

Total-evidence with fossilised birth-death model reveals the recent crown radiation of penguins

David Welch

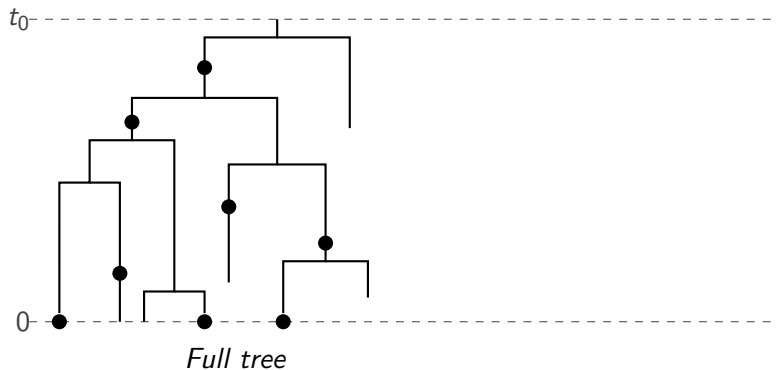
with Alexandra Gavryushkina, Tracy Heath, Daniel Ksepka,
Tanja Stadler, Alexei Drummond

Computational Evolution Group
University of Auckland
Funding from Marsden Fund, NZ
#SMBE15, Vienna
On Twitter: @PhyDyn

Headlines

- ▶ Total Evidence uses **molecular data**, **morphological data** and **fossilisation times** to estimate model parameters including dated phylogeny in a Bayesian framework.
- ▶ **Fossilised birth-death process** models speciation and fossil sampling.
- ▶ One sample can be the direct ancestor of another: these are called **sampled ancestors**.
- ▶ The crown age of penguins (i.e., tmrca of all extant species) 10-15m yr (cf other estimates of 20m+)
- ▶ Information comes from including all stem fossils.
- ▶ The method is implemented in BEAST2 (packages **SA** and **morph-models**)

Fossilised birth-death model Stadler 2010, Heath et al 2014



Time of origin t_{or}

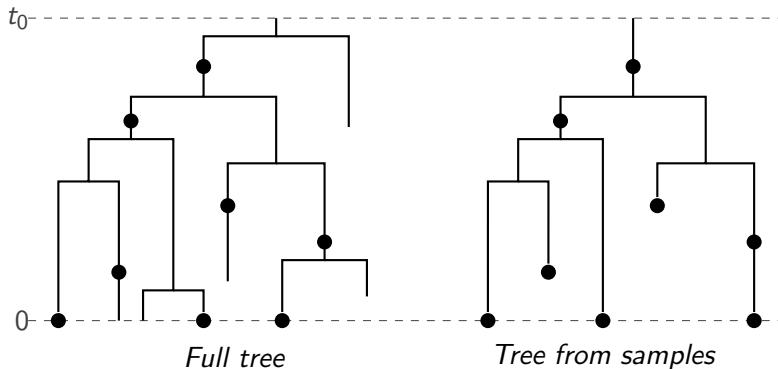
Birth rate λ

Death rate μ

Sampling rate ψ

Probability of sampling at present ρ

Fossilised birth-death model Stadler 2010, Heath et al 2014



Time of origin t_{or}

Birth rate λ

Death rate μ

Sampling rate ψ

Probability of sampling at present ρ

Use a different parametrisation for fossilised birth death model

Instead of λ , μ , and ψ common to use:

net diversification rate $d = \lambda - \mu$

turnover rate $\nu = \frac{\mu}{\lambda}$

sampling proportion $s = \frac{\psi}{\mu + \psi}$

Use Lewis Mk model for morphological trait evolution

Lewis 2001

Lewis Mk model is like Jukes Cantor: fixed number of possible traits, all substitutions at same rate.

Either: Fix number of traits to be the same for each column

Or: Partition columns into groups of columns that share same number of observed traits

Account for fact that data omits constant characters (Mkv model)
or not

Gamma variation across sites **or** no rate variation across sites

Strict clock **or** relaxed clock

Use a Bayesian inference framework to estimate all model parameters

Use Markov chain Monte Carlo to sample from the posterior distribution:

$$\begin{aligned} f(\mathcal{G}, \theta, \eta | D, \tau) &\propto f(D, \tau | \mathcal{G}, \theta, \eta) f(\mathcal{G}, \theta, \eta) \\ &= f(D | \mathcal{G}, \theta) f(\tau | \mathcal{G}) f(\mathcal{G} | \eta) f(\eta) f(\theta), \end{aligned}$$

where

- D morphological and molecular data
- τ fossil age ranges
- \mathcal{G} dated phylogeny
- η all fossilised birth death process parameters
- θ all substitution and clock model parameters.

All extant penguins (19 species)

36 fossil penguins

Morphological data for all taxa (extant and fossil)

Date ranges for all fossils

Sequences data from extant species (5 loci, 8145bp in total)

Fix sampling probability at present $\rho = 1$ as all modern penguins sampled.

Try to find best trait model by maximising marginal likelihood

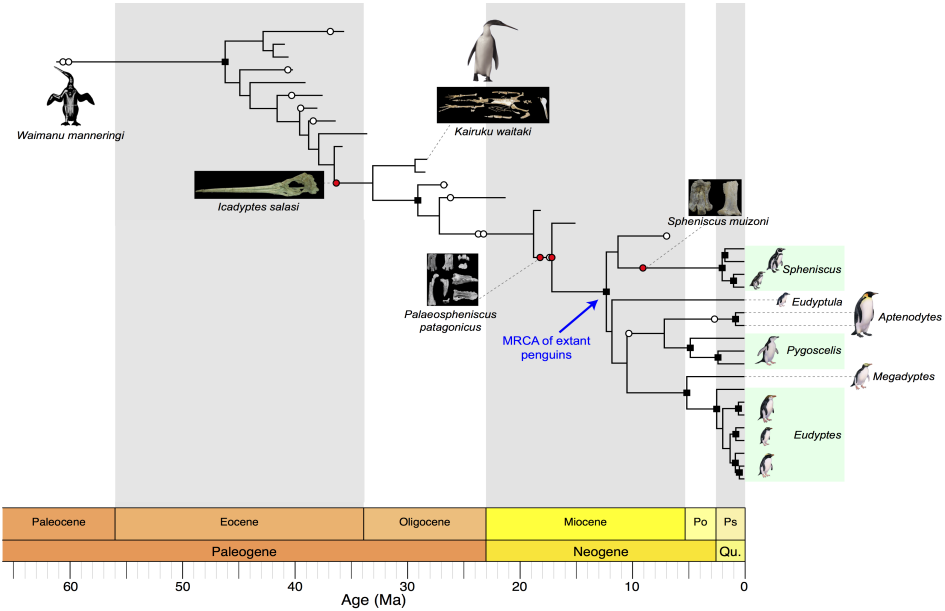
| Partition traits | Gamma variation | Missing constant columns | Clock | Priors on Parameters | Log Marginal likelihood |
|------------------|-----------------|--------------------------|---------|----------------------|-------------------------|
| | | Mk | Strict | d, ν, s | -2727 |
| | yes | Mk | Strict | d, ν, s | -2715 |
| yes | | Mk | Strict | d, ν, s | -2006 |
| yes | yes | Mk | Strict | d, ν, s | -1968 |
| yes | | Mk | Strict | λ, μ, ψ | -2005 |
| yes | | Mkv | Strict | d, ν, s | -1996 |
| yes | yes | Mkv | Strict | d, ν, s | -1854 |
| yes | yes | Mkv | Relaxed | d, ν, s | -1845 |

Try to find best trait model by maximising marginal likelihood

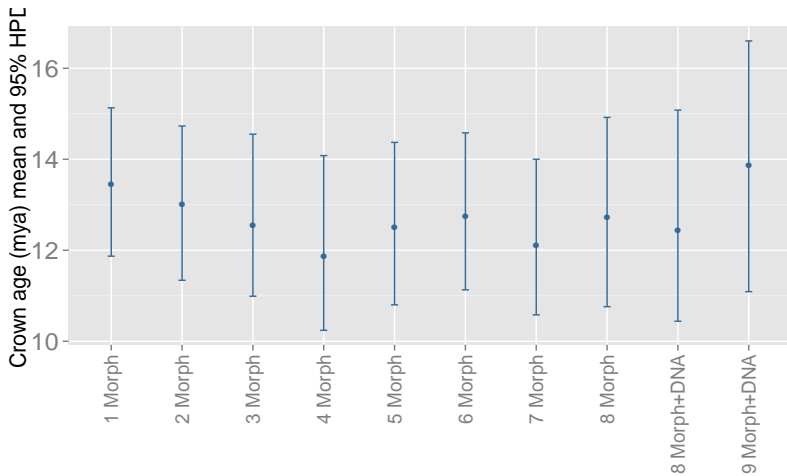
| Partition traits | Gamma variation | Missing constant columns | Clock | Priors on Parameters | Log Marginal likelihood |
|------------------|-----------------|--------------------------|---------|----------------------|-------------------------|
| | | Mk | Strict | d, ν, s | -2727 |
| | yes | Mk | Strict | d, ν, s | -2715 |
| yes | | Mk | Strict | d, ν, s | -2006 |
| yes | yes | Mk | Strict | d, ν, s | -1968 |
| yes | | Mk | Strict | λ, μ, ψ | -2005 |
| yes | | Mkv | Strict | d, ν, s | -1996 |
| yes | yes | Mkv | Strict | d, ν, s | -1854 |
| yes | yes | Mkv | Relaxed | d, ν, s | -1845 |

But all models produce similar trees and ages.

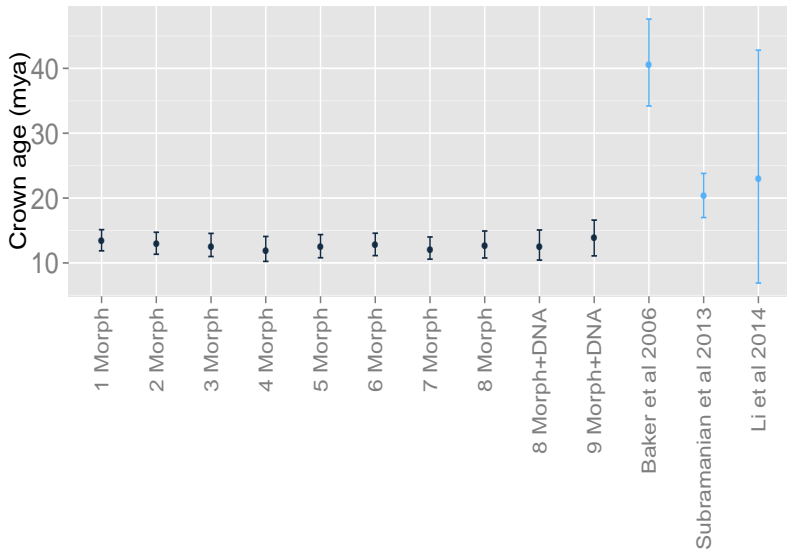
Maximum clade credibility tree (with sampled ancestors)



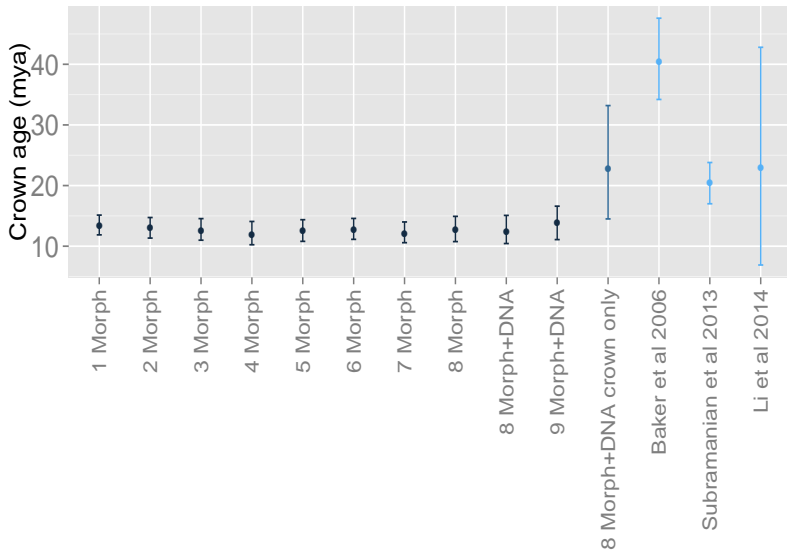
The tmrca of today's penguins: crown radiation time



The tmrca of today's penguins: crown radiation time



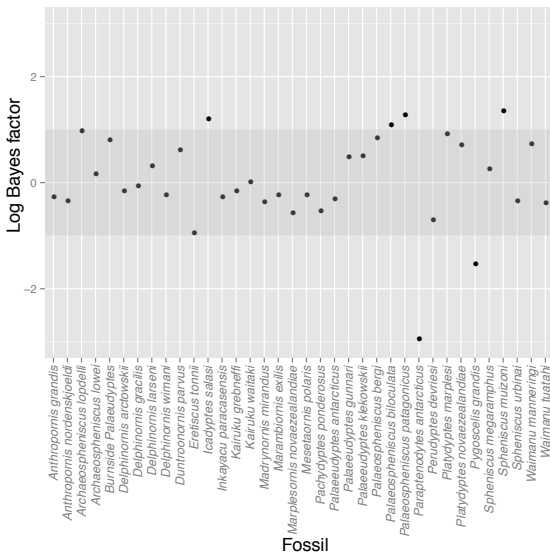
The tmrca of today's penguins: crown radiation time



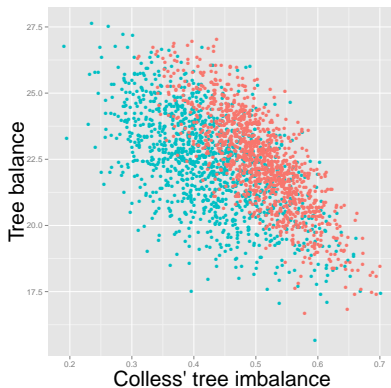
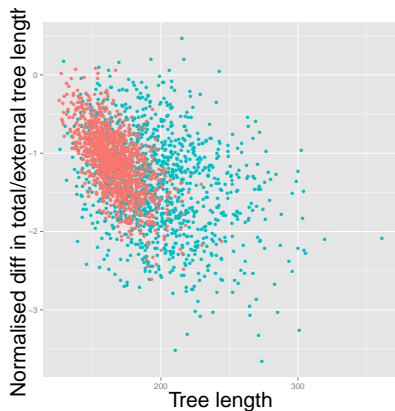
Summary

- ▶ Fossilised birth death model provides excellent framework for total evidence dating
- ▶ Using many stem fossils can greatly improve estimates of crown age
- ▶ More work to come on summarising sampled ancestor trees and including raw fossil data
- ▶ **Software:** packages morph-models and SA for Beast2
beast2.org
- ▶ **Preprint:** Bayesian total evidence dating reveals the recent crown radiation of penguins, Gavryushkina et al,
arxiv.org/abs/1506.04797

Which fossils are have strong evidence for being sampled ancestors?



How well does our model fit? A posterior predictive analysis



Red points are from the posterior, Blue points from the predictive