~~Gene trees and species trees~~

SNAPP models

David Bryant
Dept. Mathematics and Statistics
University of Otago
Dunedin, New Zealand

Species

A species is a group of individuals which are equally likely to interbreed with each other, and which don't interbreed with anyone else

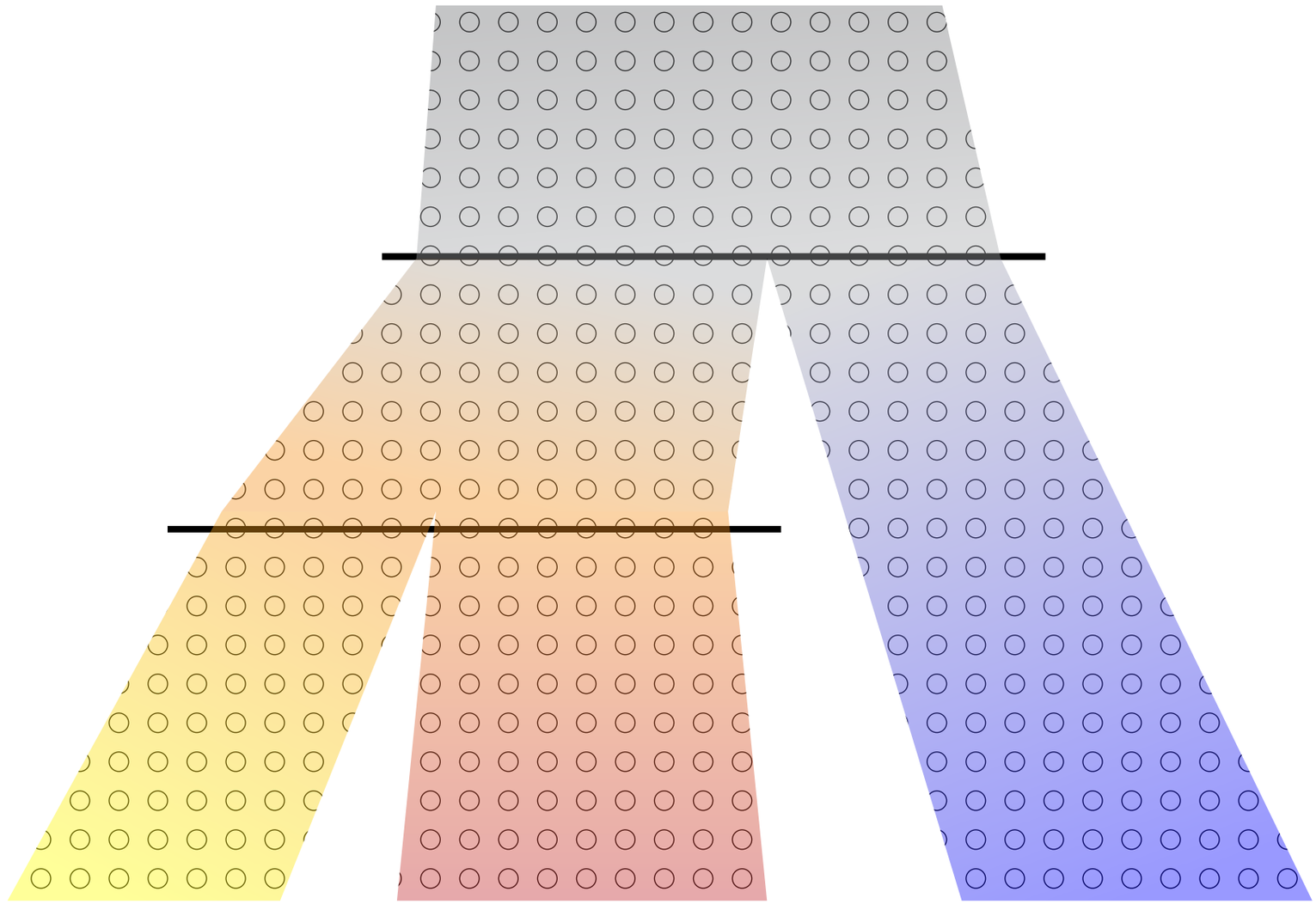
Question for later: what is wrong with this?

Species Tree

A representation of a history of a species which, over time, splits and splits again into new descendent species, eventually giving rise to the set of species under study

Question for later: what is wrong with this?

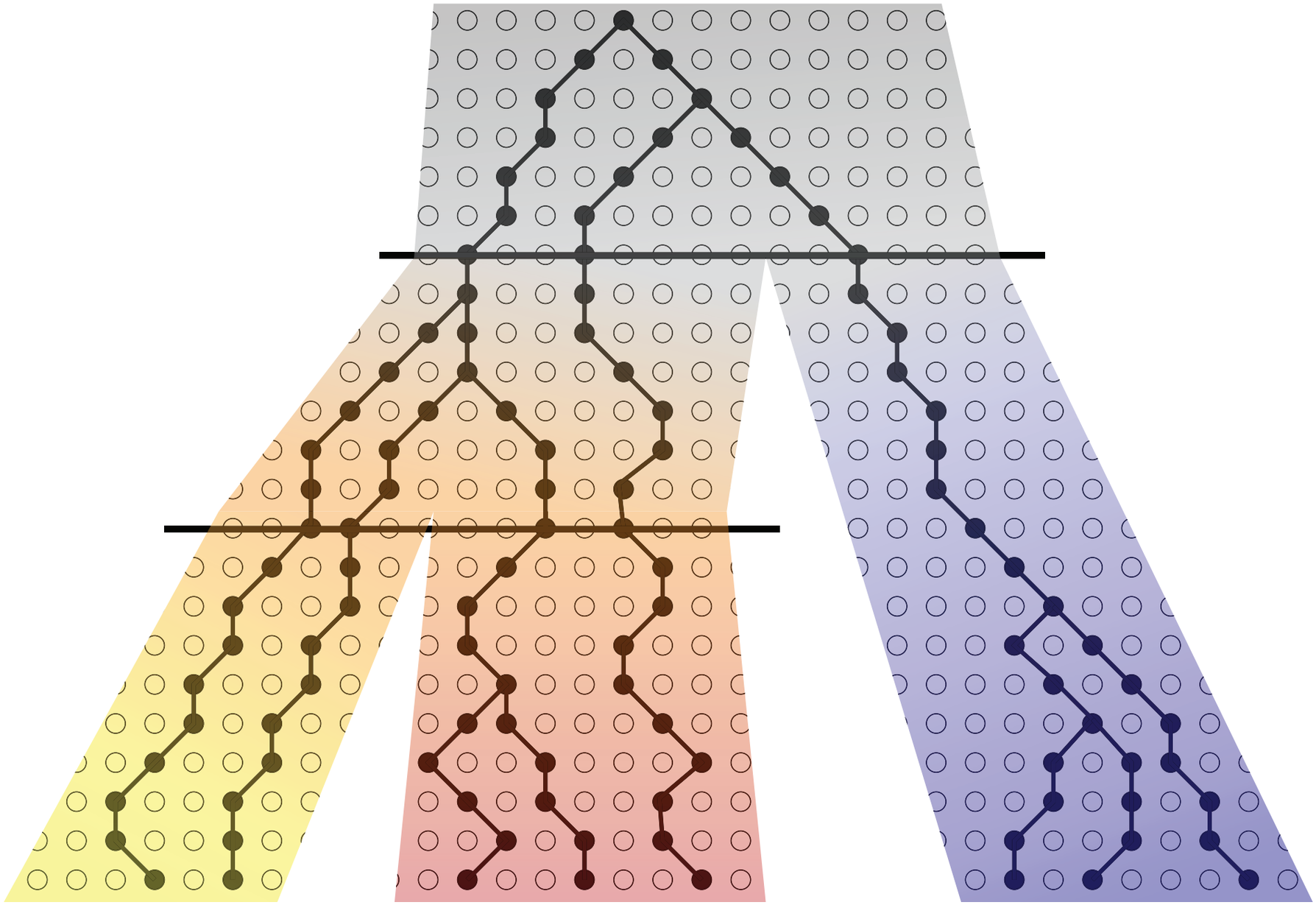
Species tree



Gene Tree

A tree representing the history of inheritance of copies of a particular gene (or locus) among the sampled individuals.

Gene tree

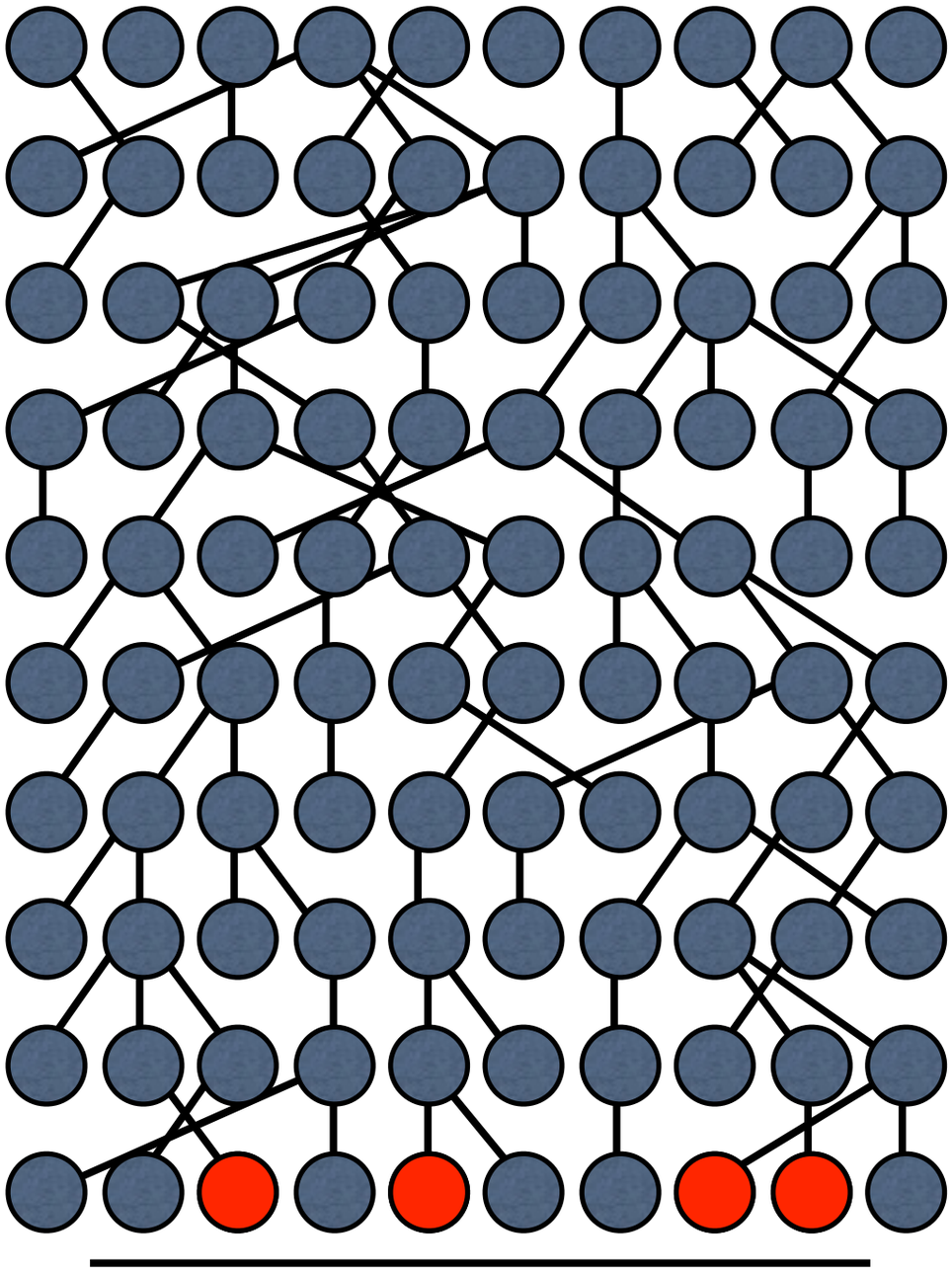


david bryant 2017

Linkage disequilibrium

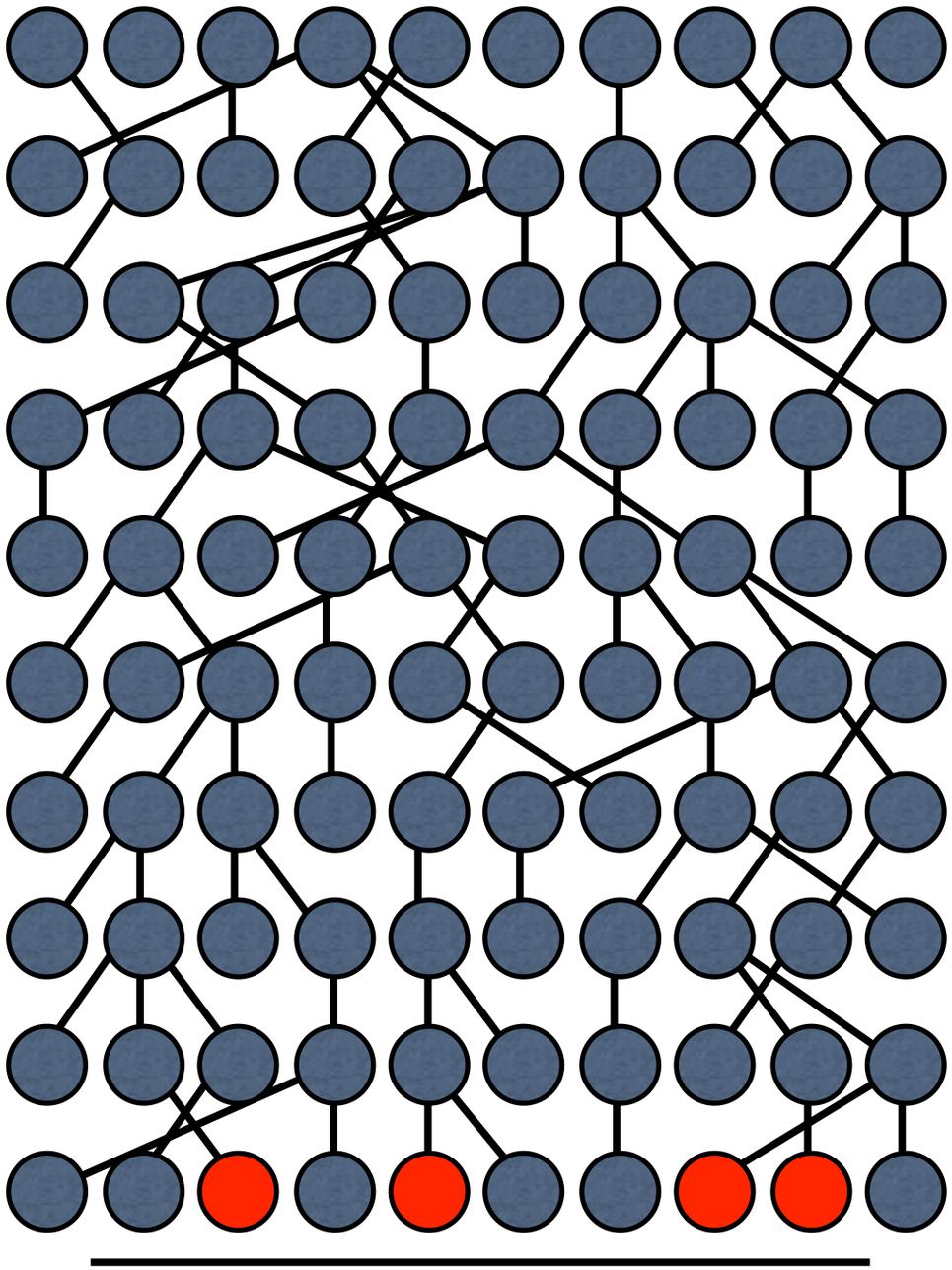
- A chromosome in a child can contain genes inherited from different parents. Hence different genes/loci can have different gene trees.
- When genes are inherited together we say they are linked. When they are inherited independently we say they are unlinked.
- Linkage disequilibrium (LD) is a correlation-based measure of the extent to which genes/loci are linked.

Forward time models

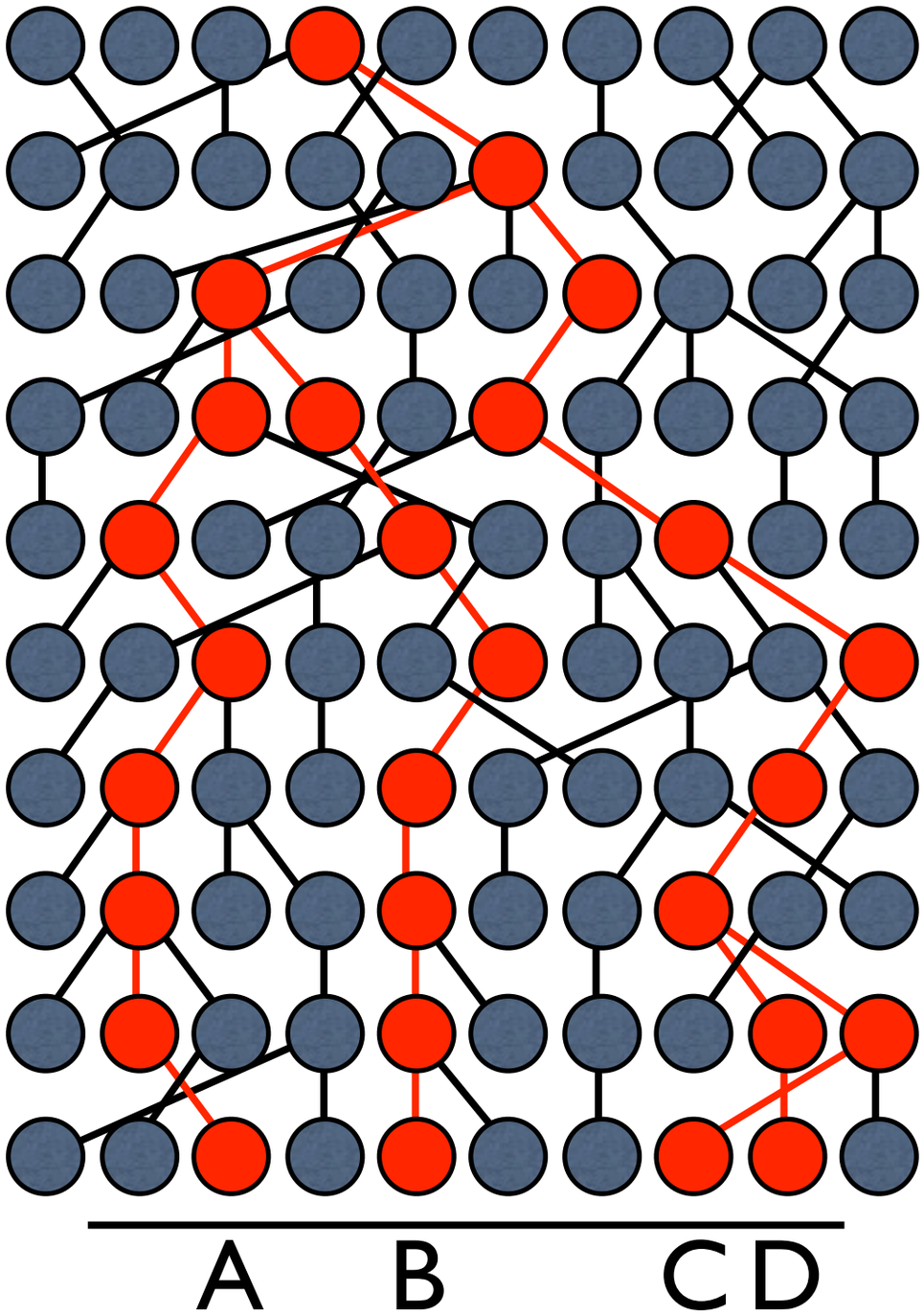


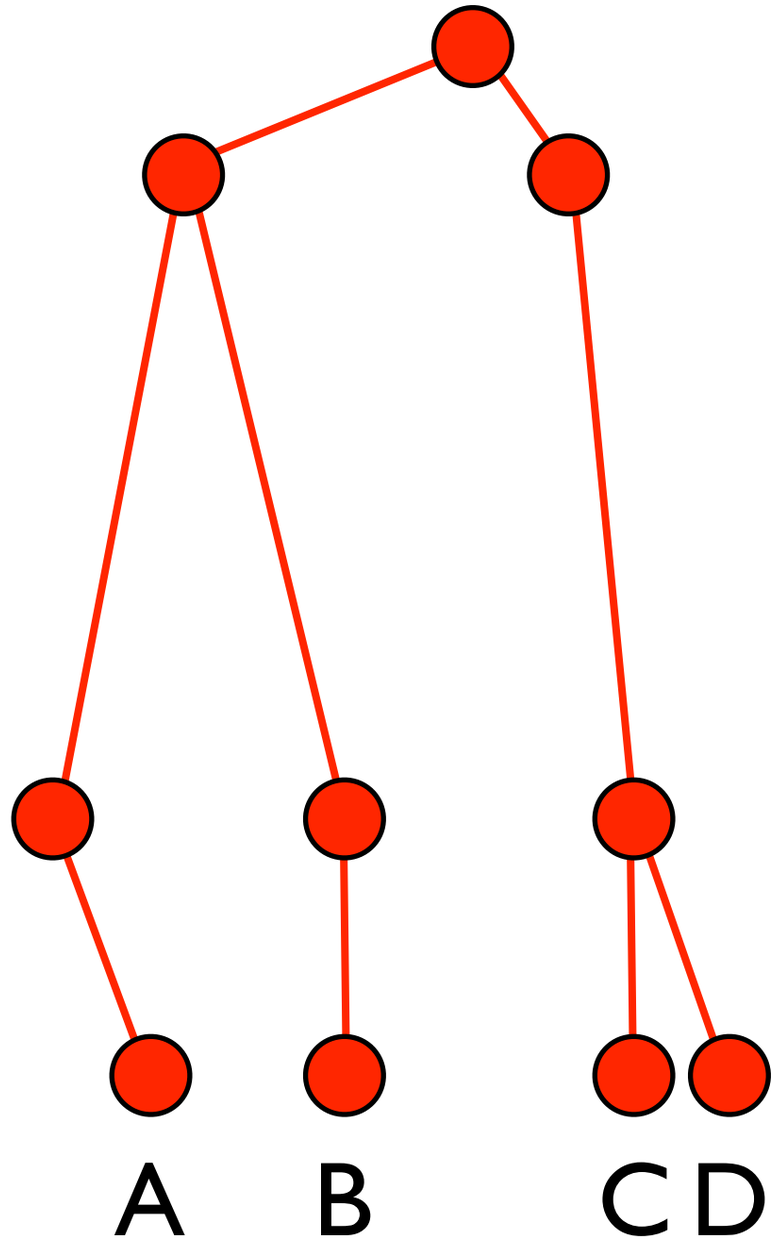
Present day samples

Wright-Fisher model



Reversing the Wright Fisher

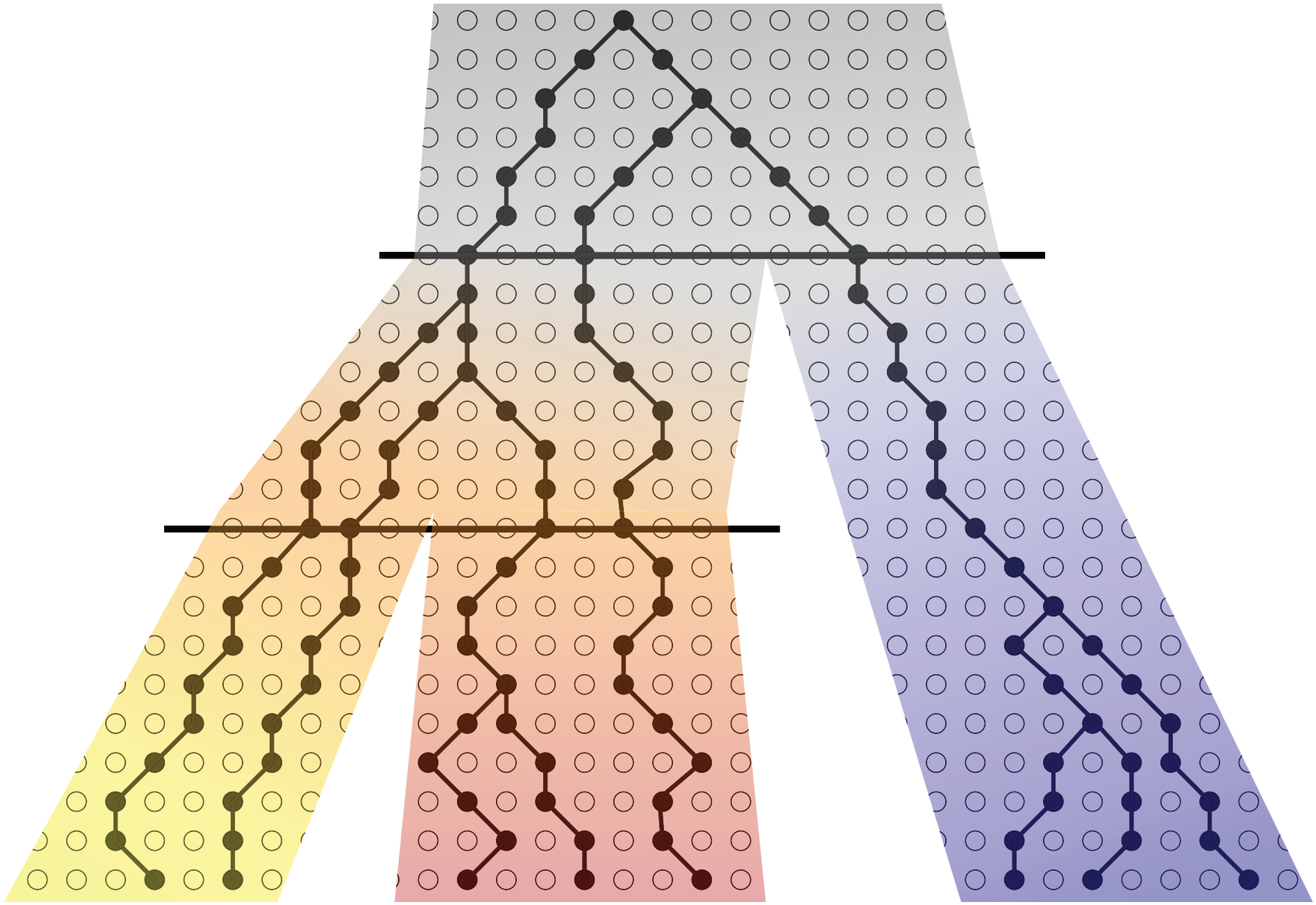




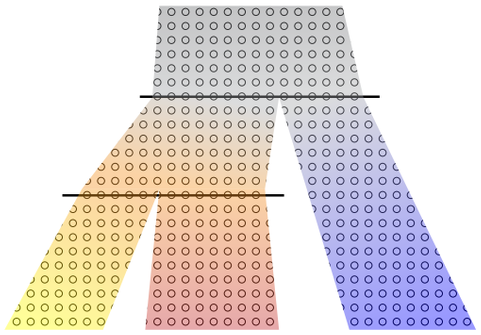
Coalescent Process - one population

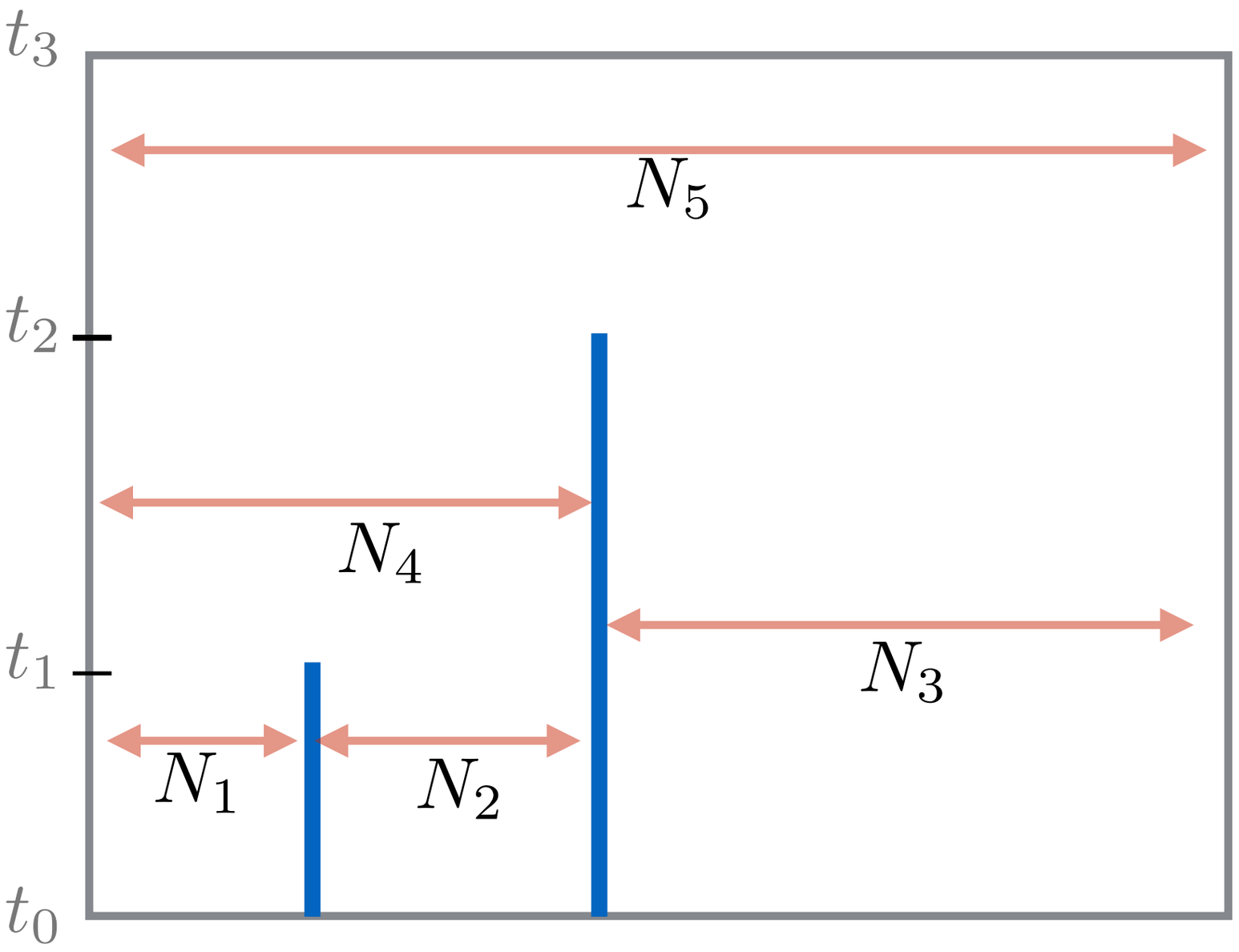
- Consider the number of distinct ancestral lineages as a death process going backwards in time.
- The event of two lineages meeting at a common ancestor is called a coalescence.
- Two lineages coalesce at rate 1/(2N) (backwards in time)
- k lineages coalesce at rate (k(k-1)/2) x 1/(2N)

Multispecies coalescent



Multispecies coalescent





Just like the single population coalescence except that

- Lineages on either side of a barrier can't coalesce
- Other pairs coalesce with rate proportional to 1/pop size.

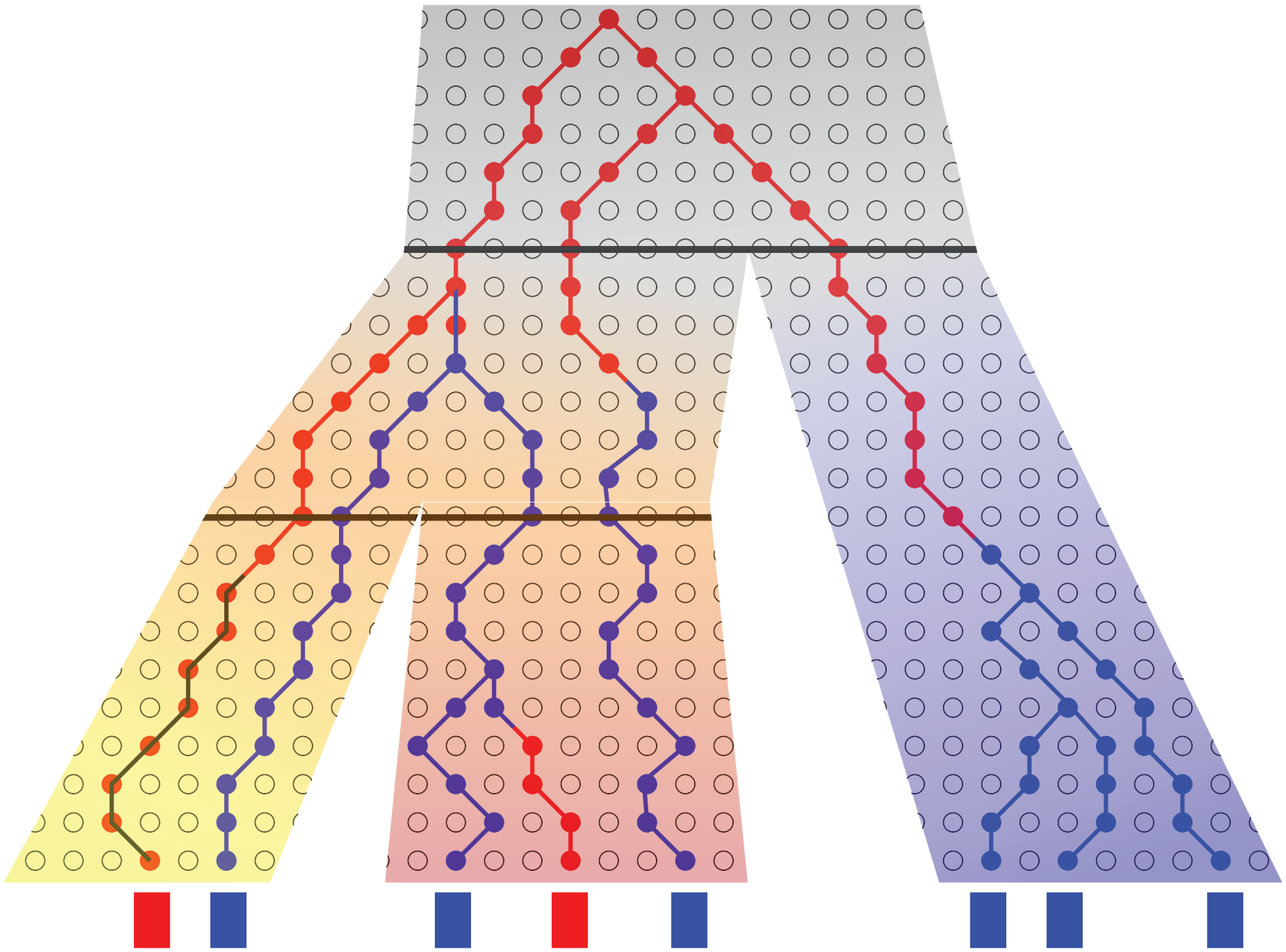
How should the N’s (population sizes) be related?

Models for species size

- The probability of reproducing
- Diversity within the species
- Q:- what is the population? What are we actually modelling? **
- Should population sizes add up over the tree / after each divergence? (N1 + N2 = N4 …) ***
- How do extant population sizes relate to ancestral?

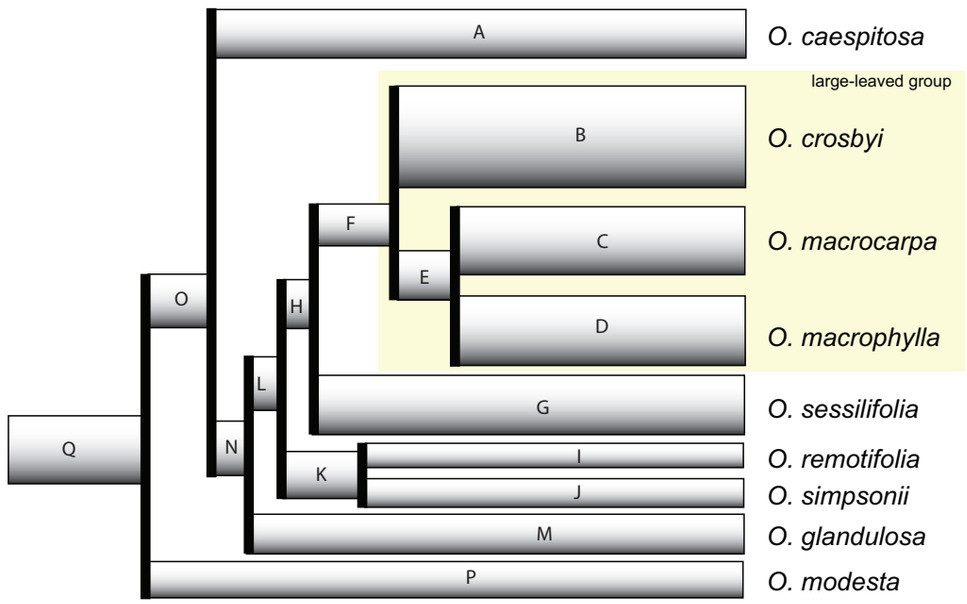
- Mechanisms for estimating ancestral population sites (perhaps look at coalescent trees as source of information about these?)
- Ancestral population sizes at least one…
- Differs depending on mechanism of speciation
- Varying change in population size along a branch, which might continue across multiple branches

Modelling (binary) SNPs



SNAPP Computations

- SNAPP uses a range of numerical and analytical tricks to compute the likelihood of a single SNP conditional on the species tree (with parameters).
- This contrasts to most gene tree / species tree approaches which either
  - sample gene trees for each locus/gene/SNP, or
  - use some kind of approximation.
- We still use MCMC to sample the species trees.

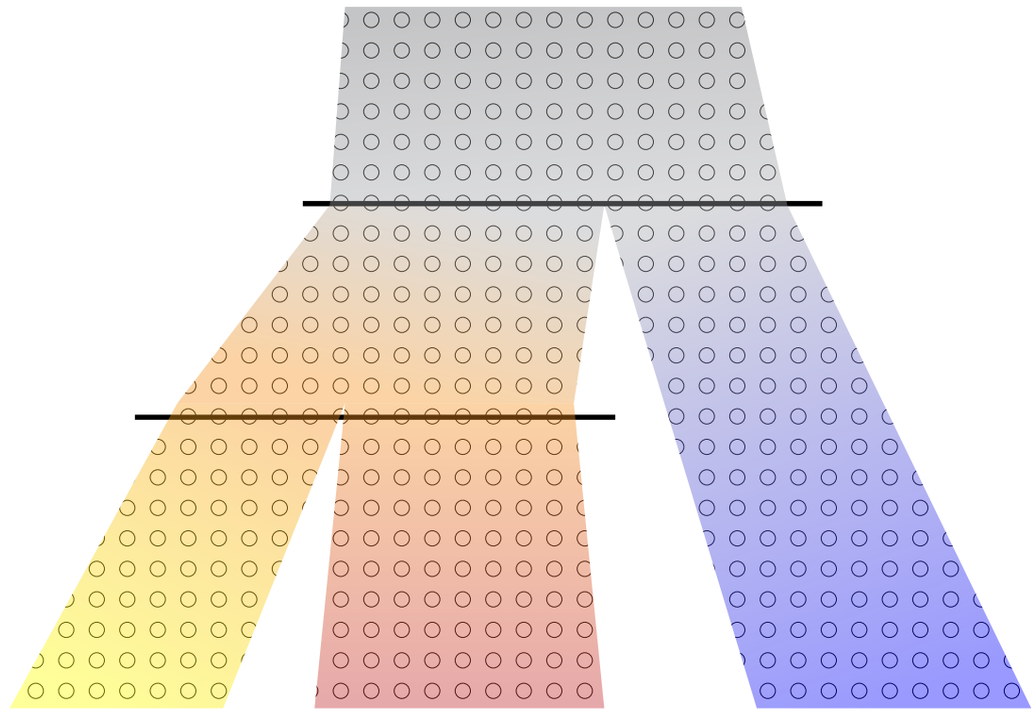


You can use SNAPP with…

- **Unlinked binary markers:**
  - SNP data
  - Full genome data (with subsampling)
  - AFLP data
  - RadSeq (one SNP per restriction site)
- **Do not use SNAPP to analyse**
  - Full genes (use *Beast or competitors)
  - Every site in a genome
  - Markers with > 2 alleles

Setting up the model (1): species trees

- Times between branching / divergences
- Tree shape
- Population sizes



Time

- Time is measured in **mutations per site**
- The mutation rate is over all sites, including constant sites (which is what we usually measure)
- The mutation rate is assumed constant over the tree

- If you want to convert SNAPP time to **numbers of generations**, divide by the mutation rate (in mutations per generation)
- If you want to convert this to **calendar time,** divide by the generation time

Species tree prior

- We implement a Yule prior, which has a single parameter λ
- λ is the rate at which lineages split

Which lambda should I use???

- 1/λ which is the expected number of mutations (per site) between divergences.
- If μ is the mutation rate, and g the generation time, then

$$(1/\lambda)\ (g/\mu)$$

is the average time between divergences

(there are a lot of prior estimates for this for multiple species, see, e.g. Coyne and Orr).

Population size

- We use the standard parameter θ for population size.
- In a diploid populatio

$$\theta = 4N\mu$$

  where N is the effective population size

- Independent gamma priors on each θ (user specifies mean and variance or gamma parameters)
- Note that this is effective population size… bottlenecks have a big effect

Mutation rate

- Since time is measured in mutations per site, we choose backward and forward mutation rates u and v so that that the expected rate

  $$2uv / (u+v)$$

  is one.

  - With binary non-directional SNP data, it makes most sense to use u=v=1
  - If there is a clear directionality to the states, (e.g. presence absence for AFLP; mutant vs wildtype) to might make sense to estimate u and v
  - In principle, it would be possible to incorporate nucleotide frequencies for (biallelic) DNA data, but this hasn't been done yet!

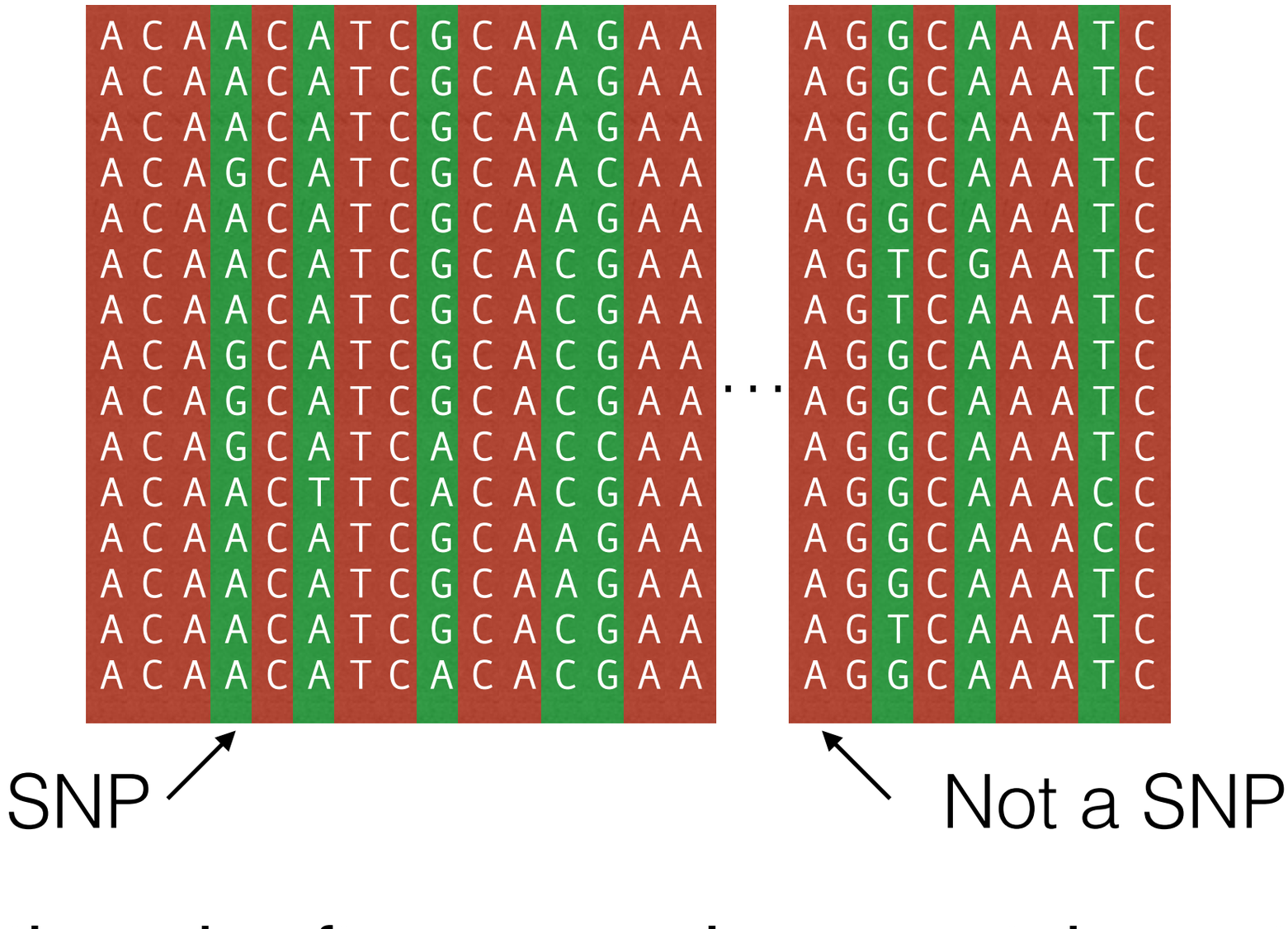
Come up with an important shortcoming of this model

Shortcoming

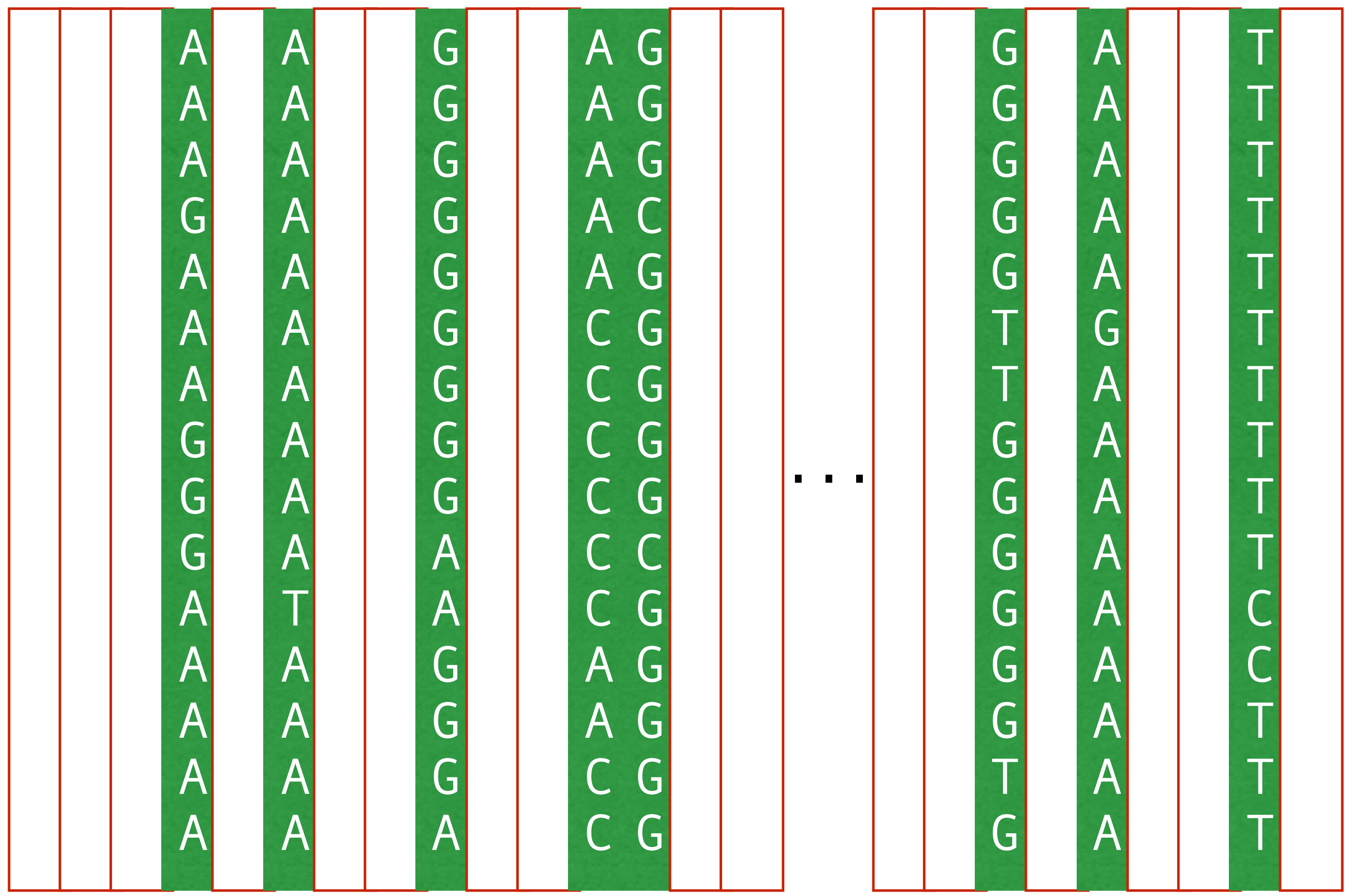
Ascertainment bias

SNP = Single Nucleotide Polymorphism

Polymorphism = Marker which varies in the population(s)



In practice there is often a complex ascertainment process



Censored data: some (constant) patterns are concealed, but we know how many there are.

Truncated data (what we're getting)

Truncated data: some (constant) patterns are completely removed from the data. We don't know how many were removed.

cf A. Gelman, *JASA*, 2004

Conditional likelihood

Instead of using the probability of a pattern (given a tree etc.)

$$P(\text{AACCAACCCCA} \mid T)$$

we use the *conditional probability,* where we condition on the pattern being non-constant

$$\frac{P(\text{AACCAACCCCA} \mid T)}{P(\text{??????????} \mid T)}$$

P(observed |T)

Conditional problems

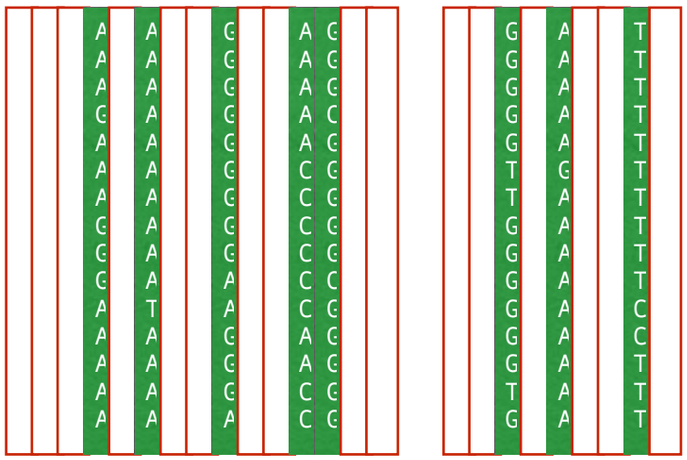
1. Conditional probability of a pattern being observed is often difficult to compute.
2. The full likelihood function is easier to approximate.
3. Removing constant sites makes it far harder to infer divergence dates and population sizes.

Use non-polymorphic sites?

- If you have SNP data or AFLP data then the ‘use non-polymorphic data’ should be set to off.
- If you have a random (sparse) sample of genomic data, and you included non-polymorphic (constant) sites, then this option should be set to on.
- Coming soon: better ways to incorporate information about genome size and proportion of segregating sites…

Walk like a Bayesian

We observed *n* variable patterns



but we know these were selected from N patterns in total, where N is unknown.

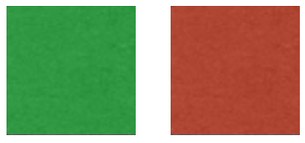
So we sample from the posterior distribution of N. Heck, we might as well sample from the truncated data while we're at it…

Manipulating conditionals

P( | N,T)

is just (as usual) a multinomial distribution.

To get

P( N|T)

we can use Bayes' rule, but we need a prior for N.

Conditional likelihoods revisited

Conditional likelihoods correspond to prior P(N) proportional to 1/N

What does this say about the use of conditional likelihoods and hidden assumptions about the data?

What to use as a prior?

OPEN MODELLING PROBLEM

Take a strategy for SNP ascertainment and selected and determine an appropriate model/prior for the “effective” number of SNPs

In the absence of convincing models, we commit the standard Bayesian transgression of choosing a convenient (general) distribution (negative Binomial).

Sequences or SNPs?

Rannala and Yang (2017)

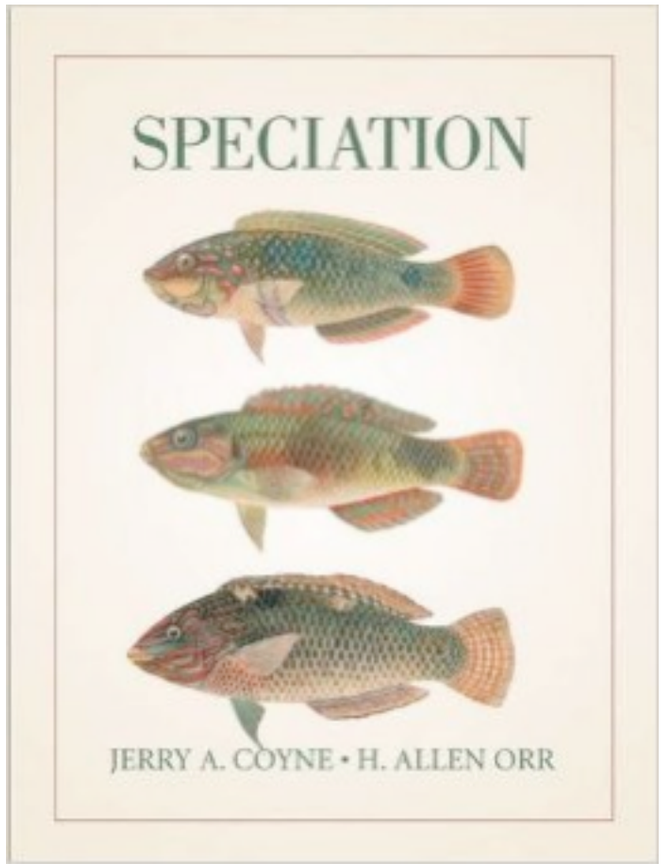
However, a drawback of such methods [like SNAPP] is that SNPs provide little information about branch lengths in the gene trees and the power may be reduced in comparison with sequence-based methods.

1. If we ignore mixing and modelling, SNAPP is a most powerful method for these data
2. It is not clear (and I don't believe it has been properly tested) that 100 genes are more powerful then 1 million SNPs…

What is a species?

What is a “species”?

Diversity of species concepts



Coyne and Orr list 19 distinct species concepts…

…other authors list even more…

…but I've found it really hard to find anything I can encode mathematically.

Biological species concept

Species are groups of interbreeding natural populations that are reproductively isolated from such other groups. (Ernst Mayr, 1995)

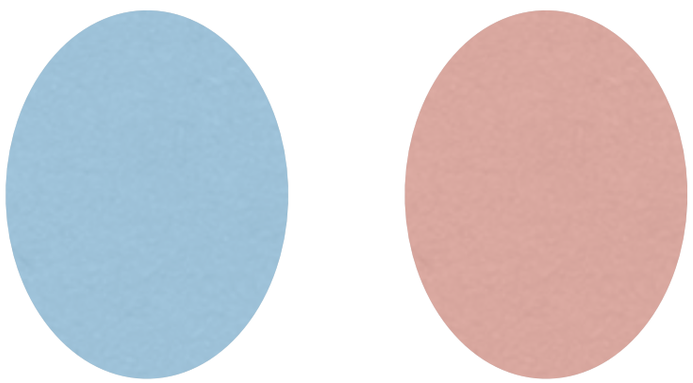
What can prevent two individuals from reproducing successfully?

What is a “species”?

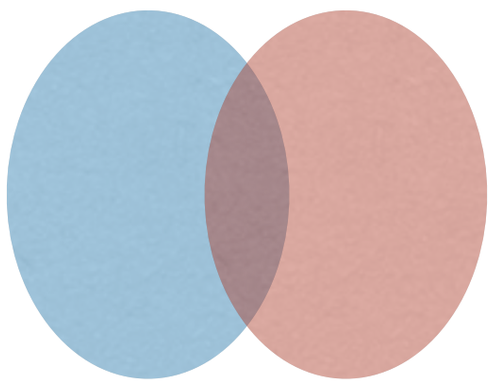
Reproductive isolating barriers

- Lack of cross-attraction between different groups
- Different habitats
- Different individuals breed at different times
- Different interaction with pollinators
- Mechanical incompatibility
- Partial or complete self-fertilisation
- Not doing it right
- Incompatibility with gametes (sperm/pollen)
- Hybrids suffer lower viability and lack of ecological niche
- Hybrids less attractive
- Hybrids fail to develop properly
- Hybrid sterility

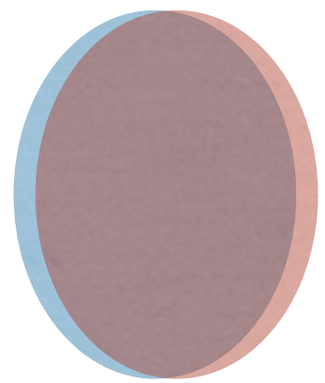
Biogeography



Allopatric Speciation



Parapatric Speciation

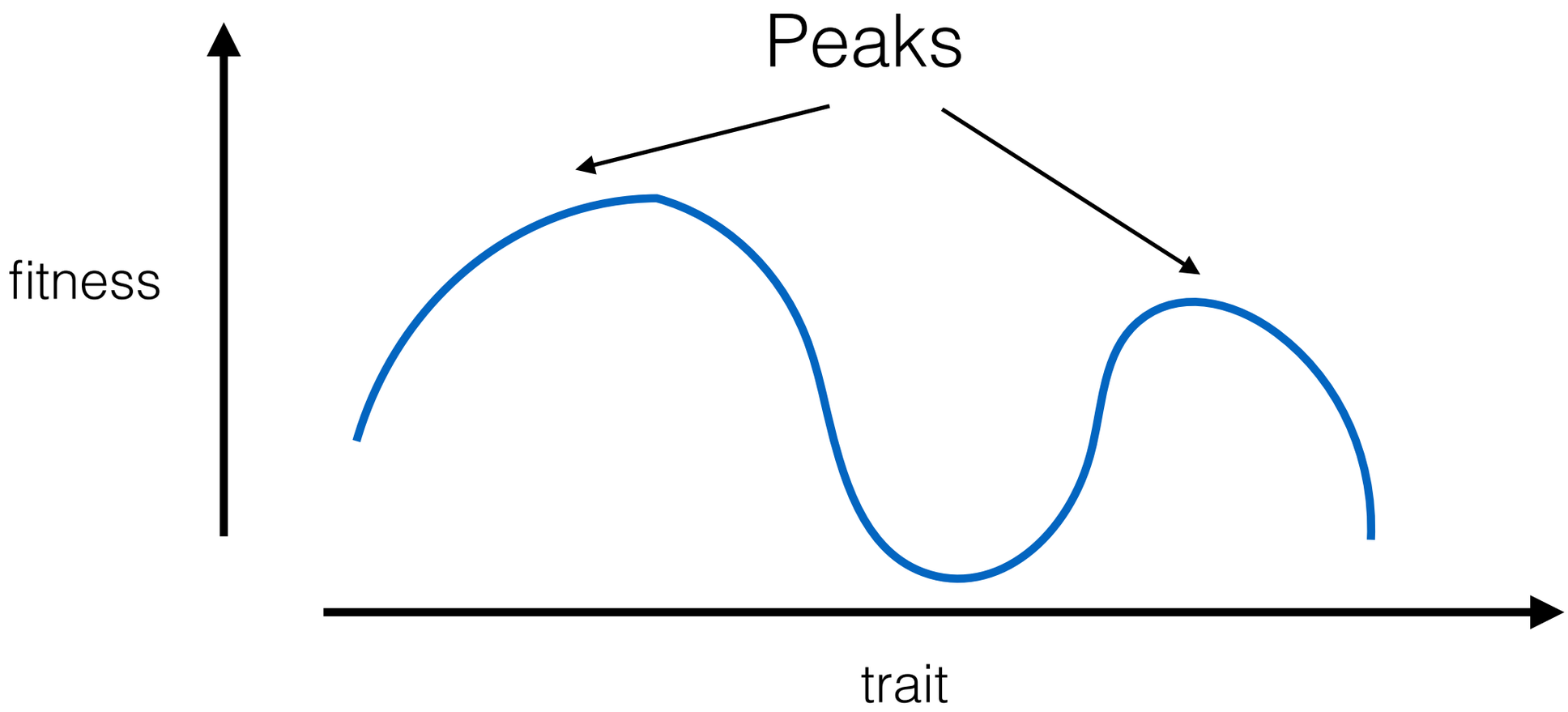


Sympatric Speciation

How isolated do populations need to be to form species?

Selection landscapes

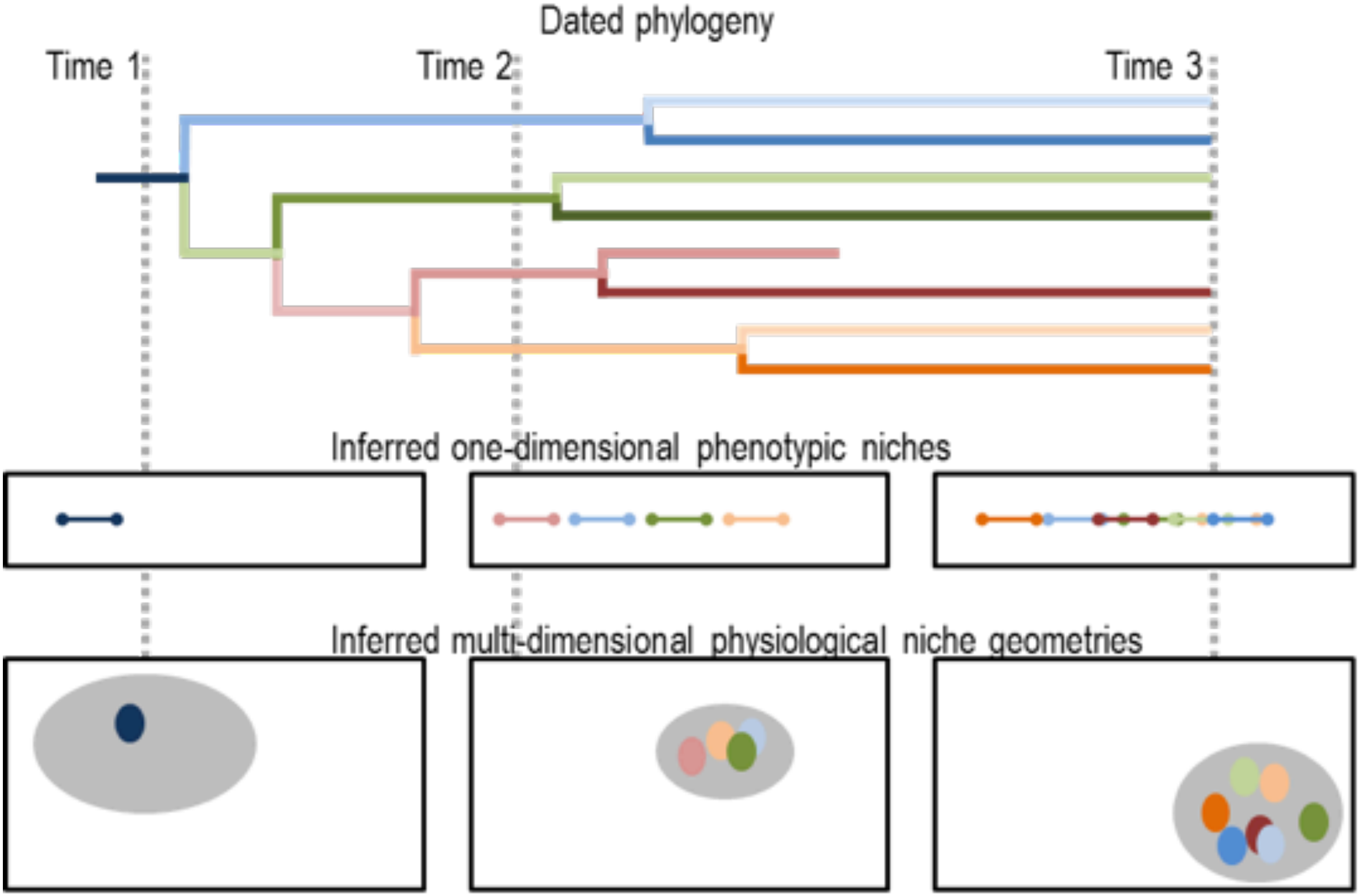
- Though it is contended, it is apparent that selection plays a significant role in many speciations



Towards a quantitative model for speciation

- Every individual is identified according to a huge collection of ‘traits’, or equivalently as a point in a high dimensional ‘trait space’
- Traits could include
  - Morphological
  - Ecological niche
  - Behavioural
  - Geographic
  - …
- The probability that two individuals reproduce successfully is a decreasing function of their distance (i.e. isolation rather than barriers).

Niche evolution



david bryant 2017

Physiological niche models

State variable

Physiological process

Environmental forcing variable

Environmental dependency

Change in state variable process

The sum of the biomasses in the root and shoot compartments is simulated for each location where presence or absence of the species has been observed. The parameters describing the environmental response functions are estimated by maximising the likelihood of observing the data given the parameters.

Shoot

Biomass

Carbon

Nitrogen

Biomass growth

Biomass loss

Carbon transport

Nitrogen transport

Root

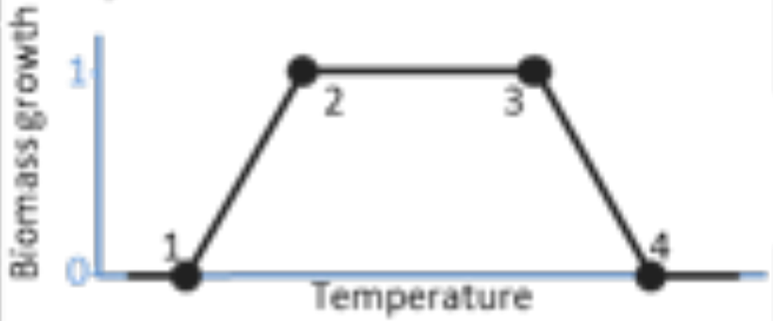
Biomass

Carbon

Nitrogen

Biomass growth

Each environmental dependency is described by a response function. In this example four parameters describe how temperature scales the biomass growth rate:



Environmental data are drawn from geographical data bases. Monthly means of these variables are used to force the model

Light

Temperature

Soil moisture

Soil Nitrogen

Carbon uptake

Nitrogen uptake

Opinion is cheap

1. There has been a shift towards more mathematically complex models of species tree, often of a macro-evolutionary flavour
2. I'd like to see a shift away from a focus solely on phylogenetics (tree+times+thetas) and more towards trying to infer the nature of the speciations at the nodes
3. Comparative method approaches are a start, but there are scary problems of logical dependencies. Perhaps forward-time trait based models could provide a practical alternative.
4. Which is hard.