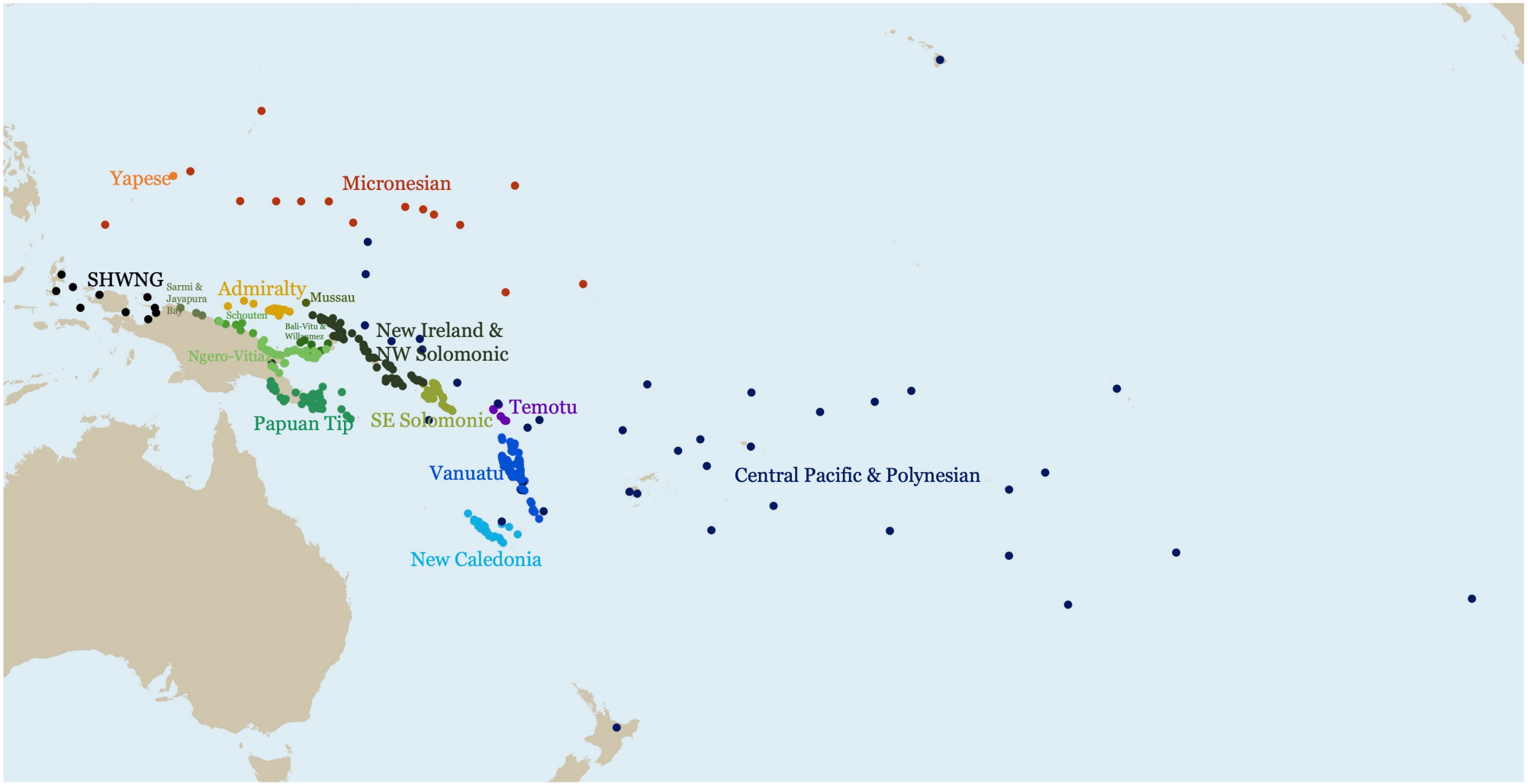
Model misspecification and overestimation of phylogenetic root age in linguistic and biological datasets

Benedict King, Aymeric Hermann, Mary Walworth, Simon Greenhill & Russell D Gray

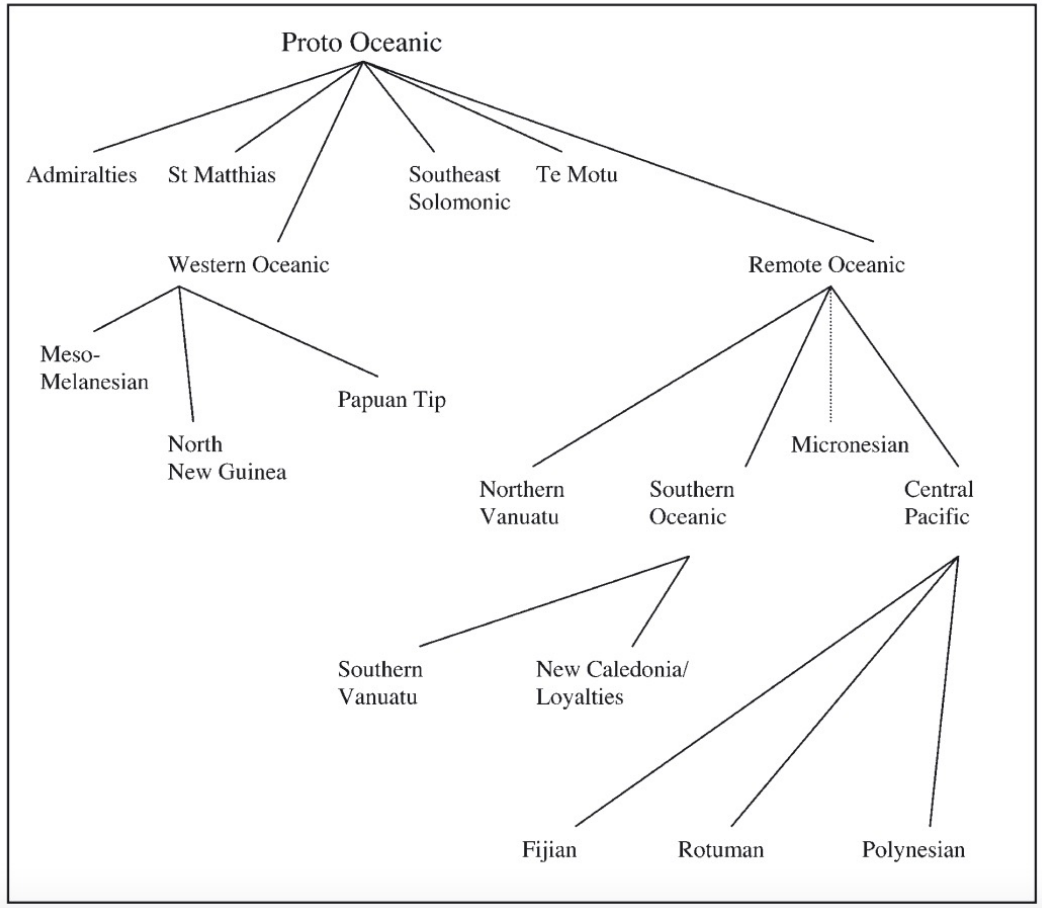
Talk outline

Phylo-purgatory

Oceanic languages



Aims



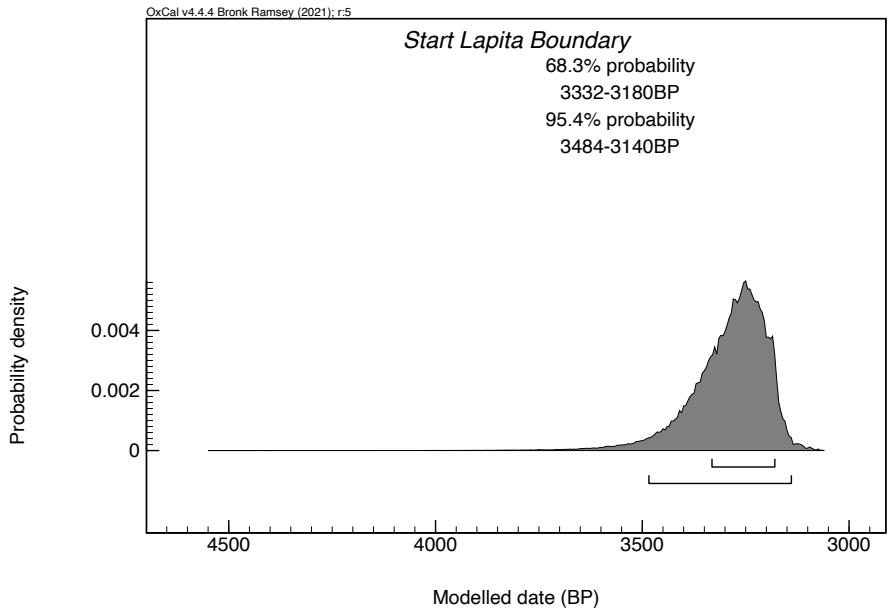
> *Despite the relatively clear evidence for Oc itself, which has been known at least since Dempwolff (1927), the higher-level subgroups of Oc have been extremely elusive*
>
> Blust (1998)

- Sort out subgrouping
- Timing and order of migration

Oceanic phylogeny set-up

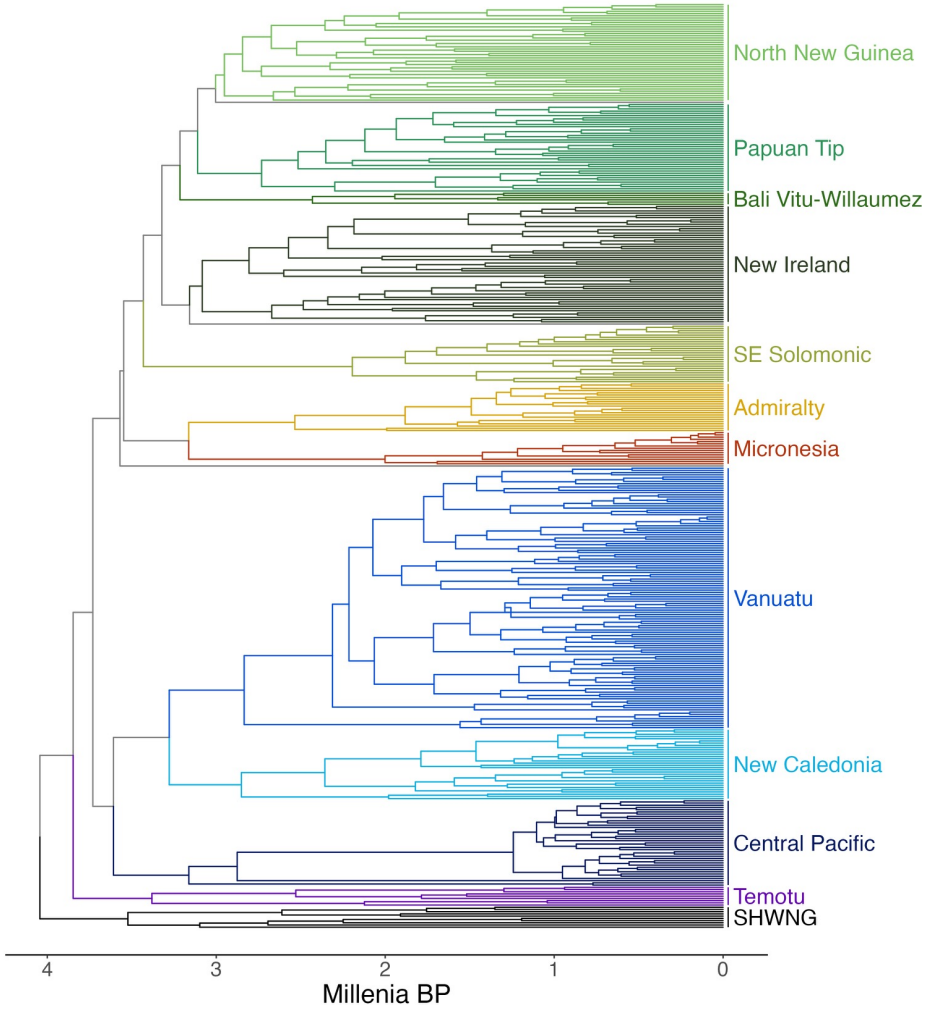
Austronesian

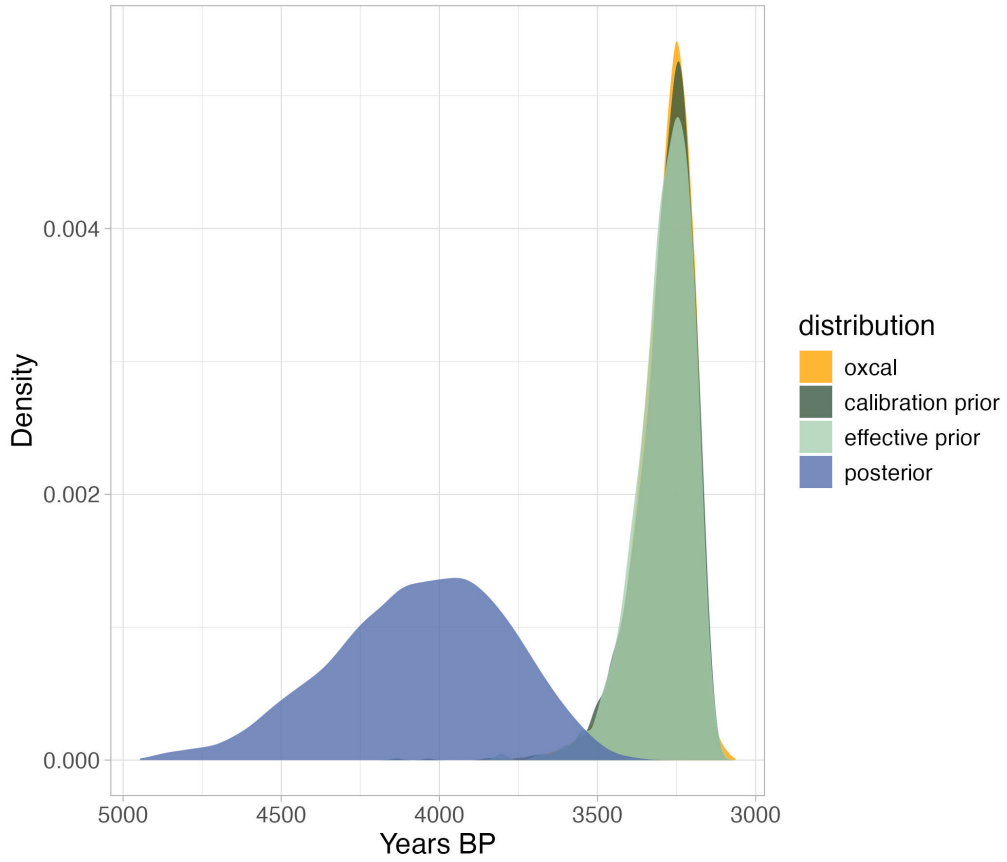
Basic Vocabulary Database



- Basic vocabulary data for 417 Oceanic languages
- 3 node calbrations based on archaeological dates
- Analysed in beast2 using the Birth-Death Skyline model

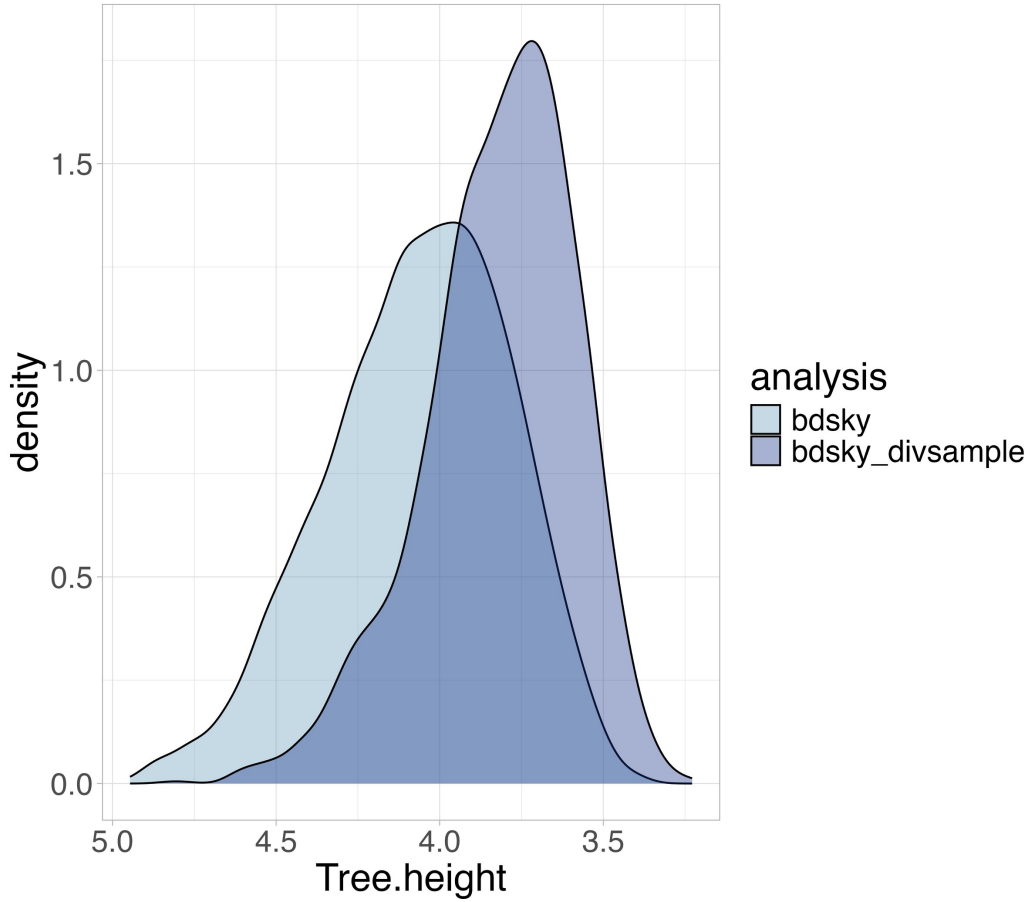
Oceanic phylogeny results





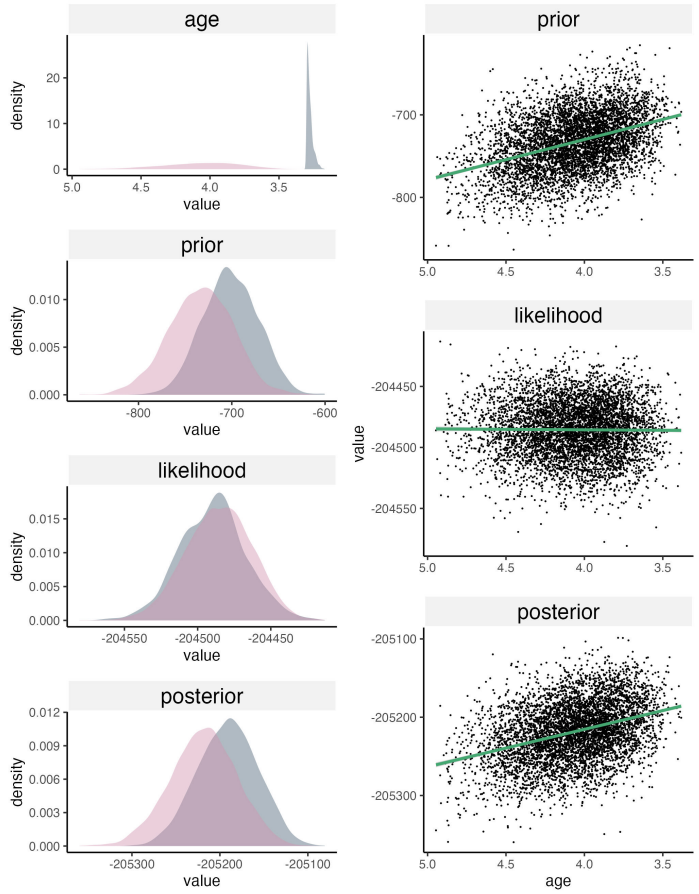
Deep Root Attraction

- Can be caused by diversified sampling
- Often the effective prior does not match the calibration density
- Does not explain Oceanic results

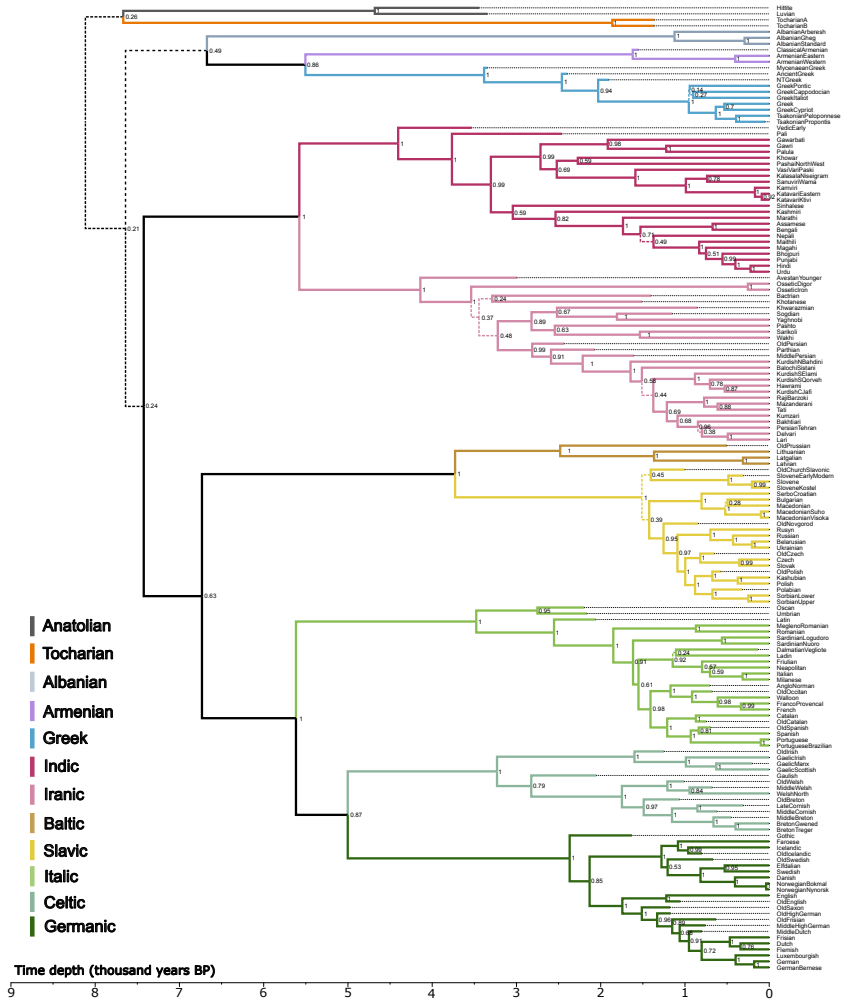


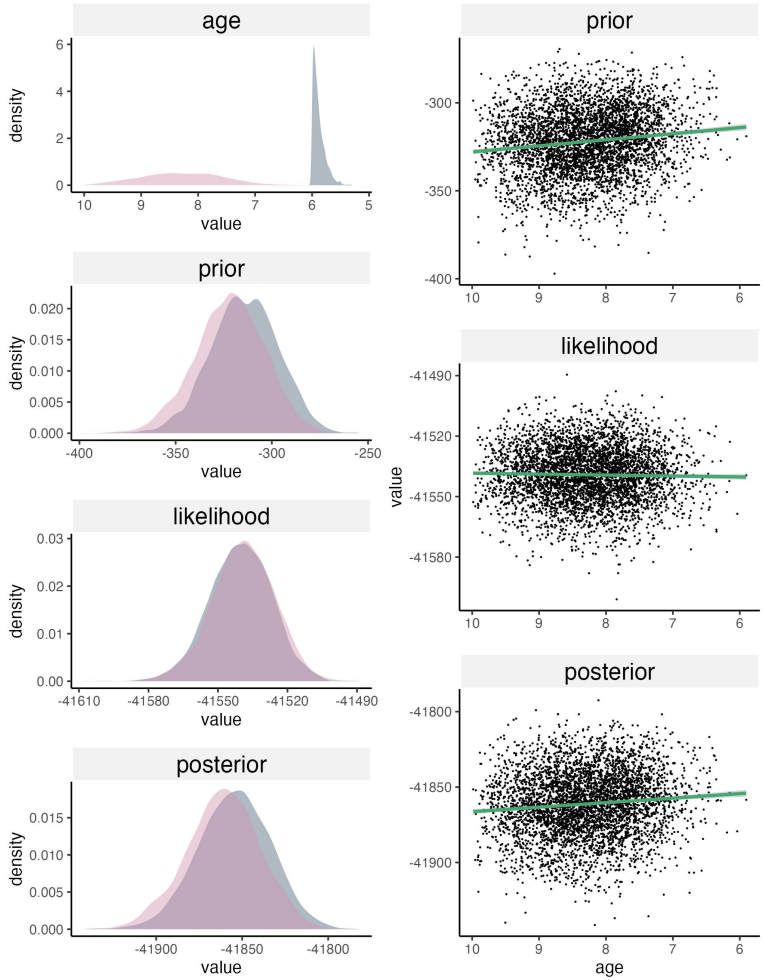
Does the data support older ages?

- Constrain a maximum age of 3300 BP
- Likelihood is identical regardless of age

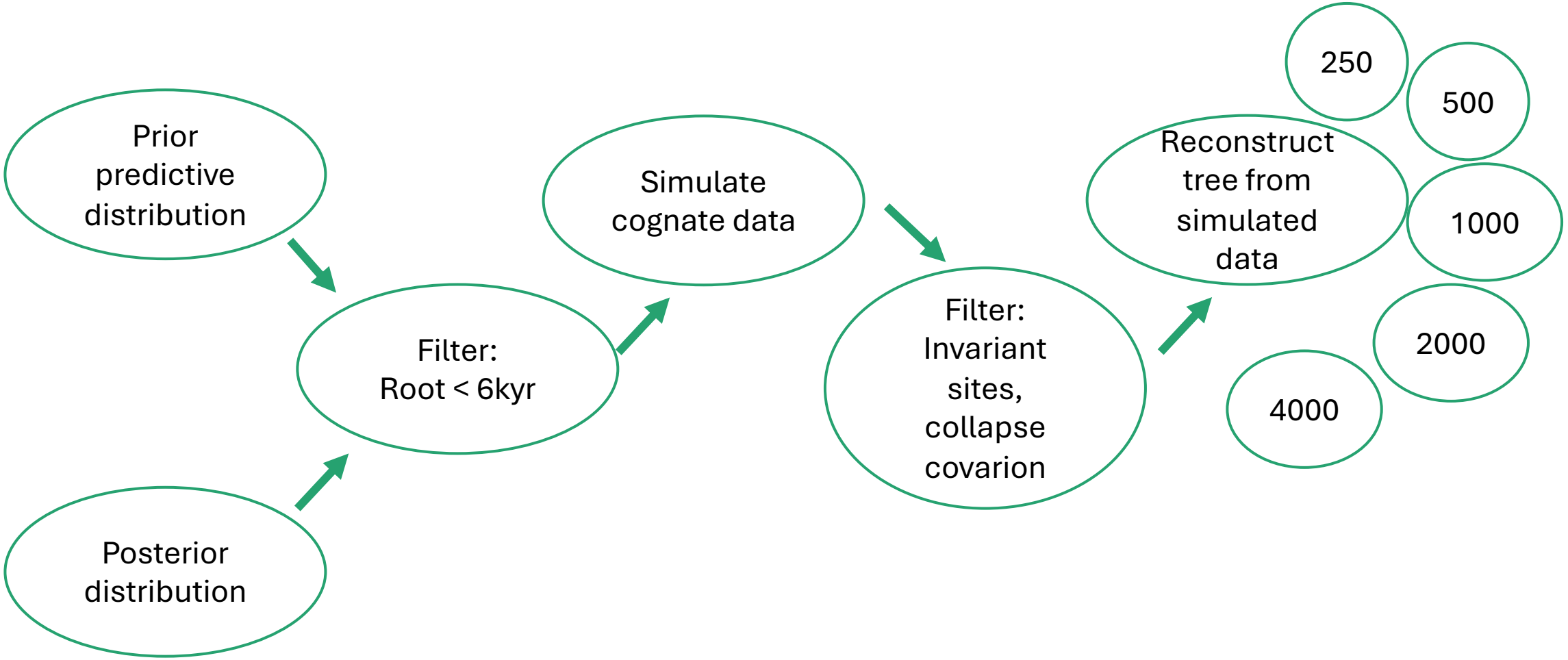


What about other language families?

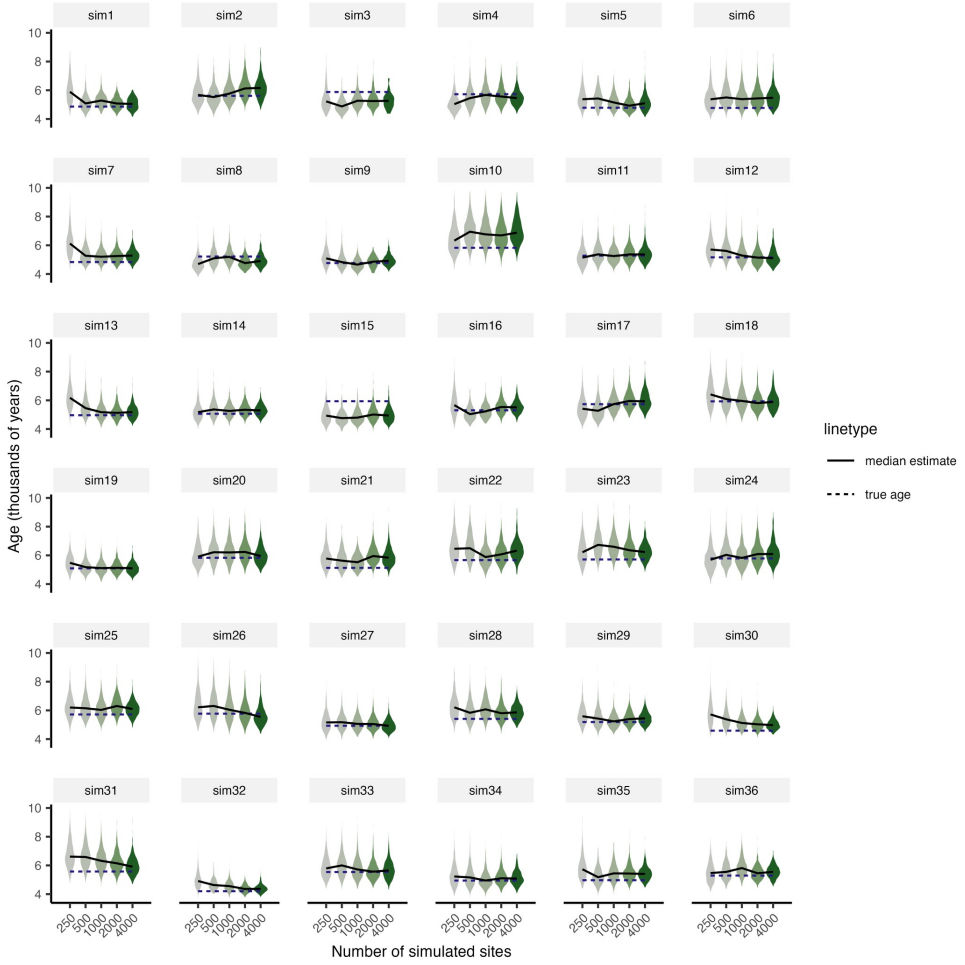


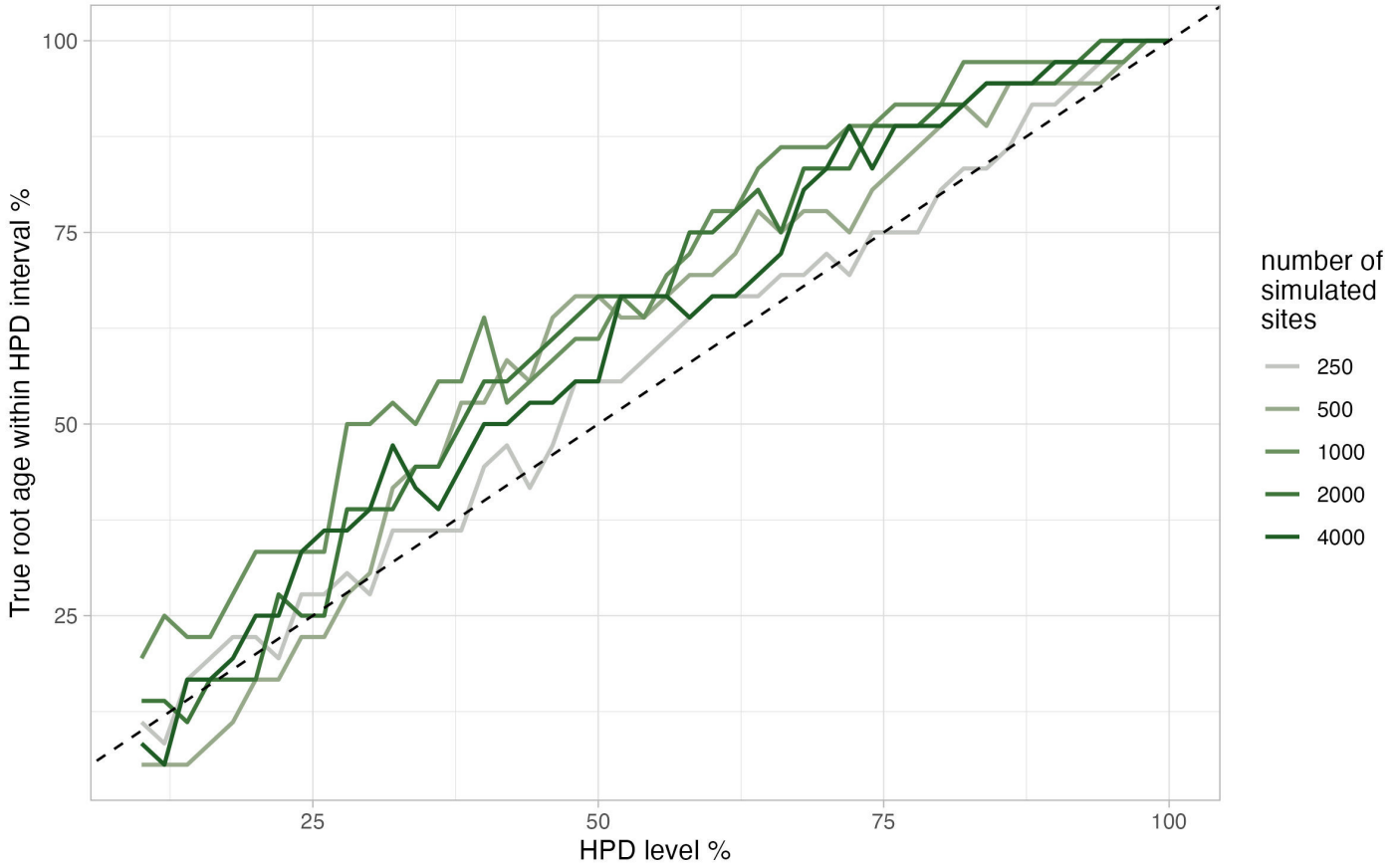


Can the methods get young trees?

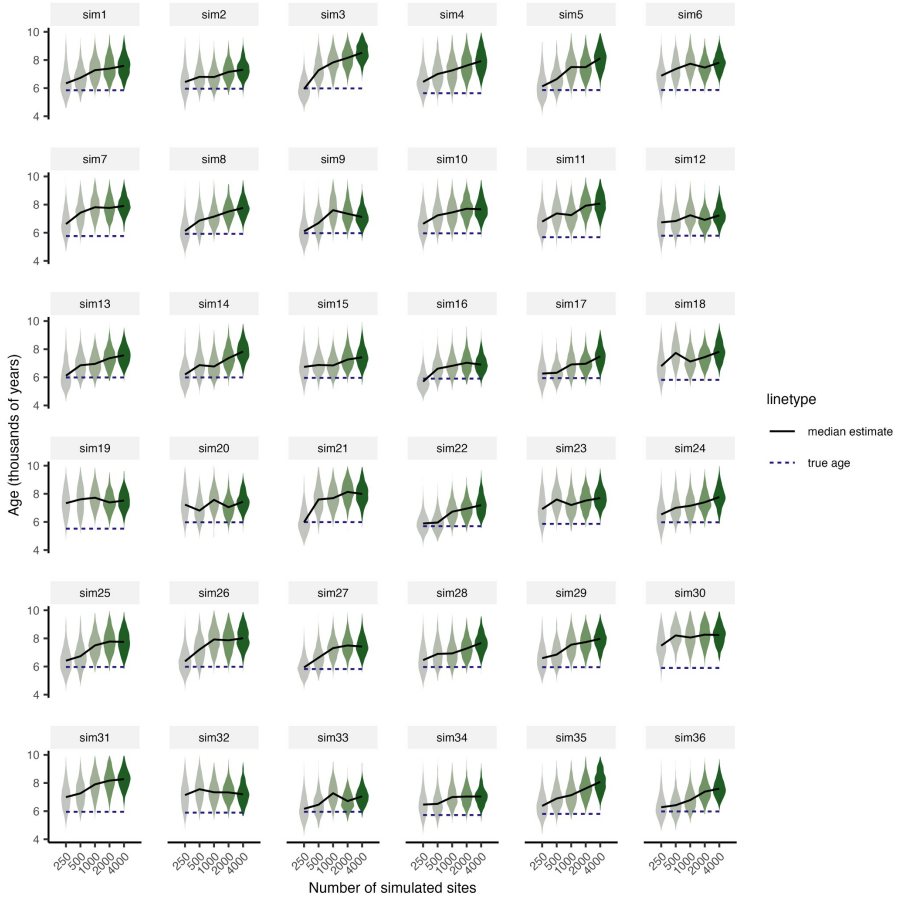


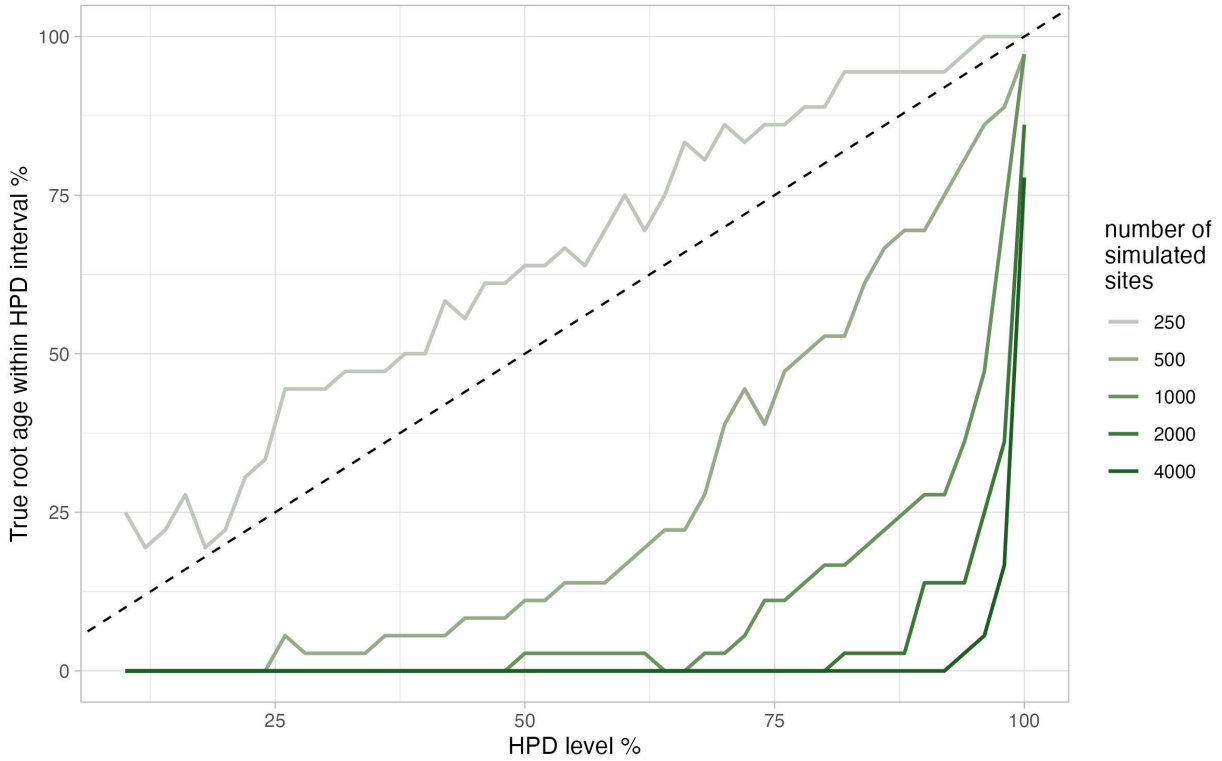
Results – trees from the prior



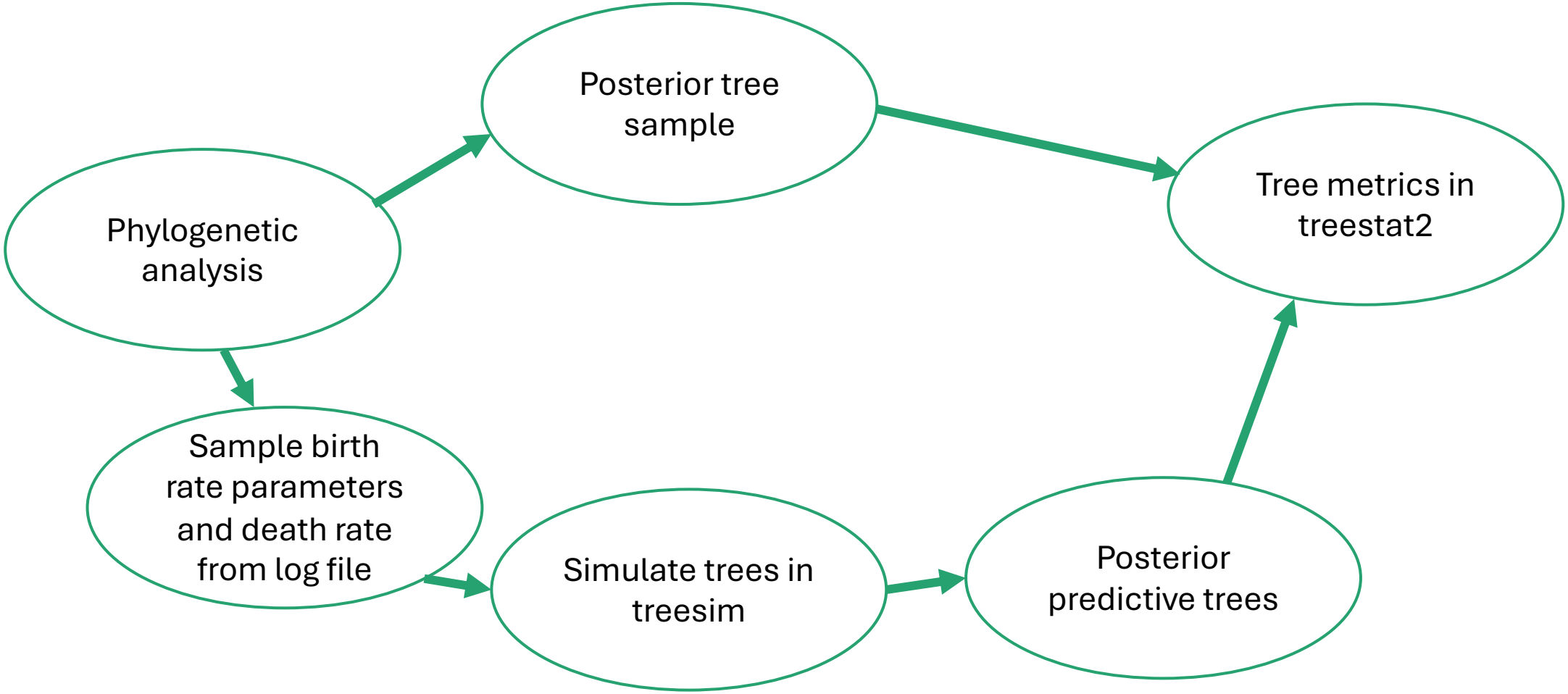


Results – trees from the posterior



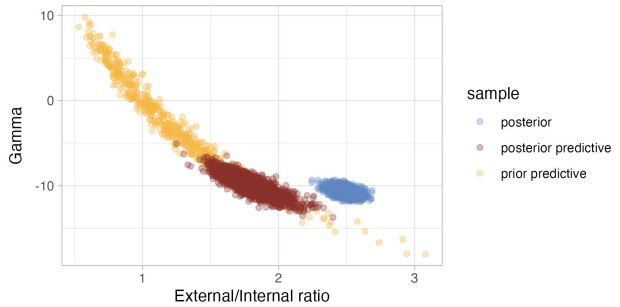


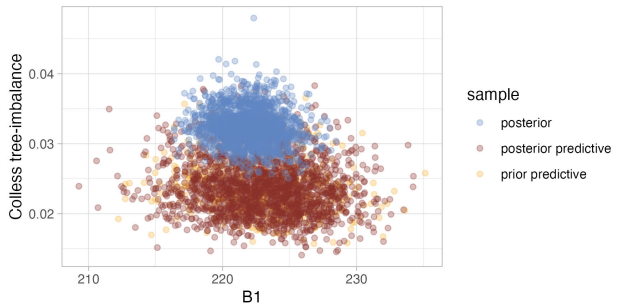
Posterior tree fit

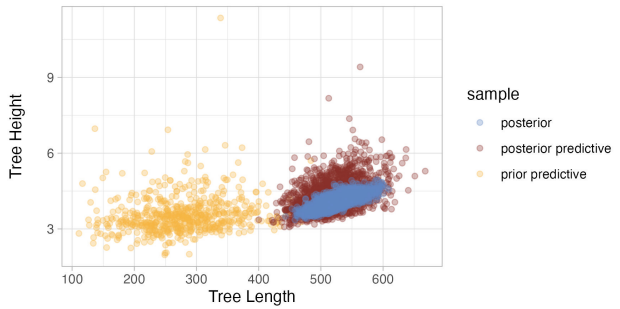


Posterior predictive results

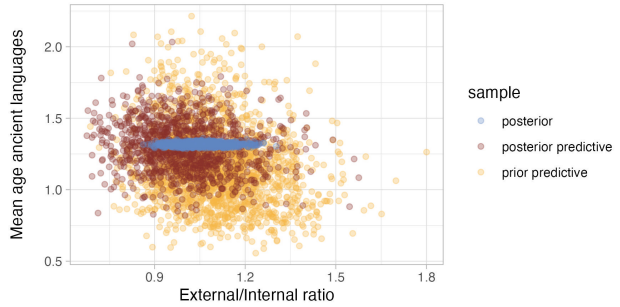
Oceanic

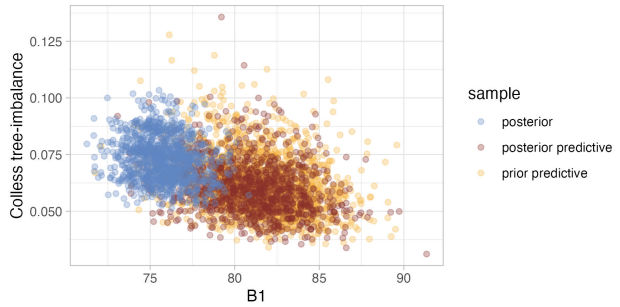


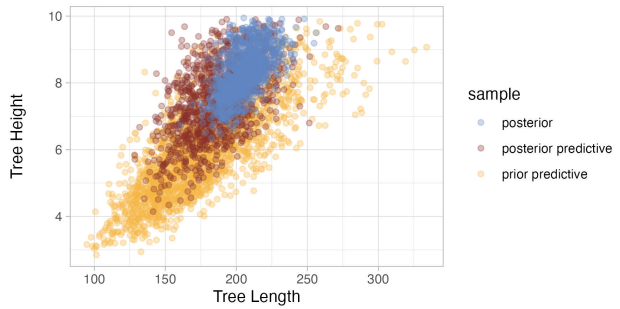




IE







Is this a linguistics problem?

"*It must be the covarion model*"

"*It's the way the cognates are collected*"

"*Biological models don't work in Linguistics*"

Do biological models work in Biology?

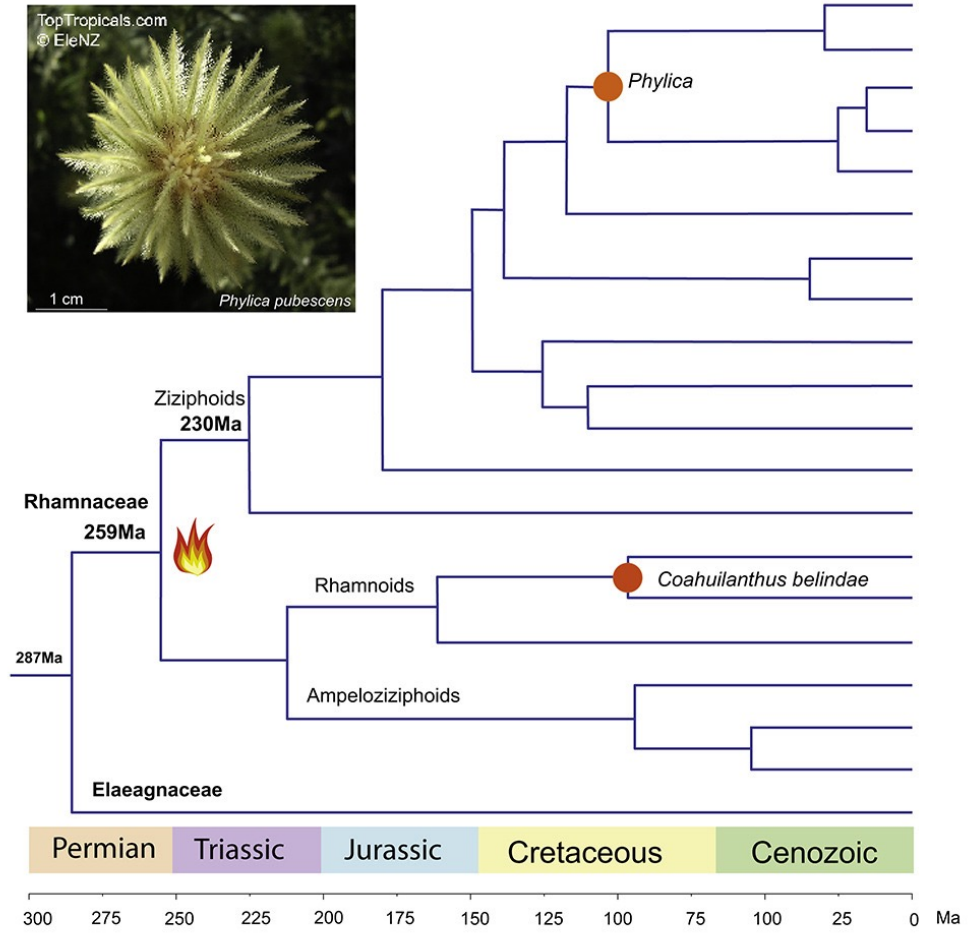
iScience

CellPress
OPEN ACCESS

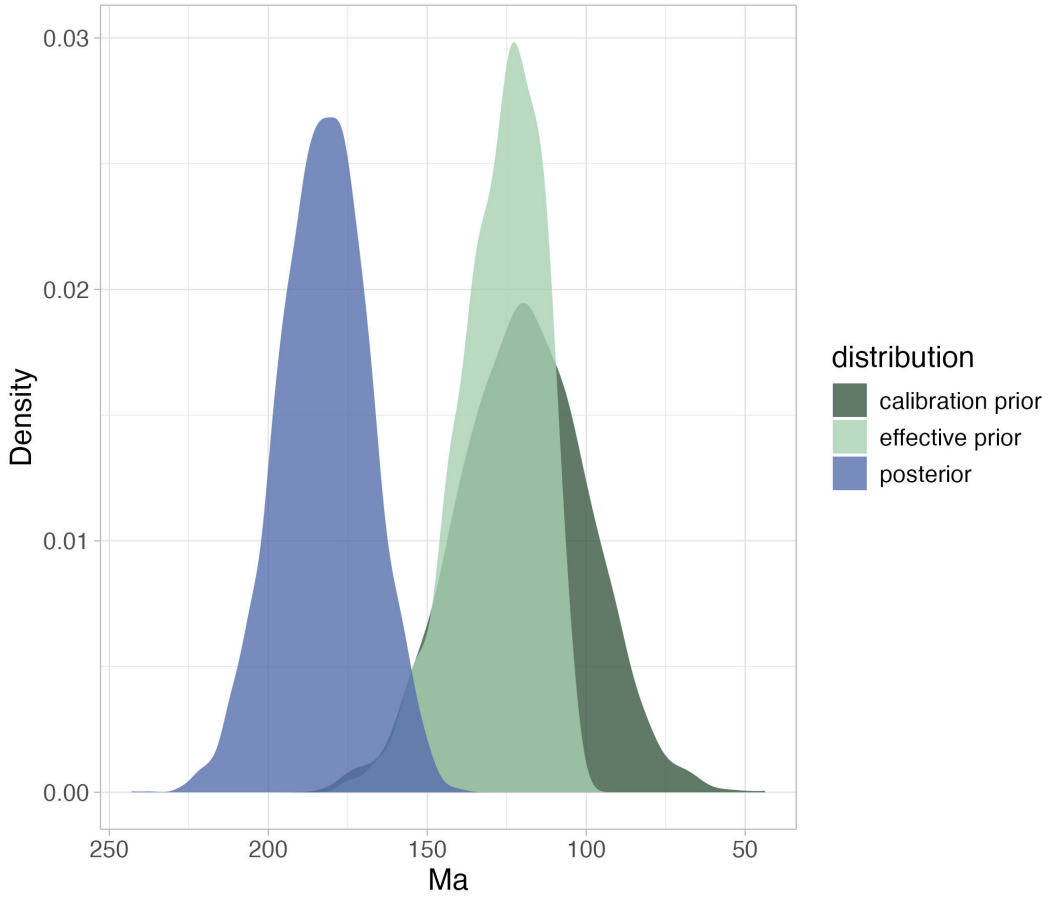
Article

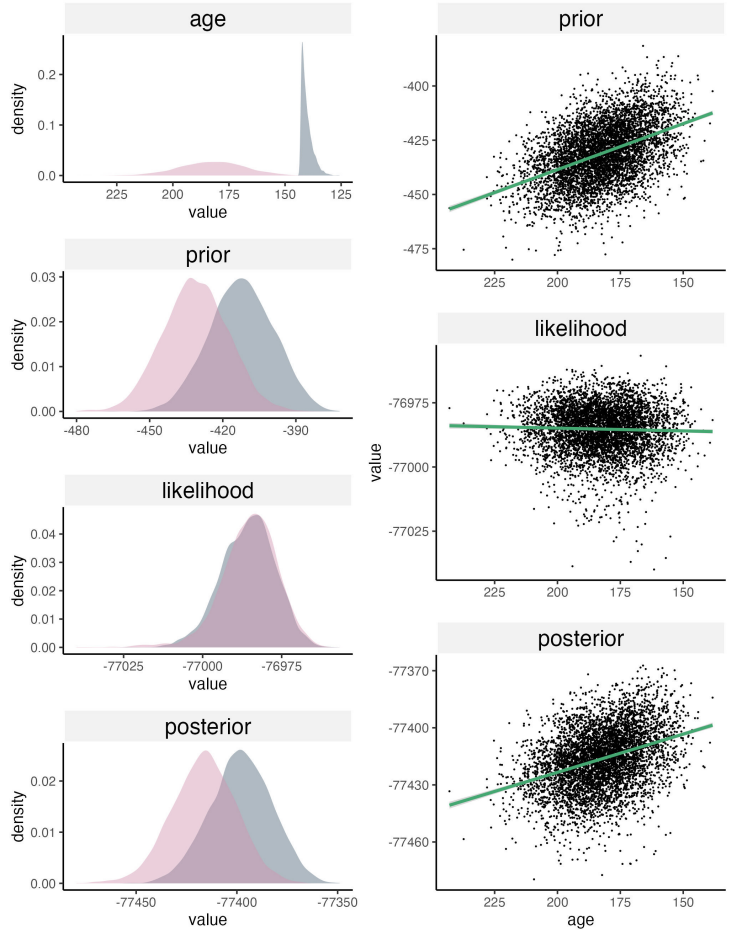
Ancient Rhamnaceae flowers impute an origin for flowering plants exceeding 250-million-years ago

He and Lamont 2022

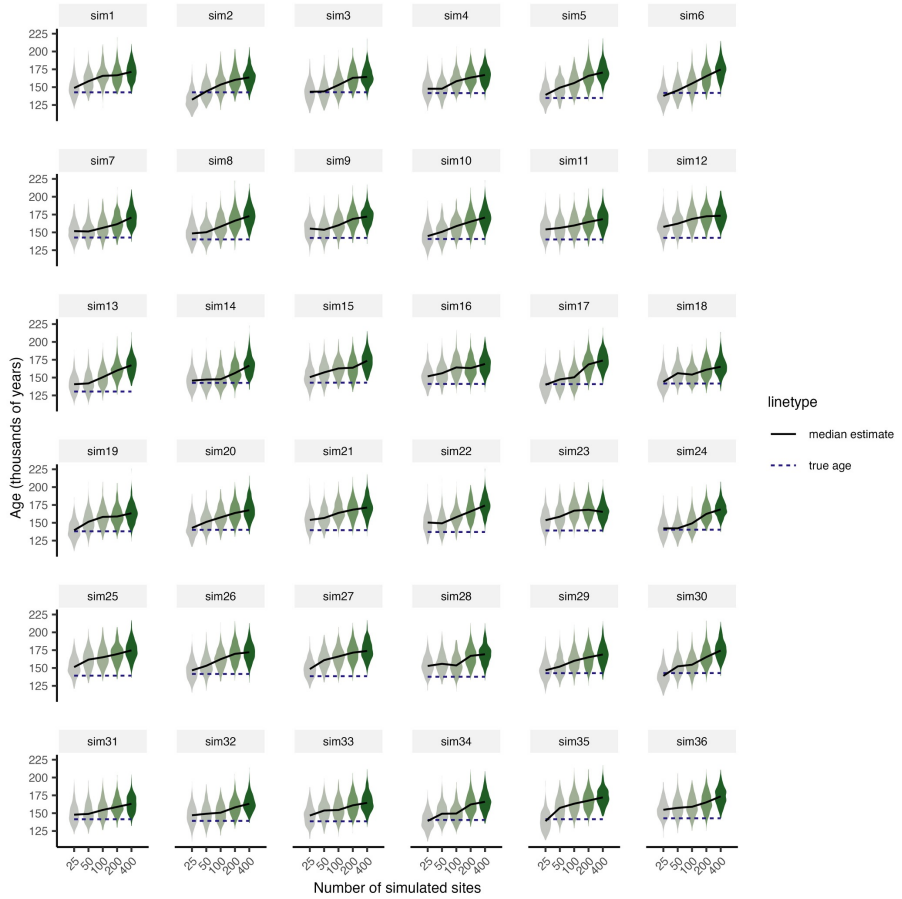


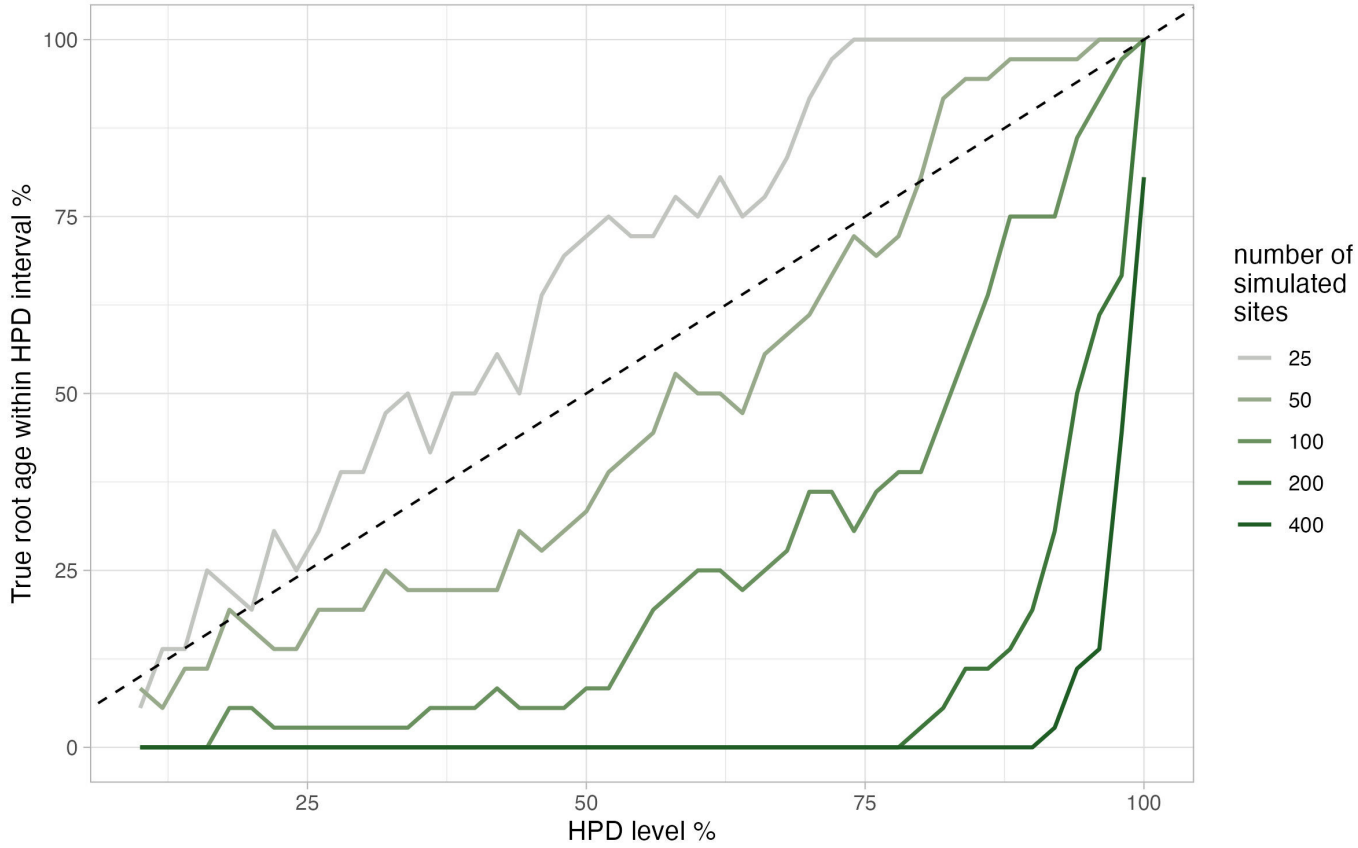
Ancient dates and flat likelihood, again



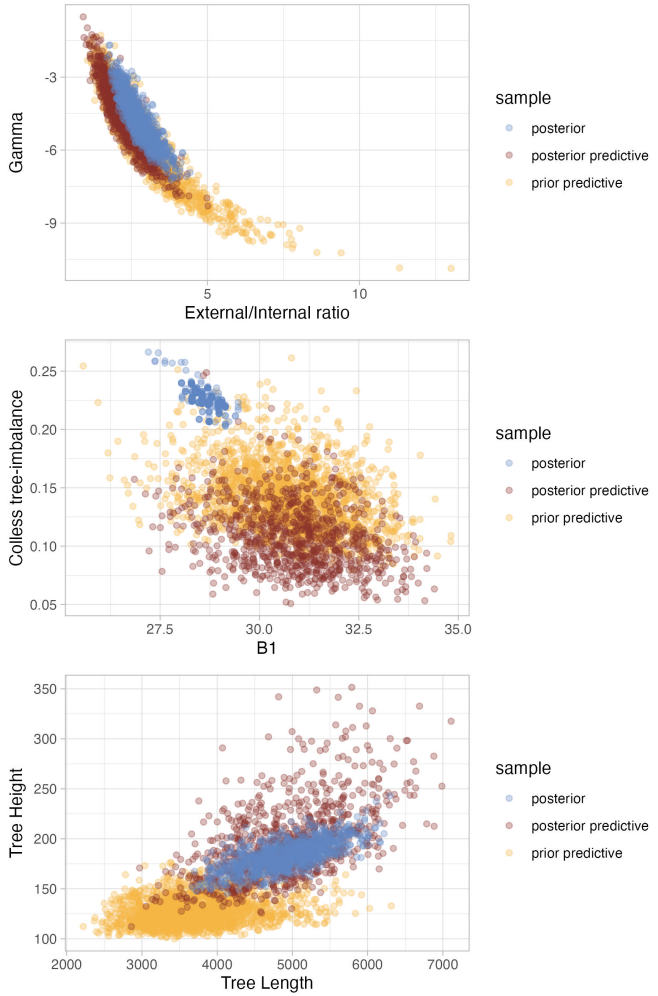


Simulations on posterior trees

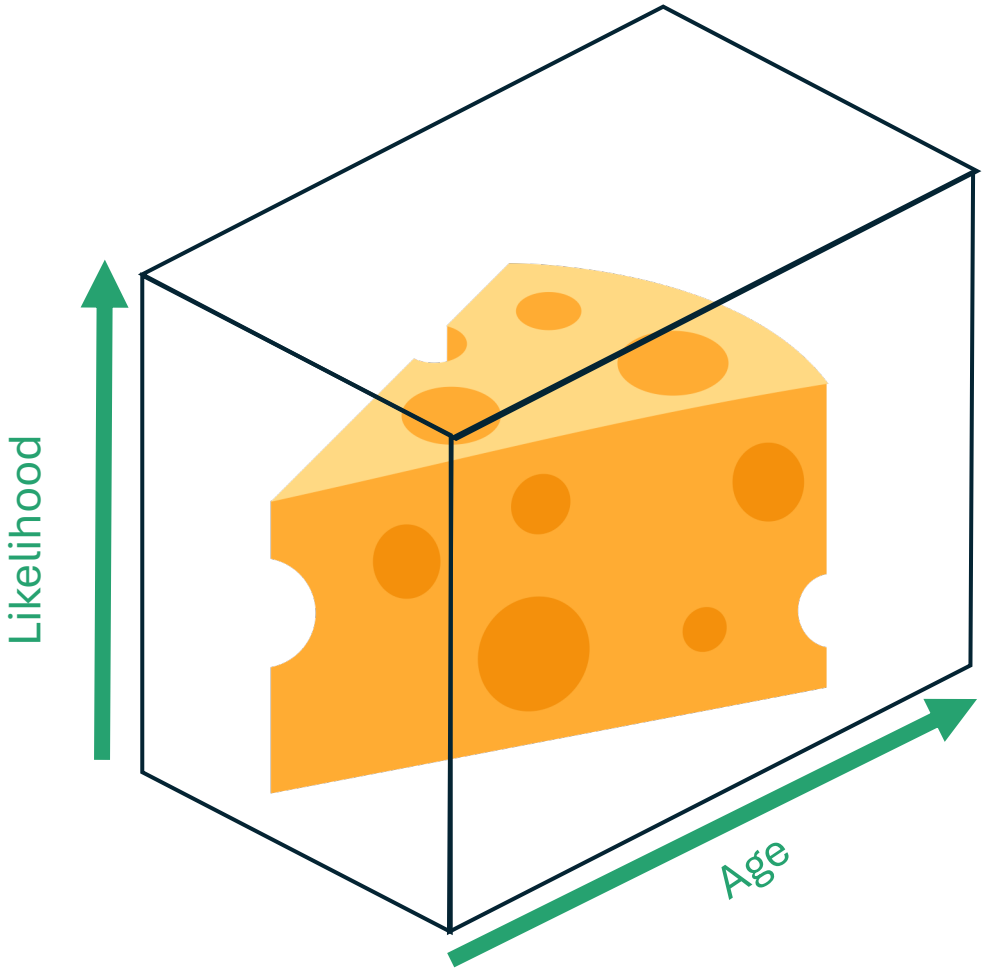


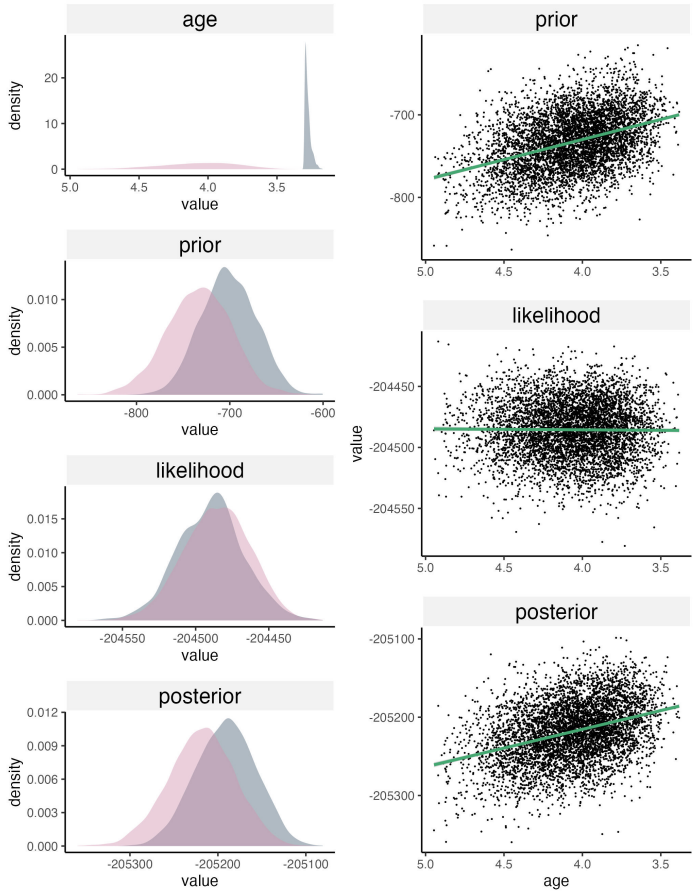


Rhamnaceae posterior predictive

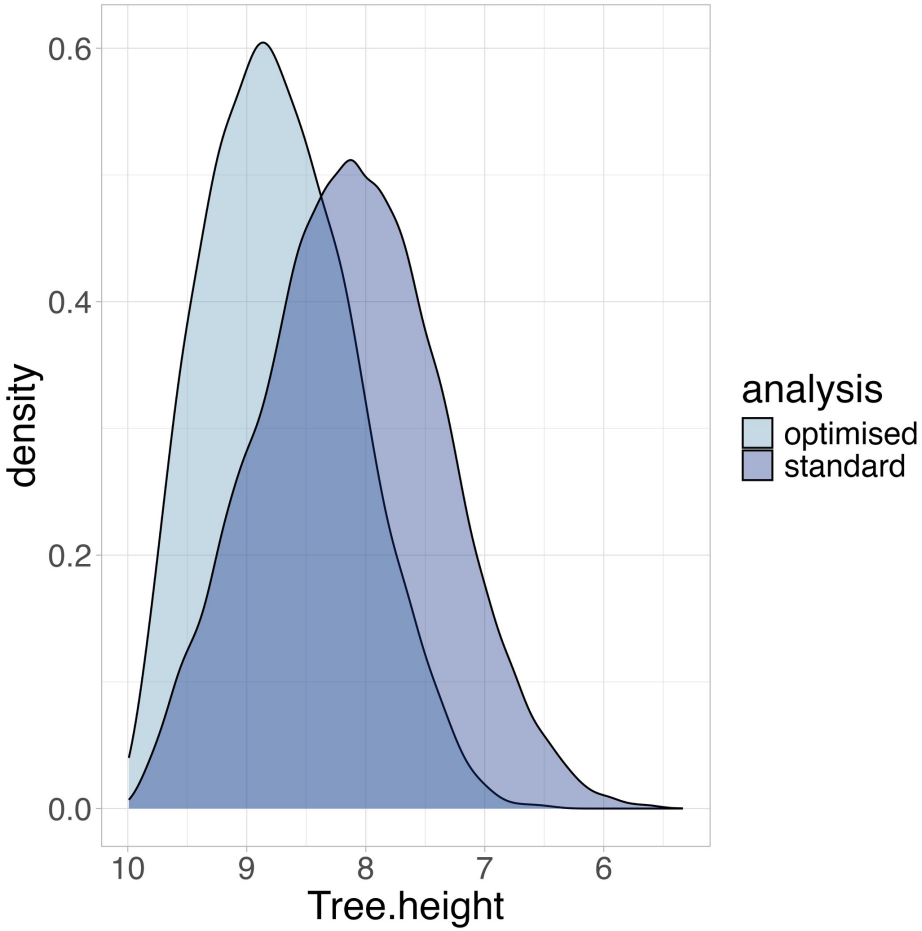


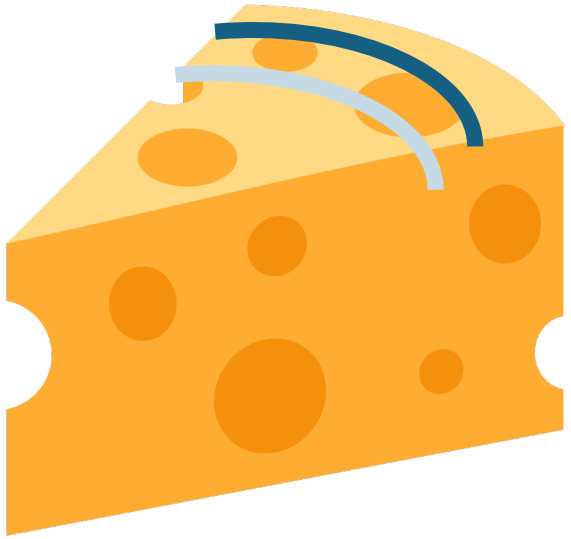
The expanding likelihood wedge

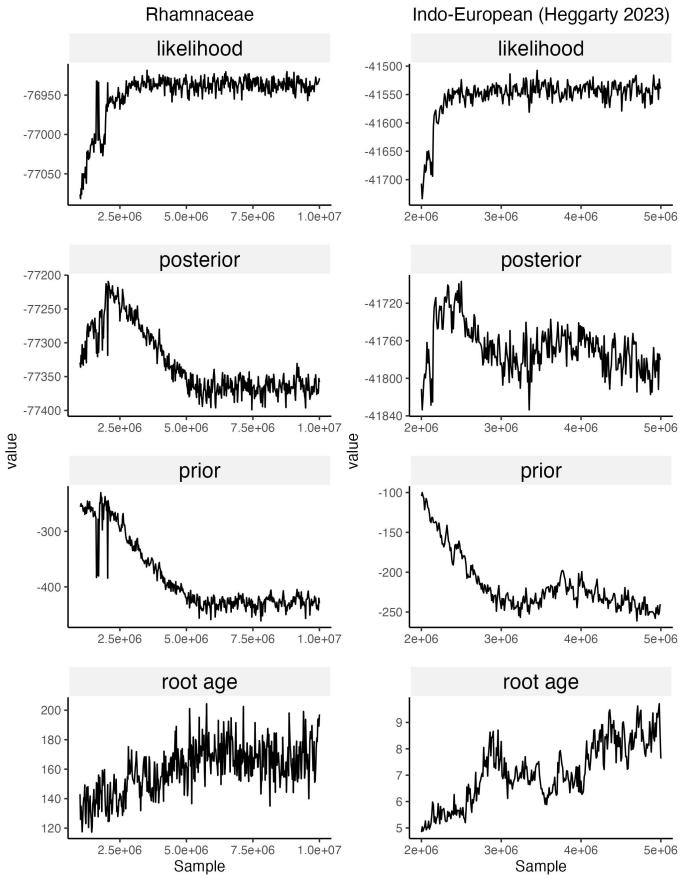




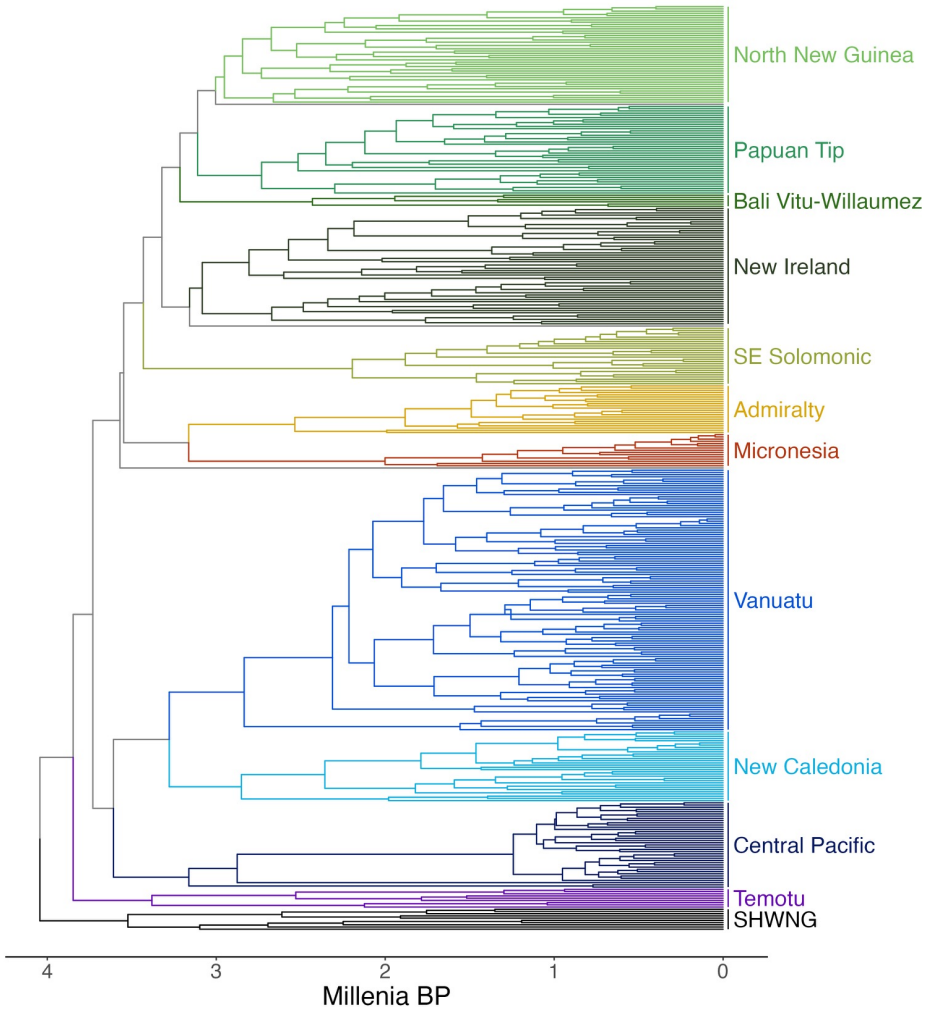
Sensitivity to software version



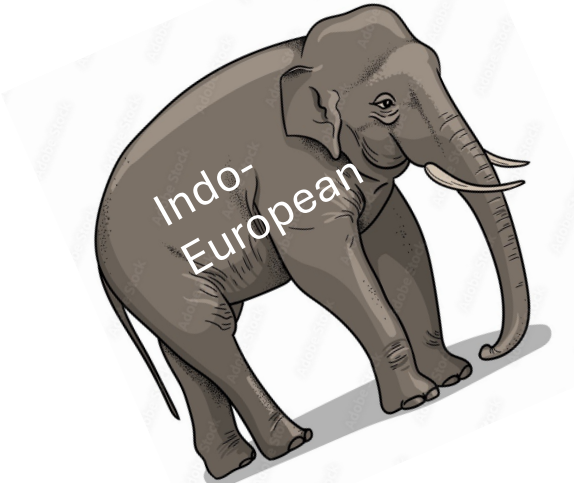




What can we say about Oceanic



- Results so far suggest that root age could be over-estimated
- Cannot draw conclusions about the early migration
- Alternative tree prior (CLADS) has not been successful



Conclusions

- Model mis-specification of tree priors could render molecular clocks uninformative
- This occurs across different calbration set-ups (tip and node) and data type (linguistic, DNA)
- Not clear how widespread