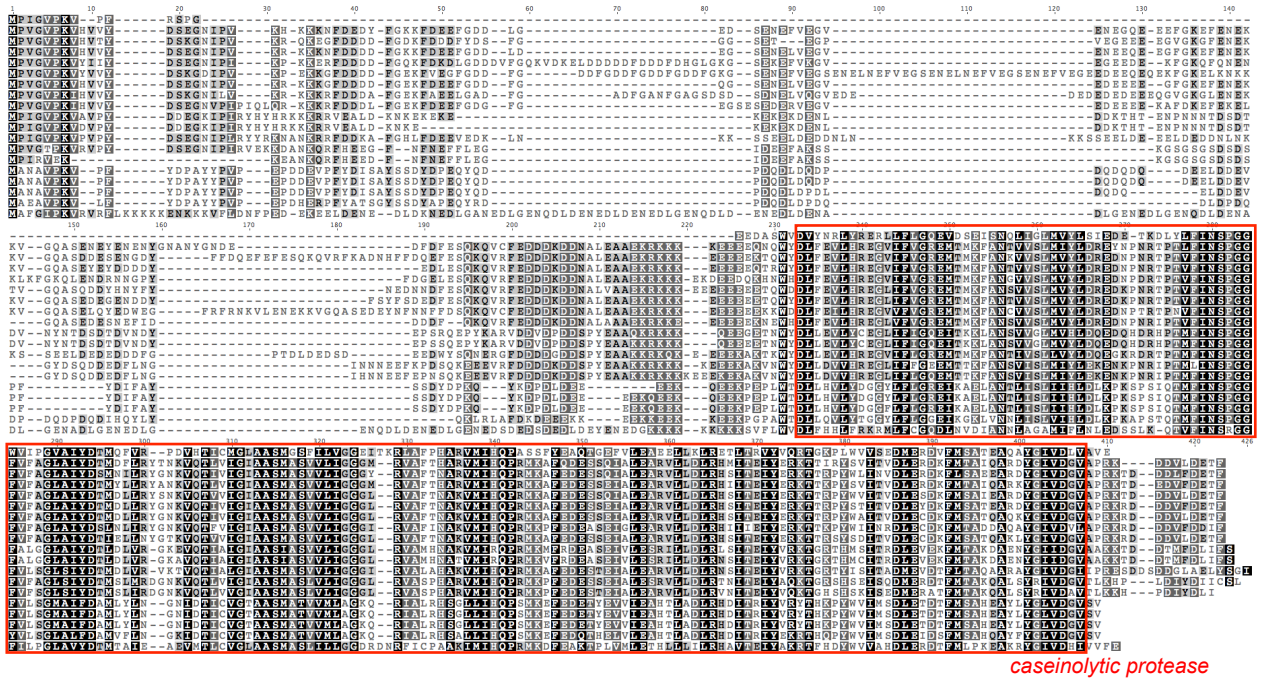

Figure S14. Nucleotide alignment of the conserved domain sequences of the plastid-encoded *clpP* gene. Exons are annotated on *Arabidopsis thaliana*. The C to U editing site at alignment position is highlighted in pink.

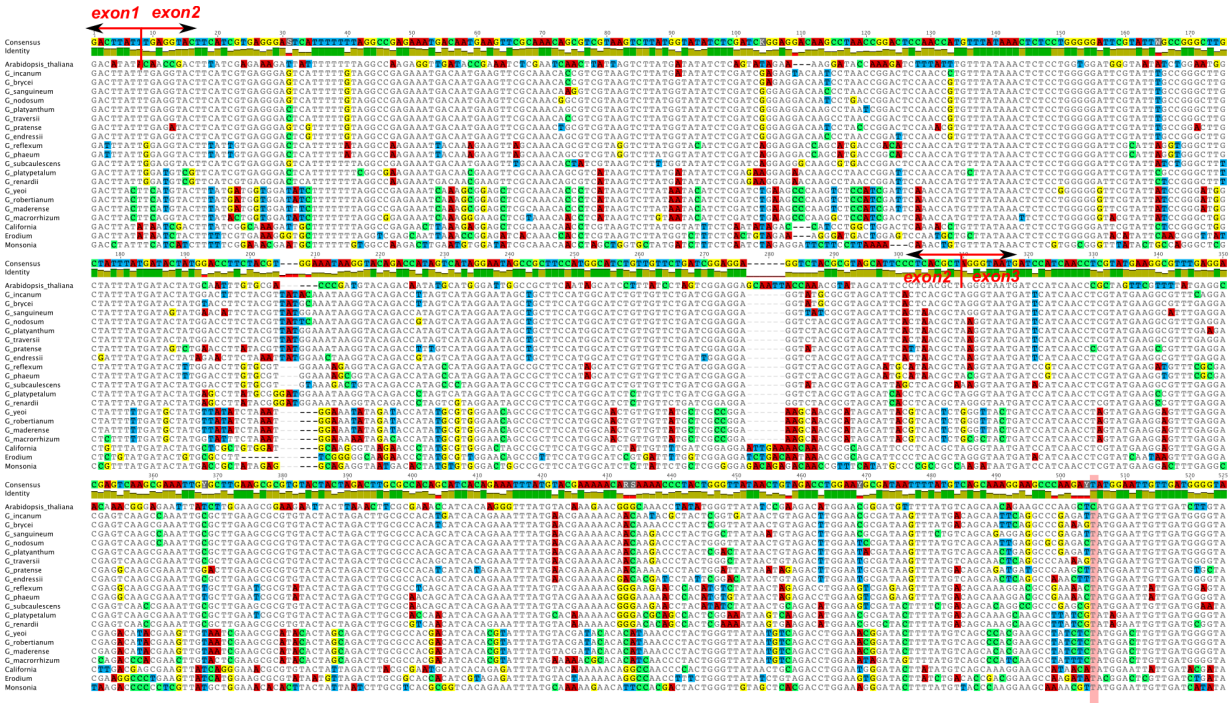

Figure S15. Correlation between nonsynonymous (d_N) and synonymous (d_S) substitution rates of *clpP* (a) and *accD* (b) genes. Boxplot distribution of the values of d_N and d_S for *clpP* and *accD* genes between divergent groups (D, red) and other angiosperms (O, blue) that contain canonical *clpP* and *accD* genes. The box represents values between quartiles, solid lines extend to minimum and maximum values, outliers are shown as circles and horizontal lines in boxes show median values. Significance of fit was evaluated by Pearson's correlation and pairwise Wilcoxon rank sum tests in the R package. (a) Linear regression analyses included divergent groups (red circles) and other angiosperms (blue triangles). In particular, Geraniaceae data points are highlighted in closed red circles. (b) Linear regression analyses included four branches leading to divergent lineages (see Fig. S15, closed red circles) and other angiosperms (blue triangle). *Geranium*, *California*, and *Pelargonium* terminal branch values and the value of the internal branch were marked with red open circles.

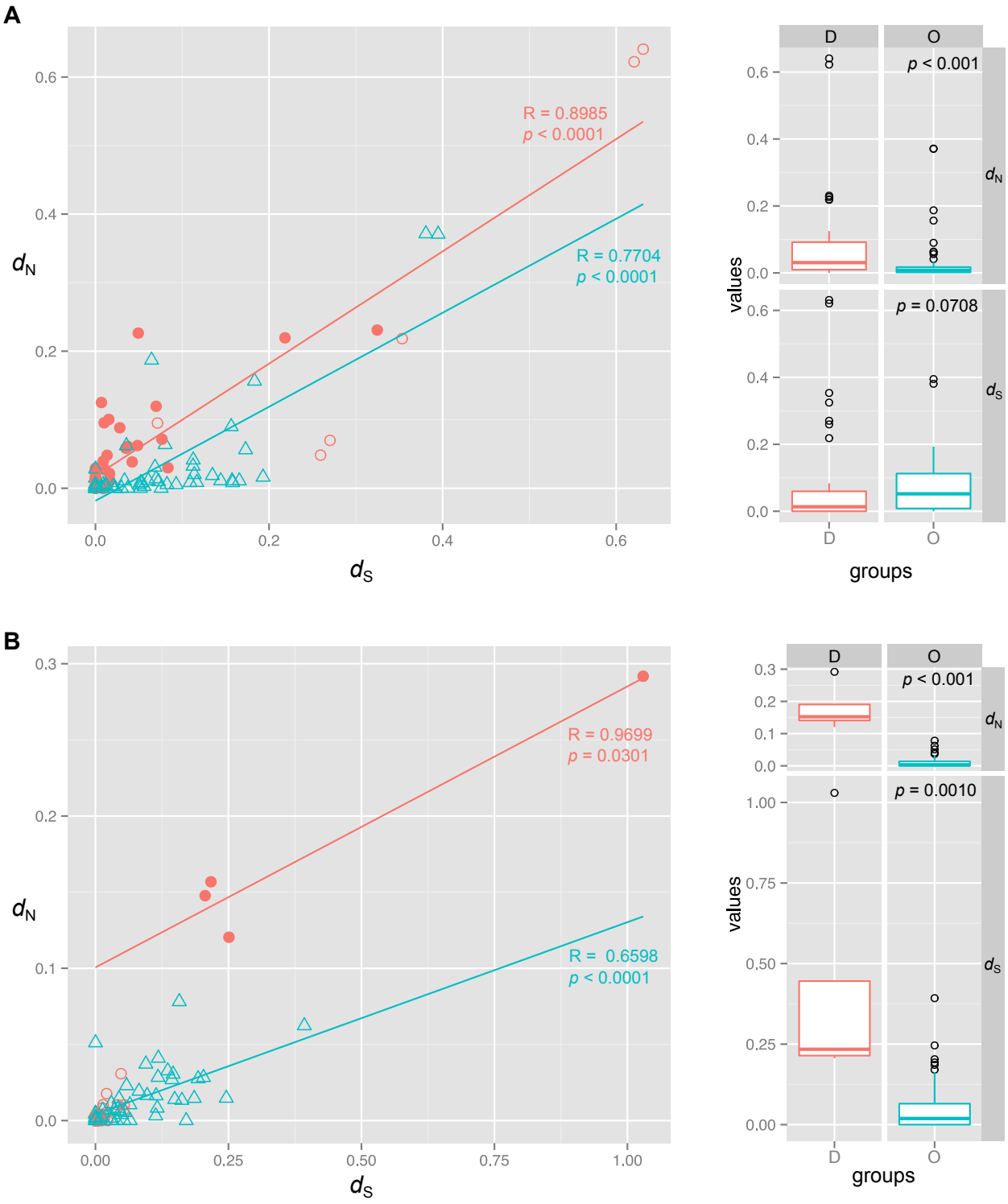

Figure S16. Phylograms of plastid *a ccD* gene showing nonsynonymous (d_N) and synonymous (d_S) substitution rates. Rectangles indicate the conserved domain (acetyl-CoA carboxylase beta subunit D; pink) in the predicted ORFs. Arrowheads indicate the lineages with a disrupted *accD* gene. *C*. = *California*, *G*. = *Geranium*, *F*. = *Francoa*, *M*. = *Melianthus*, *P*. = *Pelargonium*, *V*. = *Viviania*. Scale bar indicates the number of substitutions per site.

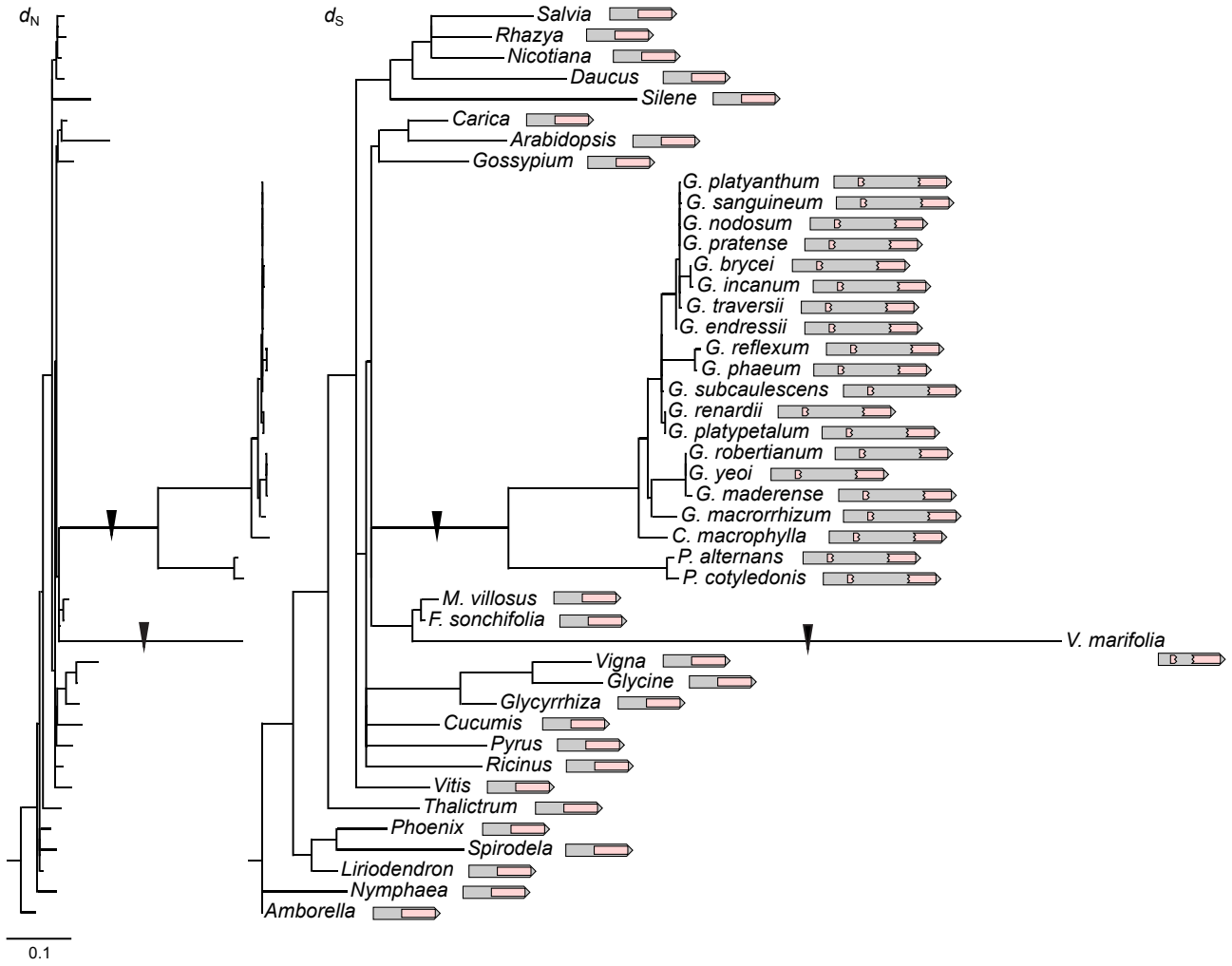

Figure S17. Characterization of the nuclear-encoded acetyl-CoA carboxylase beta subunit (*accD*) genes of *Hypseocharis* and *Monsonia*. (a) Architecture of the nuclear-encoded *accD* gene from *Hypseocharis bilobata* and *Monsonia emarginata*. Pink boxes indicate the conserved domain of the acetyl-CoA carboxylase subunit β (Lee et al. 2004), green boxes in N-terminus indicate a transit peptide.
(b) Amino acid sequence alignment of the nuclear-encoded *accD* regions from *H. bilobata* and *M. emarginata.* Colors indicate subunits of ACC corresponding to the schematic diagram in (a).

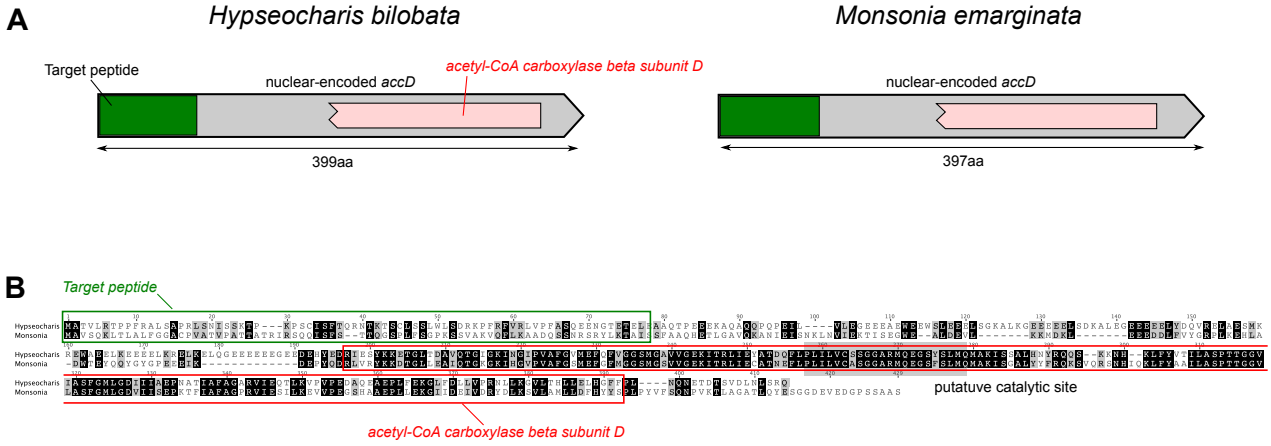

Figure S18. Nuclear acetyl-CoA carboxylase (ACC) divergence among selected angiosperms. (a) Maximum likelihood tree showing nonsynonymous (d_N) and synonymous (d_S) substitution rates for the nuclear ACC homologs. All trees are drawn to the same scale. *G.* = *Geranium*. The numbers '1' and '2' after each species in the phylograms represent paralogs for ACC. (b) Boxplot distribution of the values of d_N and d_S for *Geranium* ACC orthologs (ACC1 and ACC2). The box represents values between quartiles, solid lines extend to minimum and maximum values, outliers are shown as circles and horizontal lines in boxes show median values. Significance of fit was evaluated by Pearson's correlation and pairwise Wilcoxon rank sum tests in the R package. Scale bar indicates the number of substitutions per site.

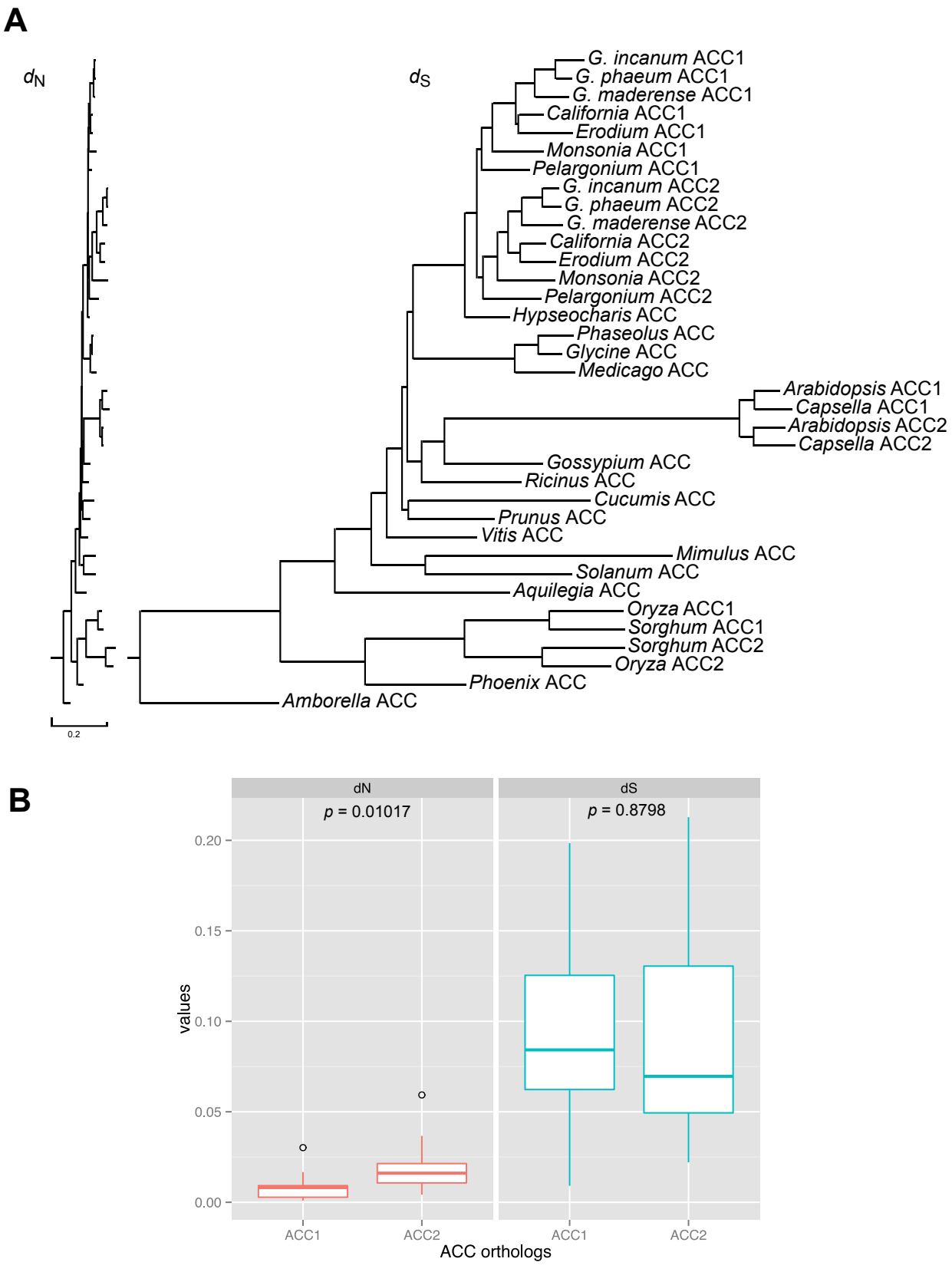

Figure S19. Chronogram of Geraniaceae divergence times. Times shown are the median age estimates from the BEAST analysis. Blue bars indicated 95% highest posterior density intervals in age estimates. *G*. = *Geranium*. The numbers '1' and '2' after each species in the phylograms represent paralogs for ACC. Ma = million years ago.