Supplementary Material

SimPhy: Phylogenomic Simulation of Gene, Locus and Species Trees

Diego Mallo, Leonardo de Oliveira Martins and David Posada

List of Figures

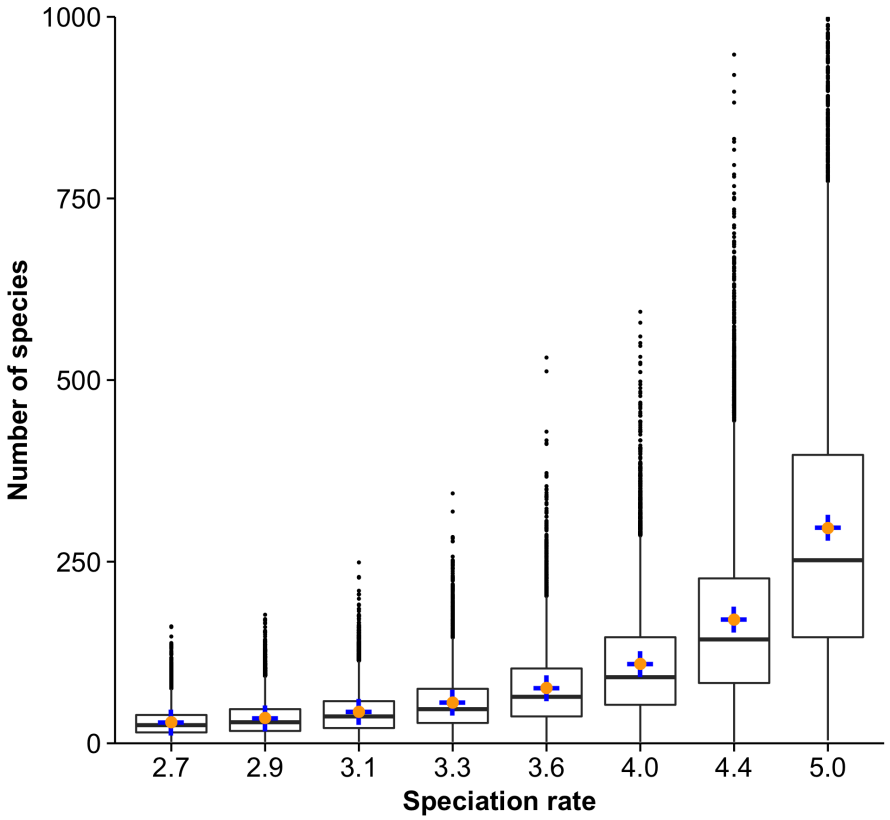

Fig. S1. Validation of *SimPhy*'s species tree simulation (variable number of species).Boxplots describe the distribution of the number of species generated by 10000 simulation replicates across different speciation rates (speciations/1M generations) given a fixed tree height (1M generations). Expected theoretical values are indicated with a blue cross, while the observed average values are depicted with an orange dot. For representation purposes, extreme values in the rightmost boxplot are not shown.

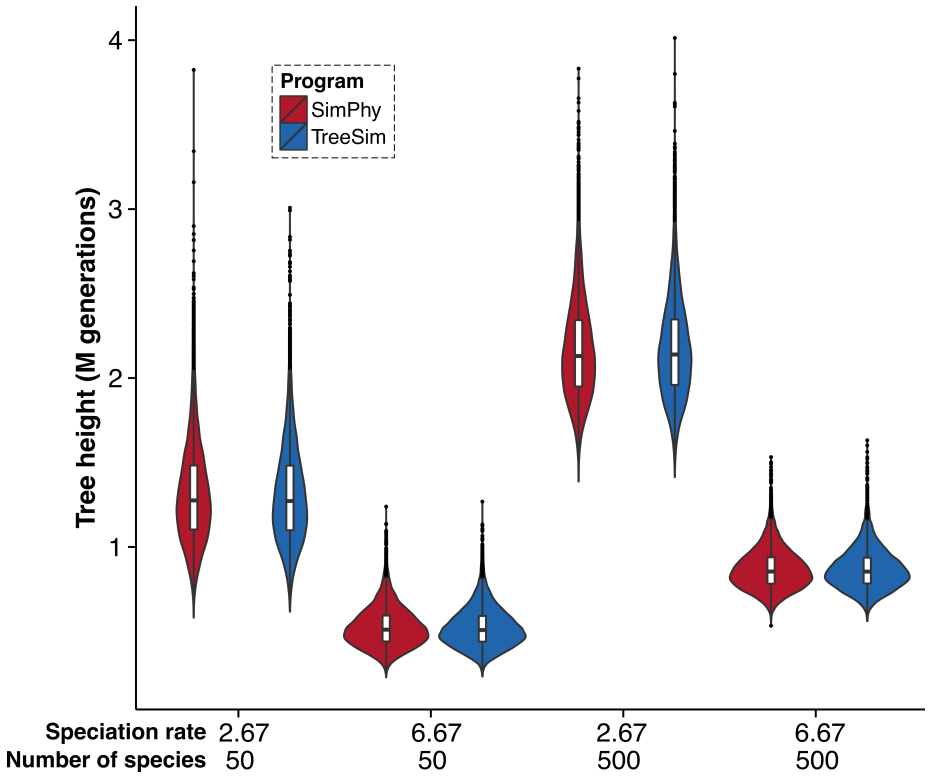

Fig. S2. Validation of *SimPhy*'s species tree simulation (variable tree height).Violin plots (kernel density curve with a boxplot inside) describe the distribution of 10000 species tree heights simulated by *SimPhy* (red) and TreeSim (blue) (speciations/1M generations).

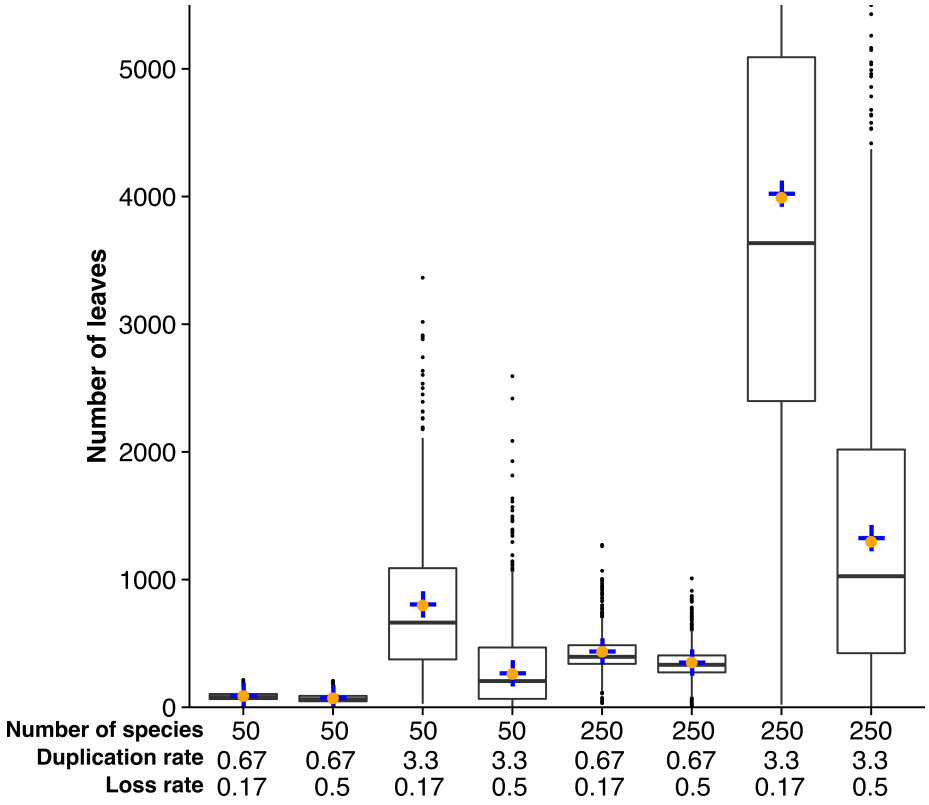

Fig. S3. Validation of *SimPhy*'s locus tree simulation under a GDL model.Boxplots describe the distribution of the number of locus tree leaves generated by 10000 simulation replicates across different duplication rates (speciations/1M generations), loss rates (relative to the duplication rate) and number of species. Expected theoretical values are indicated with a blue cross, while the observed average values are depicted with an orange dot. For representation purposes, some extreme values are not shown.

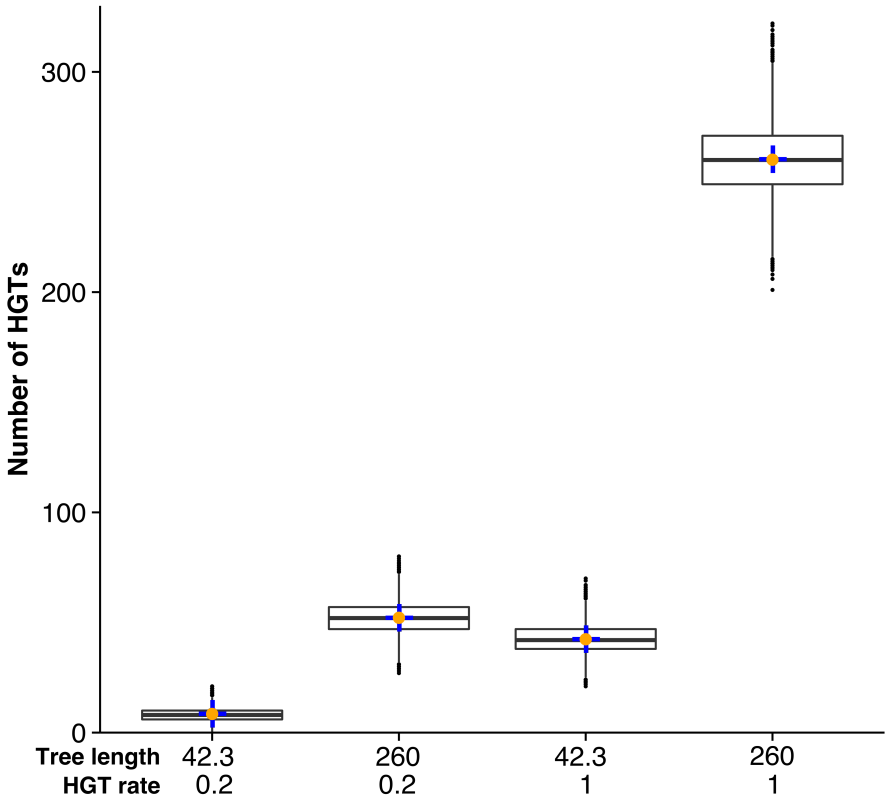

Fig. S4. Validation of *SimPhy*'s locus tree simulation under an HGT model.Boxplots describe the distribution of the number of HGT events per locus tree generated by 10000 simulation replicates across different HGT rates (transfer/1M generations), and tree lengths (M generations). Expected theoretical values are indicated with a blue cross, while the observed average values are depicted with an orange dot.

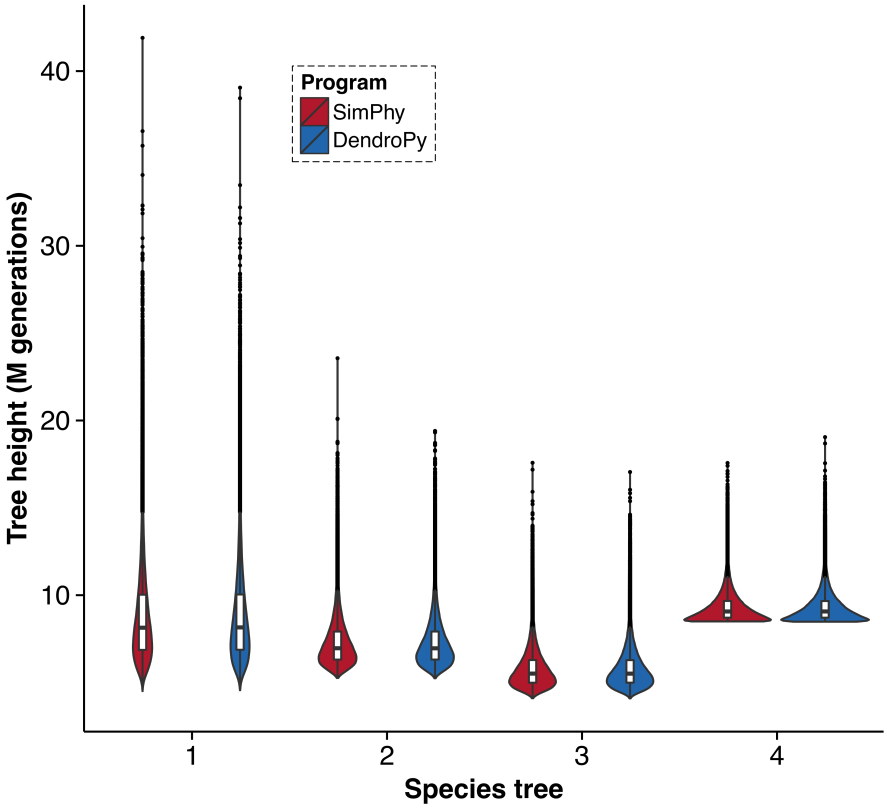

Fig. S5. Validation of *SimPhy*'s bounded multispecies coalescent sampling. Violin plots describe the distribution of 10000 gene tree heights simulated by *SimPhy* (red) and DendroPy (blue) under four different species trees.

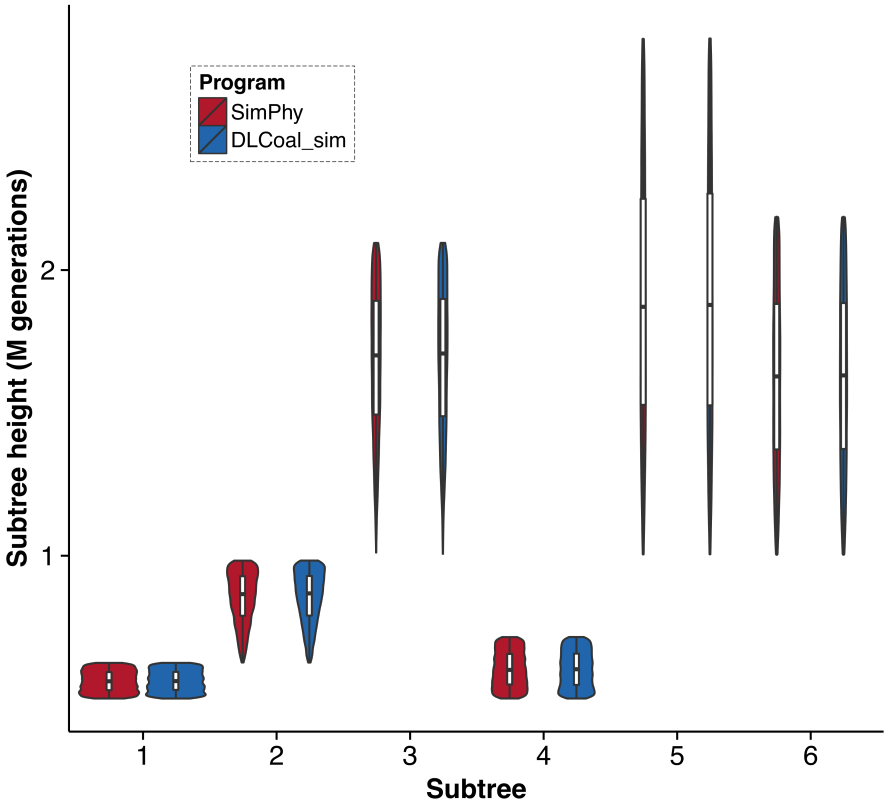

Fig. S6. Validation of *SimPhy*'s bounded multilocus coalescent sampling. Violin plots describe the distribution of 10000 gene subtree heights simulated by *SimPhy* (red) and DLCoal_sim (blue) under six different bounded locus subtrees.

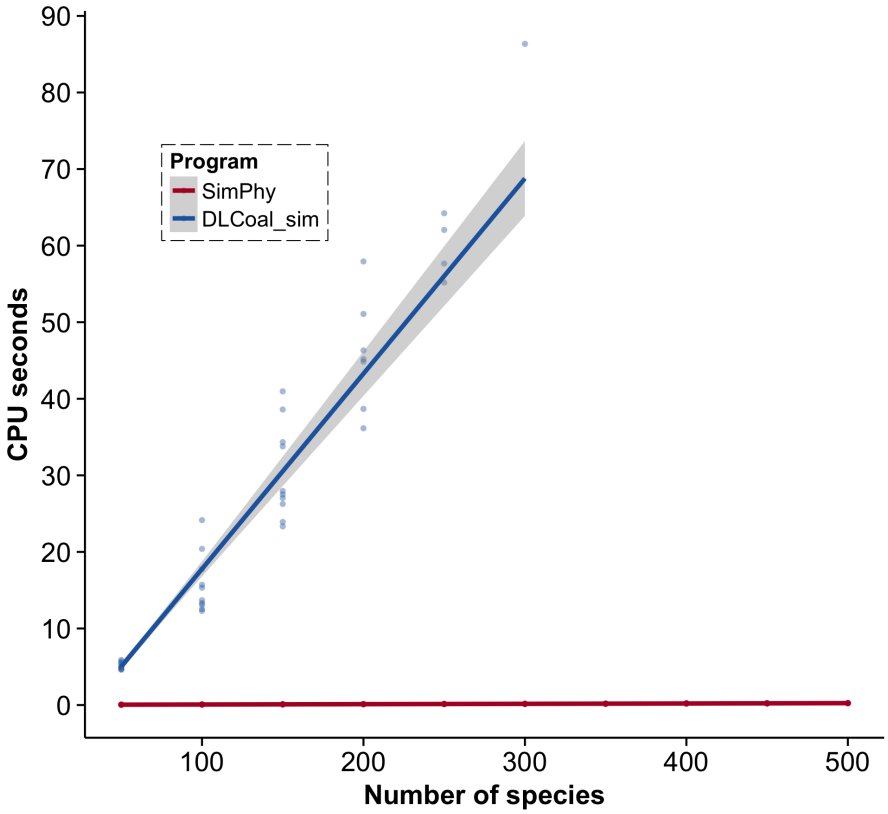

Fig. S7. Running time comparison between *SimPhy* (red) and DLCoal_sim (blue) under an ILS model. One hundred gene trees were simulated along 100 locus trees with different numbers of species. A generalized linear model with a Gamma error distribution and the identity function as link was fitted to each data series. DLCoal_sim was unable to run with more than 300 species. Note that the execution time of *SimPhy* also includes the species tree simulation, while DLCoal_sims execution time does not.

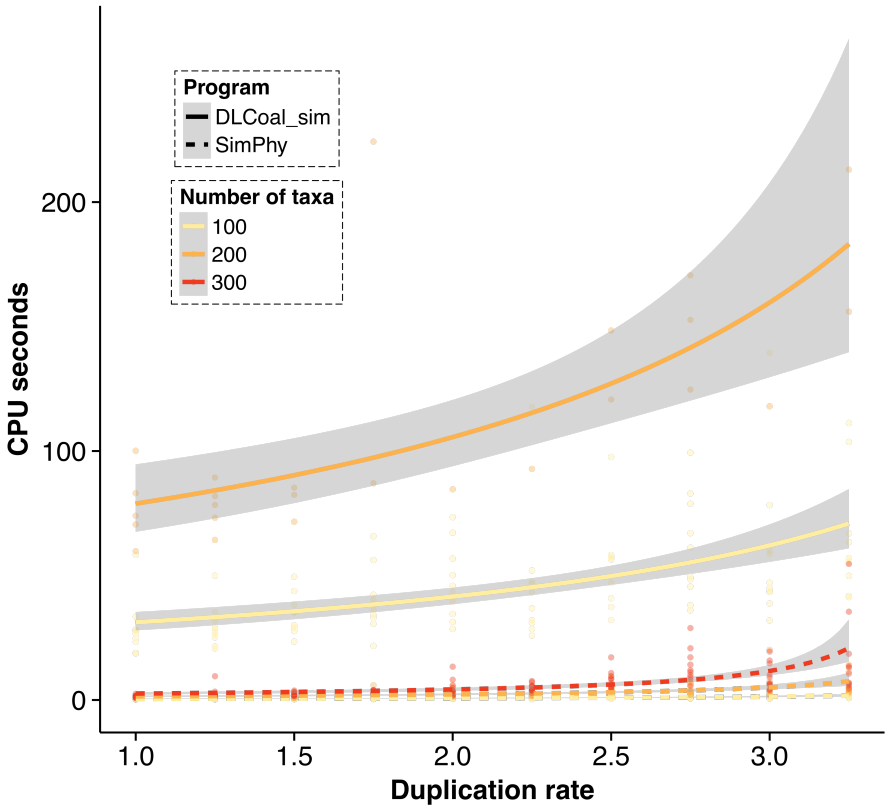

Fig. S8. Running time comparison between *SimPhy* (dashed lines) and DLCoal_sim (solid lines) under an GDL+ILS model. One hundred gene trees were simulated along 100 locus trees with different numbers of species and duplication rates (duplications/1M generations). A generalized linear model with a Gamma error distribution and the identity function as link was fitted to each data series. DLCoal_sim was unable to run with more than 300 species.