Species delimitation using SNAPP

Huw Ogilvie[1,2]

[1]Research School of Biology
Australian National University

[2]Centre for Computational Evolution
University of Auckland

2018 Workshop on Population and
Speciation Genomics, Český Krumlov

Part I

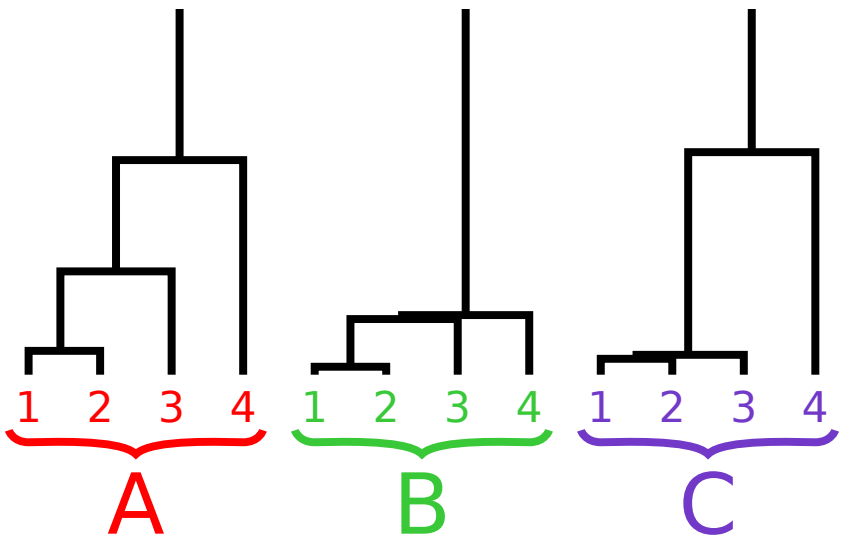
SNAPP

About SNAPP

- SNAPP is a multispecies coalescent (MSC) method
- is a Bayesian MSC method (implemented in BEAST2)
- can be used with Bayes factors for species delimitation

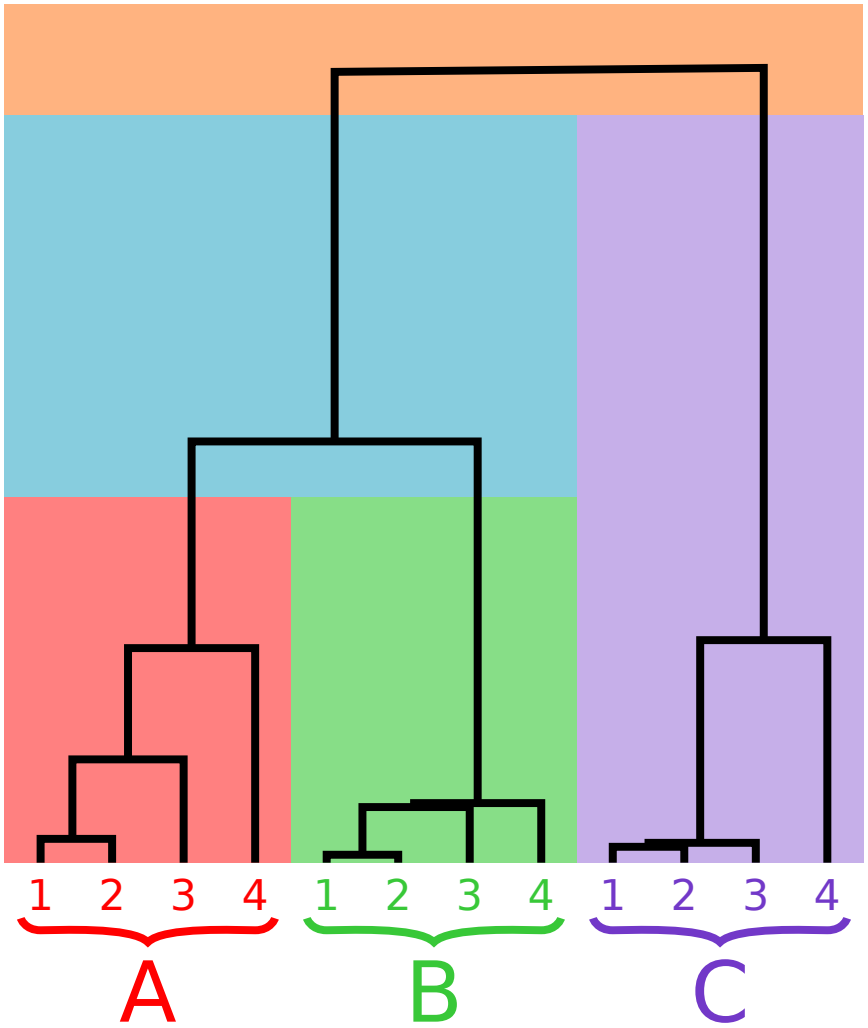
Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular biology and evolution*, 29(8), 1917–1932.

The (Kingman) coalescent



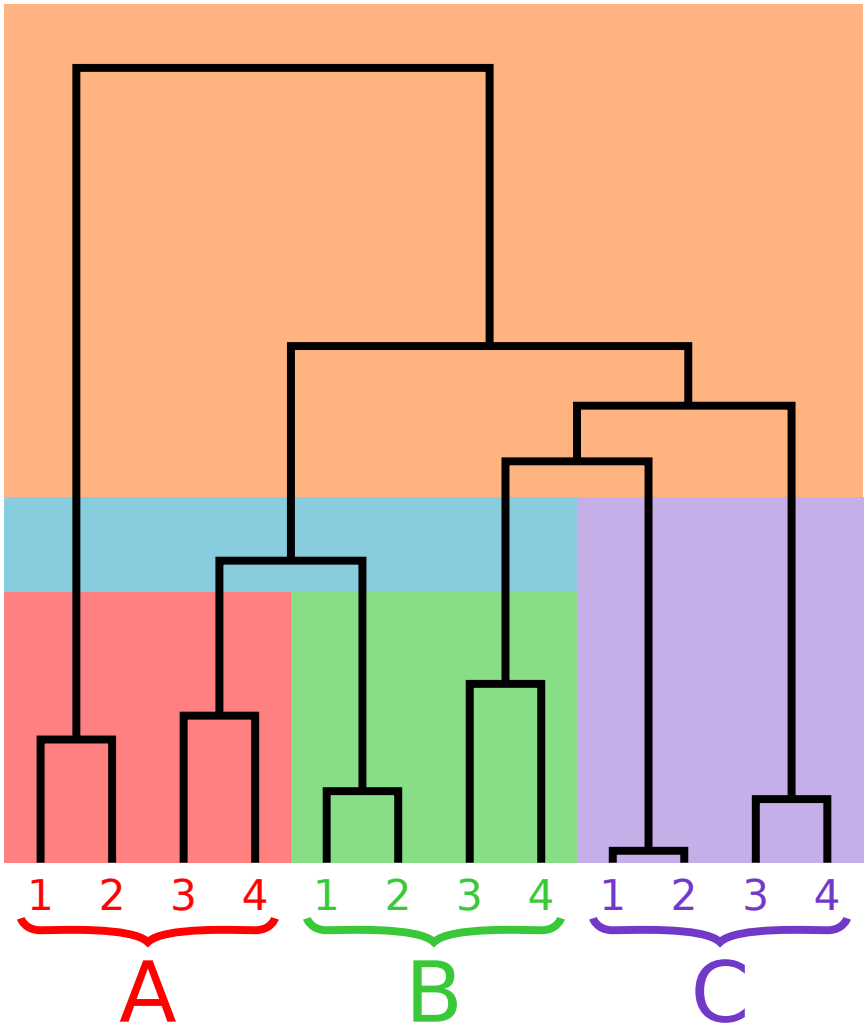
- Models the evolution of orthologous loci
- Applies to a single population
- Backwards in time
- Coalescent rate inversely proportional to $N_e g$ (the effective population size N_e scaled by generation time g)

The multispecies coalescent



- A separate coalescent process applies to each branch
- Assumes speciation is instantaneous
- Assumes no gene flow between populations

The multispecies coalescent



- A separate coalescent process applies to each branch
- Assumes speciation is instantaneous
- Assumes no gene flow between populations
- Incomplete lineage sorting (ILS) is associated with large $N_e g$ and shorter branches

Bayesian MSC inference

You may be familiar with *multilocus* MSC methods such as BPP or StarBEAST2. They are based on this formula:

$$P(S, \theta|D) = \frac{\prod_i P(D_i|G_i) \cdot P(G_i|S, \theta) \cdot P(S, \theta)}{P(D)}$$

$P(S, \theta|D)$ The posterior probability of the species tree topology S and divergence times and effective population sizes θ

$P(D|G_i)$ The phylogenetic likelihood of a gene tree G_i

$P(G_i|S)$ The coalescent likelihood of the species tree

$P(S, \theta)$ The prior probability of the species tree topology S and divergence times and effective population sizes θ

$P(D)$ The marginal likelihood of our model

Bayesian MSC inference

You may be familiar with *multilocus* MSC methods such as BPP or StarBEAST2. They are based on this formula:

$$P(S, \theta|D) = \frac{\prod_i P(D_i|G_i) \cdot P(G_i|S, \theta) \cdot P(S, \theta)}{\cancel{P(D)}}$$

$P(S, \theta|D)$ The posterior probability of the species tree topology S and divergence times and effective population sizes θ

$P(D|G_i)$ The phylogenetic likelihood of a gene tree G_i

$P(G_i|S)$ The coalescent likelihood of the species tree

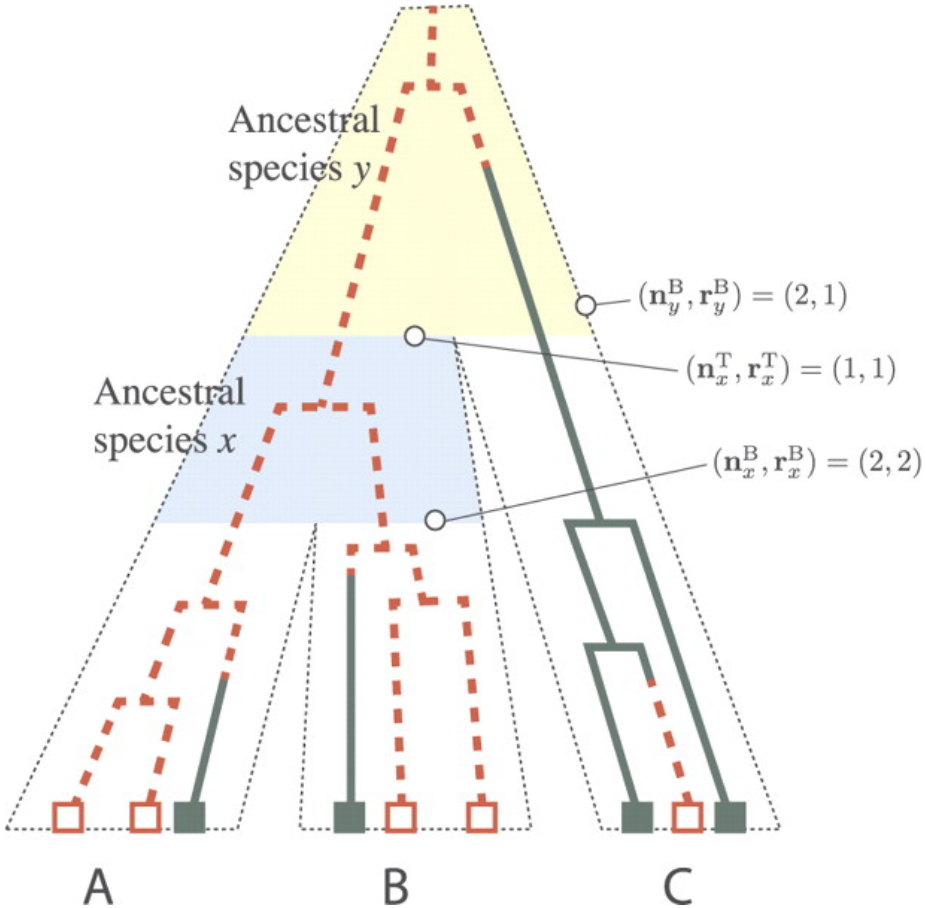
$P(S, \theta)$ The prior probability of the species tree topology S and divergence times and effective population sizes θ

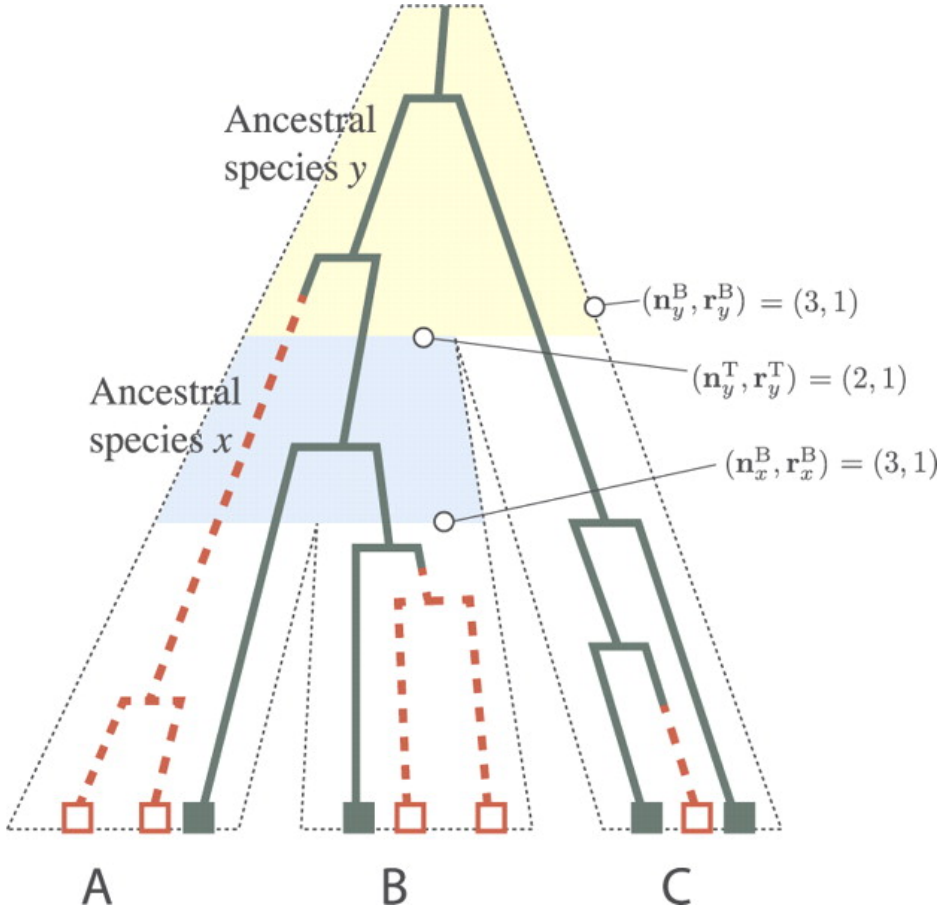
$P(D)$ The marginal likelihood of our model

SNAPP

- “SNP and AFLP Phylogenies” – Bayesian *biallelic* MSC method!
- For younger species trees, only one mutation is observed for most polymorphic sites, so nuclear data can be approximated as biallelic.
- Analytical integrates over: $\prod_i P(D_i|G_i) \cdot P(G_i|S)$

Integrating over gene trees





Bryant *et al.* 2012

Advantages of SNAPP

Multilocus Assumes no recombination within each locus

SNAPP Each locus is a single nucleotide

Multilocus Scales poorly, difficult to use with many loci

SNAPP Can be used with a large number of unlinked sites

	Topologies	Divergence times	Population sizes	Coalescence times
Multilocus	101	29	59	8900
SNAPP	1	29	59	0

Part II

Bayes factors

Testing models

Approach	Relative model fit
Method	Bayes factors
Question	How much closer to the truth is model 1 vs. model 2

Approach	Absolute model fit
Method	Posterior predictive simulations
Question	How close to the truth is model 1

What's in a model?

- The species tree process (birth-death)
- The priors on birth-death parameters – λ, ν
- The gene tree process (multispecies coalescent)
- The priors on coalescent parameters – $N_e g$
- The substitution model (e.g. HKY+G)
- The priors on HKY+G parameters – κ, α, μ
- **The assignment of individuals to species**

Deriving Bayes factors I

Bayes' rule is often written as:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

But in practice is usually:

$$P(\theta|D, M) = \frac{P(D|\theta, M) \cdot P(\theta|M)}{P(D|M)}$$

$P(D|M)$ is the marginal *likelihood*, and using Bayes' rule we can turn likelihoods into probabilities!

Deriving Bayes factors II

Absolute probability intractable because of $P(D)$ (again!):

$$P(M|D) = \frac{P(D|M) \cdot P(M)}{P(D)}$$

But when calculating relative fit (Bayes factor):

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1) \cdot P(M_1)}{\cancel{P(D)}} \cdot \frac{\cancel{P(D)}}{P(D|M_2) \cdot P(M_2)}$$

Then $P(D)$ cancels out:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1) \cdot P(M_1)}{P(D|M_2) \cdot P(M_2)}$$

Evaluating Bayes factors

If our belief is that $P(M_1) = P(M_2)$:

$$2 \ln B_{12} = 2 \ln \frac{P(M_1|D)}{P(M_2|D)} = 2 \ln \frac{P(D|M_1) \cdot \cancel{P(M_1)}}{P(D|M_2) \cdot \cancel{P(M_2)}}$$
$$= 2(\ln P(D|M_1) - \ln P(D|M_2))$$

$2 \ln(B_{12})$	Support for M_1 over M_2
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very strong

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Hang on

Didn't I hear that the marginal likelihood $P(D)$ is really hard to calculate, which is what motivated the development of MCMC?

...

...

Yes.

Calculating the marginal likelihood

Remember that the marginal likelihood normalizes $P(D|\theta) \cdot P(\theta)$:

$$P(D) = \int_{\theta} P(D|\theta) \cdot P(\theta) \, \mathrm{d}\theta$$

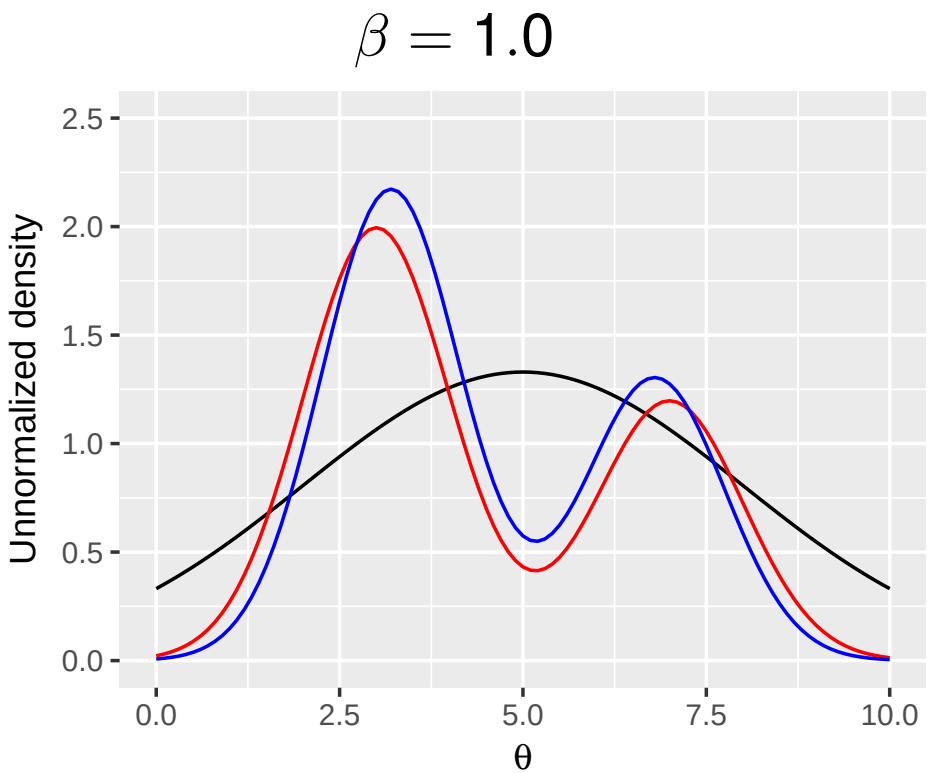
Which can be solved by computing the expected value of the likelihood $P(D|\theta)$ when sampling from the prior distribution:

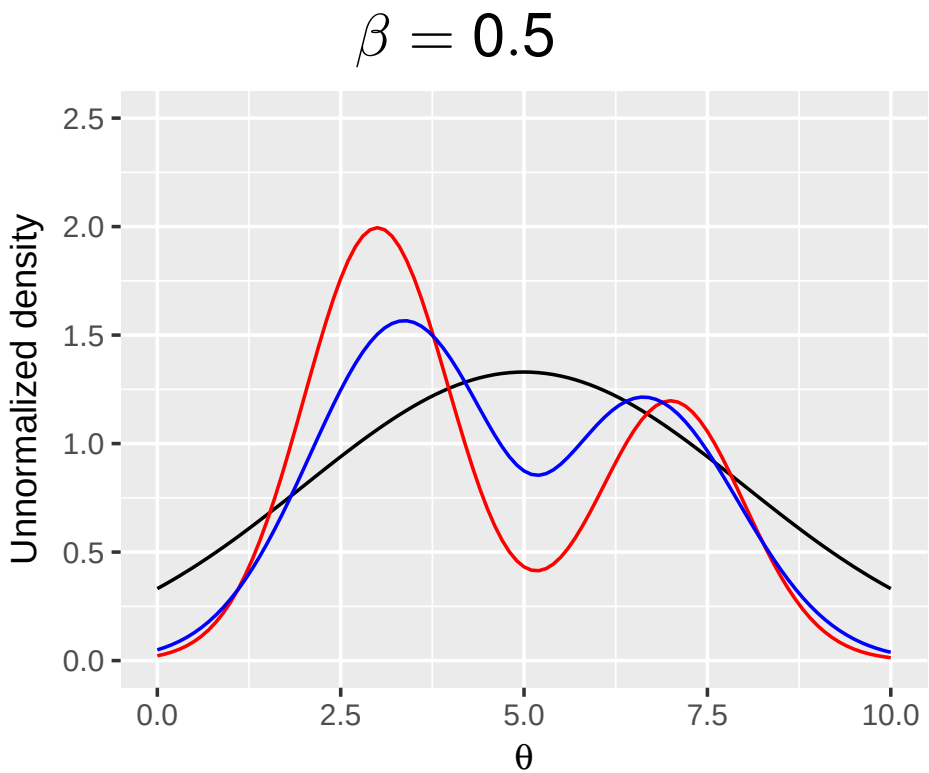
$$P(D) = E[P(D|\theta)]$$

`https://darrenjw.wordpress.com/2013/10/01/marginal-likelihood-from-tempered-bayesian-posteriors/`

Power posteriors

Using MCMC, E(likelihood) will not be well sampled under the prior. But we can sample a series of intermediates using power posteriors:

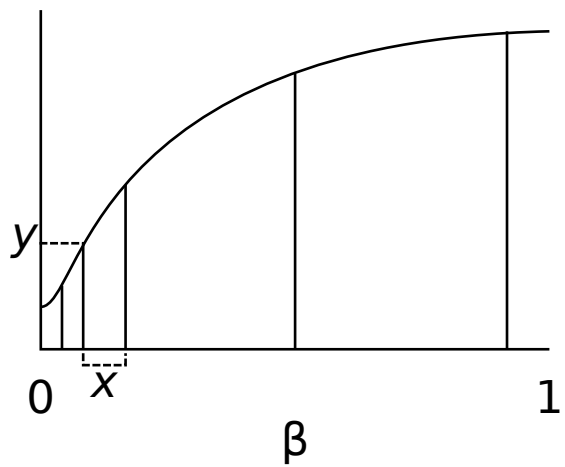




$$P(\theta|D) \propto P(D|\theta)^{\beta} \cdot P(\theta)$$

Stepping-stone sampling

$$P(D) = E[P(D|\theta)] = \prod_{i=0}^{N-1} E_i[P(D|\theta)^{\beta_{i+1}-\beta_i}]$$



$x = \beta_{i+1} - \beta_i$ and $y = E[P(D|\theta)]$

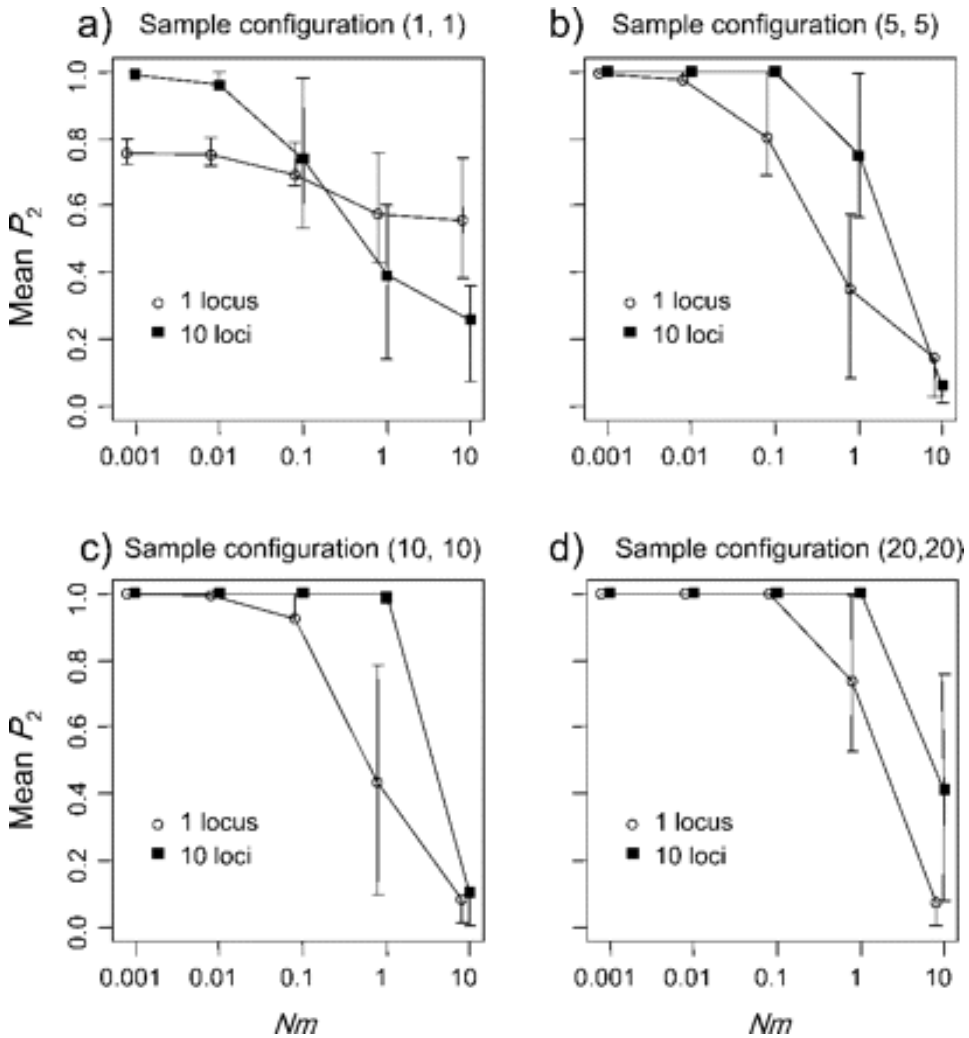
Part III

Species delimitation

Everything is great

- Now we can calculate marginal likelihoods
- Therefore we can calculate Bayes factors
- Therefore we can compare species delimitation probabilities
- Does this mean we can delimit species?

What is actually going on



Zhang, C., Zhang, D. X., Zhu, T., & Yang, Z. (2011). Evaluation of a Bayesian coalescent method of species delimitation. *Systematic biology*, 60(6), 747-761.

Species concepts

Biological	Interbreeding (natural reproduction resulting in viable and fertile offspring)
Isolation	Intrinsic reproductive isolation (absence of interbreeding between heterospecific organisms based on intrinsic properties, as opposed to extrinsic [geographic] barriers)
Recognition	Shared specific mate recognition or fertilization system (mechanisms by which conspecific organisms, or their gametes, recognize one another for mating and fertilization)
Ecological	Same niche or adaptive zone (all components of the environment with which conspecific organisms interact)
Evolutionary	Unique evolutionary role, tendencies, and historical fate
Cohesion	Phenotypic cohesion (genetic or demographic exchangeability)
Hennigian	Ancestor becomes extinct when lineage splits
Monophyletic	Monophyly (consisting of an ancestor and all of its descendants; commonly inferred from possession of shared derived character states)
Genealogical	Exclusive coalescence of alleles (all alleles of a given gene are descended from a common ancestral allele not shared with those of other species)
Diagnosability	Diagnosability (qualitative, fixed difference)
Phenetic	Form a phenetic cluster (quantitative difference)
Clustering	Form a genotypic cluster (deficits of genetic intermediates; e.g., heterozygotes)

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic biology*, 56(6), 879-886.

Alternatives

BPP: Uses reversible jump MCMC to integrate over the space of species assignment and delimitation

STACEY: Uses a time threshold to delimit species, combined with a "lumpy" prior on the species tree

Tracer: Implements harmonic mean estimation (HME) of the marginal likelihood, which has been called the "Worst Monte Carlo Method Ever"

`http://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/`

Question I

What impact does the prior distribution on the speciation rate λ have on marginal likelihood estimates? If the prior distribution favors faster values of λ, how could this change the Bayes factors?

Question II

SNAPP can estimate a forward (zero to one) and reverse (one to zero) mutation rate. How should these rates be set when used with nucleotide data?