Concatenation vs Coalescence

Complications to phylogenetic inference with multiple genes

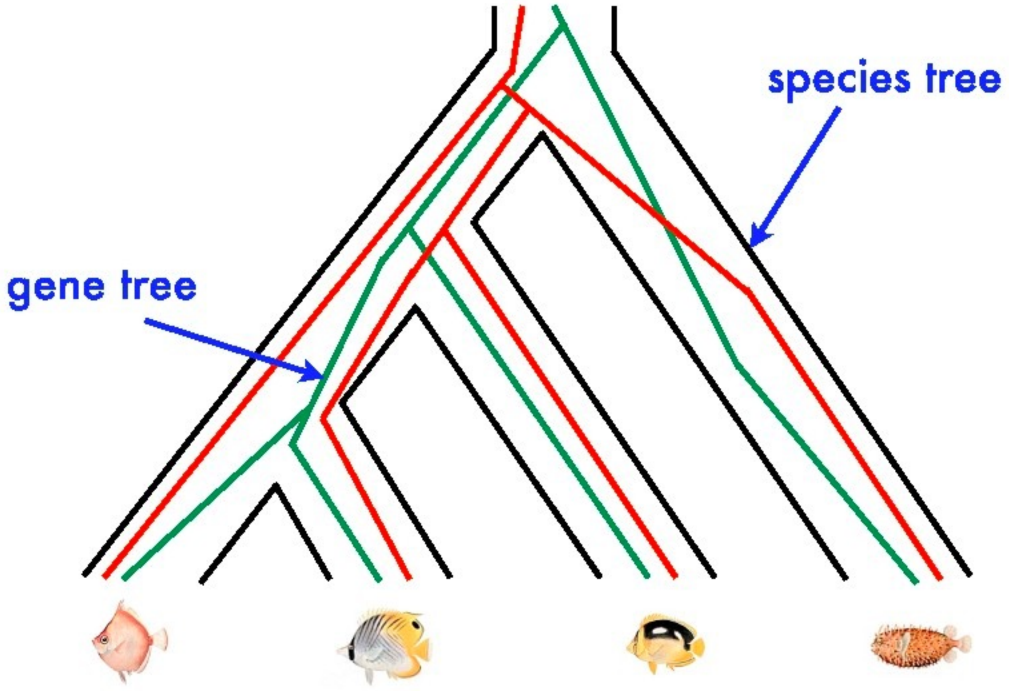
“Hard” problems in phylogenetics

- Long branch attraction (Felsenstein Zone)
  - homoplasy overwriting true signal (saturation)
- Gene duplication, gene extinction or paralogous sampling
  - Thinking you are analyzing homologous genes but some specimens have different gene copies
- Incomplete lineage sorting (ILS)
  - Gene trees differ from species tree due to variation shared among species/lineages
- Horizontal gene transfer
  - Hybridization
  - Introgression

- ALL HAVE TO DO WITH GENE TREES NOT MATCHING SPECIES TREES DUE TO NATURAL PHENOMENON

Gene trees vs Species tree

- **Species tree** represents true relationships among **species**
- **Gene tree** represents the evolutionary history of the **gene**
    - We can use genes and gene trees to infer species trees and relationships**.**



So what do we do when we have multiple loci?

- We are trying to infer species relationships from genes.
- What do we do?
  - Concatenate?
  - Coalescence?
  - Other approaches?

Concatenation

- What can it do?

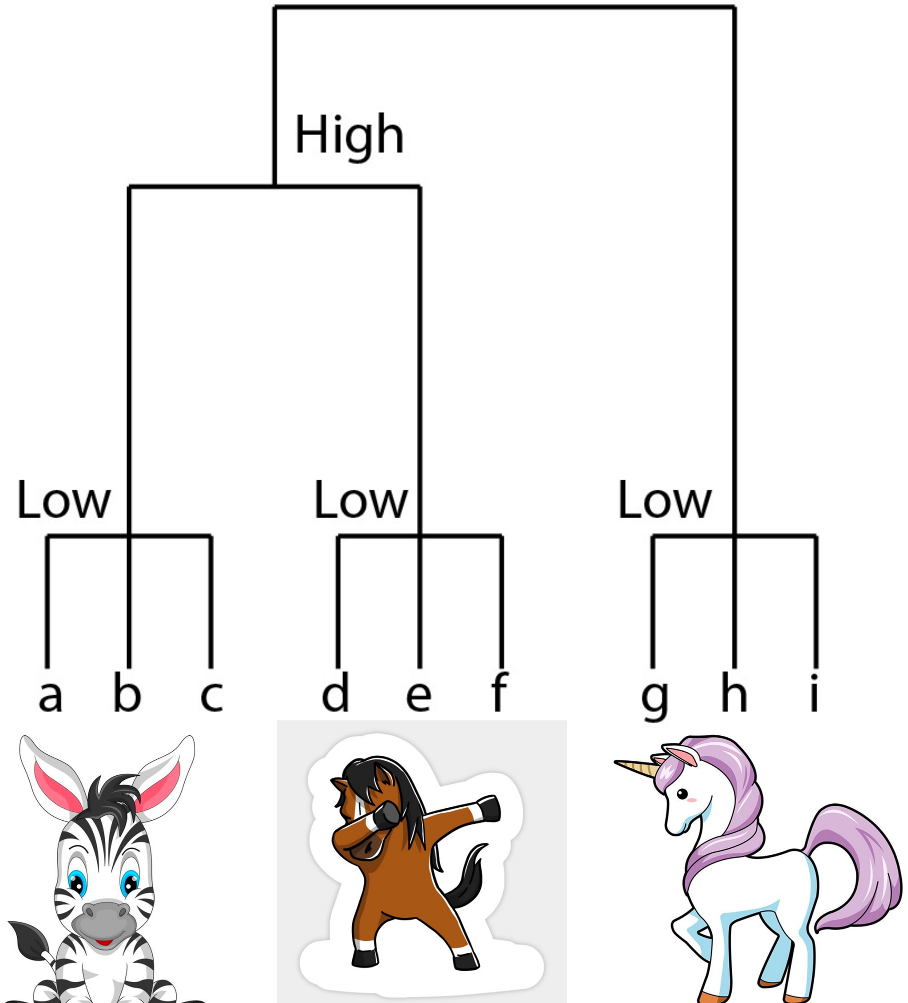
- We have 2 genes one slow (nuclear) one fast (mitochondrial)

- Nine taxa in the same family (Equidae)

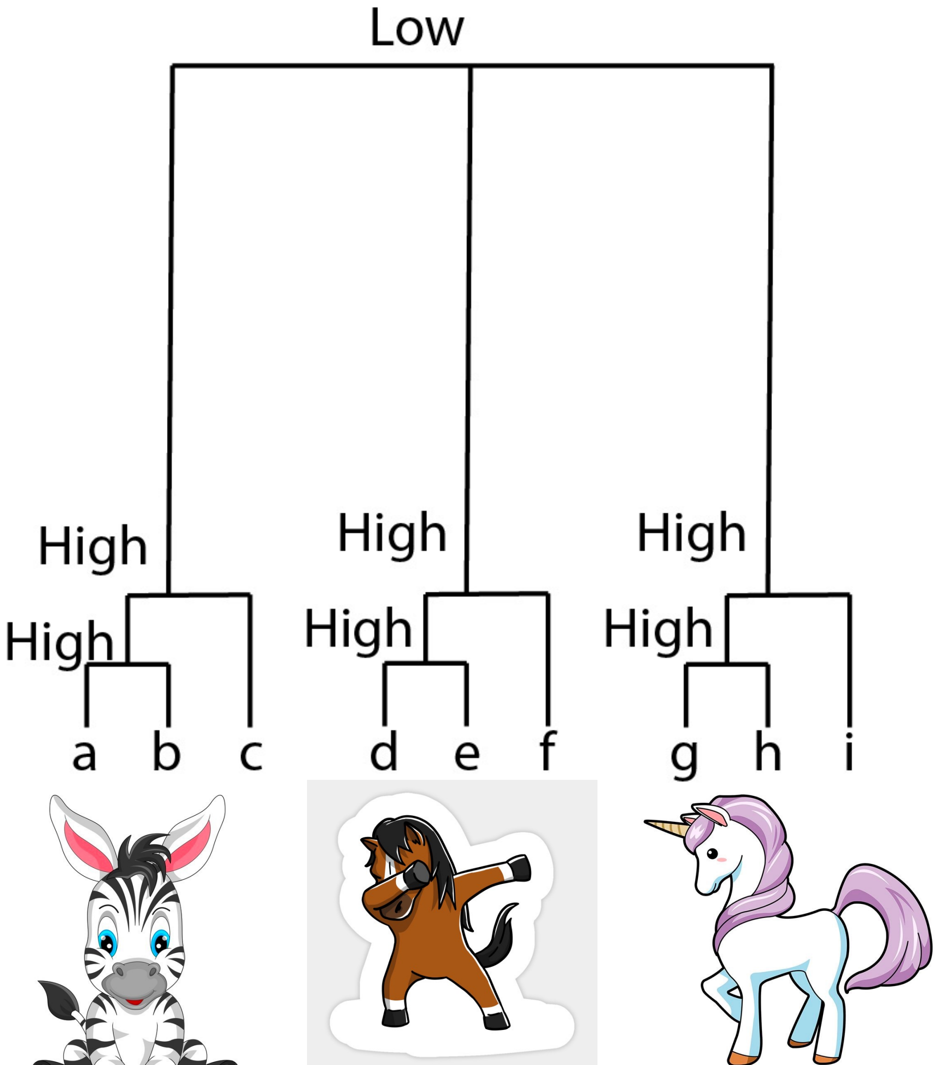
- What would that look like as gene trees and concatenated

Gene trees

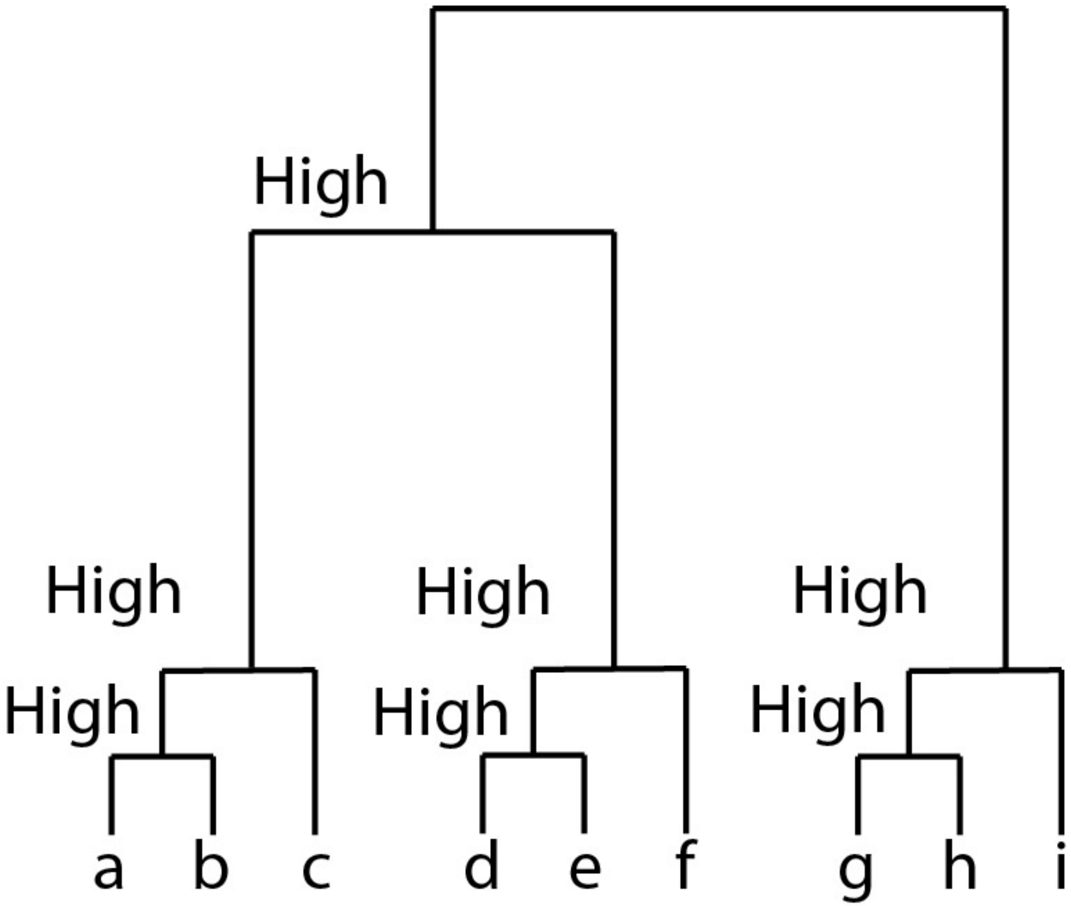
Slow gene



Fast gene



Concatenated tree



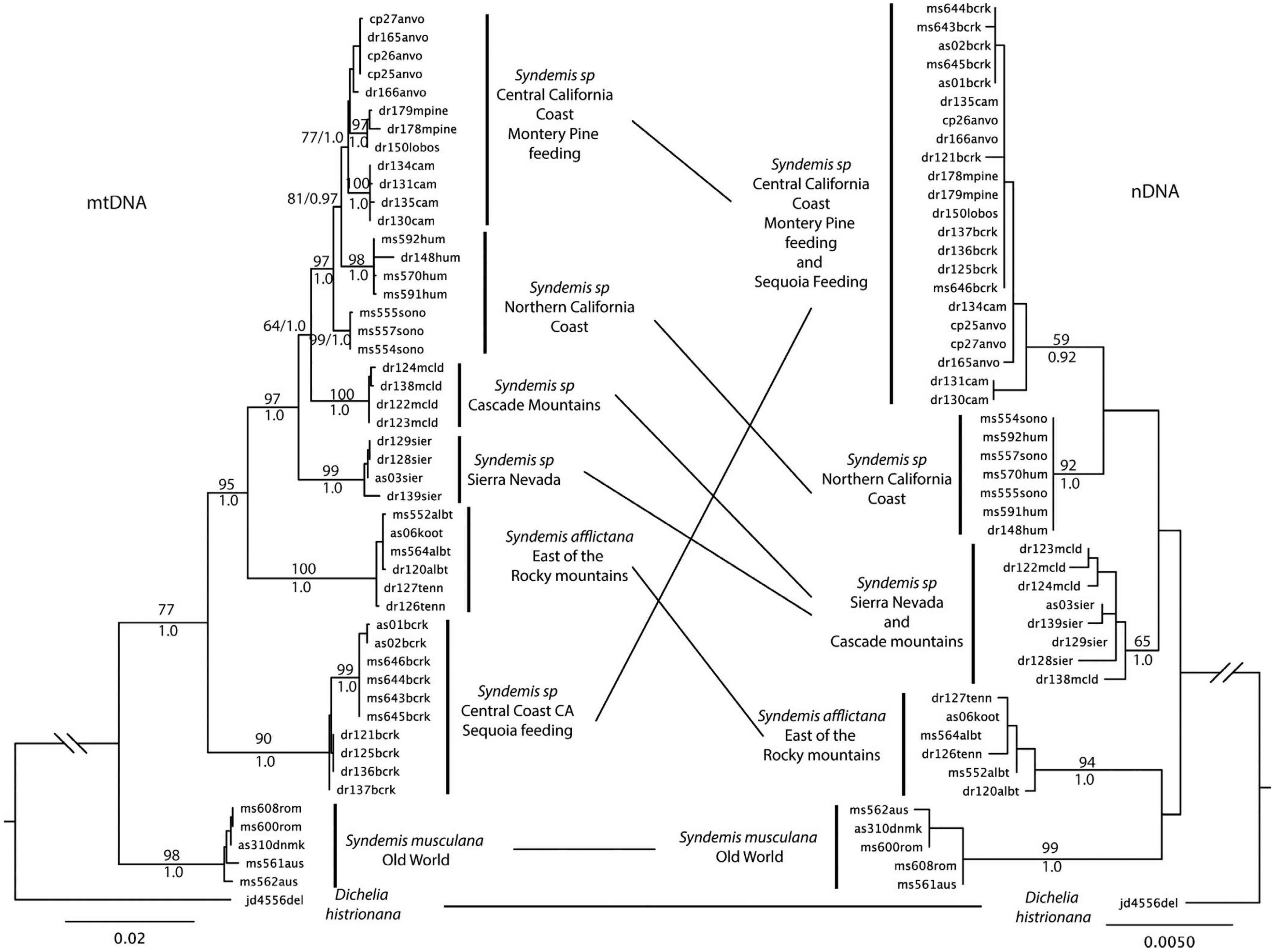
Problems with concatenation

- Genes can have entirely different evolutionary histories (mitochondrial vs nuclear)? Is it appropriate to concatenate?

Syndemis

- Tanglegram
- mtDNA/nDNA
- Different

- No concat



Rubinoff et al. 2017

Problems with concatenation

- Genes can have entirely different evolutionary histories (mitochondrial vs nuclear)? Is it appropriate to concatenate?

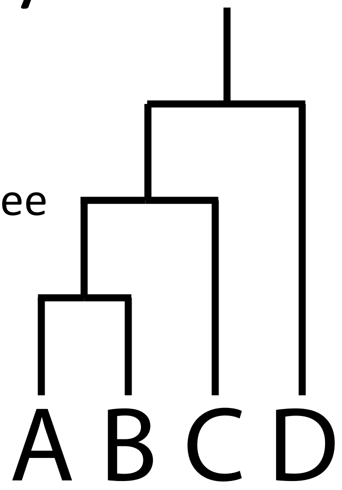
- Assumes no hybridization

- Assumes no incomplete lineage sorting

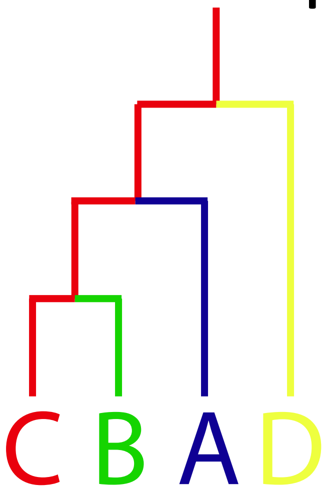
- Can give high support for wrong relationships (ILS/hybridization)

Hybridization vs incomplete lineage sorting

Species Tree



Gene Tree

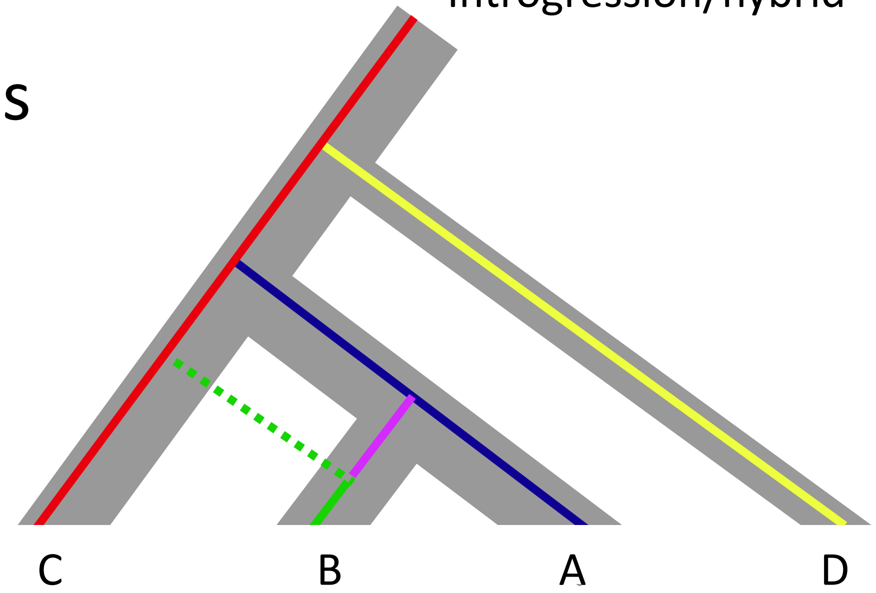


- Only difference is

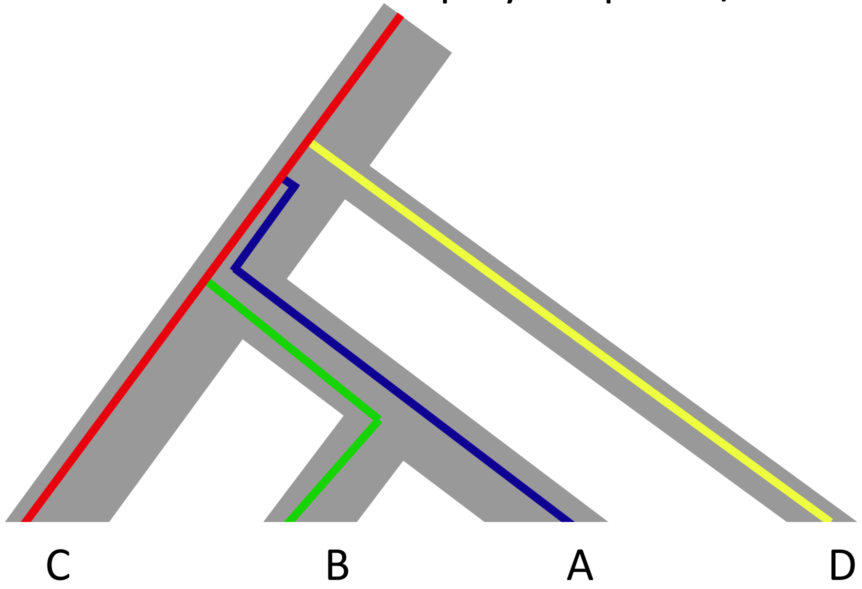
Branch lengths

ILS = longer

Hybrid=shorter



Introgression/hybrid

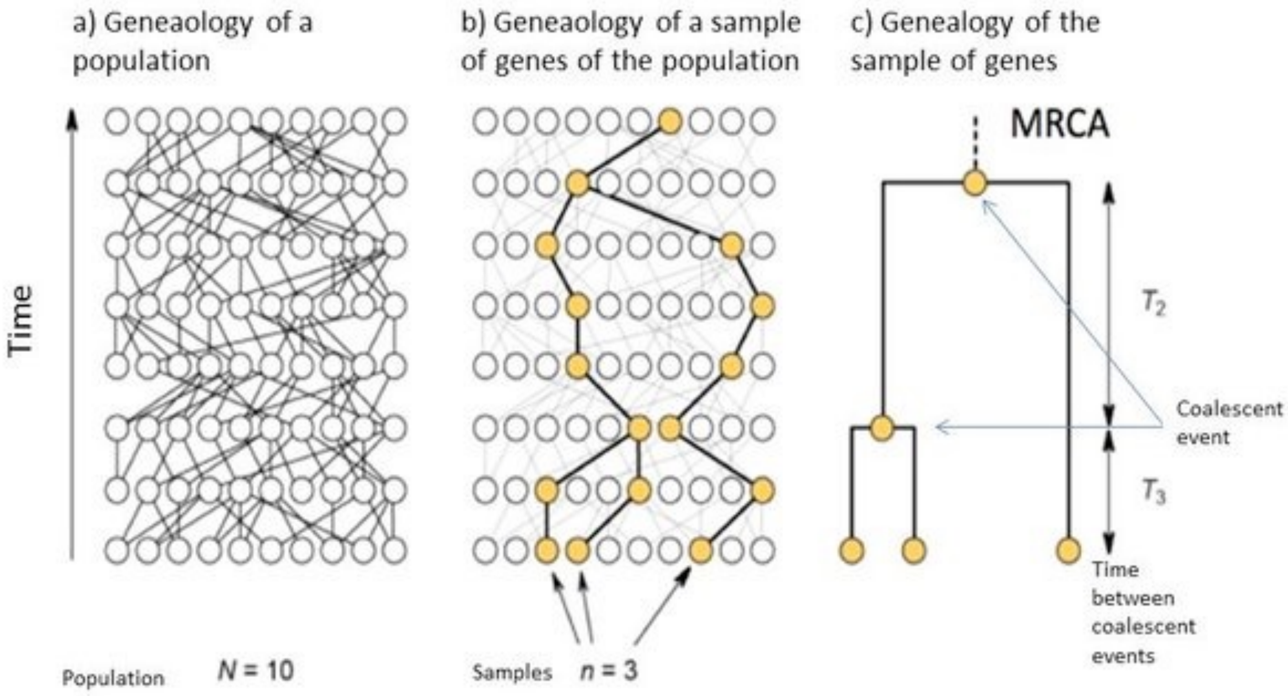


Ancestral polymorphism/ILS

Coalescence can account for ILS

- What is coalescence theory?

- Population genetic theory
  - Where alleles are sampled at random through discrete points in time.


a) Geneaology of a population
b) Geneaology of a sample of genes of the population
c) Genealogy of the sample of genes
Time
MRCA
T_2
T_3
Coalescent event
Time between coalescent events
Population $N = 10$
Samples $n = 3$

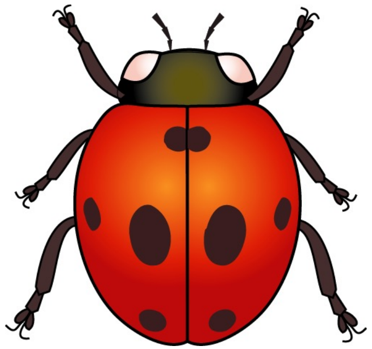
The Wright-Fisher Population Model

- Assumptions
    - Constant population size
    - Discrete and non-overlapping generations
    - Random mating (= panmixia)
    - Equal sex-ratio
    - Diploid
    - One locus
    - No Recombination

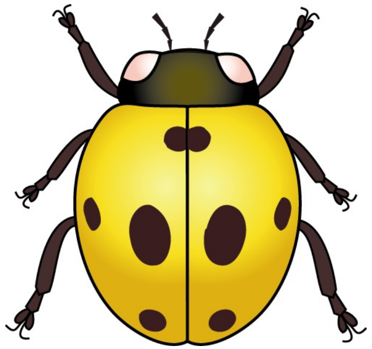
The Wright-Fisher Population Model

Consider a biallelic gene in a diploid organism

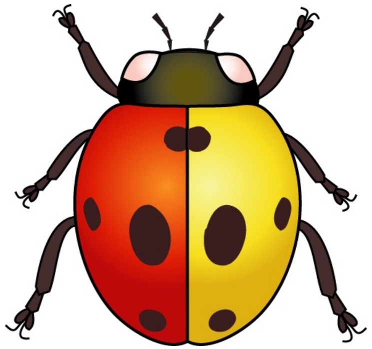
The wings of ladybeetles are colored to represent the alleles carried by each individual



2 Red Alleles



2 Yellow Alleles

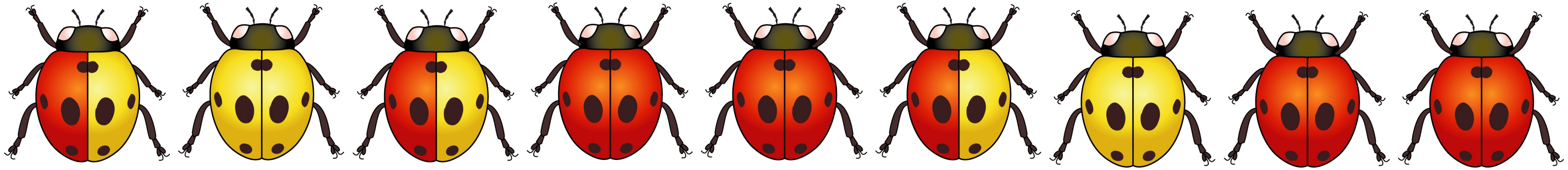


1 Red 1 yellow

The Wright-Fisher Population Model

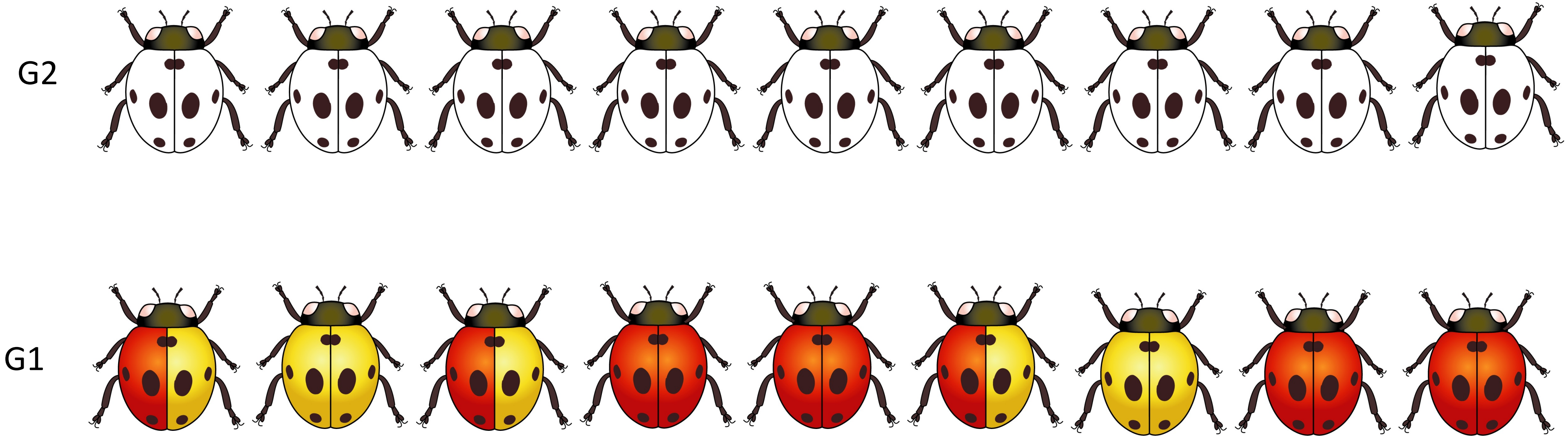
- Start with a population of size N
- N=9

G1



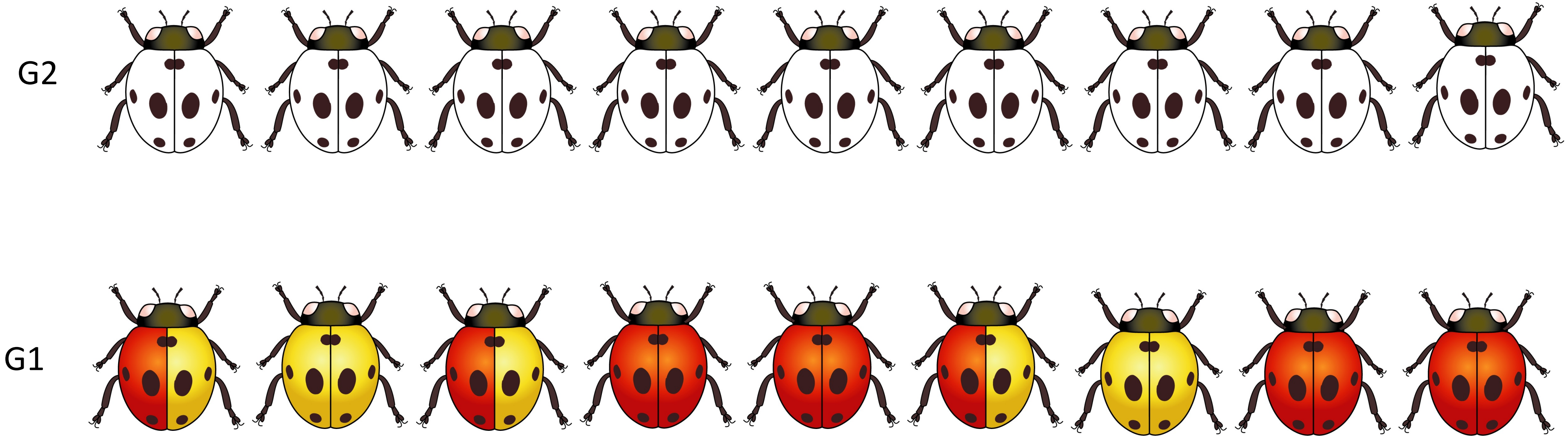
The Wright-Fisher Population Model

- Start with a population of size N
- N=9
- As soon as an individual dies it is replaced by a new offspring, so the population size remains constant



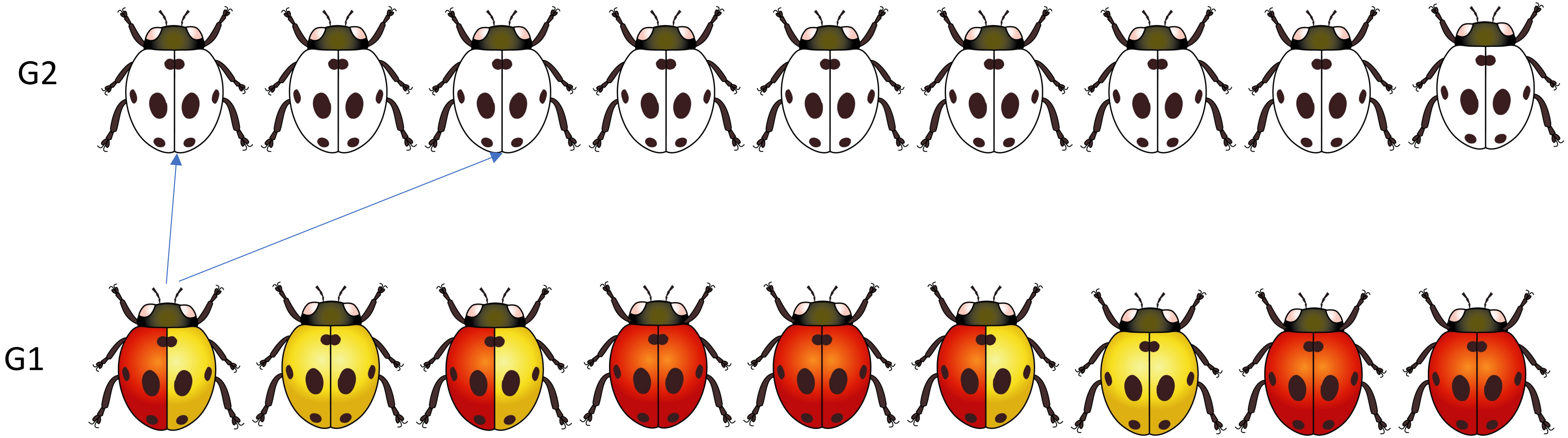
The Wright-Fisher Population Model

- Mating is random so phenotype of next generation is drawn randomly from last generation



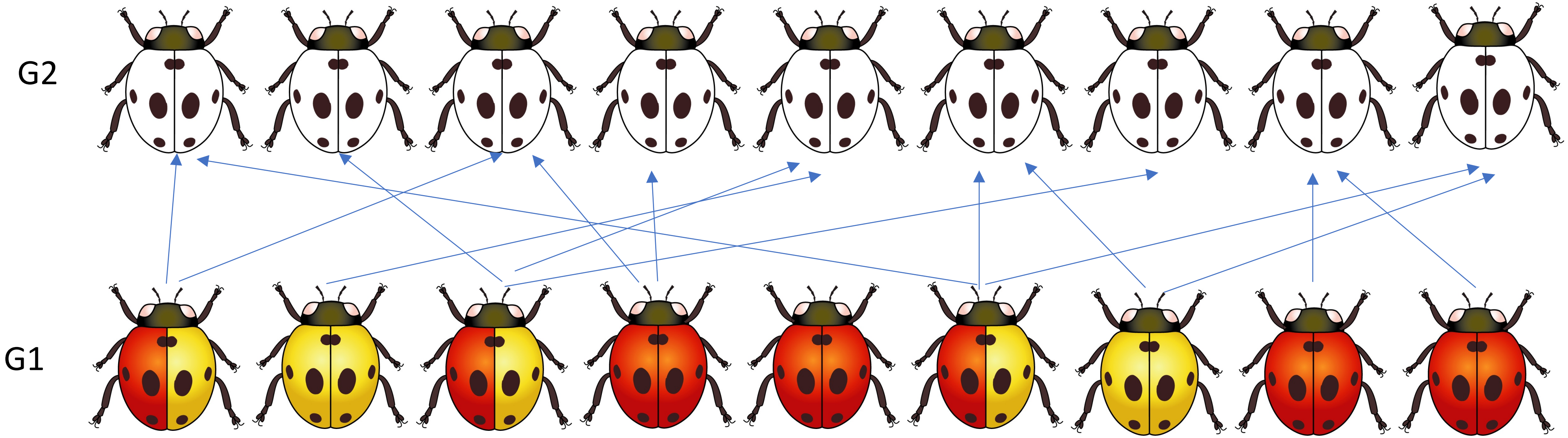
The Wright-Fisher Population Model

- Mating is random so phenotype of next generation is drawn randomly from last generation



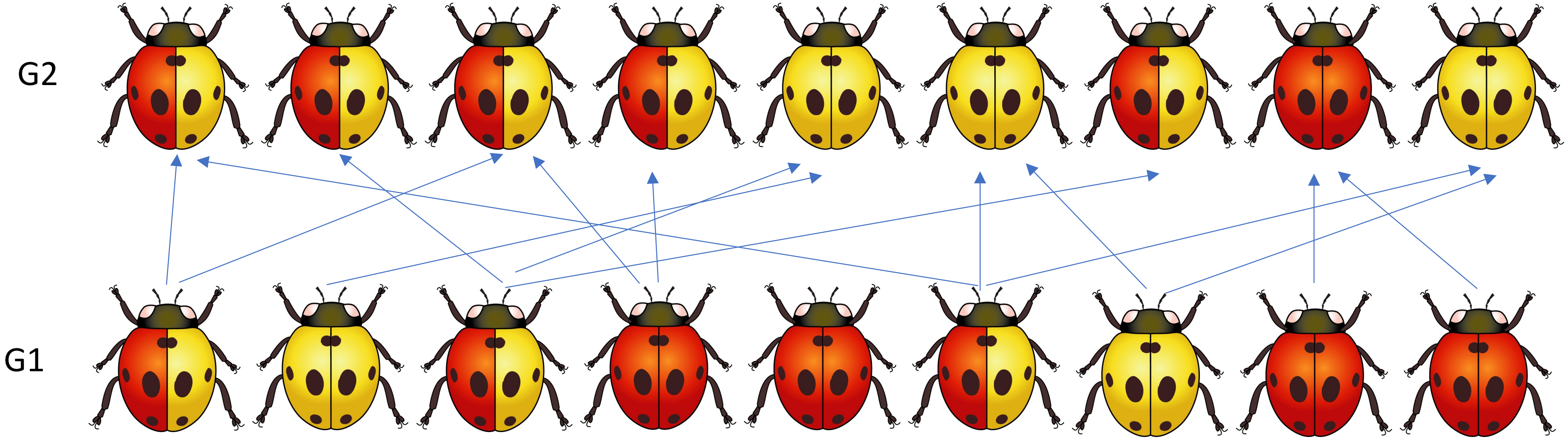
The Wright-Fisher Population Model

- Mating is random so phenotype of next generation is drawn randomly from last generation



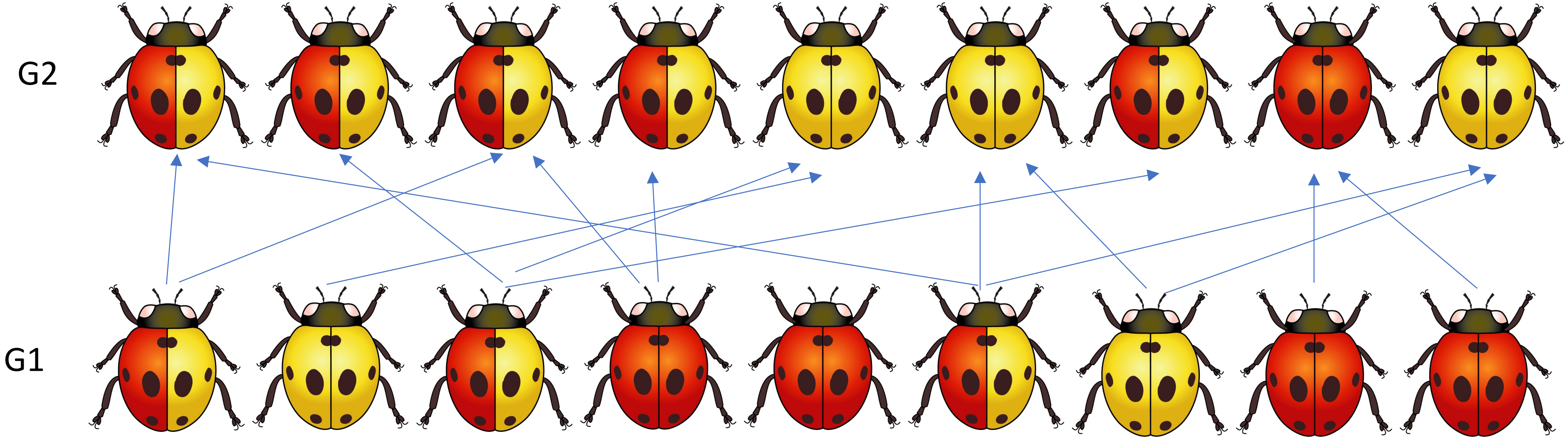
The Wright-Fisher Population Model

- Mating is random so phenotype of next generation is drawn randomly from last generation



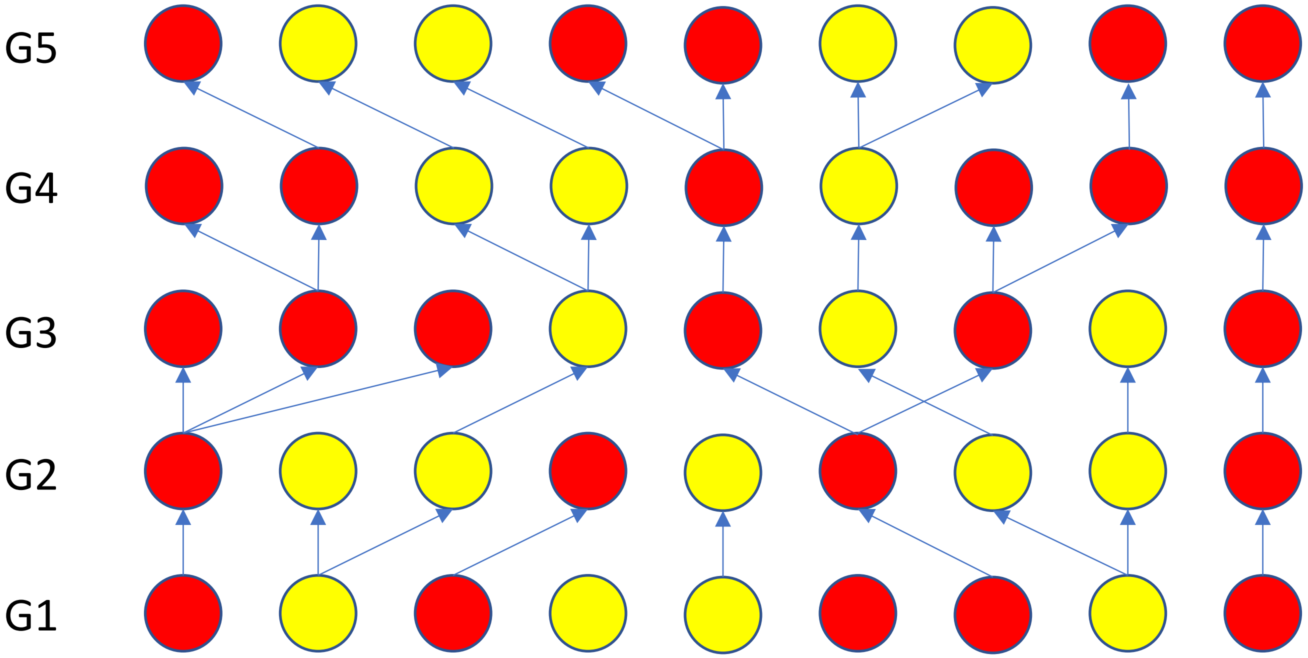
The Wright-Fisher Population Model

- Allele frequencies change through time
- Simple genetic drift



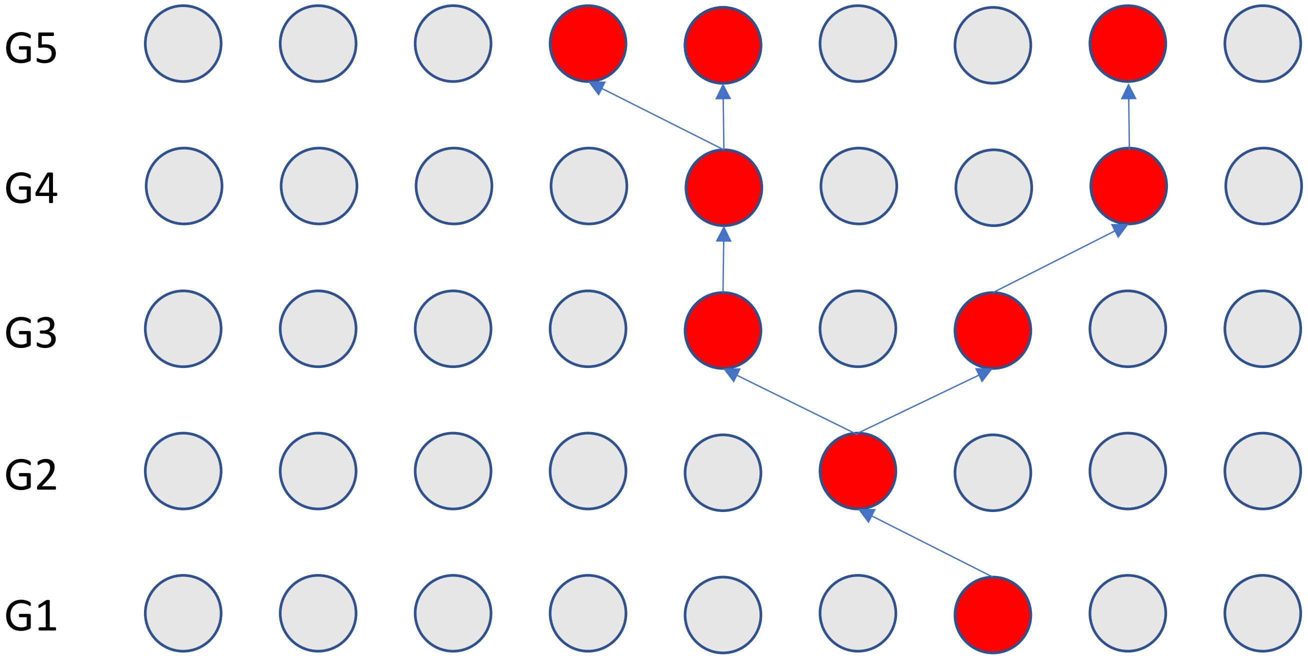
The Wright-Fisher Population Model

- Instead of lady beetles they can be modeled this way



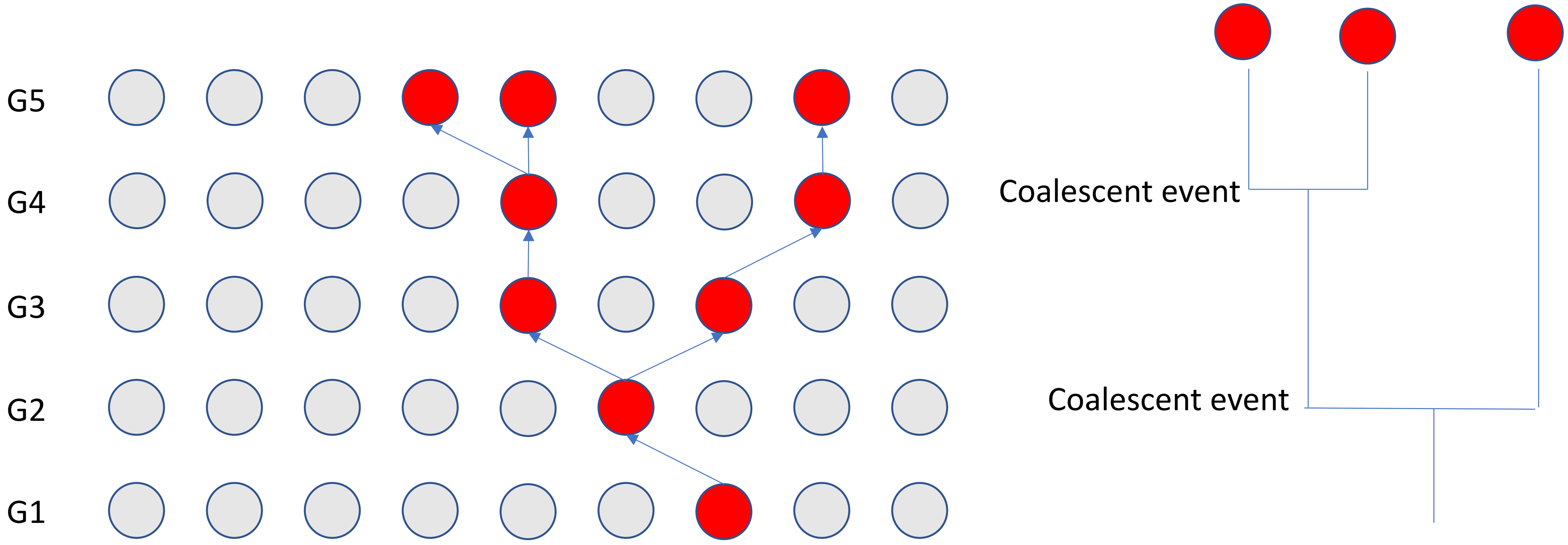
The Wright-Fisher Population Model

- Instead of ladybird beetles they can be modeled this way



The Wright-Fisher Population Model

- Instead of ladybird beetles they can be modeled this way

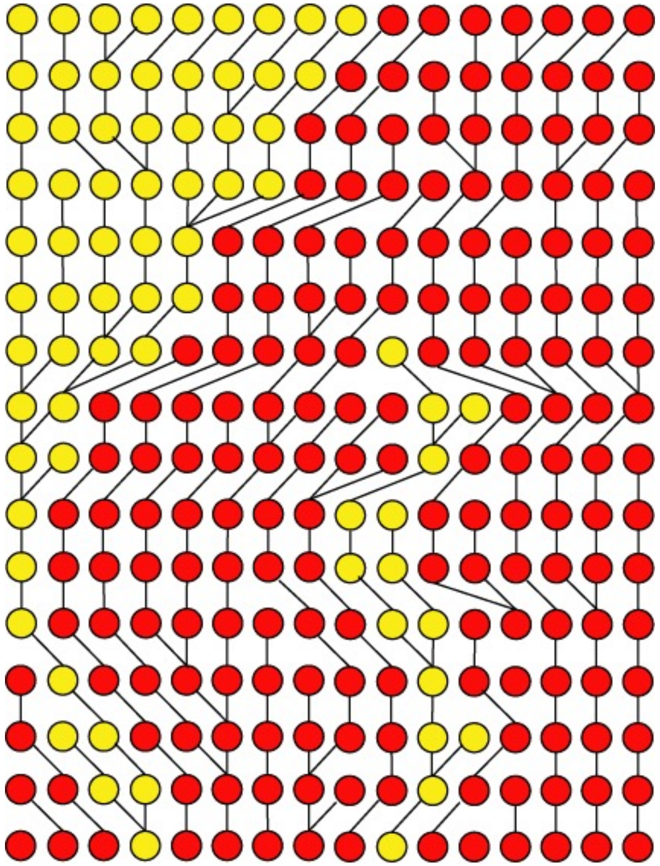


Time to coalescence = 2*Ne generations

The coalescent

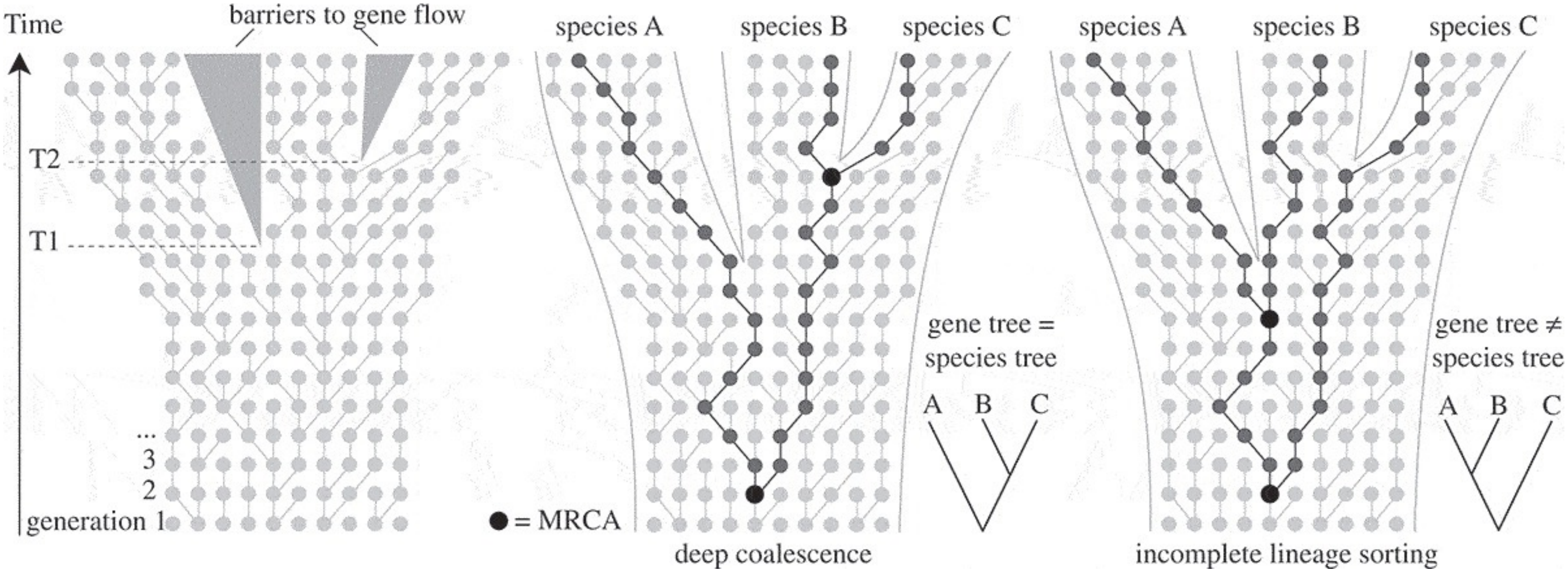
- It is a model of the distribution of coalescent events on a gene genealogy

- Based on a sample of extant gene copies and a model of evolution for a gene
- The coalescent can estimate population genetic parameters associated with coalescent events
  - Ne
  - selection



Multispecies coalescent model (MSC)

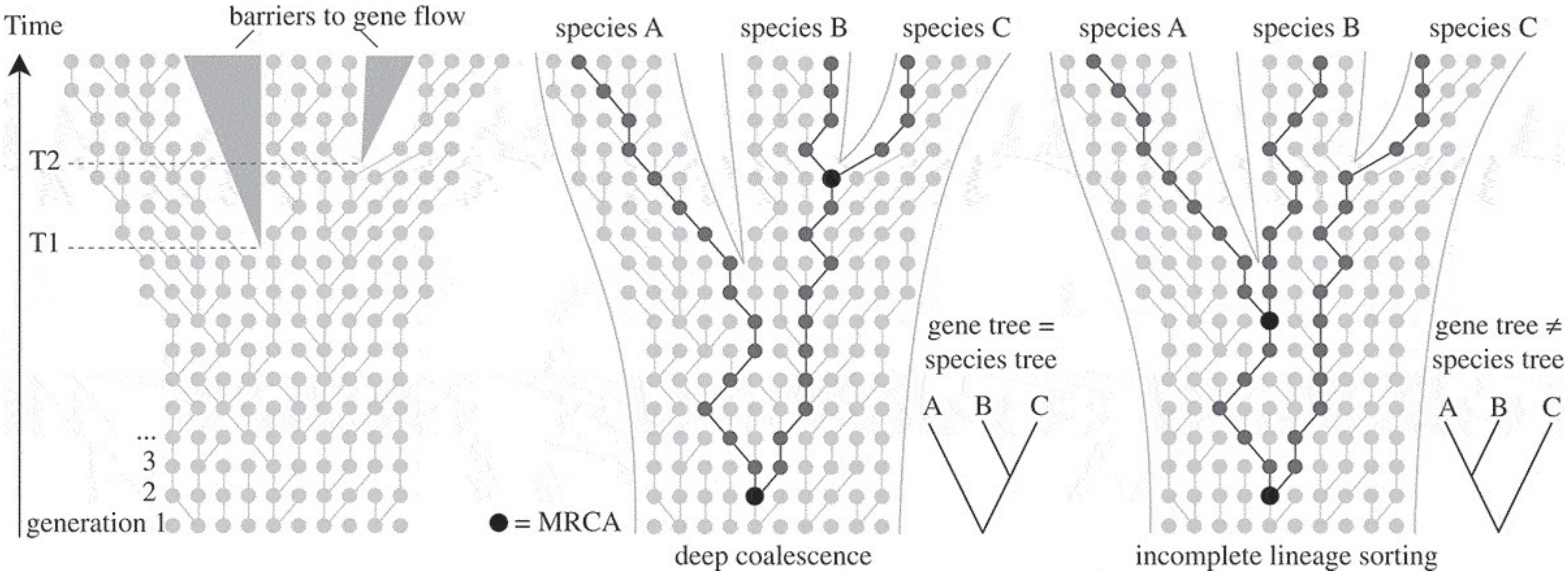
- Coalescent theory provides the link between phylogenetic models and the underlying population genetics.



Leliaert et al 2014, DNA-based species delimitation in algae

MSC model

- Able to account for ILS



Leliaert et al 2014, DNA-based species delimitation in algae

MSC model

- Attempting to model possible coalescent histories for genes given a species tree

- Coalescent histories are independent and random

- Coalescent events have to occur on species branches but can go back further in time than species divergences

So what do you need

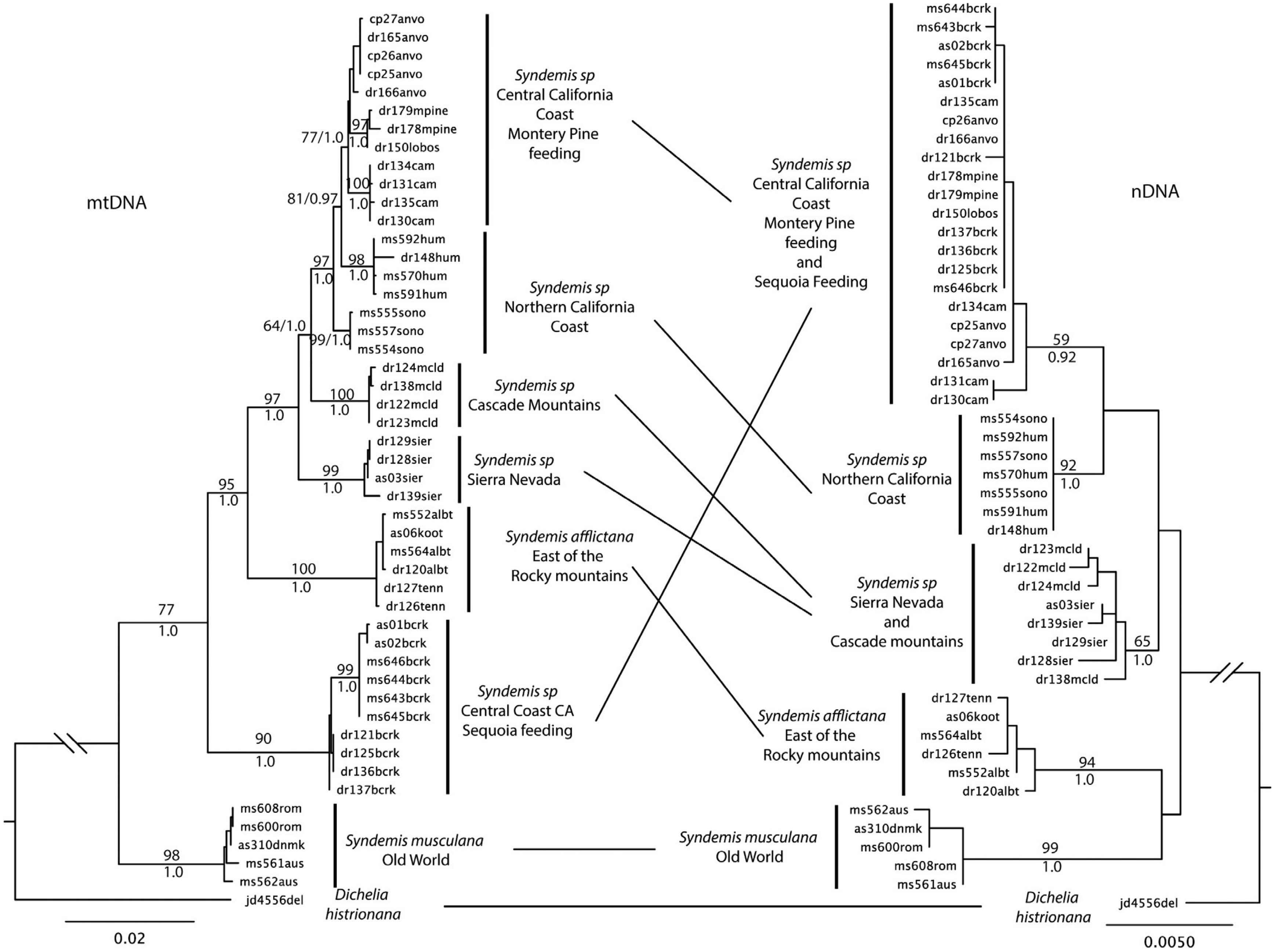
- At least a one gene copy /snp for each ”species” but multiple gene copies /snp are better.

- Multiple genes/snps

Syndemis

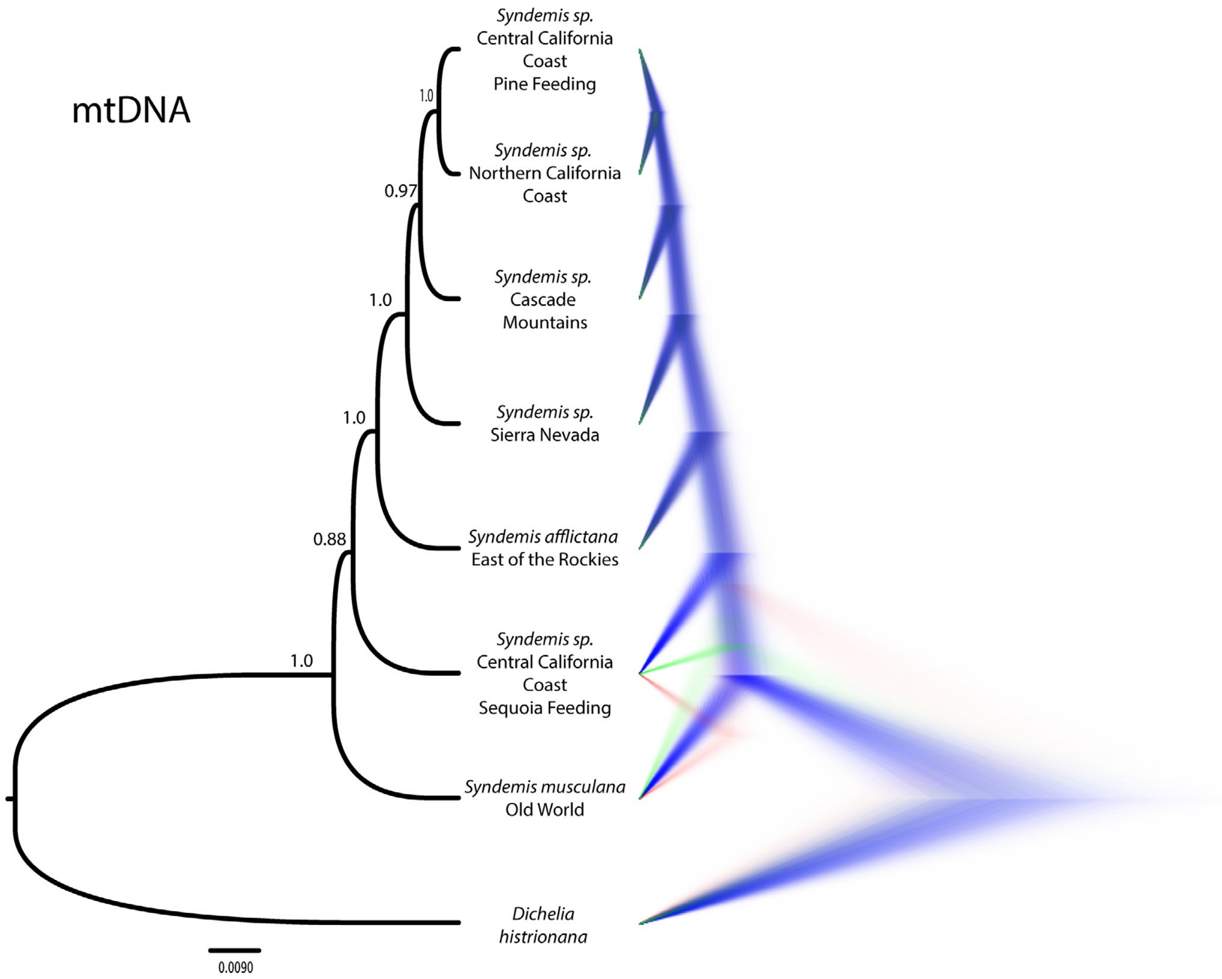
- mtDNA/nDNA
- Different

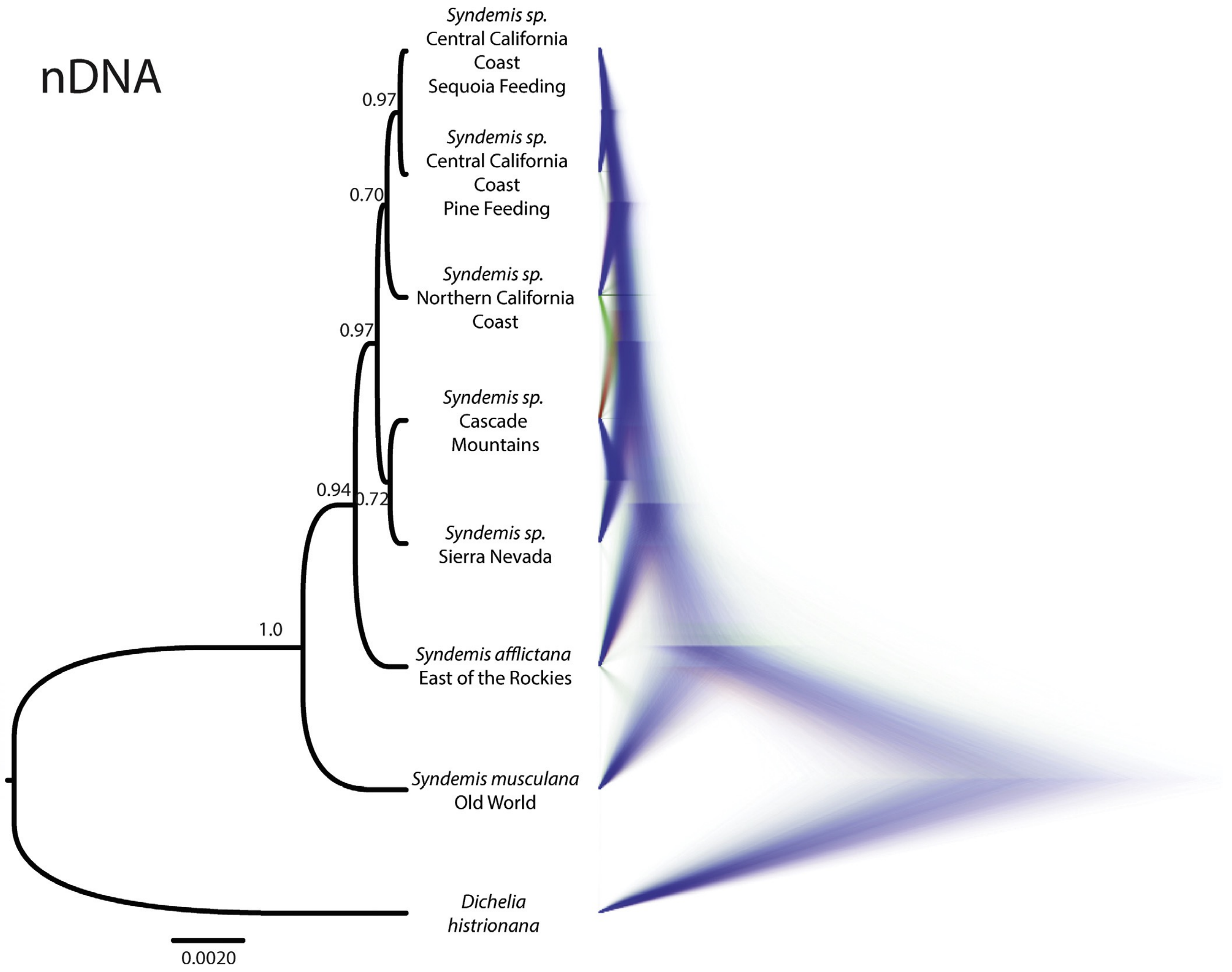
- No concat



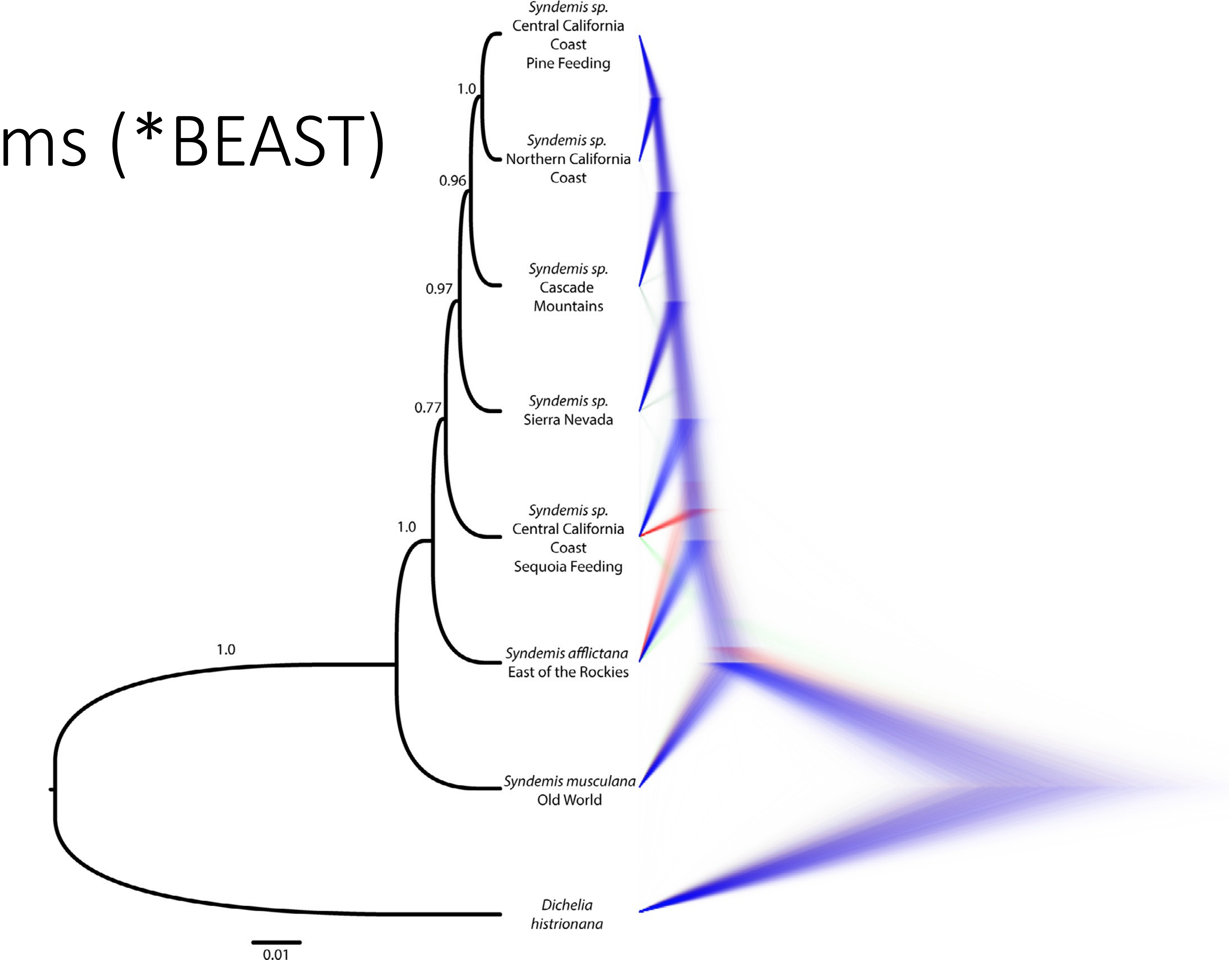
Rubinoff et al. 2017

Cloudograms (*BEAST)

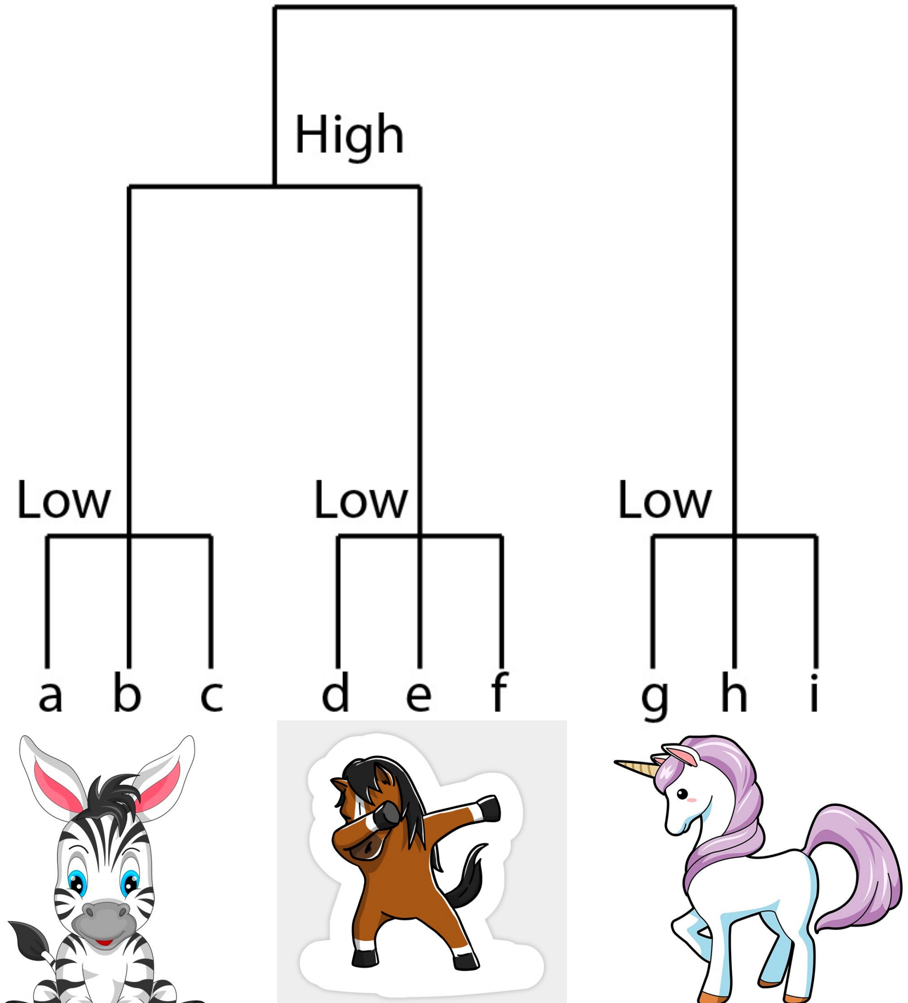

mtDNA
Syndemis sp. Central California Coast Pine Feeding
Syndemis sp. Northern California Coast
Syndemis sp. Cascade Mountains
Syndemis sp. Sierra Nevada
Syndemis afflictana East of the Rockies
Syndemis sp. Central California Coast Sequoia Feeding
Syndemis musculana Old World
Dichelia histrionana
1.0
0.97
1.0
1.0
0.88
1.0
0.0090


nDNA
Syndemis sp. Central California Coast Sequoia Feeding
Syndemis sp. Central California Coast Pine Feeding
Syndemis sp. Northern California Coast
Syndemis sp. Cascade Mountains
Syndemis sp. Sierra Nevada
Syndemis afflictana East of the Rockies
Syndemis musculana Old World
Dichelia histrionana
0.97
0.70
0.97
0.94
0.72
1.0
0.0020

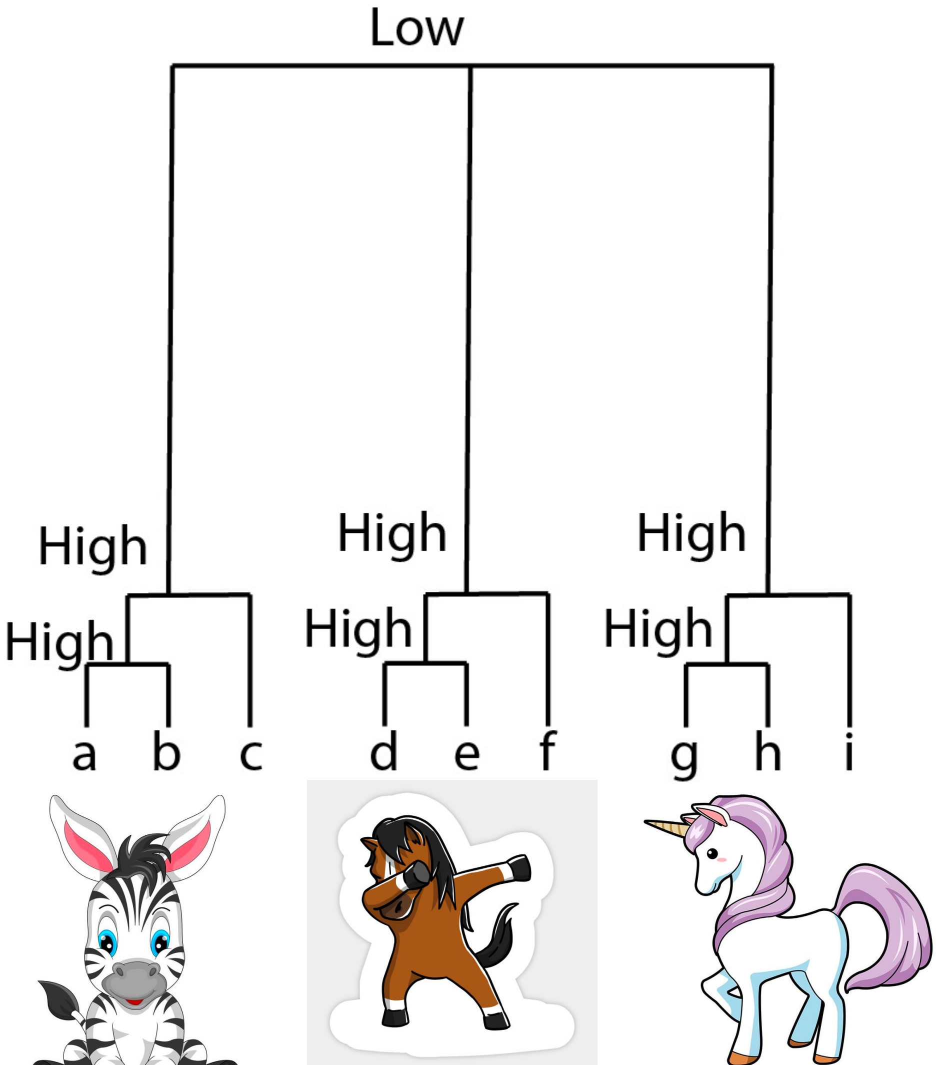
Cloudograms (*BEAST)



Gene trees

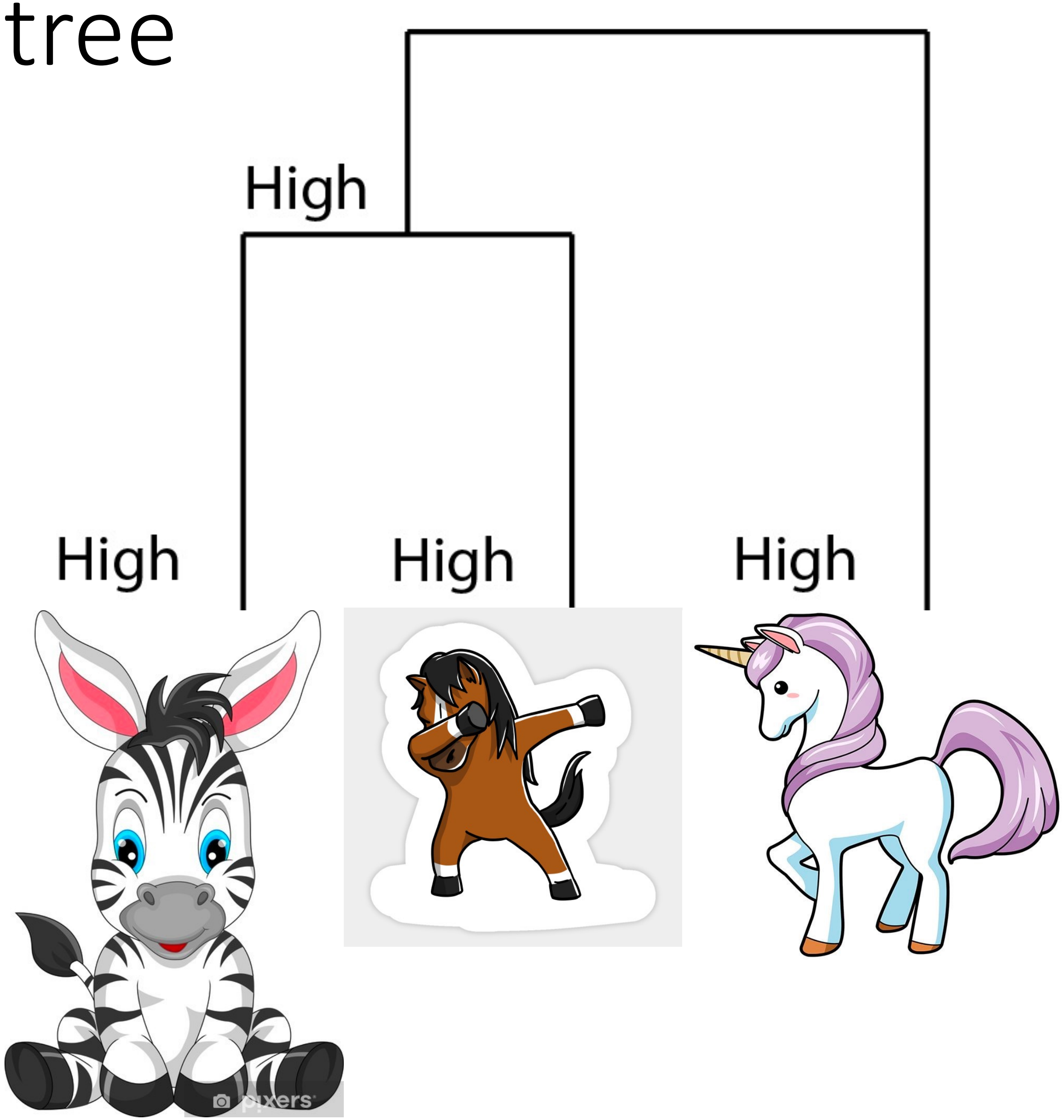


Slow gene



Fast gene

Species tree



Methods of MSC

- Short cut methods (likelihood)
  - Calculate individual gene trees through (IQTREE)
  - Then use these trees to calculate species trees
  - The trees become the data!

- More accurate when based on large number of loci (Phylogeneomic datasets)

- Astral is the most popular program

Methods of MSC

- Bayesian Estimation of Species Trees
  - Co-estimate species tree and gene trees all at once
    - P (species tree | P(gene trees|Data))
  - Computationally intensive
    - Can only run on a limited taxa and data ~20 species and 100 genes

- *BEAST is probably the most popular.