Introduction to the Multispecies Coalescent

Will Freyman

IB200, Spring 2016

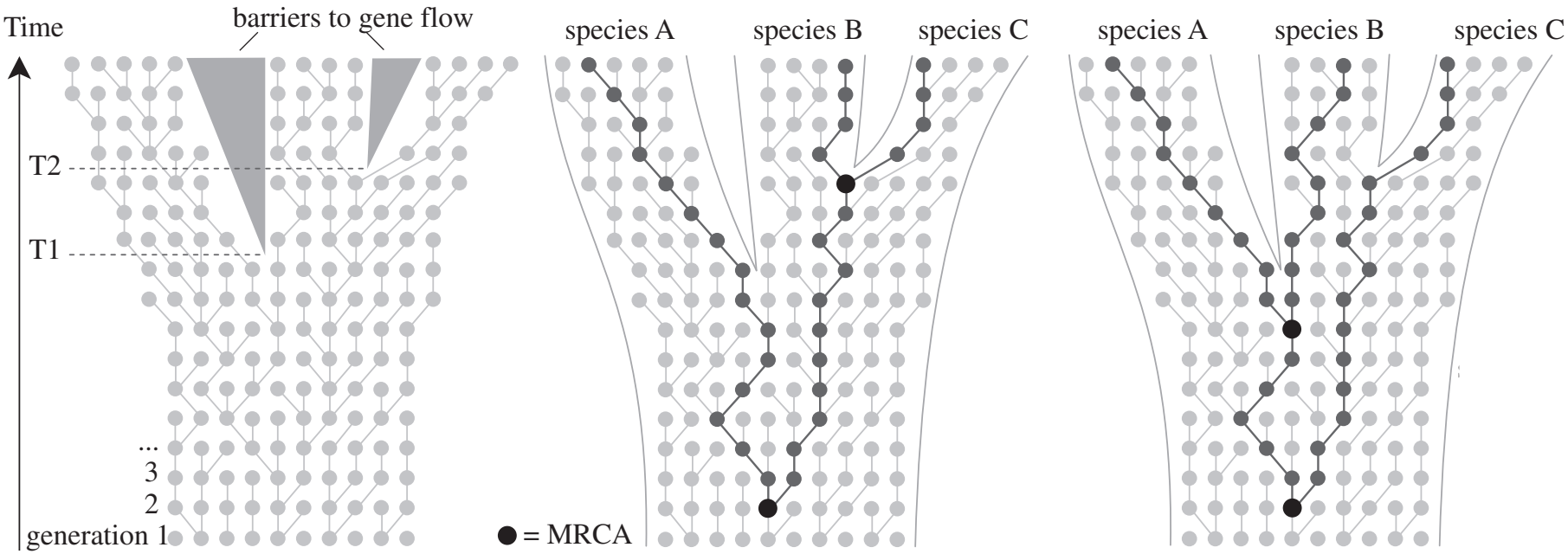


Image: Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European journal of phycology, 49(2), 179-196.

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

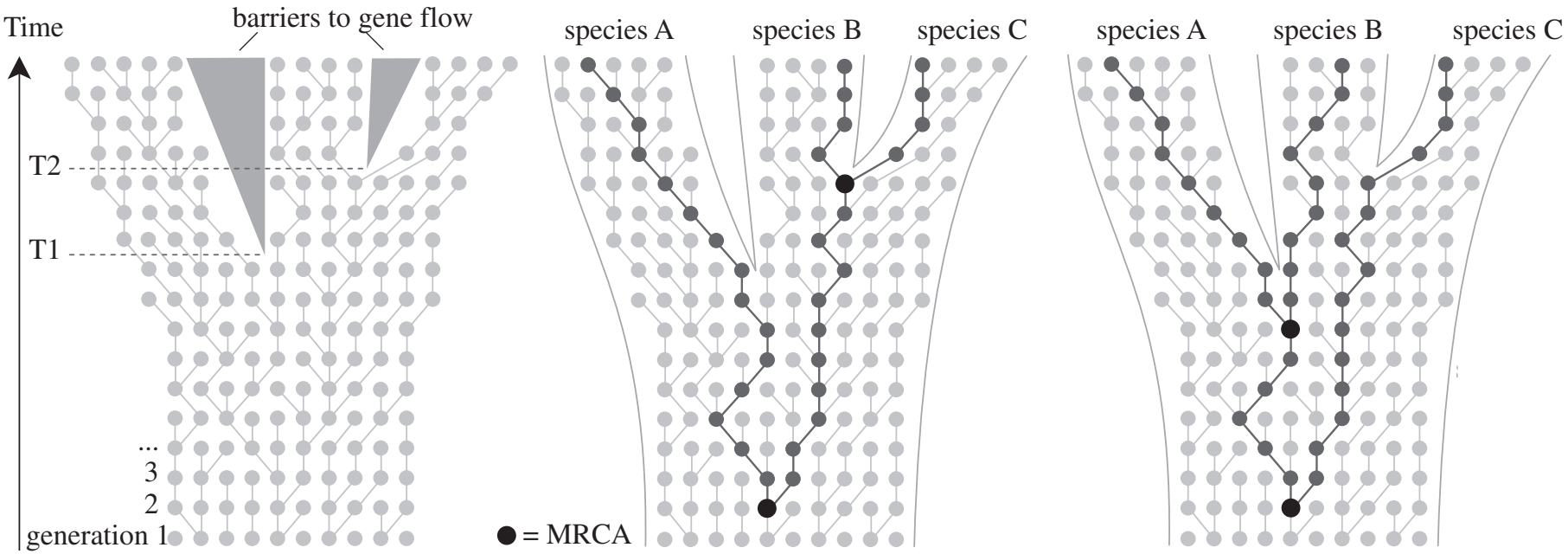


Image: Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European journal of phycology, 49(2), 179-196.

Coalescent theory allows us to model sources of gene tree-species tree incongruence that arise due to population level processes, and make estimates about demographic parameters.

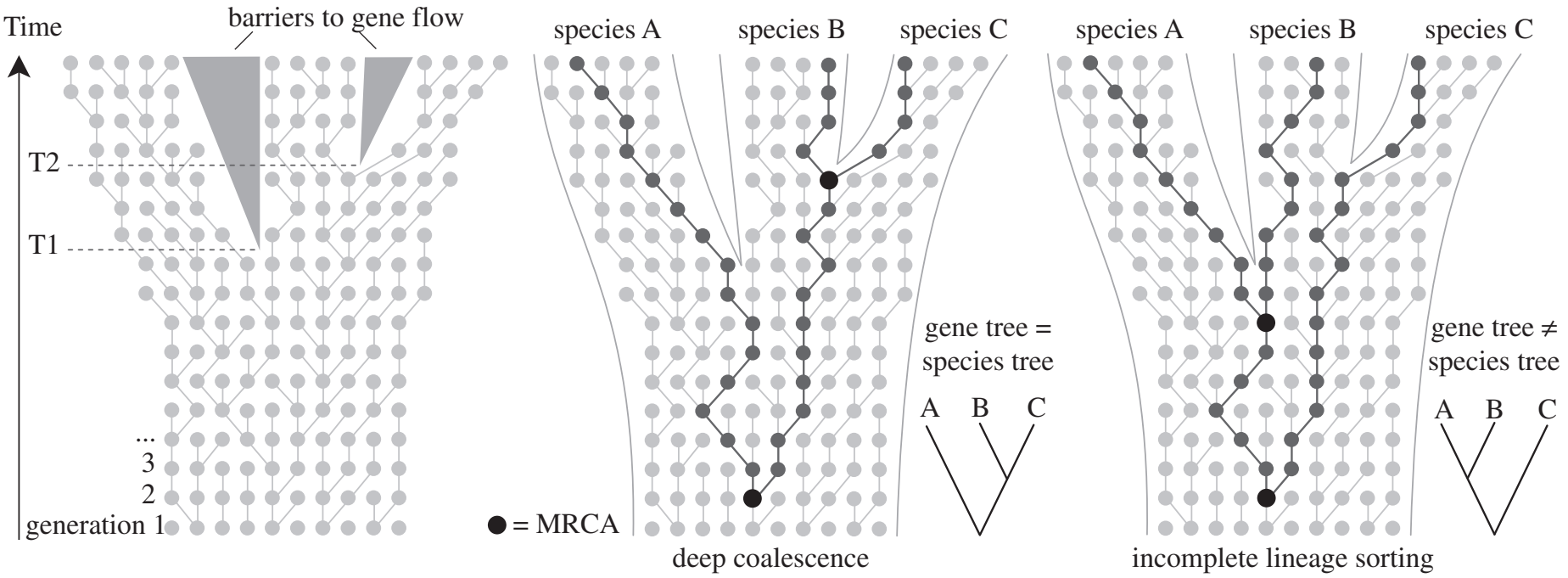


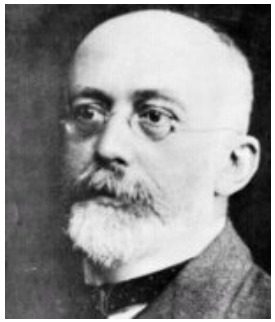
Image: Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European journal of phycology, 49(2), 179-196.

Classic population genetic models:

These models predict changes in allele frequencies forward in time according to neutral drift.

Hardy-Weinberg (1908) assumptions:

- Infinite population size
- Non-overlapping generations
- Random mating
- No selection
- No population structure
- No mutation

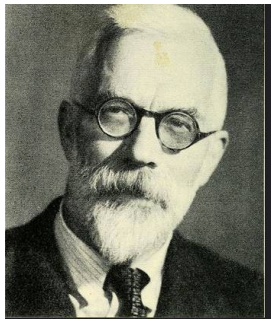


Classic population genetic models:

These models predict changes in allele frequencies forward in time according to neutral drift.

Wright-Fisher (1930) assumptions:

- **Finite** population size
- Non-overlapping generations
- Random mating
- No selection
- No population structure
- No mutation



Classic population genetic models:

These models predict changes in allele frequencies forward in time according to neutral drift.

Moran (1958) assumptions:

- Finite population size
- **Overlapping** generations
- Random mating
- No selection
- No population structure
- No mutation

Forward-time Wright-Fisher Model:

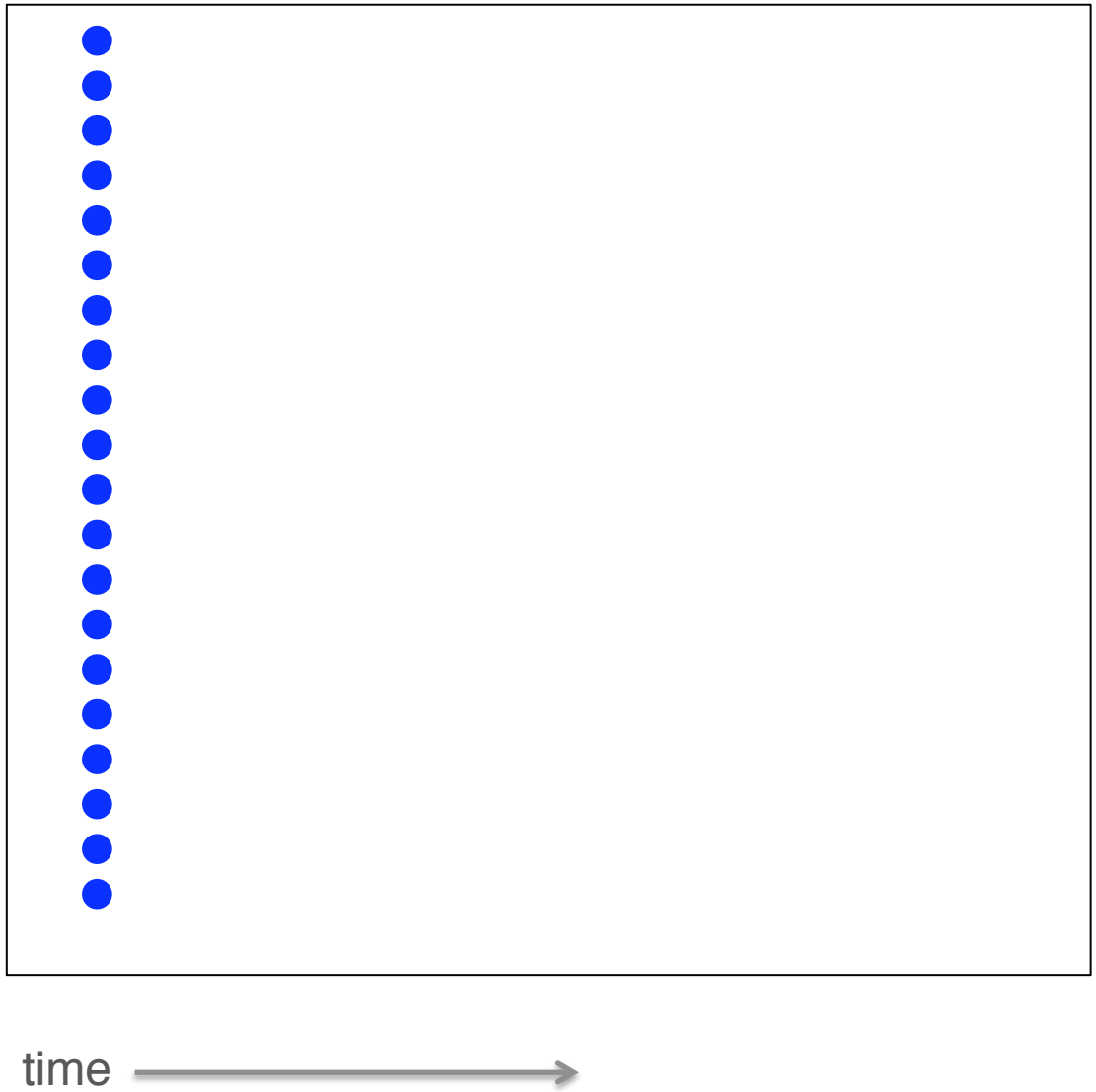


Image: Laura Kubatko's slides

Forward-time Wright-Fisher Model:

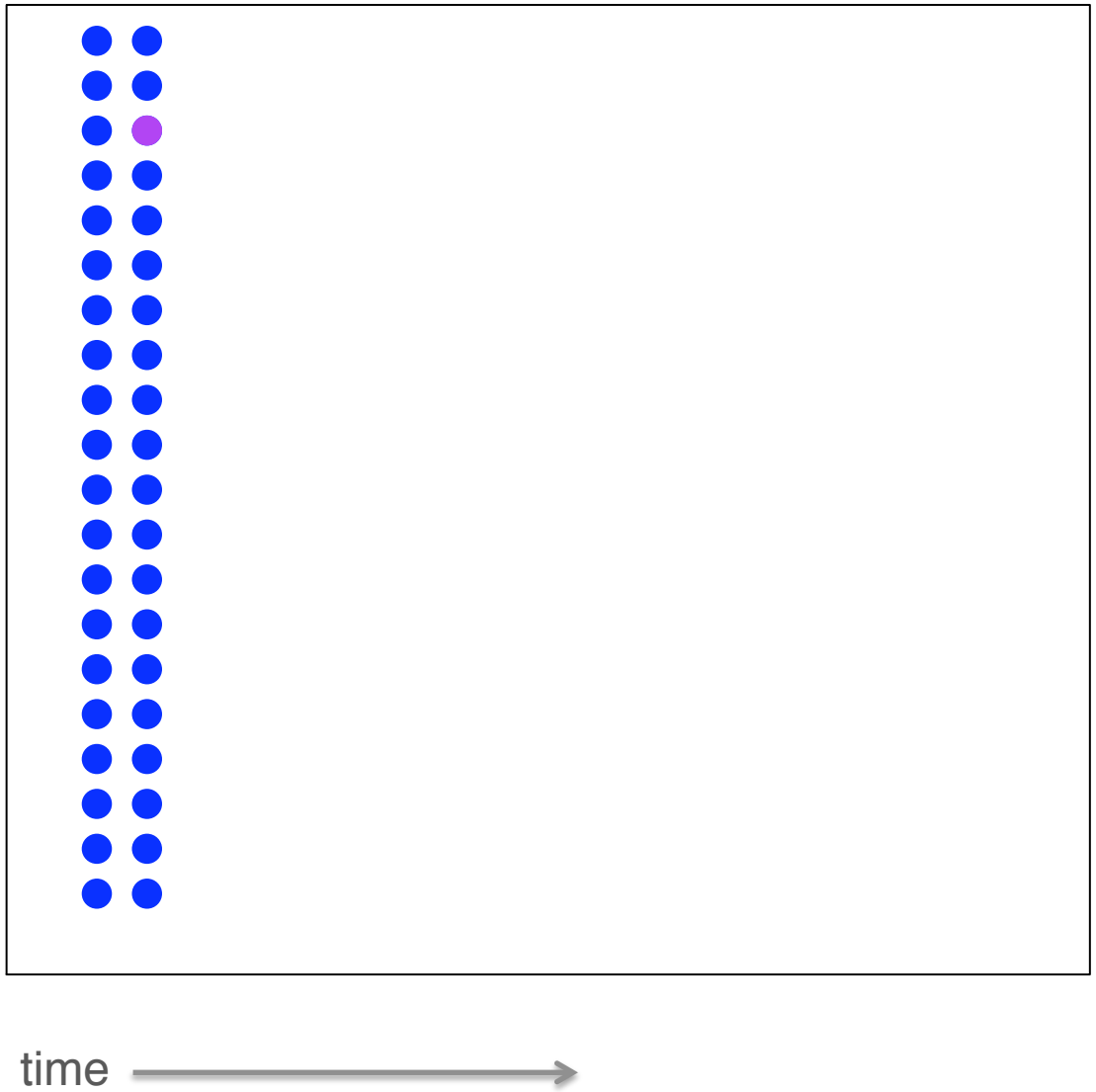


Image: Laura Kubatko's slides

Forward-time Wright-Fisher Model:

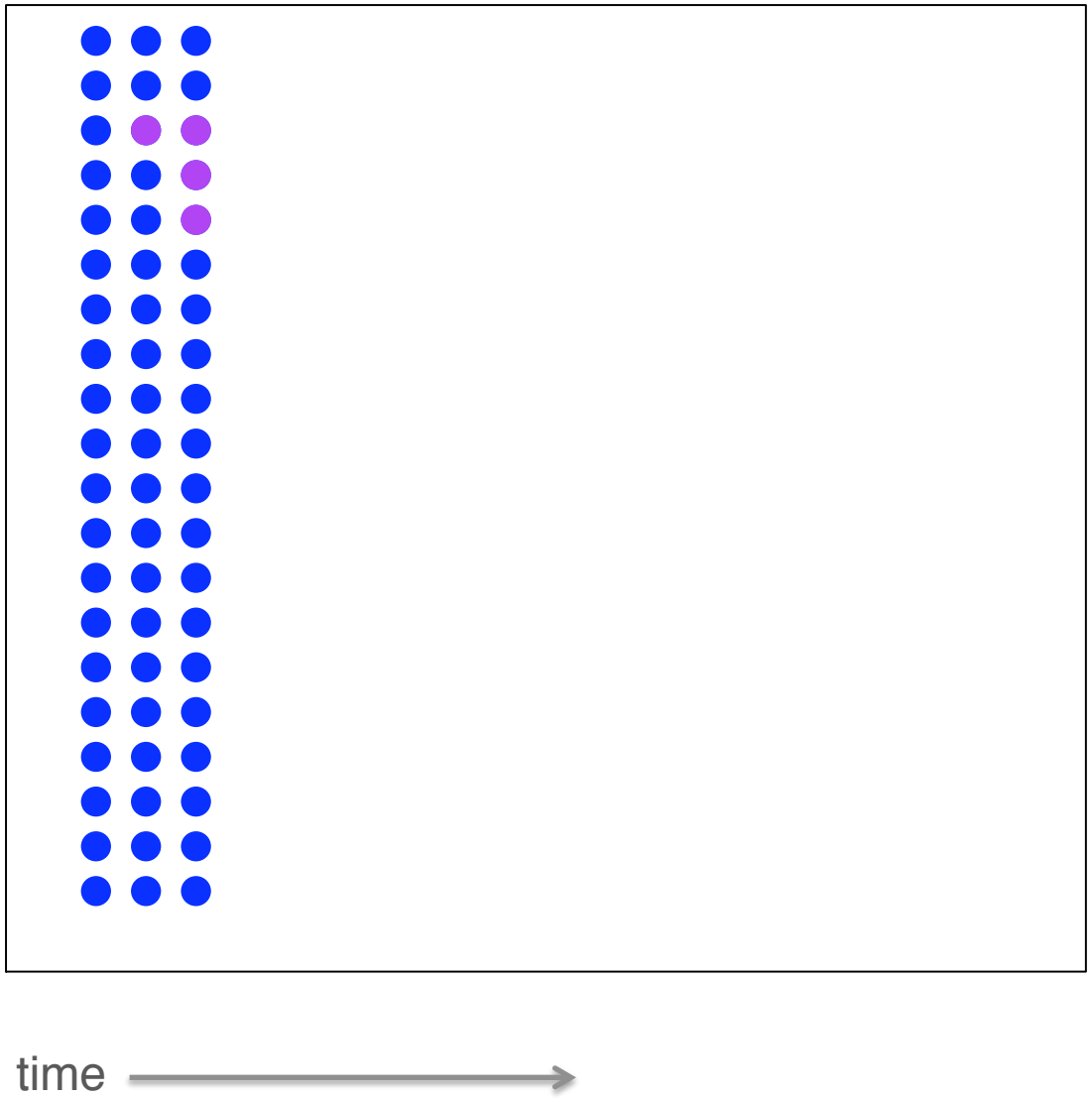
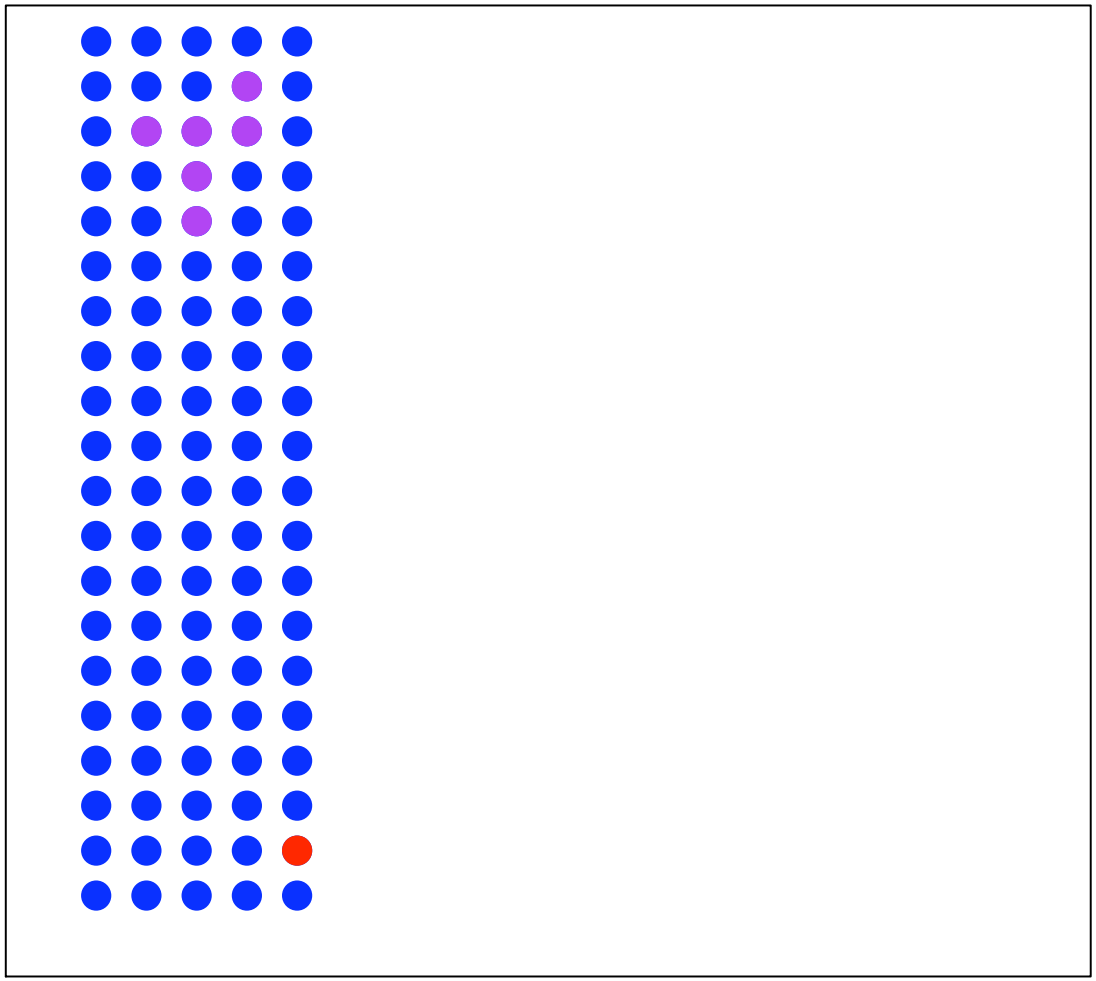


Image: Laura Kubatko's slides

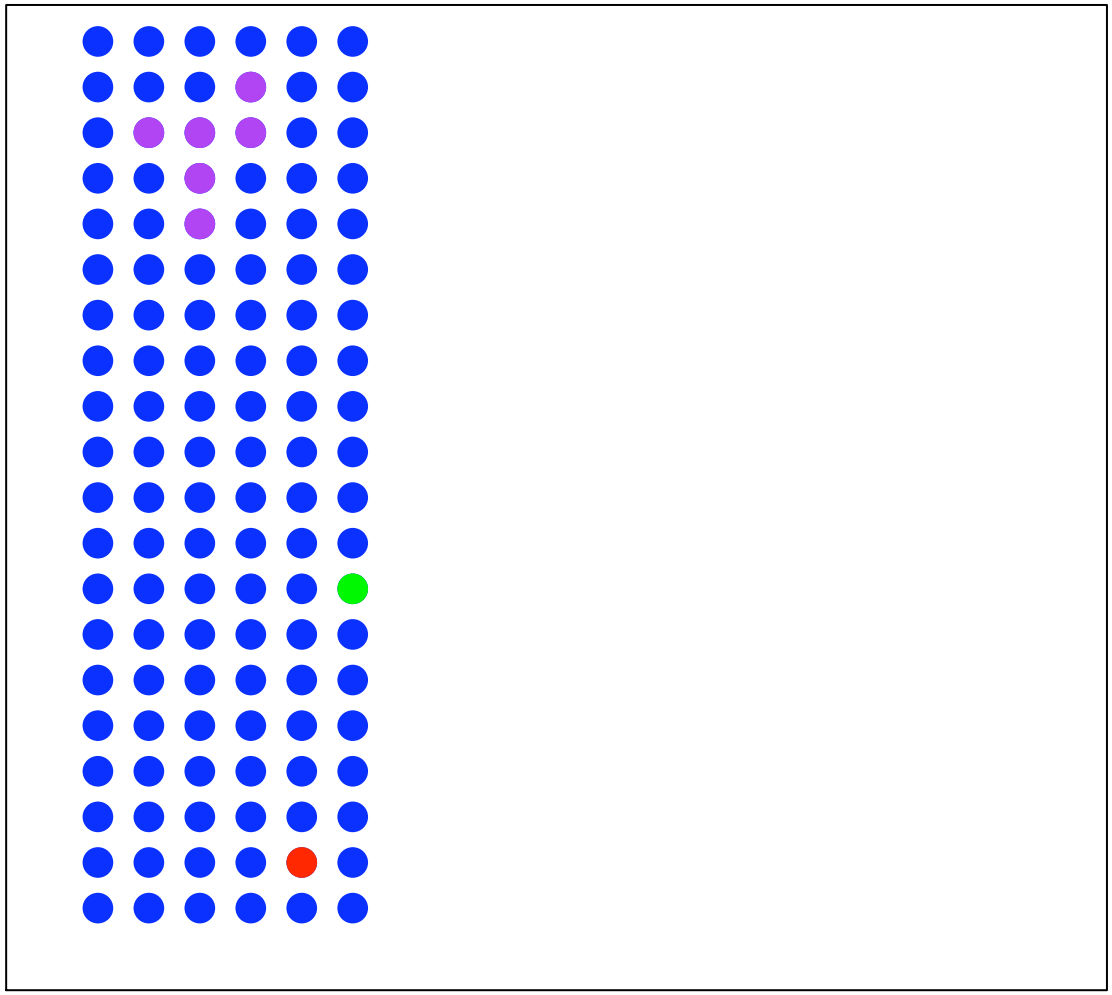
Forward-time Wright-Fisher Model:



time

Image: Laura Kubatko's slides

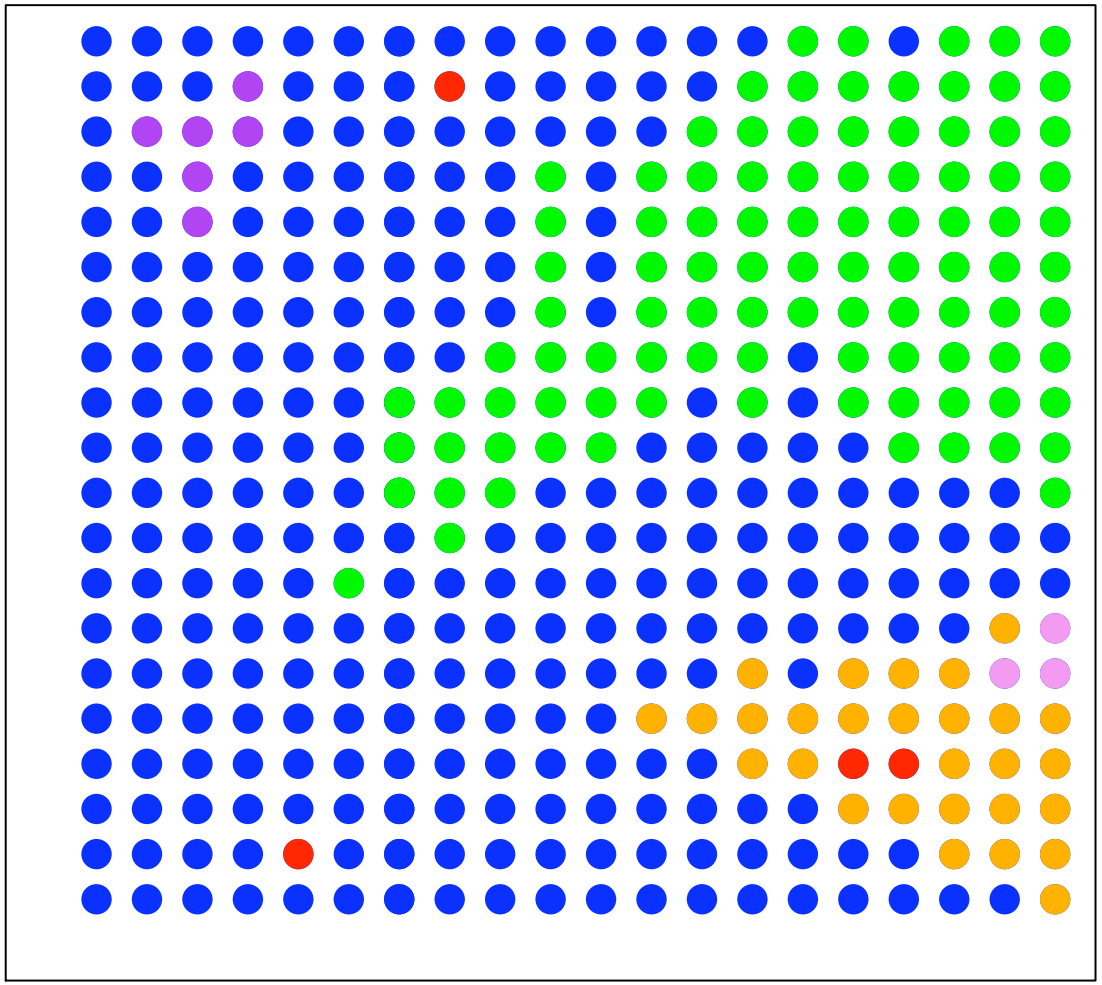
Forward-time Wright-Fisher Model:



time

Image: Laura Kubatko's slides

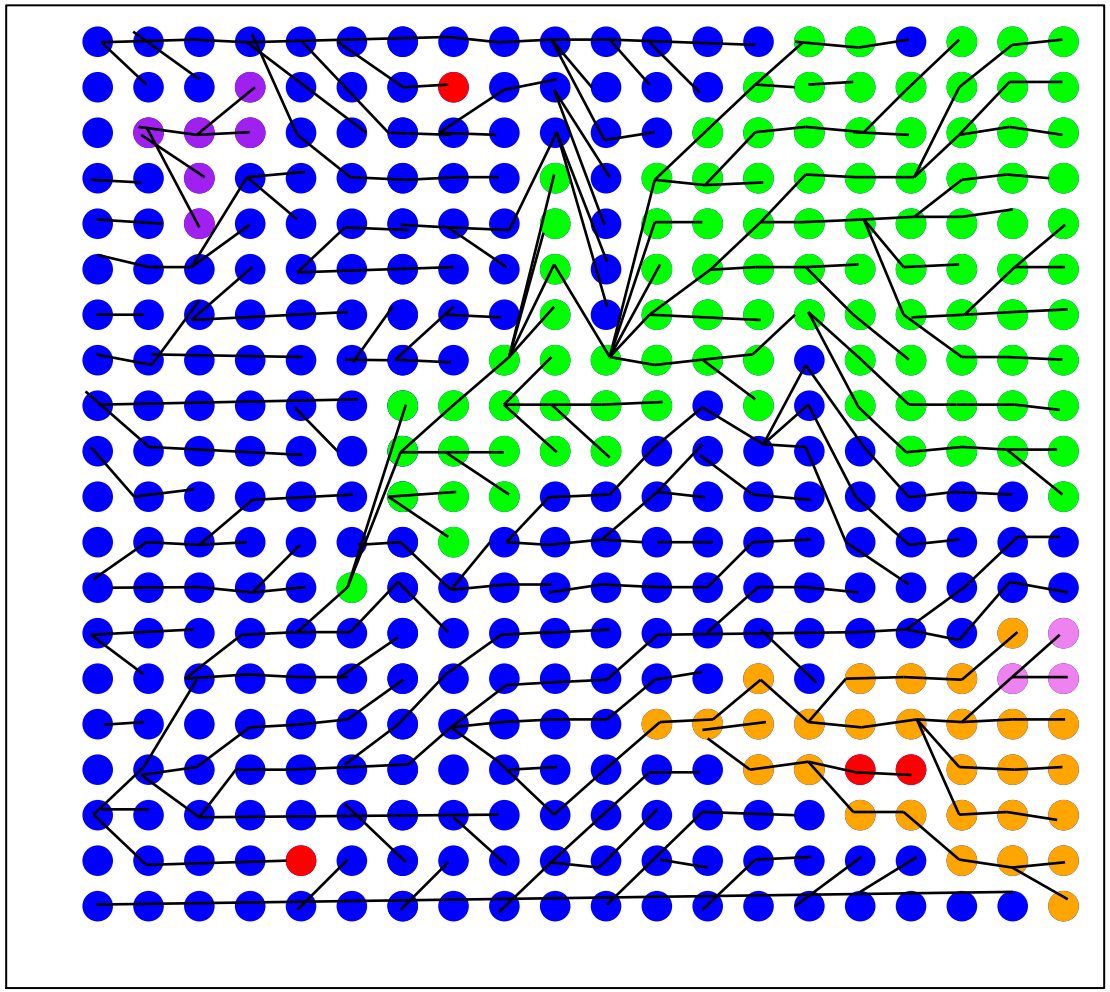
Forward-time Wright-Fisher Model:



time

Image: Laura Kubatko's slides

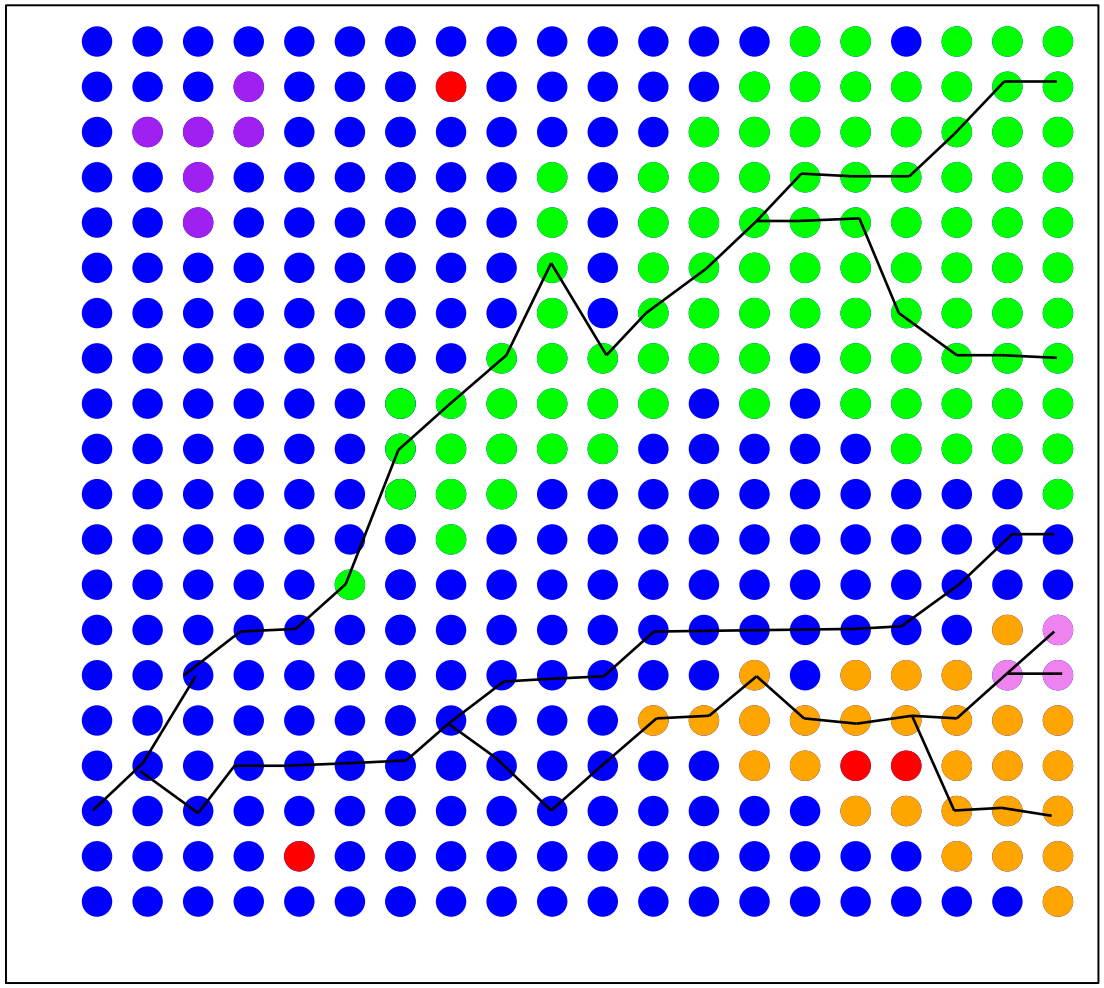
Forward-time Wright-Fisher Model:



time

Image: Laura Kubatko's slides

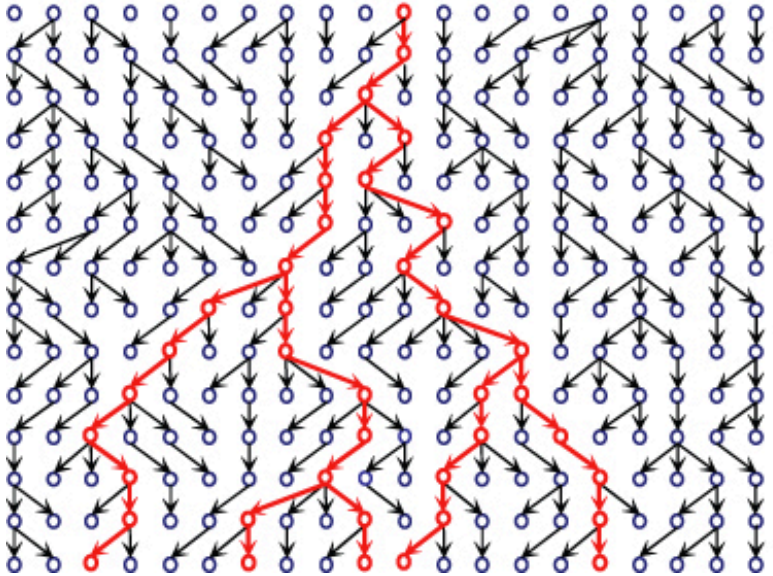
Coalescent theory: looking at the same process backwards in time



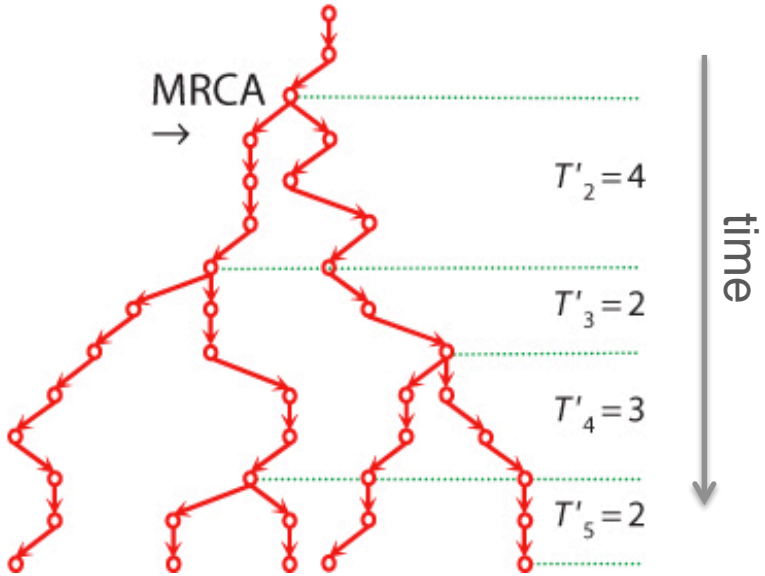
time

Image: Laura Kubatko's slides

(a) Fisher-Wright model



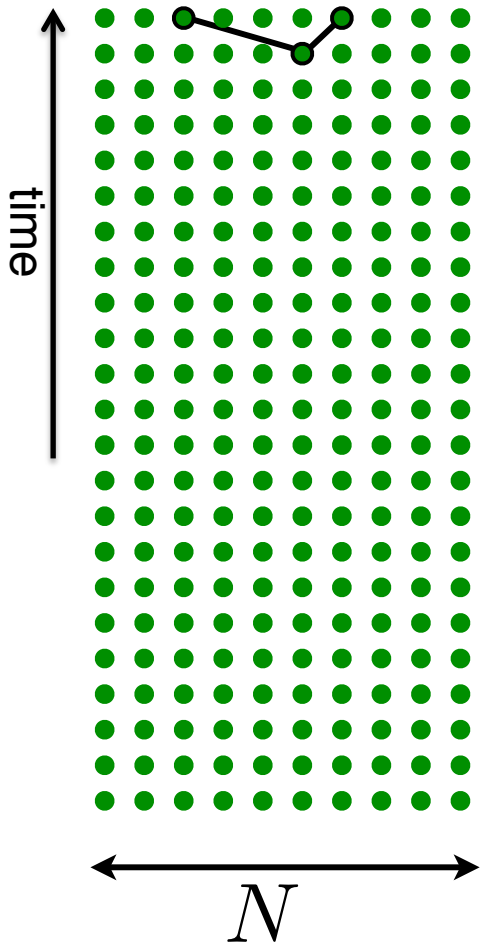
(b) Gene tree with coalescent times



John Kingman's coalescent theory (1982):

- Finite population size
- Random mating
- No selection
- No population structure
- No gene flow
- No recombination

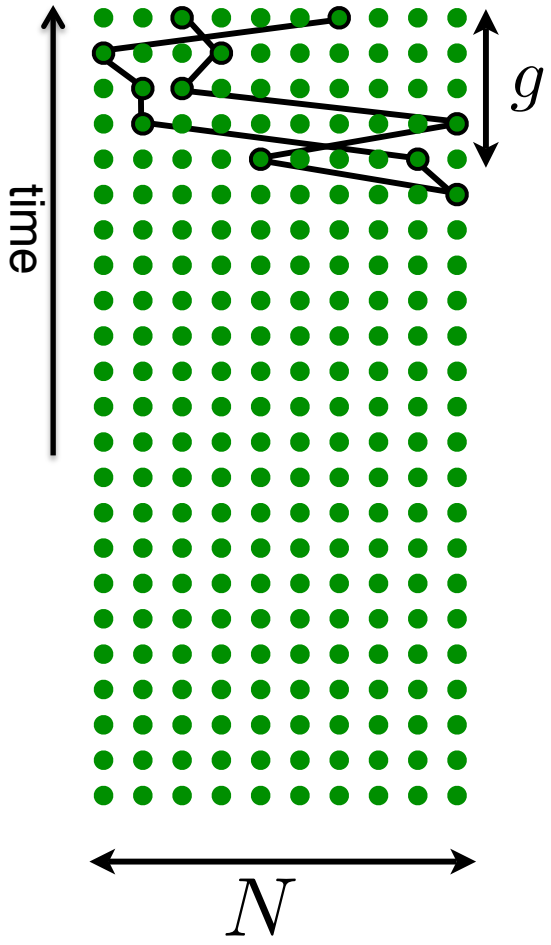
Coalescent probabilities:



- The probability of 2 lineages coalescing is the probability of them choosing the same ancestor: $\frac{1}{N}$

Image: Bob Thomson's slides

Coalescent probabilities:



- Probability that coalescence occurs g+1 generations back:
  - Probability of no coalescence for g generations

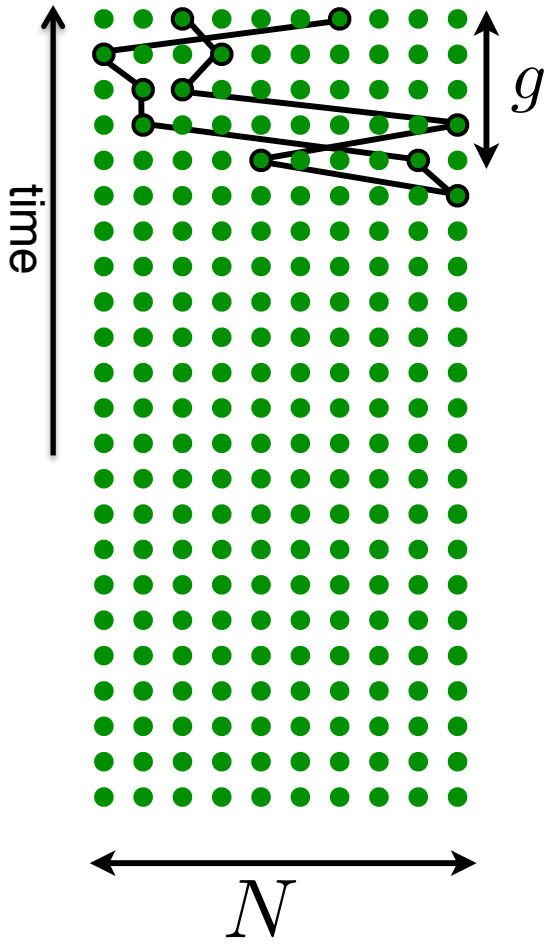
$$(1-\frac{1}{N})\times(1-\frac{1}{N})...=(1-\frac{1}{N})^g$$

  - followed by coalescence $\frac{1}{N}$

$$=\frac{1}{N}(1-\frac{1}{N})^g$$

Image: Bob Thomson's slides

Coalescent probabilities:

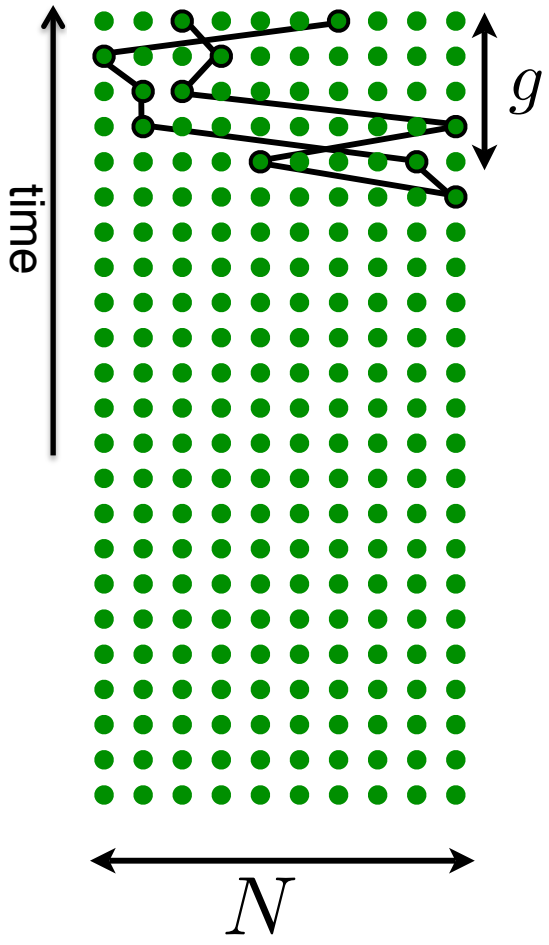


$$= \frac{1}{N}(1 - \frac{1}{N})^{g}$$

- This is the geometric distribution
  - Describes the time of the first success for independent trials with probability of success p and probability of failure (1-p)
  - Rate = p or 1/N
  - Mean = 1/p or N

Image: Bob Thomson's slides

Coalescent probabilities:



- Probability of coalescence event (or success rate) among n sampled lineages is

$$\frac{\binom{n}{2}}{N}$$

- n choose 2 accounts for the variety of ways that coalescence can occur

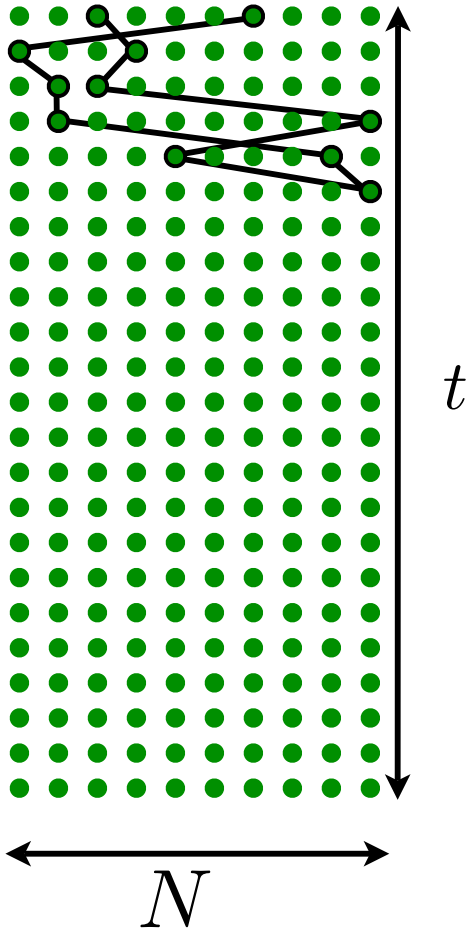
$$\frac{n!}{2!(n-2)!}$$

- Expected time becomes:

$$\frac{\binom{n}{2}}{N}(1-\frac{\binom{n}{2}}{N})^{g}$$

Image: Bob Thomson's slides

Coalescent probabilities:



- Geometric distribution is a discrete time distribution
- Continuous time version is the exponential distribution

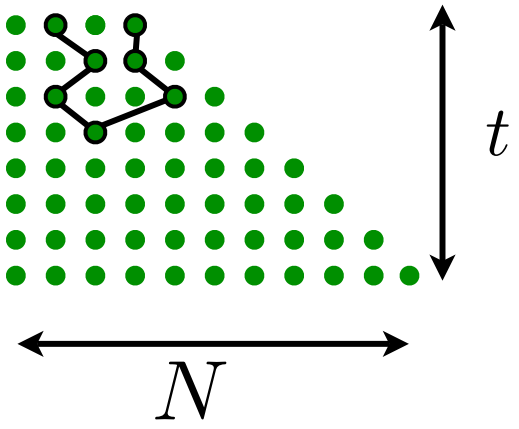
$$\lambda e^{-\lambda t}$$

- As N goes to infinity, the coalescent process converges to a continuous time markov process with instantaneous rate of coalescence:

$$\lambda = \frac{\binom{n}{2}}{N}$$

Image: Bob Thomson's slides

Coalescent probabilities:



- We've been assuming constant population size
- Instead of N, we can specify a function that describes a changing population size through time

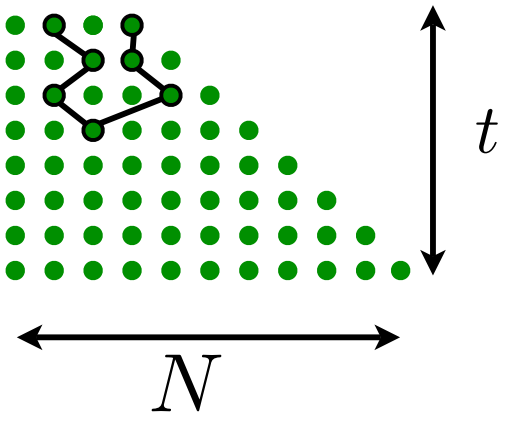
$$N \Rightarrow N(t)$$

- Our instantaneous rate of coalescence is a function of N, so we need to integrate the rate of coalescence across the function for N

$$\frac{\binom{n}{2}}{N} e^{-\frac{\binom{n}{2}}{N} t} \Rightarrow \frac{\binom{n}{2}}{N(t)} \exp\left(- \int_0^t \frac{\binom{n}{2}}{N(t)} dt \right)$$

Image: Bob Thomson's slides

Coalescent probabilities:



- We have a nice function to calculate the probability of one coalescent event occurring at time t, given a demographic function of t:

$$\mathrm{P}(t) = \frac{\binom{n}{2}}{N(t)} \exp\left(- \int_0^t \frac{\binom{n}{2}}{N(t)} dt \right)$$

- What is the probability of <u>all</u> coalescent events observed in a sample?
  - Given a demographic function and a list of coalescence times $\mathrm{L} = (0, t_n, t_{n-1}, \ldots)$
  - Each probability is independent, so take the product

$$\mathrm{P}(L|N(t)) = \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(- \int_0^t \frac{\binom{n}{2}}{N(t)} dt \right)$$

Image: Bob Thomson's slides

Coalescent probabilities:

We have now derived a likelihood function for a model that describes the probability of a coalescent history within a lineage:

$$\mathrm{P}(L|N(t)) = \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(- \int_0^t \frac{\binom{n}{2}}{N(t)} dt \right)$$

These probabilities are based on:

1. The population size
2. The times to coalescent events

Image: Bob Thomson's slides

Population size and coalescent times are related.
Given one parameter we can estimate the other.

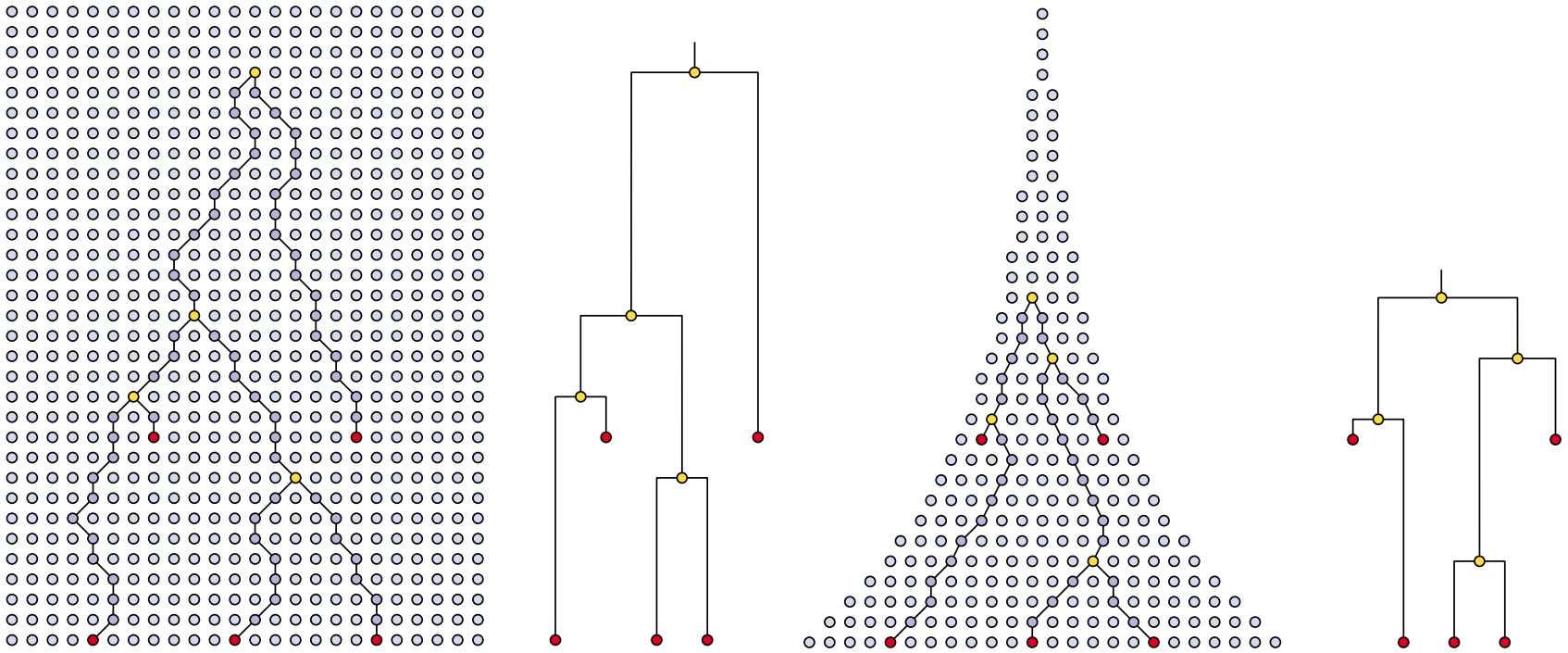


Image: Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., & Rodrigo, A. G. (2003). Measurably evolving populations. Trends in Ecology & Evolution, 18(9), 481-488.

Coalescent models usually assumes neutrality,
but can also be used to detect selection.

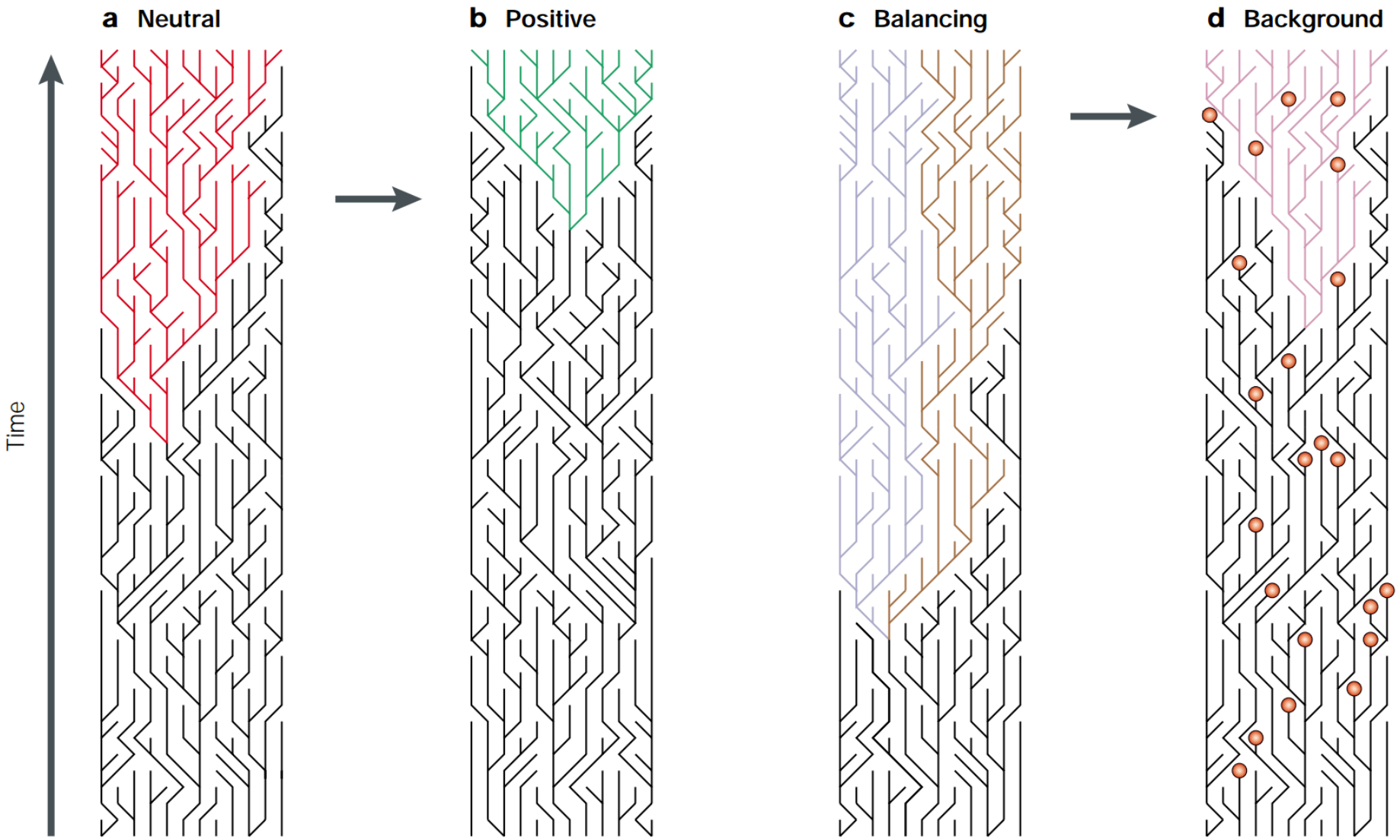
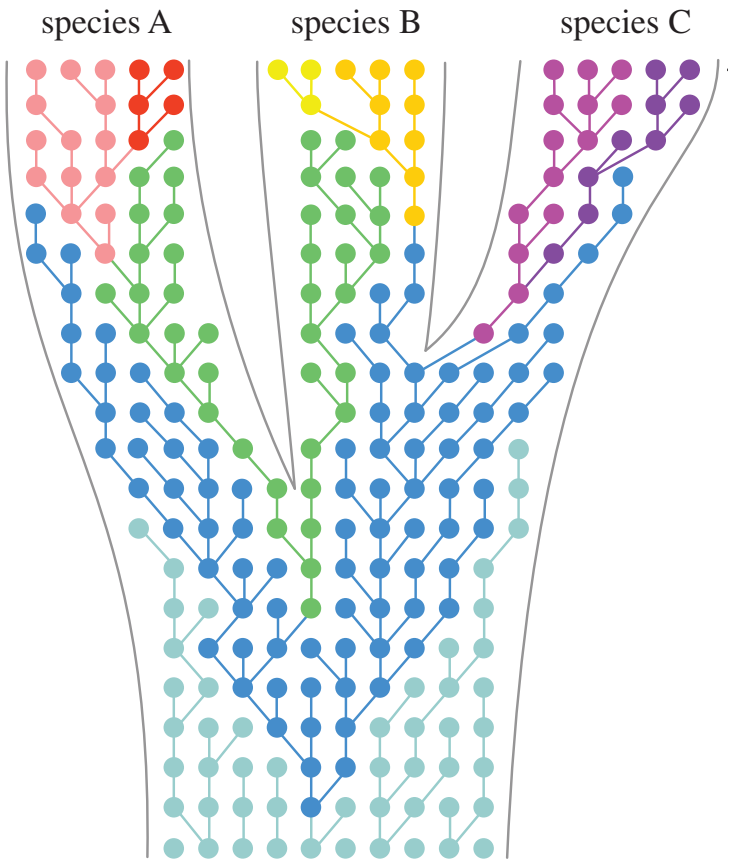


Image: Bamshad, M., & Wooding, S. P. (2003). Signatures of natural selection in the human genome. Nature Reviews Genetics, 4(2), 99-111.



Extending to the multispecies coalescent:

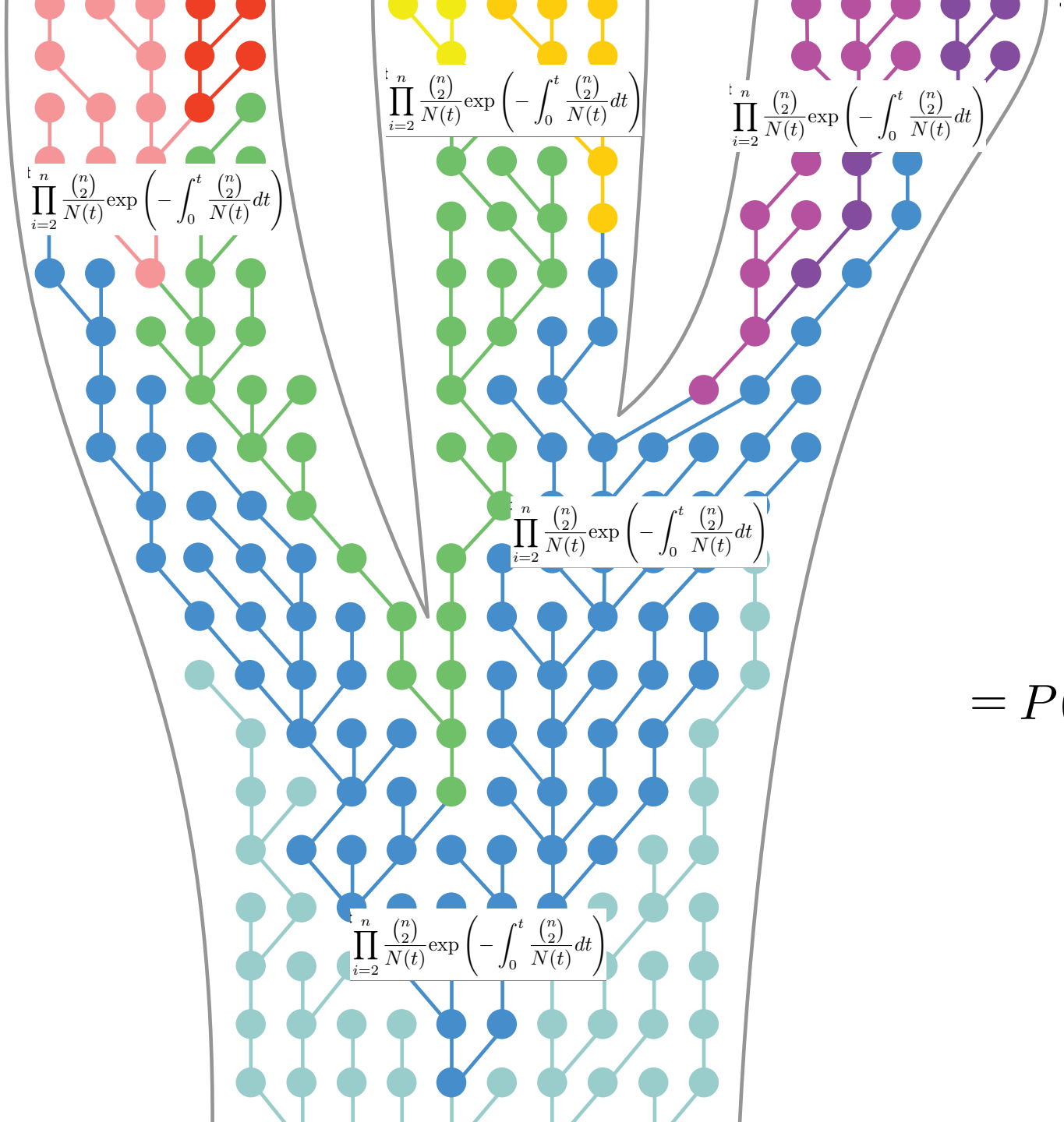
We have described a model that calculates the probability of a coalescent history in a lineage:

$$\mathrm{P}(L|N(t)) = \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(- \int_0^t \frac{\binom{n}{2}}{N(t)} dt \right)$$

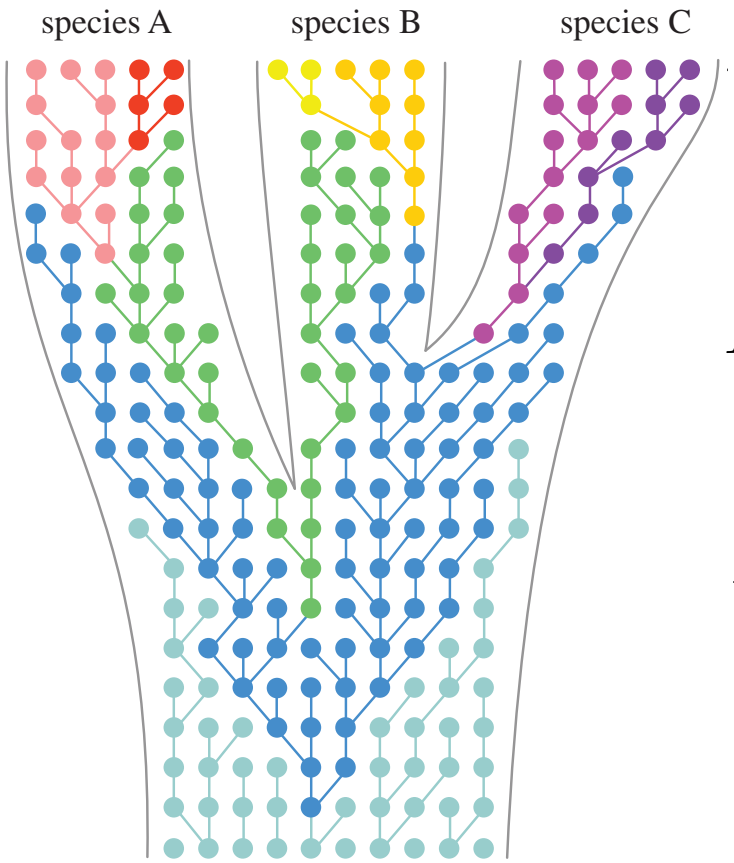
Each branch of a phylogeny is a lineage.

For every 2 alleles present within a branch, they will have some probability of coalescing or not.

So the probability of a gene tree given a species tree is the product of the coalescent histories for all branches of the tree…..


$\prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$
$\prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$
$\prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$
$\prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$
$\prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$

$$= P(g_i|S)$$



Therefore the full multispecies coalescent is given by:

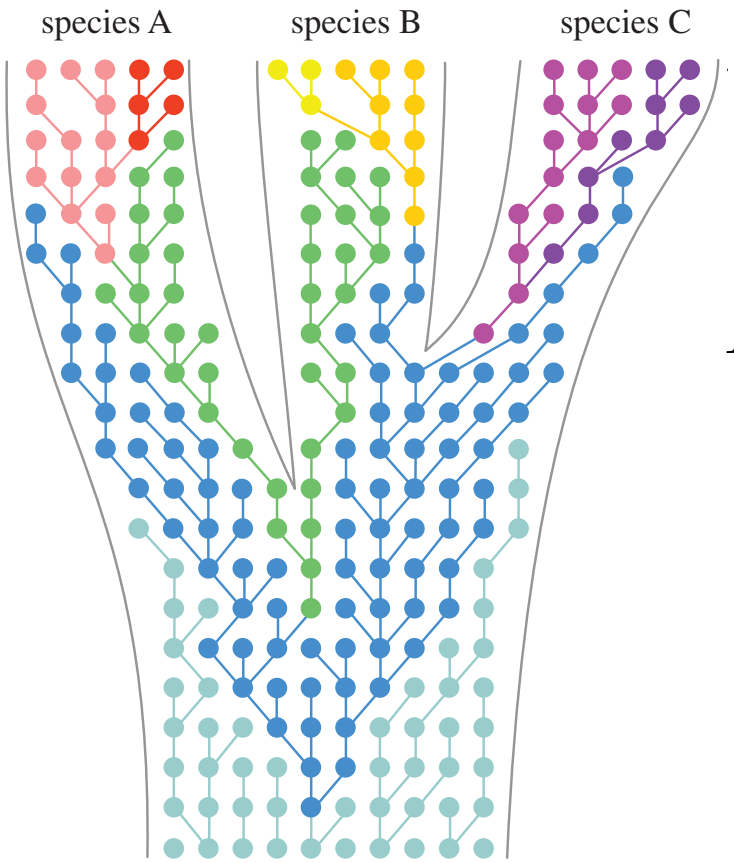
$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

$P(d_i|g_i)$ = standard likelihood for gene tree and a sequence alignment

$P(g_i|S)$ = coalescent likelihood for gene tree and species tree

$P(S)$ = prior for species tree (uniform, birth-death, Yule)

$P(D)$ = marginal likelihood

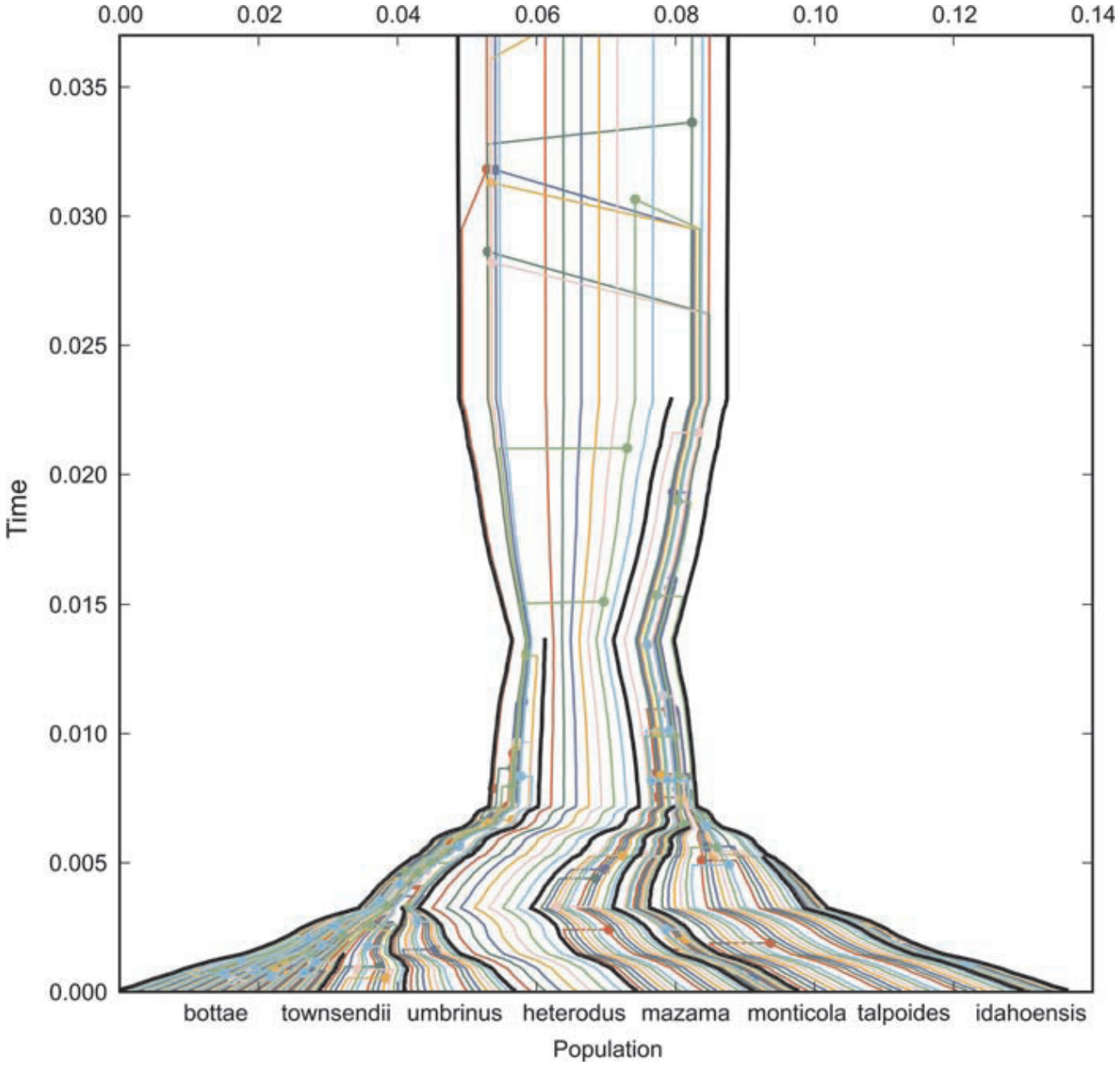


Therefore the full multispecies coalescent is given by:

$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

Remember:

- Only models incongruence due to incomplete lineage sorting
- No horizontal gene flow or introgression
- Assumes no selection, no recombination



Uses of the coalescent:

inferring demographic history of western pocket gophers…

Heled, J., & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. Molecular biology and evolution, 27(3), 570-580.

Uses of the coalescent: Bayesian skyline plots

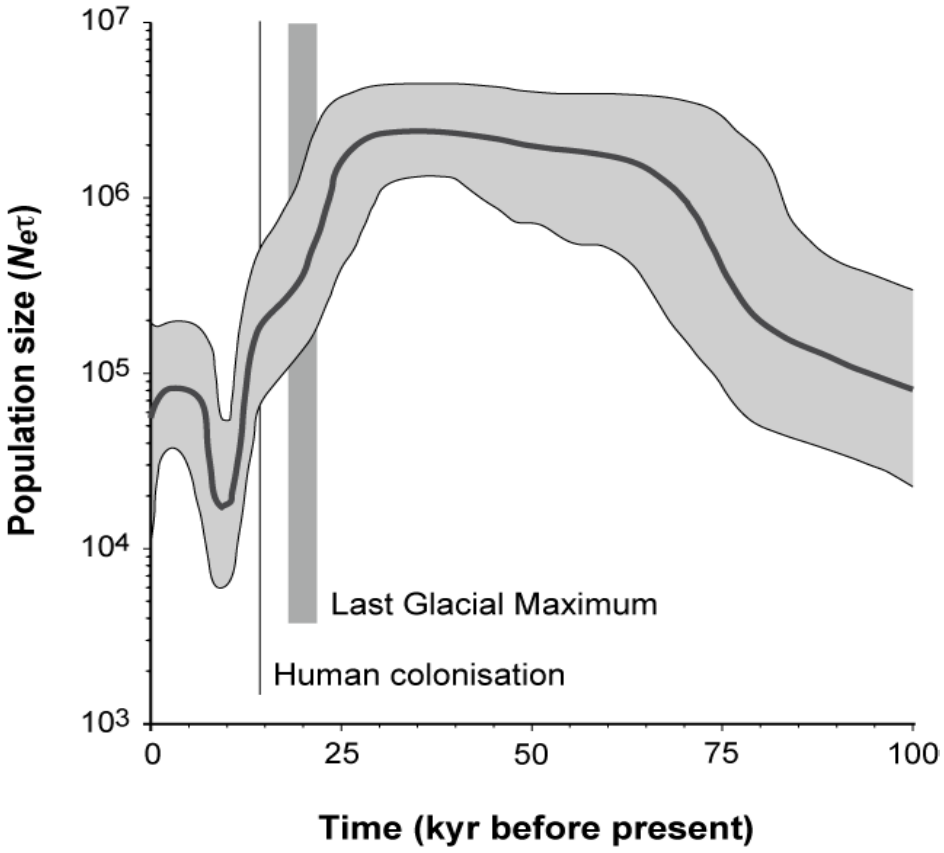
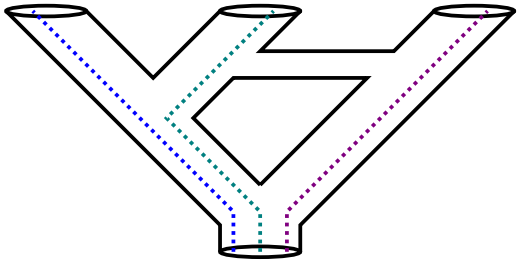


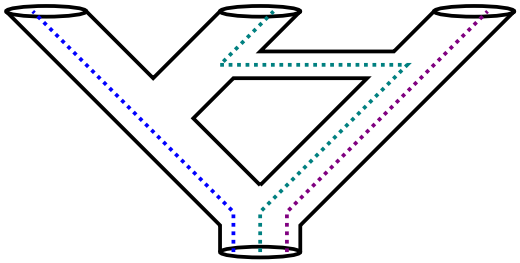
Image: Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular biology and evolution, 22(5), 1185-1192.

Future extensions of the multispecies coalescent: inferring reticulate evolution



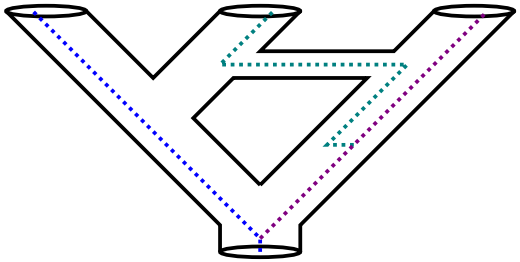
$$P(ABBA) = P(BABA)$$
$$= (1-\gamma)\frac{2N\mu}{3}\left(1-\frac{1}{2N}\right)^{t_3-t_2}$$

Scenario 1:
incomplete lineage sorting and no introgression



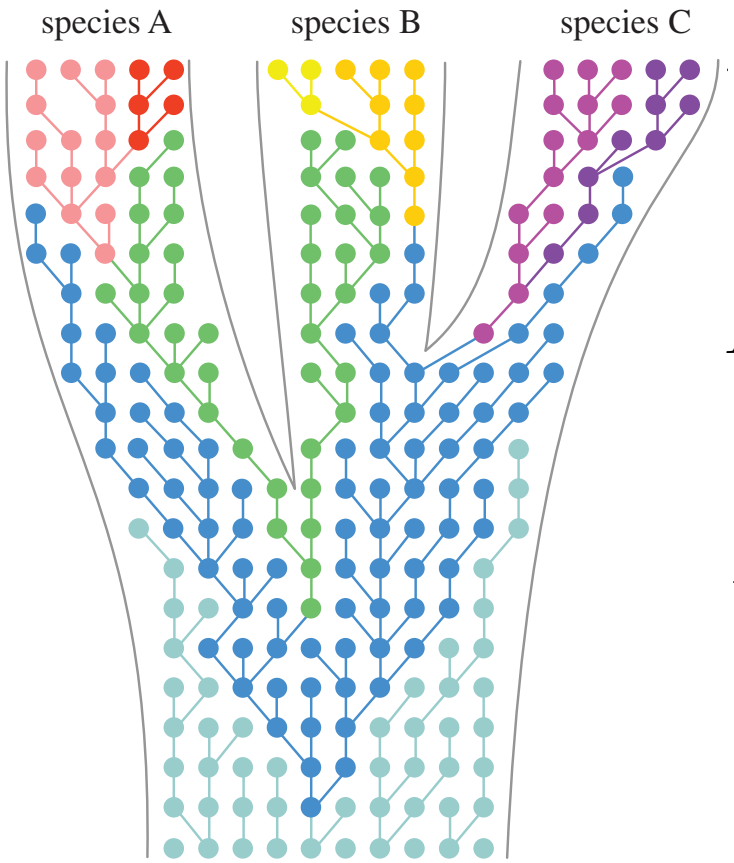
$$P(ABBA) = P(BABA)$$
$$= \frac{2N\gamma\mu}{3}\left(1-\frac{1}{2N}\right)^{t_3-t_{gf}}$$

Scenario 2:
incomplete lineage sorting and introgression



$$P(ABBA) = \gamma\mu(t_3 - t_{gf})$$
$$P(BABA) = 0$$

Scenario 3:
introgression and no incomplete lineage sorting



Today in lab we'll use RevBayes to implement the full multispecies coalescent:

$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

$P(d_i|g_i)$ = standard likelihood for gene tree given a sequence alignment

$P(g_i|S)$ = coalescent likelihood for gene tree given a species tree

$P(S)$ = prior for species tree (uniform, birth-death, Yule)

$P(D)$ = marginal likelihood