The phylogenetic dataset contains a high proportion of inapplicable codings (404/3325 = 12% of tokens), which are known to introduce error and bias to phylogenetic reconstruction (Maddison 1993; Brazeau *et al.* 2017). As such, phylogenetic search employed a new algorithm that correctly handles inapplicable data, implemented in the R package TreeSearch v0.0.8 (Brazeau *et al.* 2017). Heuristic searches were conducted using the parsimony ratchet (Nixon 1999) under equal and implied weights with a variety of concavity constants. The consensus tree presented in the main manuscript represents a strict consensus of all trees that are most parsimonious under one or more of the concavity constants (*k*) 2, 3, 4.5, 7, 10.5, 16 and 24, an approach that is known to produce higher accuracy than equal weights at any fixed level of precision (Smith 2017).

The results of phylogenetic tree search under traditional algorithms in a parsimony an Bayesian framework are included in the supplementary information; these recover the same overall relationships but differ in the placement of a small number of incidental taxa.

## Supplementary methods

Parsimony search was conducted in TNT (Goloboff *et al.* 2008) using sectorial and ratchet heuristics under equal and implied weights. We acknowledge the Willi Hennig society for their sponsorship of the TNT software.

Bayesian search was conducted in MrBayes v3.2.6 (Ronquist and Huelsenbeck 2003) using the Mk model (Lewis 2001) with a gamma parameter (lset coding=variable rates=gamma;).

Branch length was drawn from a dirichlet prior distribution, which is less informative than an exponential model (Rannala *et al.* 2012), but requires a prior mean tree length within about two orders of magnitude of the true value (Zhang *et al.* 2012); we specified the prior mean tree length to be equal to the length of the most parsimonious tree under equal weights (i.e. prset brlenspr = unconstrained: gammadir(1,0.33,1,1); αT = 1, βT = 1/(equal weights tree length / number of characters), α = *c* = 1).

Neomorphic and transformational characters (sensu Sereno 2007) were allocated to two separate partitions whose proportion of invariant characters and gamma shape parameters were allowed to vary independently (unlink shape=(all) pinvar=(all);). Neomorphic characters were not assumed to have a symmetrical transition rate – that is, the probability of the absent → present transition was allowed to differ from that of the present → absent transition, being drawn from a uniform prior (prset applyto=([number of Neomorphic partition]) symdirihyperpr=fixed(1.0);).

Two MrBayes runs were executed, each sampling eight chains for 400 000 generations, with samples taken every 250 generations. Convergence was indicated by PSRF = 1.00 and an estimated sample size of > 100 for each parameter. The first 10% of samples were discarded as burn-in, and a posterior tree topology was derived from the combined posterior sample of both runs.