Hypothesis Testing with Models of Reticulate Evolution

George P. Tiley
Royal Botanic Gardens, Kew
Methodological Advances in Reticulate Evolution
9 November 2023

Learning Goals

Expectations for gene tree variation

Deviations as evidence for gene glow

Site-based methods

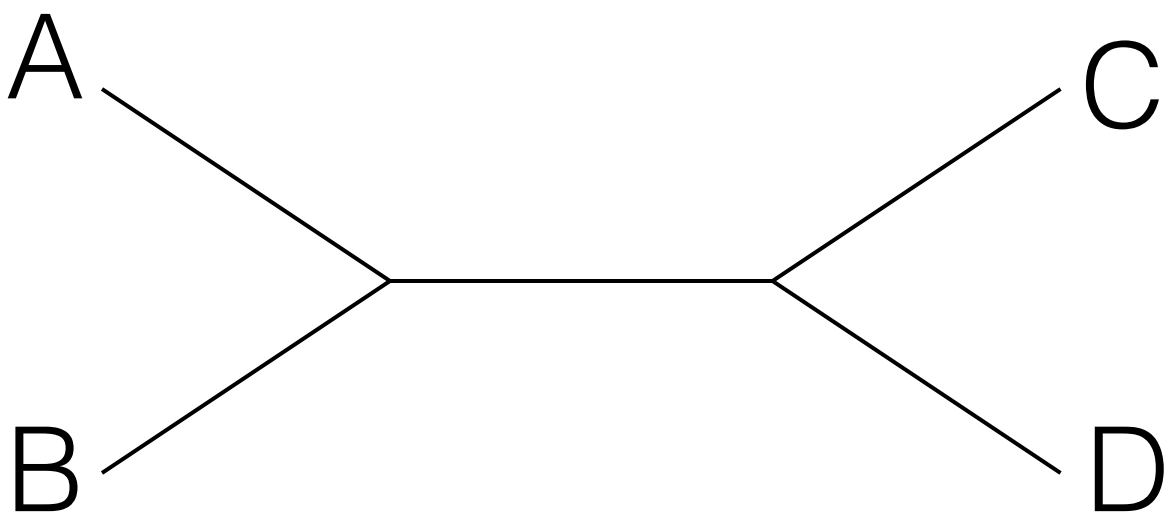
Bayes factors and marginal likelihoods

Approximate Methods

A practical example

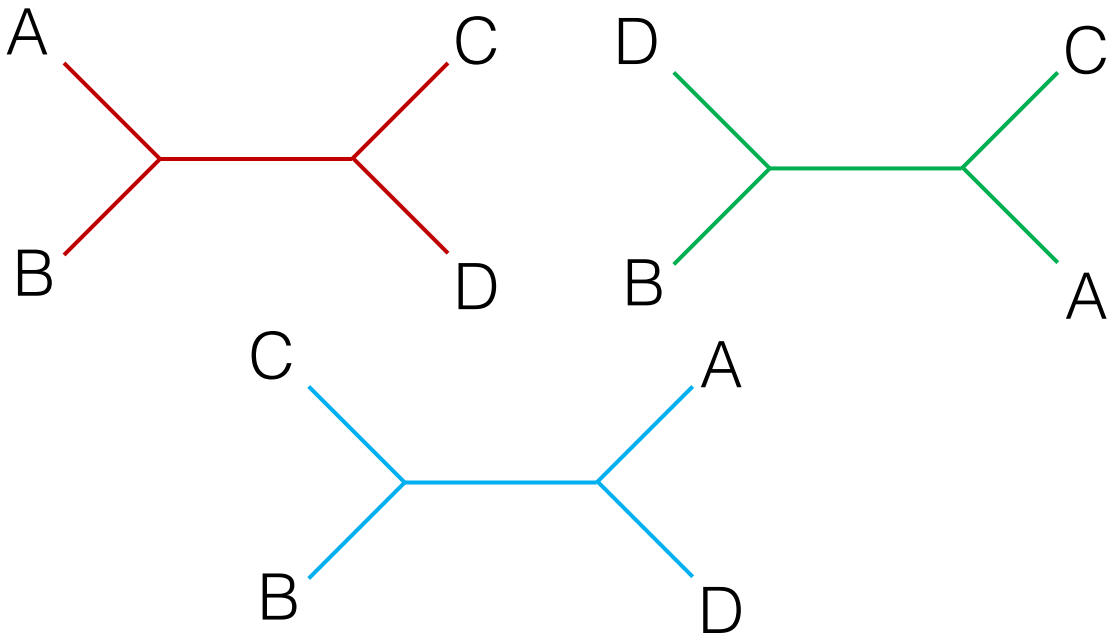
Expectations for gene tree variation

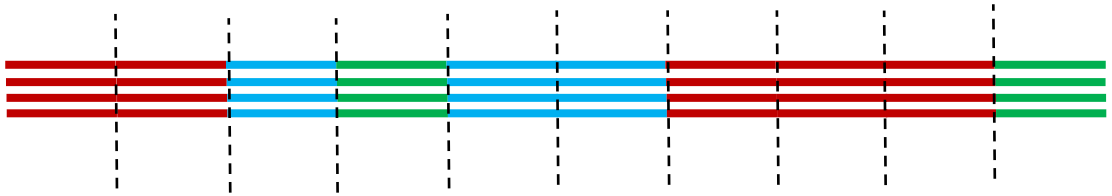
Consider a species tree



Expectations for gene tree variation

Consider a species tree
Collecting a lot of data may reveal gene tree variation

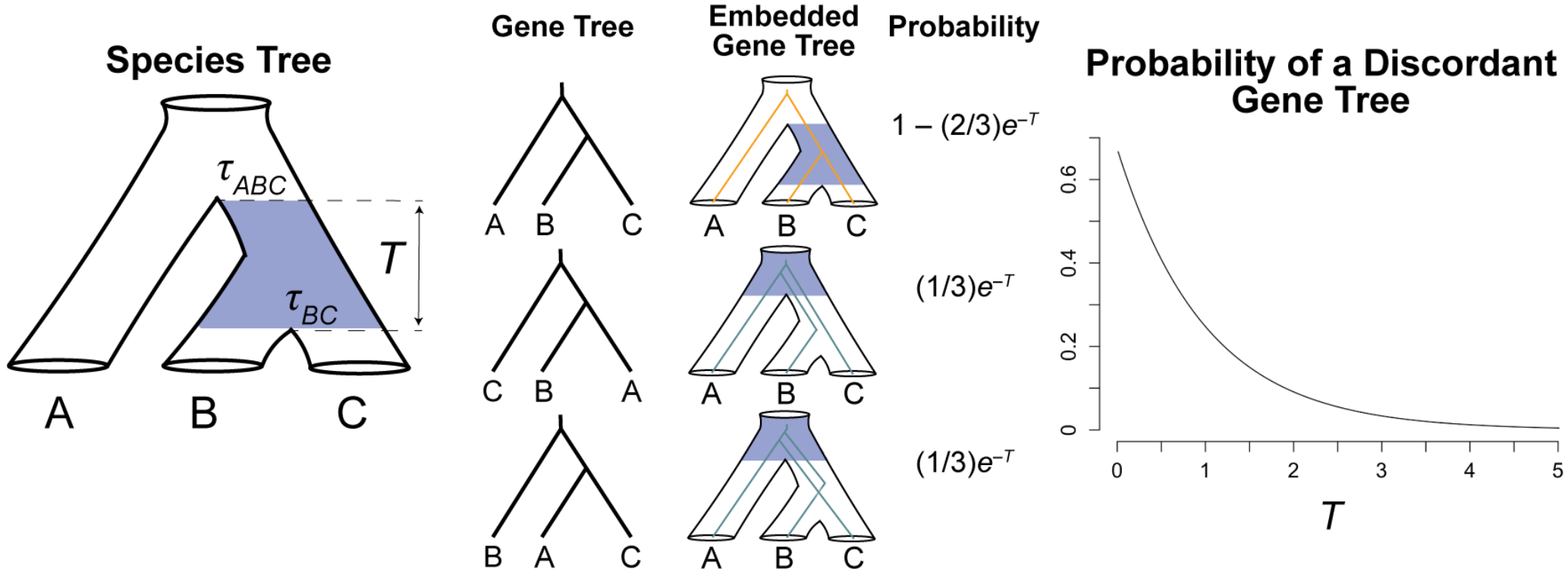




We want to explain the processes underlying variation

Expectations for gene tree variation

Incomplete lineage sorting is a good null model



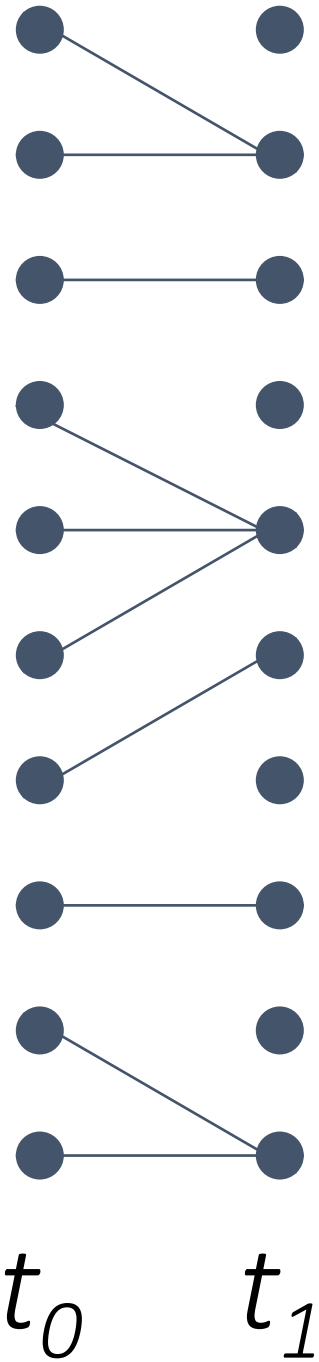
Rannala B and Yang Z. 2003. Genetics. 164:1645-1656.
Pamilo P and Nei M. 1988. Mol Biol Evol. 5:568-583.
Hudson RR. 1983. Evolution. 37:203-217.
Kingman JFC. 1982. J Appl Probab. 19A:27-43.
Ewans WJ. 1972. Theor Popul Biol. 3:87-112.

Fig. 2 — Tiley et al. 2020 Trends Genet. 36:845-856

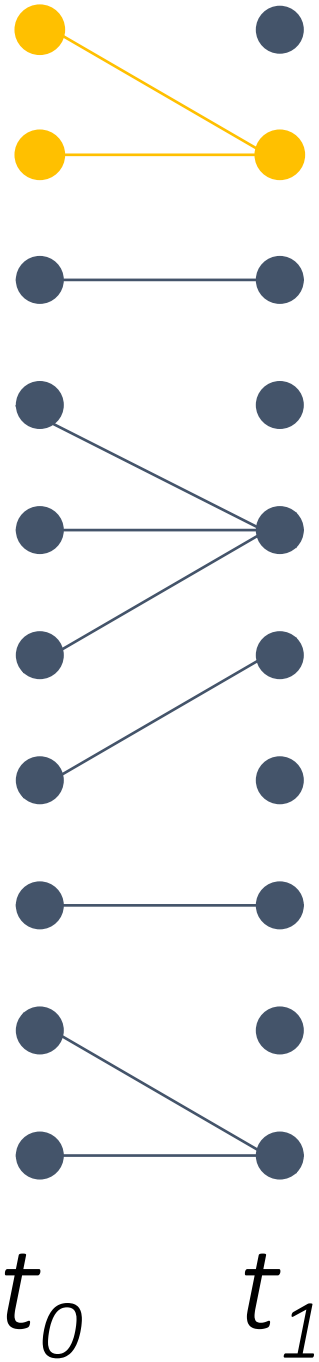
Expectations for gene tree variation

The Coalescent for a Single Population

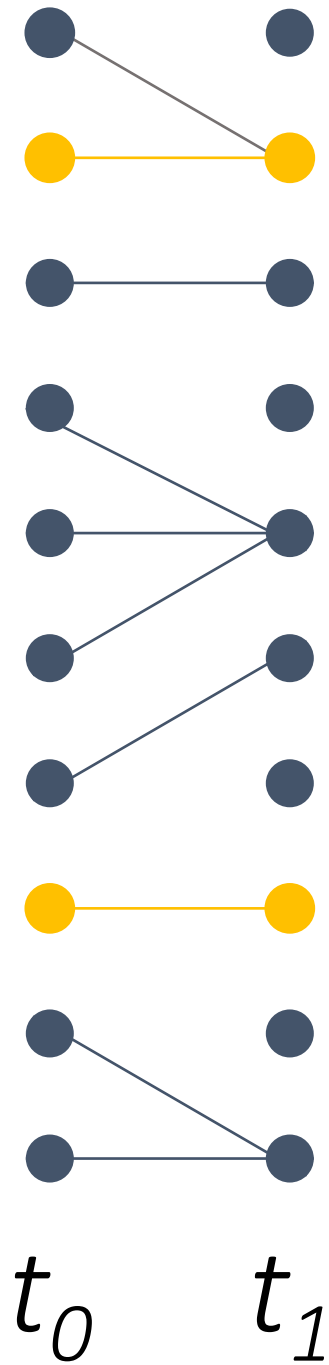
A sample of 10 individuals
in the present



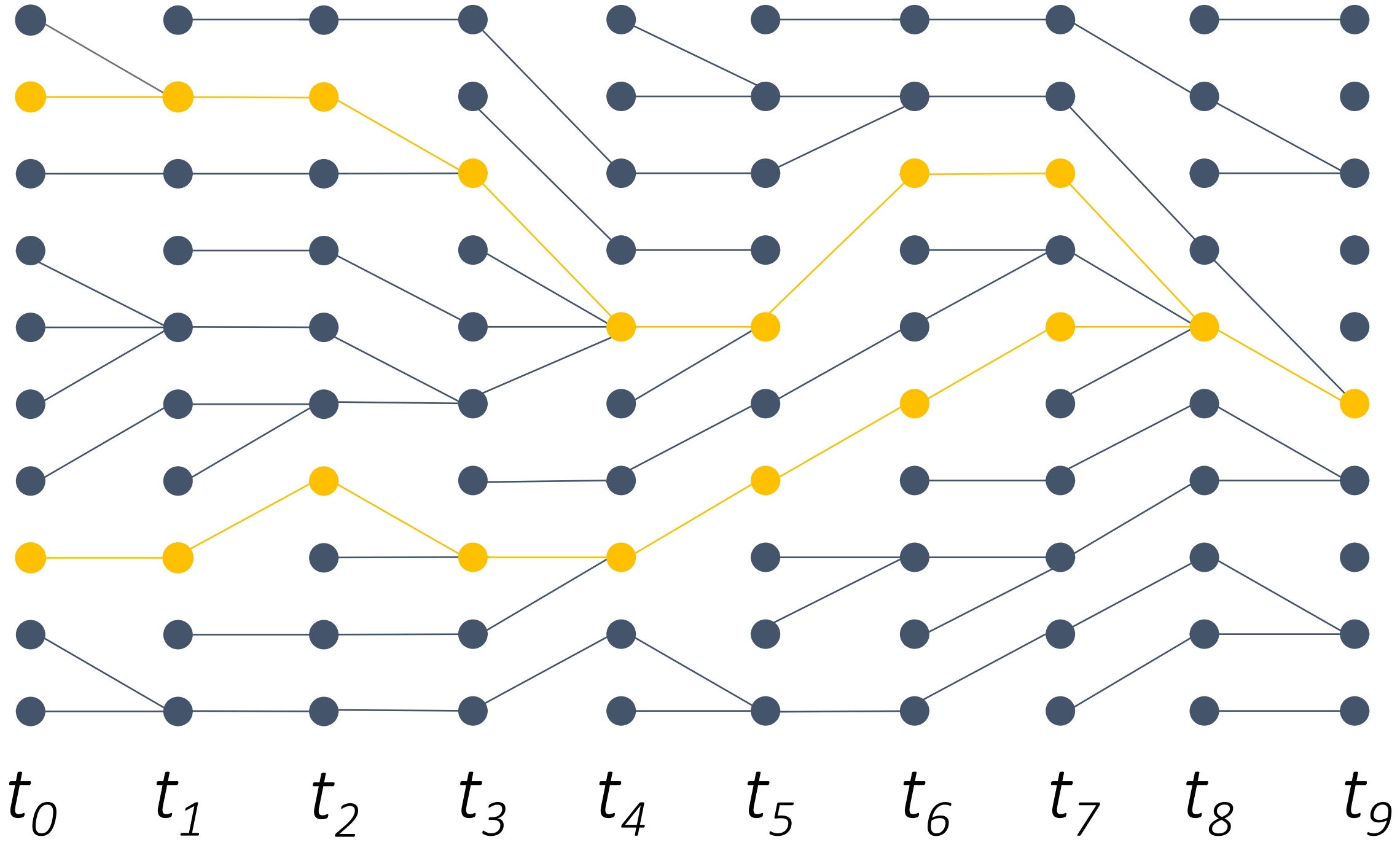
Sampled from the previous generation under Wright-Fisher

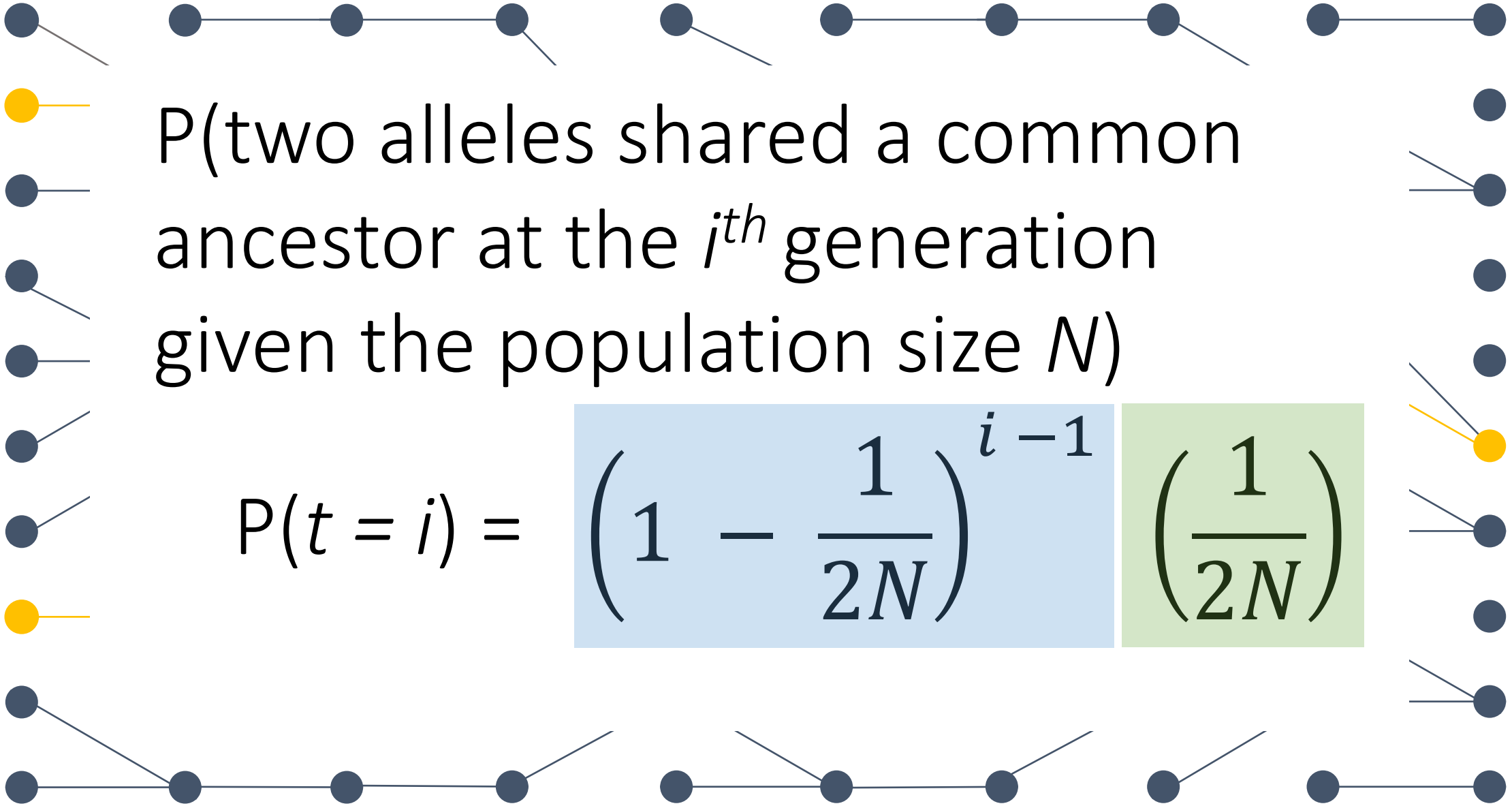


P(Two Alleles Share Ancestor 1 generation ago)
= 1/2N



P(Two Alleles **Do Not** Share Ancestor 1 generation ago)
= 1 – (1/2N)


t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9



P(two alleles shared a common ancestor at the i^{th} generation given the population size N)

$$P(t = i) = \left(1 - \frac{1}{2N}\right)^{i-1} \left(\frac{1}{2N}\right)$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

Geometric distribution with

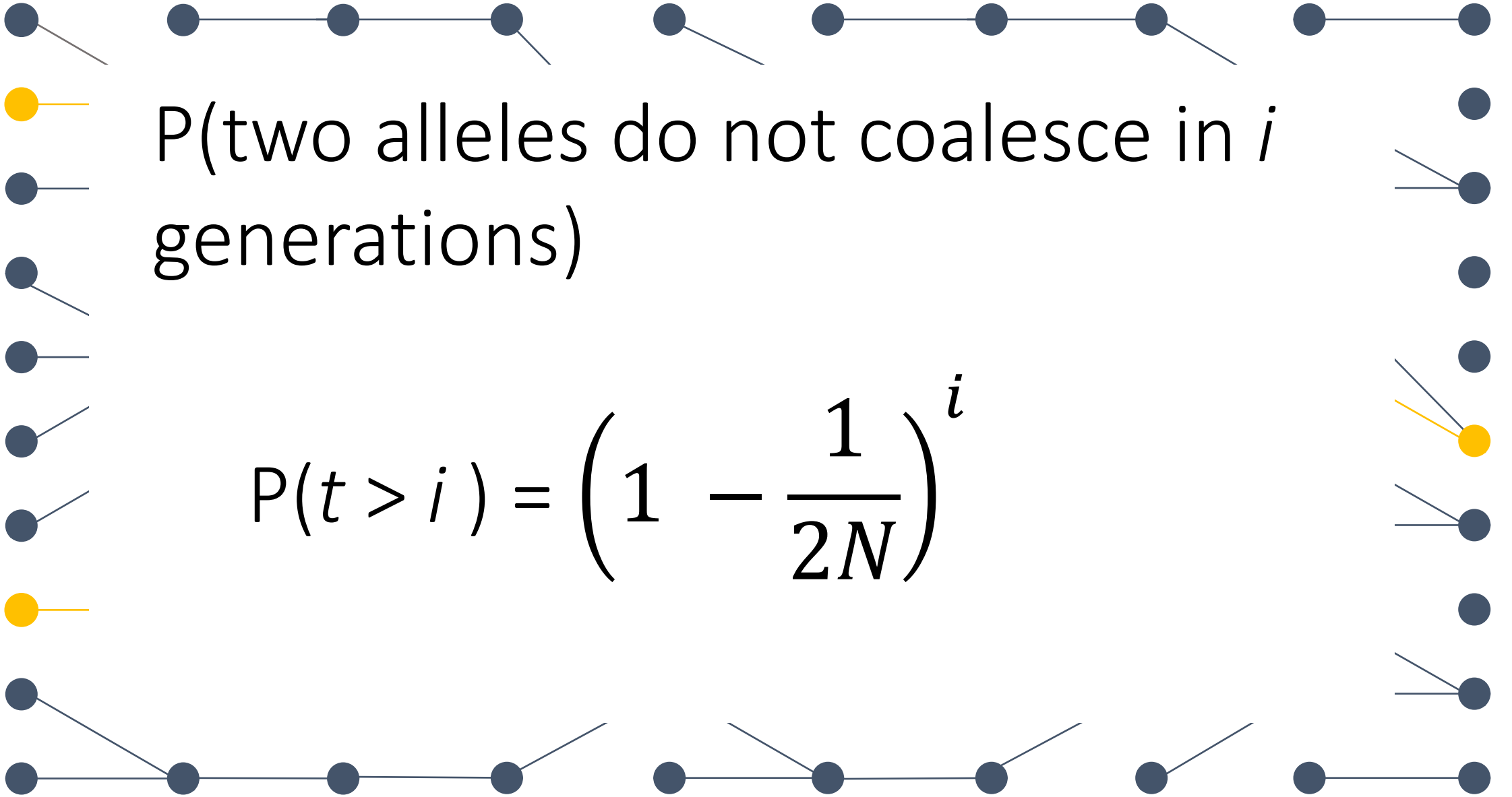
$p = 1/2N$

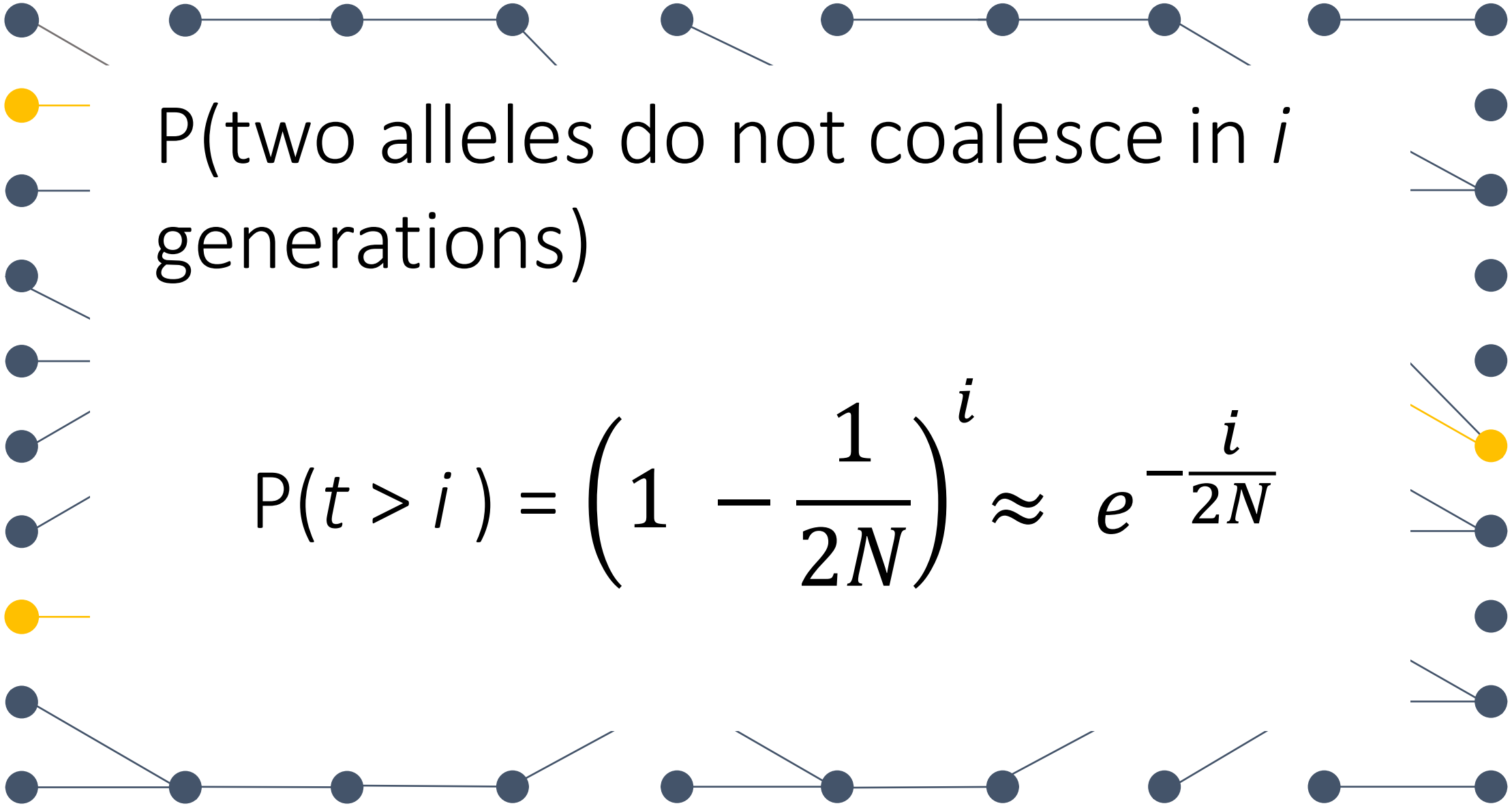
$$P(t|N) = \left(1 - \frac{1}{2N}\right)^{i-1} \left(\frac{1}{2N}\right)$$

$E[t] = 1/p = 2N$

P(two alleles do not coalesce in *i* generations)

$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i}$$

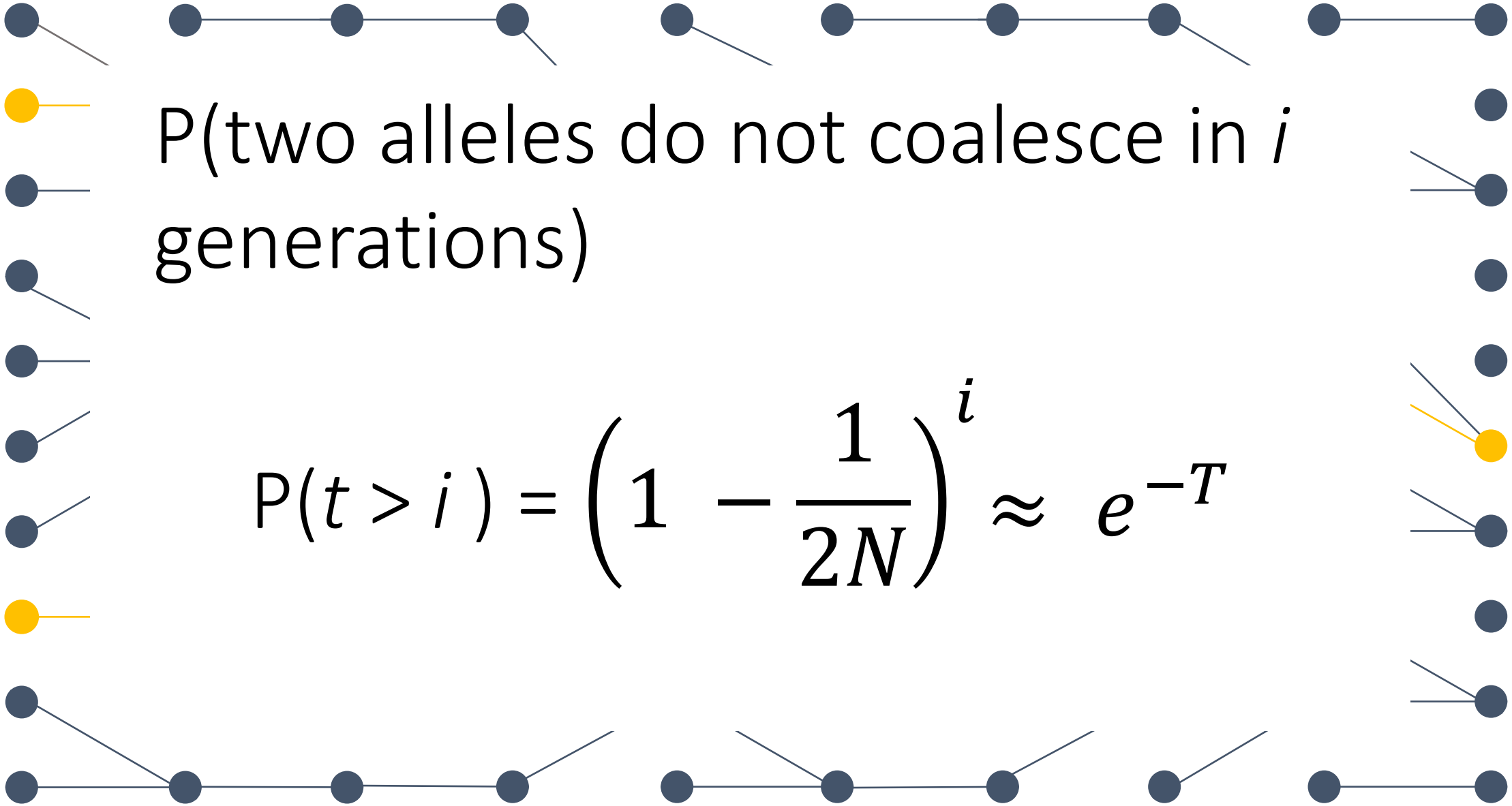




P(two alleles do not coalesce in *i* generations)

$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i} \approx e^{-\frac{i}{2N}}$$

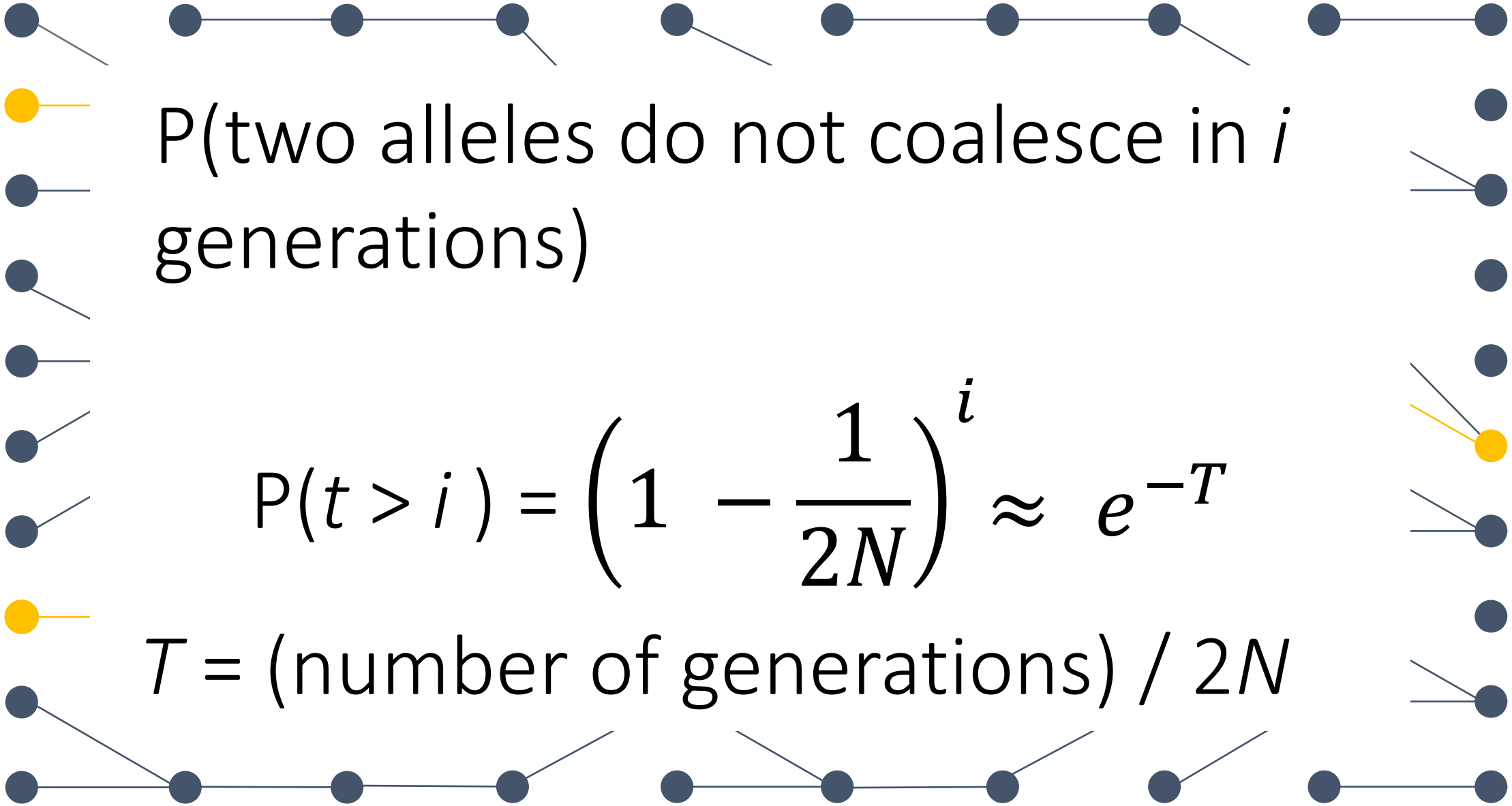
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(two alleles do not coalesce in *i* generations)

$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i} \approx e^{-T}$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(two alleles do not coalesce in *i* generations)

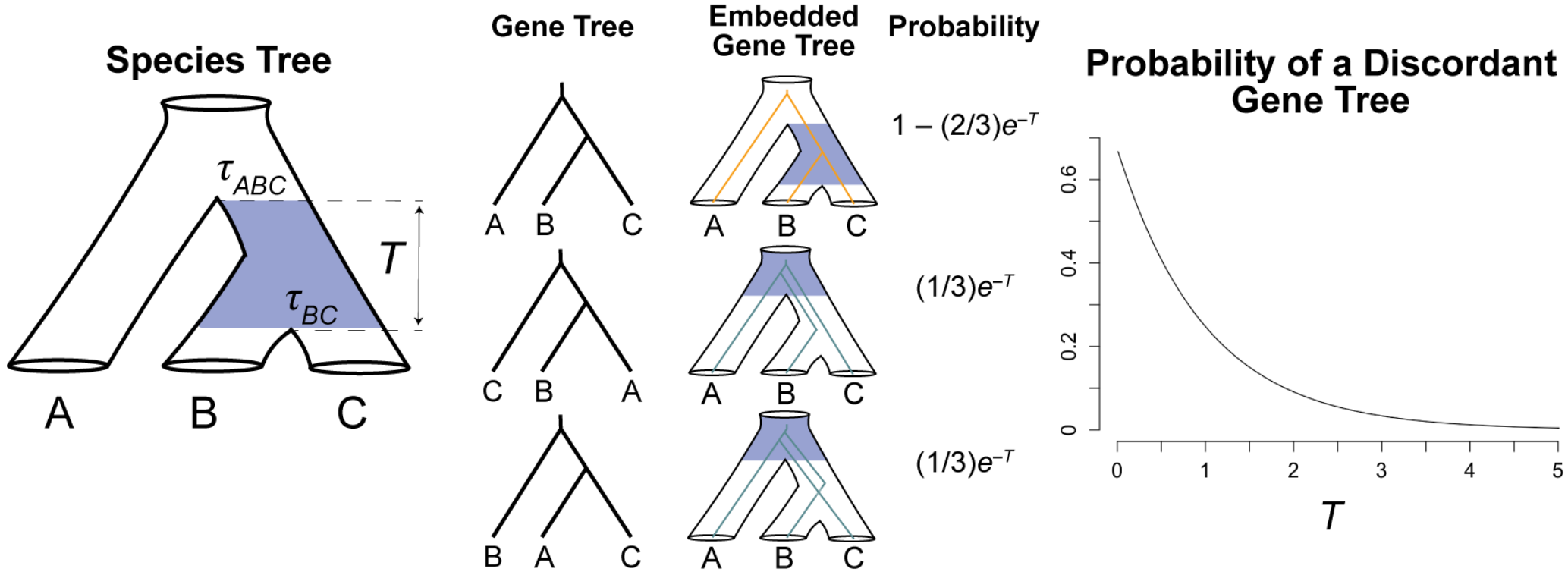
$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i} \approx e^{-T}$$

T = (number of generations) / 2*N*

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

Expectations for gene tree variation

Incomplete lineage sorting is a good null model

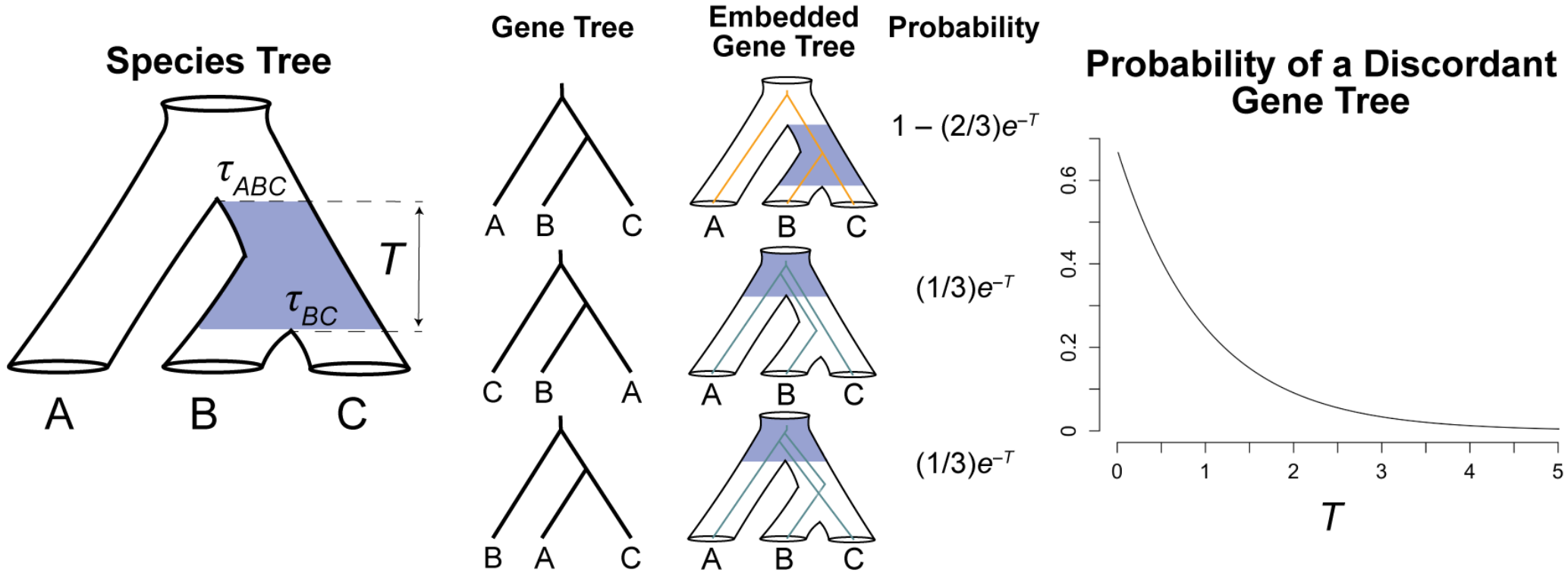


Rannala B and Yang Z. 2003. Genetics. 164:1645-1656.
Pamilo P and Nei M. 1988. Mol Biol Evol. 5:568-583.
Hudson RR. 1983. Evolution. 37:203-217.
Kingman JFC. 1982. J Appl Probab. 19A:27-43.
Ewans WJ. 1972. Theor Popul Biol. 3:87-112.

Fig. 2 — Tiley et al. 2020 Trends Genet. 36:845-856

Expectations for gene tree variation

Incomplete lineage sorting is a good null model



$$\tau = \frac{t}{2N} = \frac{\mu t}{\frac{\theta}{2}}$$

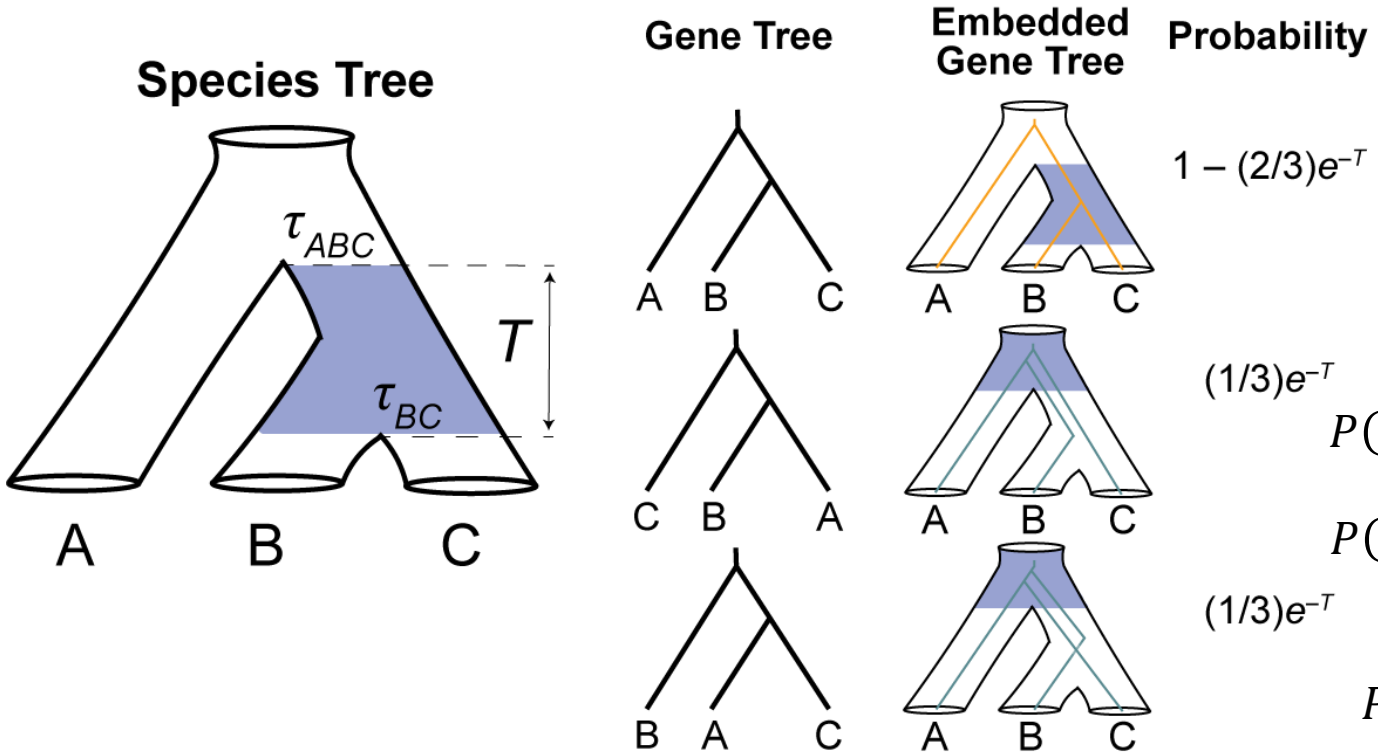
coalescent branch lengths

mutation-scaled branch lengths

Fig. 2 — Tiley et al. 2020 Trends Genet. 36:845-856

Expectations for gene tree variation

Incomplete lineage sorting is a good null model



$$P(2\ alleles\ do\ not\ coalesce) = e^{-T}$$

$$P(B\ and\ C\ do\ not\ coalesce) = e^{-T}$$

$$P(B\ and\ C\ do\ coalesce) = 1 - e^{-T}$$

$$P(B\ and\ C\ do\ not\ coalesce\ and\ A\ and\ B\ do\ first) = \frac{1}{3}e^{-T}$$

$$P(B\ and\ C\ do\ not\ coalesce\ and\ A\ and\ C\ do\ first) = \frac{1}{3}e^{-T}$$

$$P(B\ and\ C\ do\ coalesce) = 1 - e^{-T} + \frac{1}{3}e^{-T} = 1 - \frac{2}{3}e^{-T}$$

$$\tau = \frac{t}{2N} = \frac{\mu t}{\frac{\theta}{2}}$$

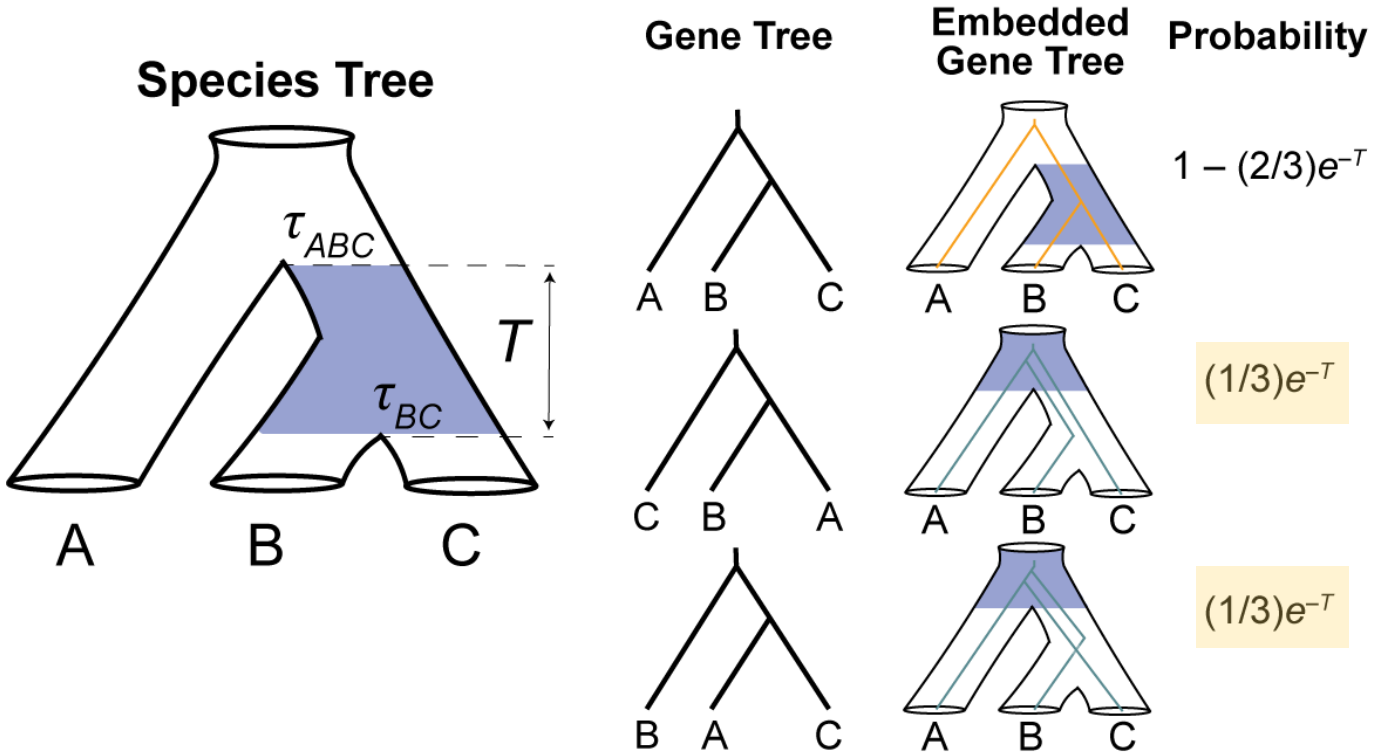
coalescent branch lengths

mutation-scaled branch lengths

Fig. 2 — Tiley et al. 2020 Trends Genet. 36:845-856

Expectations for gene tree variation

Incomplete lineage sorting is a good null model



Under a model that considers ILS as the only source of gene tree variation, non-speciodentric quartets or rooted triples should occur in equal frequencies.

$$\tau = \frac{t}{2N} = \frac{\mu t}{\frac{\theta}{2}}$$

coalescent branch lengths

mutation-scaled branch lengths

Fig. 2 — Tiley et al. 2020 Trends Genet. 36:845-856

Expectations for gene tree variation

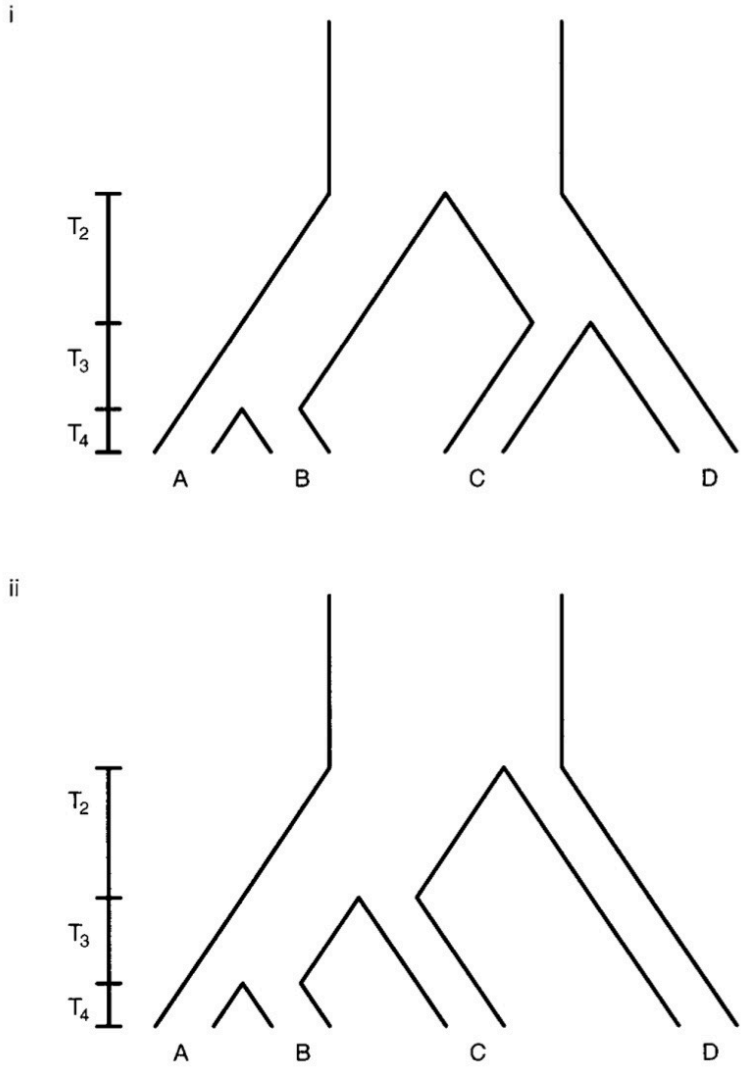


FIG. 9. The two bifurcating tree topologies that can be generated by four species. (i) Balanced tree topology. (ii) Unbalanced tree topology.

TABLE IV

Probabilities of the 15 Gene Tree Topologies When the Species Tree Topology Is ((*AB*)(*CD*))

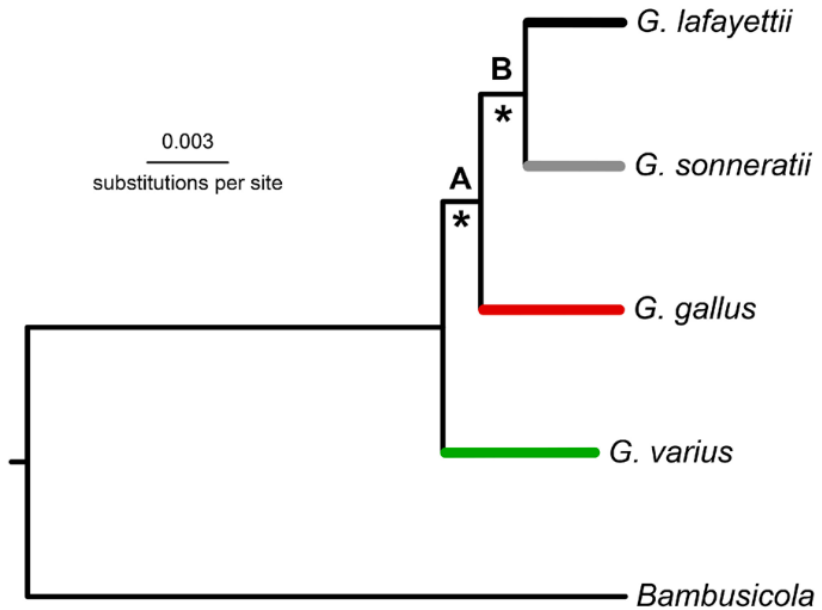
Gene tree topology	Probability	Probability at $T_3 = T_2 = 1$
$((AB)(CD))$	$g_{21}(T_3+T_2)\, g_{21}(T_2)$ $+g_{21}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{3}$ $+g_{22}(T_3+T_2)\, g_{21}(T_2)\, \frac{1}{3}$ $+g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{2}{6}\, \frac{1}{3}$	0.6867
$((AC)(BD))$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{2}{6}\, \frac{1}{3}$	0.0055
$((AD)(BC))$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{2}{6}\, \frac{1}{3}$	0.0055
$(((AB)\, C)\, D)$	$g_{21}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{3}$ $+g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.1088
$(((AB)\, D)\, C)$	$g_{21}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{3}$ $+g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.1088
$(((AC)\, B)\, D)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((AC)\, D)\, B)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((AD)\, B)\, C)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((AD)\, C)\, B)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BC)\, A)\, D)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BC)\, D)\, A)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BD)\, A)\, C)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BD)\, C)\, A)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((CD)\, A)\, B)$	$g_{22}(T_3+T_2)\, g_{21}(T_2)\, \frac{1}{3}$ $+g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0313
$(((CD)\, B)\, A)$	$g_{22}(T_3+T_2)\, g_{21}(T_2)\, \frac{1}{3}$ $+g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0313

Note. Notation is as in Fig. 9i. Values of $g_{ij}(T)$ are given by (2).

Analytical gene tree probability distributions are limited to 4-taxon rooted trees

Rosenberg. 2002 Theor. Popul. Biol. 61:225-247

An example directly using gene tree distributions



Internal branch lengths:

	Branch A	Branch B
ML concatenation	0.0014	0.0017
ASTRAL (gene trees)	0.507 cu	0.668 cu
ASTRAL (transversions)	0.425 cu	0.462 cu

Fig. 3 Whole genome phylogeny for *Gallus*. ML and the ASTRAL species trees were topologically identical. ML bootstrap support for both nodes was 100%, while both ASTRAL (estimated from gene trees or from perfect transversions) had local posterior probabilities of 1.0. Estimated branch lengths for the internal branches (A and B) are presented; the units for branch lengths are substitutions per site for ML and coalescent units for ASTRAL

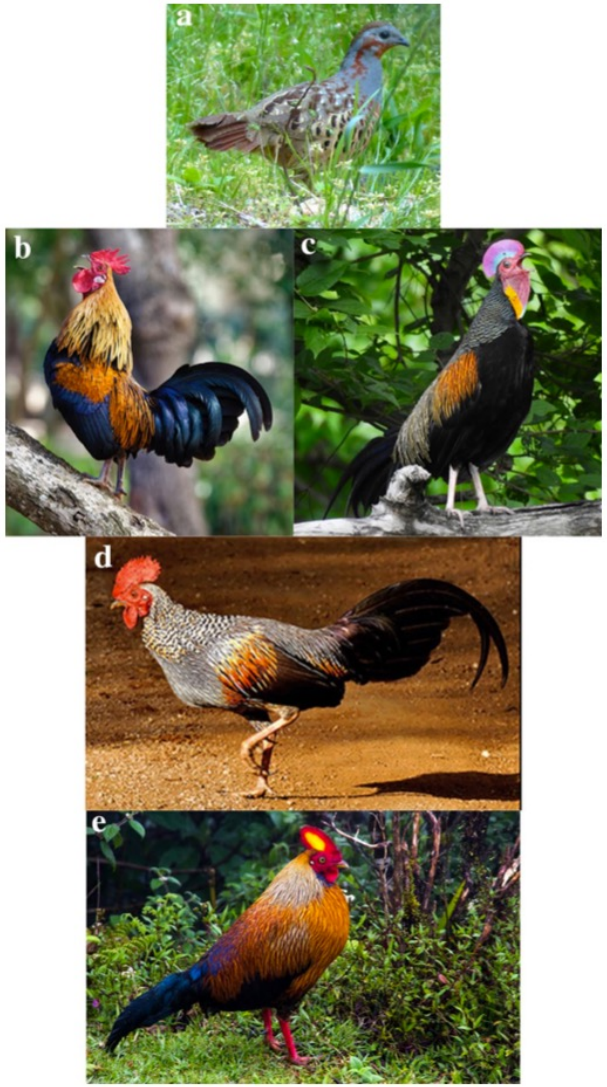
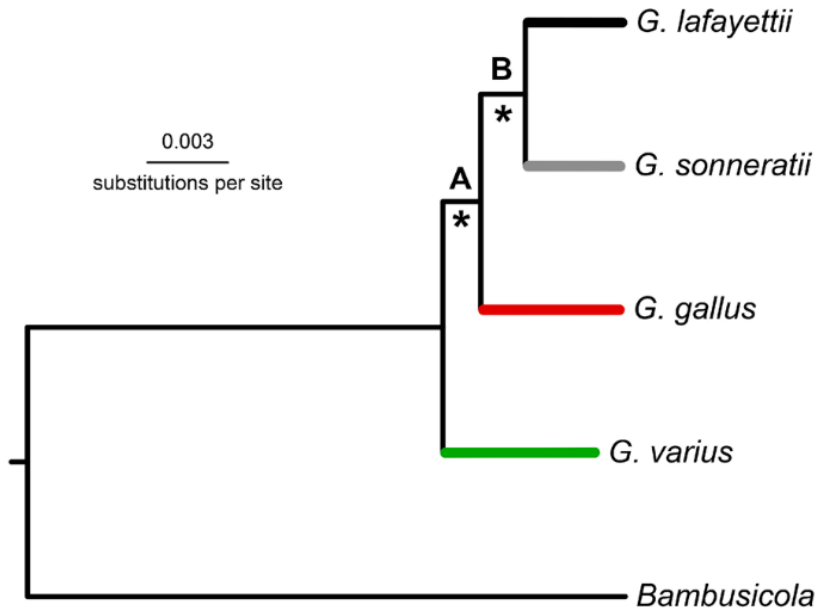


Fig. 1 Appearance of each focal species. **a** *Bambusicola thoracicus*. **b** *Gallus gallus*. **c** *G. varius*. **d** *G. sonneratii*. **e** *G. lafayettii*. For *Gallus*, only males are shown; *Bambusicola* is sexually monomorphic. Sources of each photograph are available in the Additional file 1

When working on small phylogenies, we can directly observe evidence for introgression and reject ILS alone by counting gene trees

Tiley et al. 2020. Avian Res. 11:7

An example directly using gene tree distributions



Internal branch lengths:

	Branch A	Branch B
ML concatenation	0.0014	0.0017
ASTRAL (gene trees)	0.507 cu	0.668 cu
ASTRAL (transversions)	0.425 cu	0.462 cu

Fig. 3 Whole genome phylogeny for *Gallus*. ML and the ASTRAL species trees were topologically identical. ML bootstrap support for both nodes was 100%, while both ASTRAL (estimated from gene trees or from perfect transversions) had local posterior probabilities of 1.0. Estimated branch lengths for the internal branches (A and B) are presented; the units for branch lengths are substitutions per site for ML and coalescent units for ASTRAL

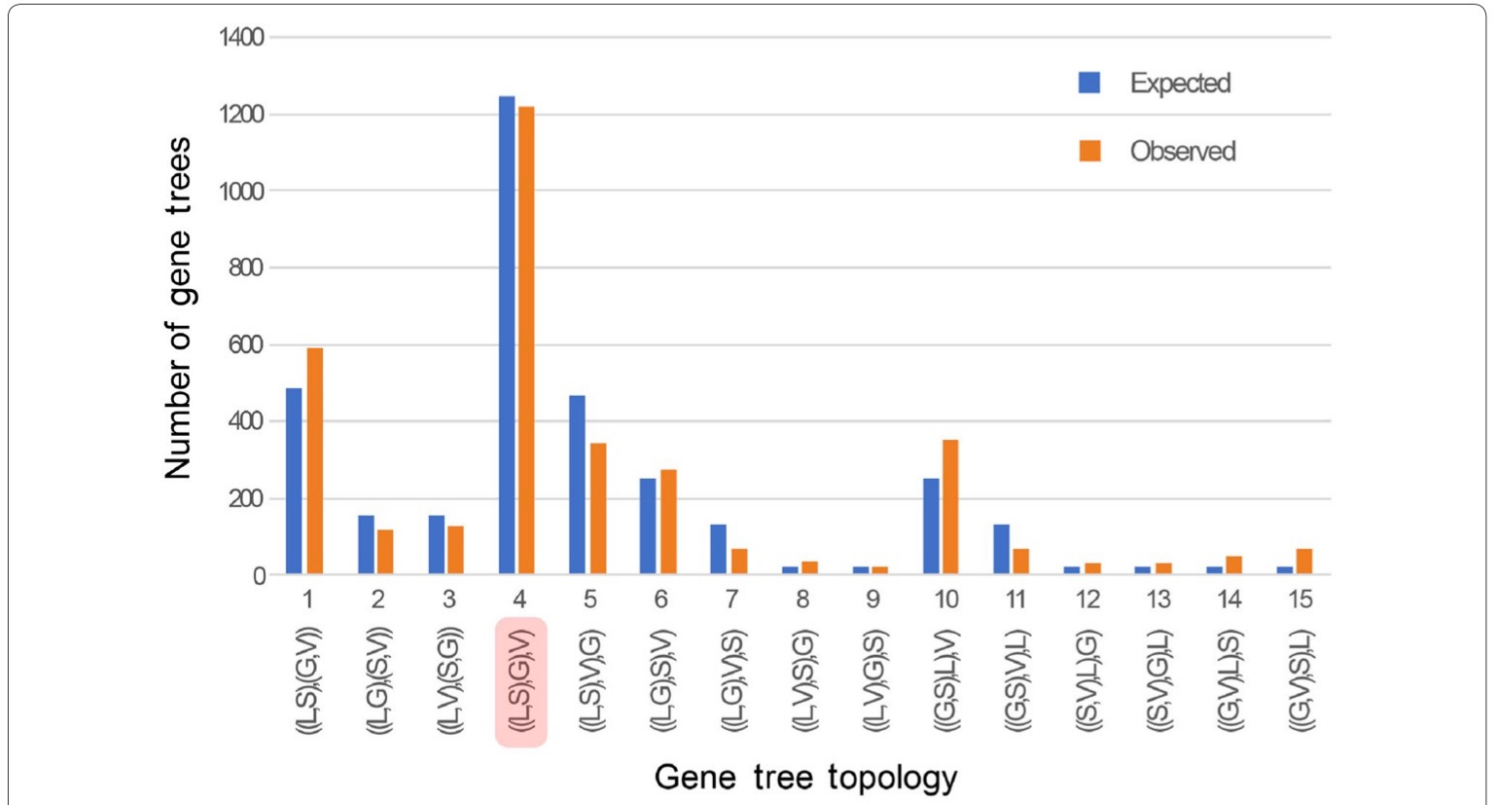
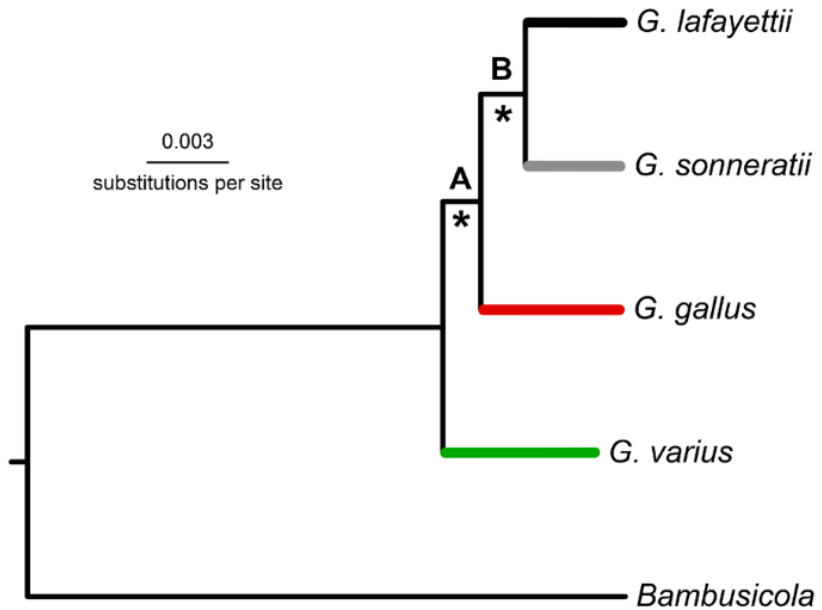


Fig. 4 The observed spectrum of *Gallus* gene tree topologies is similar to expectation given the multispecies coalescent. Expected numbers of gene trees with each topology were calculated assuming the topology and branch lengths that were estimated by ASTRAL when gene trees were used as input. The topology for each gene tree is presented as a rooted newick string using a single letter (first letter of the species name) to indicate each *Gallus* species. The speciodendric gene tree is shaded in pink

Tiley et al. 2020. Avian Res. 11:7

An example directly using gene tree distributions



Internal branch lengths:

	Branch A	Branch B
ML concatenation	0.0014	0.0017
ASTRAL (gene trees)	0.507 cu	0.668 cu
ASTRAL (transversions)	0.425 cu	0.462 cu

Fig. 3 Whole genome phylogeny for *Gallus*. ML and the ASTRAL species trees were topologically identical. ML bootstrap support for both nodes was 100%, while both ASTRAL (estimated from gene trees or from perfect transversions) had local posterior probabilities of 1.0. Estimated branch lengths for the internal branches (A and B) are presented; the units for branch lengths are substitutions per site for ML and coalescent units for ASTRAL

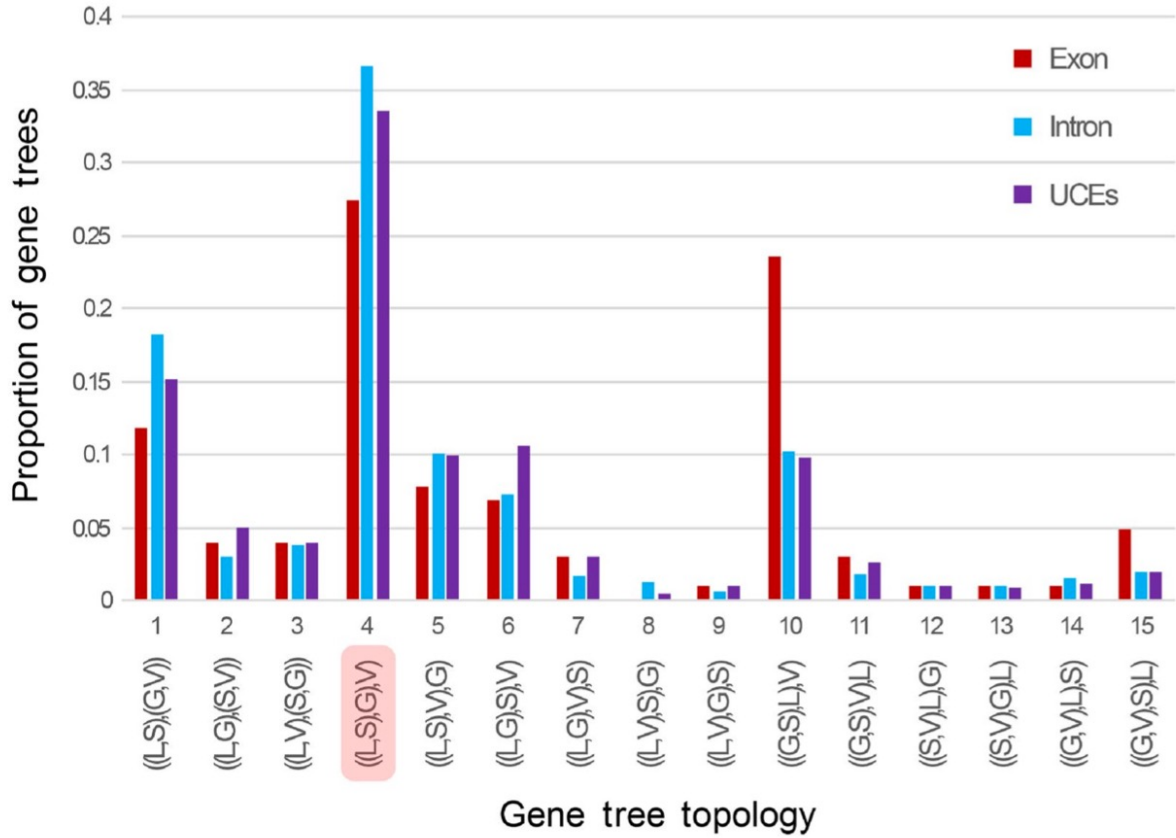
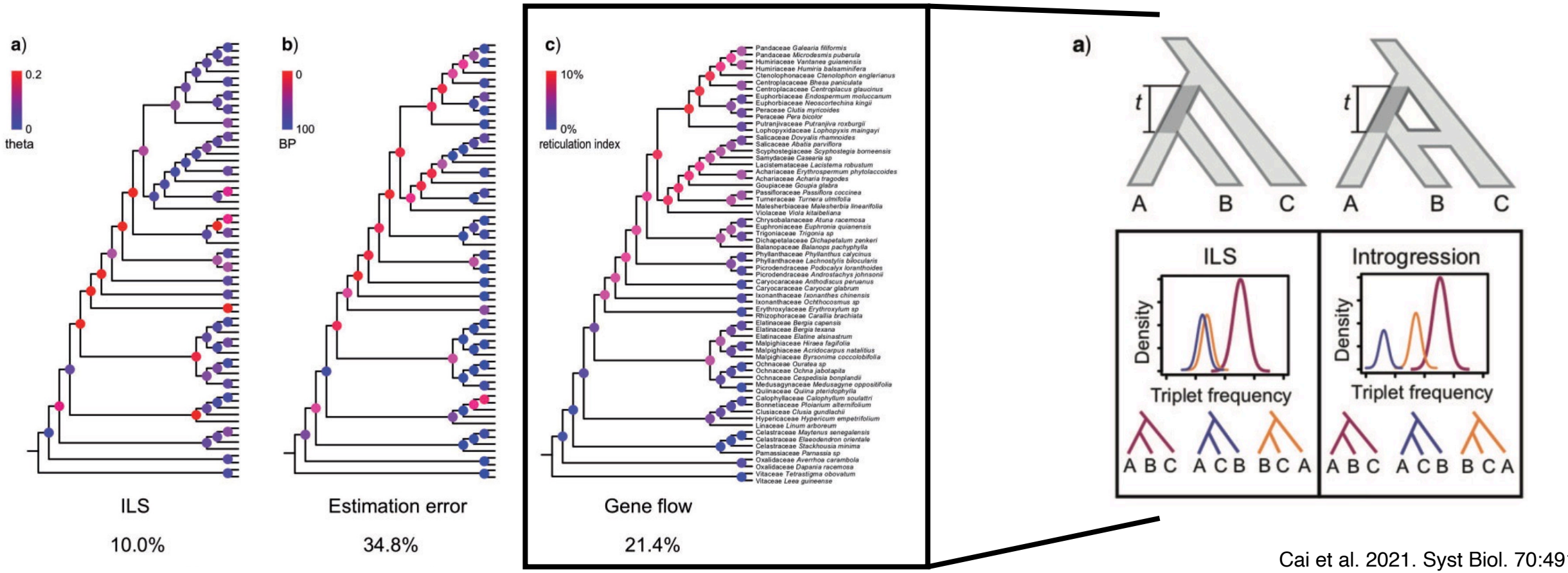


Fig. 5 Observed spectra of gene tree topologies for each data type. The gene trees spectra are expressed as proportions for the three data types that have a sufficient number of gene trees. The order of the topologies for the individual gene trees presented is identical to Fig. 4

Splitting larger problems into rooted triples

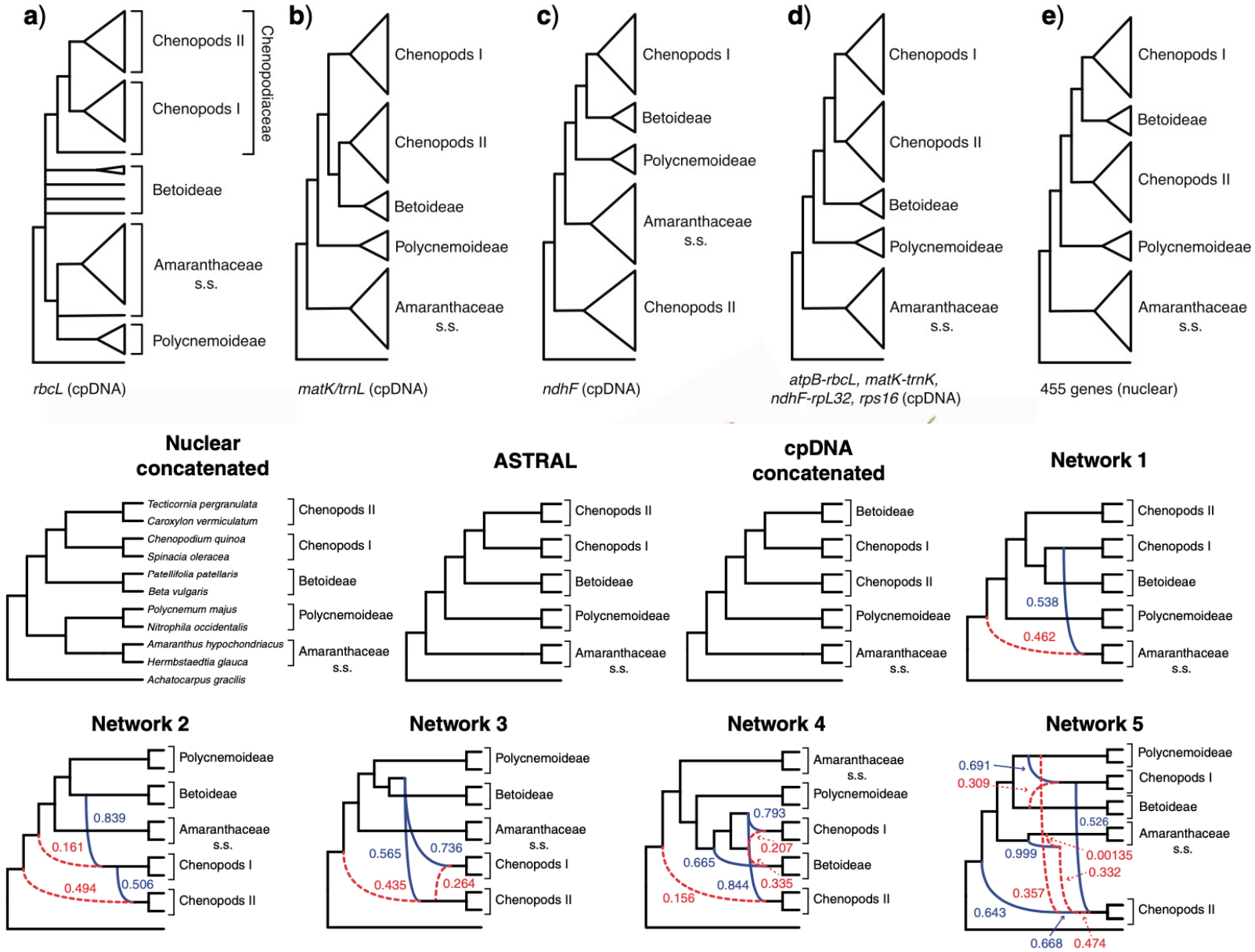
Instead of making strong assumptions about the null distribution for a test statistic, it can be simulated. This appears beneficial as other sources of gene tree variation could cause false positives for gene flow.



Cai et al. 2021. Syst Biol. 70:491-507

Splitting larger problems into rooted triples

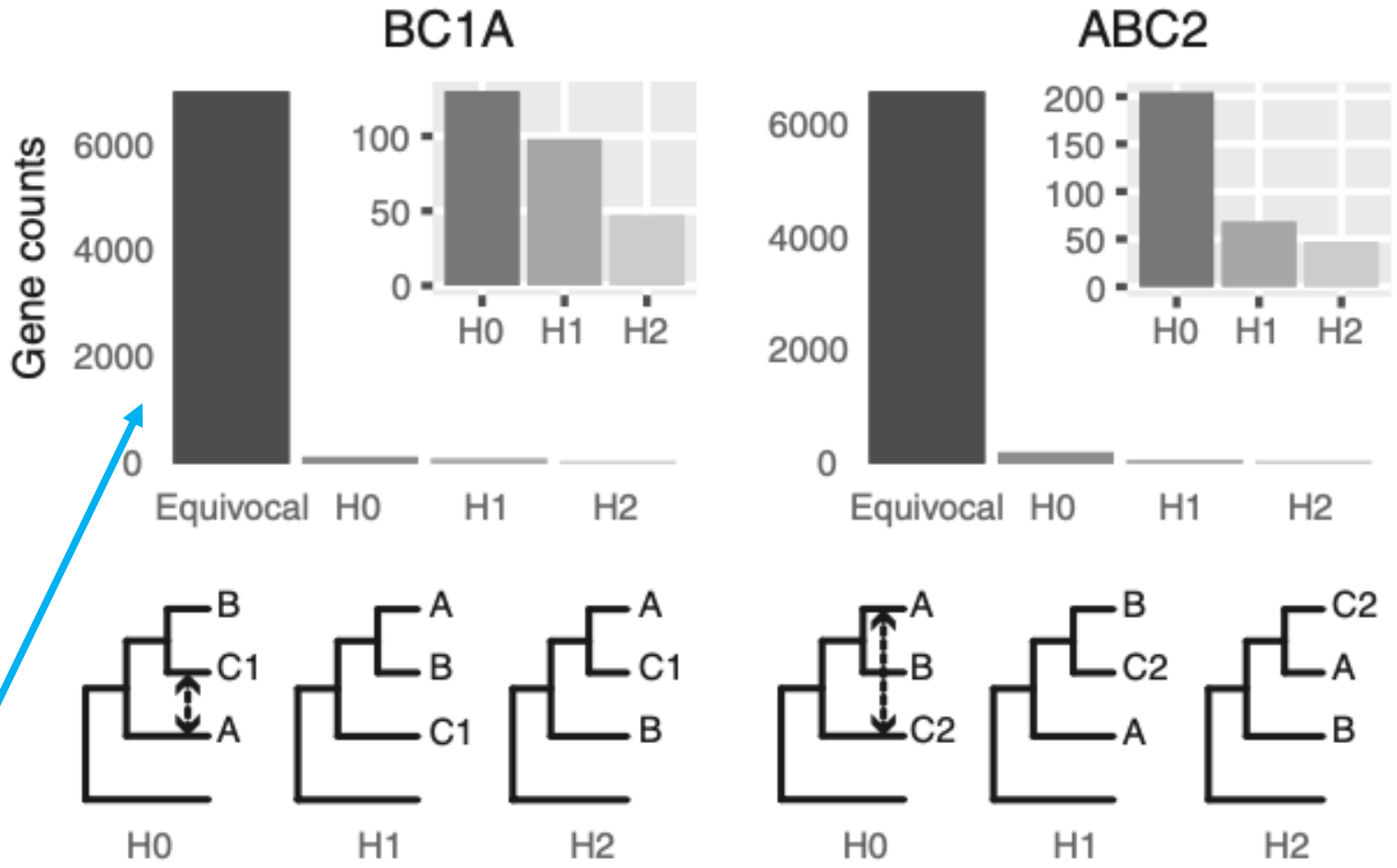
Networks are difficult estimation problems and interpreting them is not always straightforward. A battery of triplet-based tests can instill confidence in some hypotheses.



Morales-Briones et al. 2021. Syst Biol. 219-235

Splitting larger problems into rooted triples

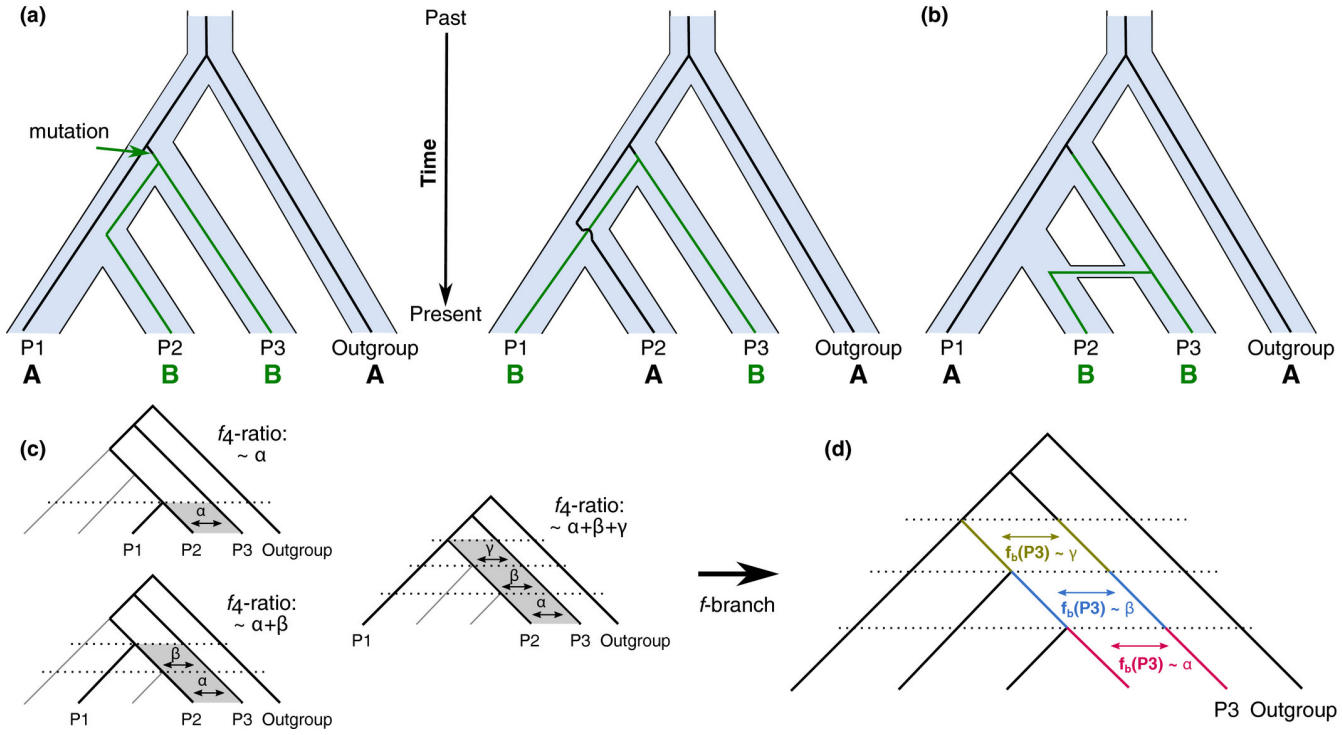
Still testing for inequalities in the minor splits but using an approximately unbiased (AU; Shimodaira 2002) test to only look at gene trees providing strong evidence for a competing topology.



Many individual gene trees may carry little information with respect to a specific topology

Morales-Briones et al. 2021. Syst Biol. 219-235
Shimodaira. 2002. Syst Biol. 51:492-508.

Relationship to some site-pattern methods



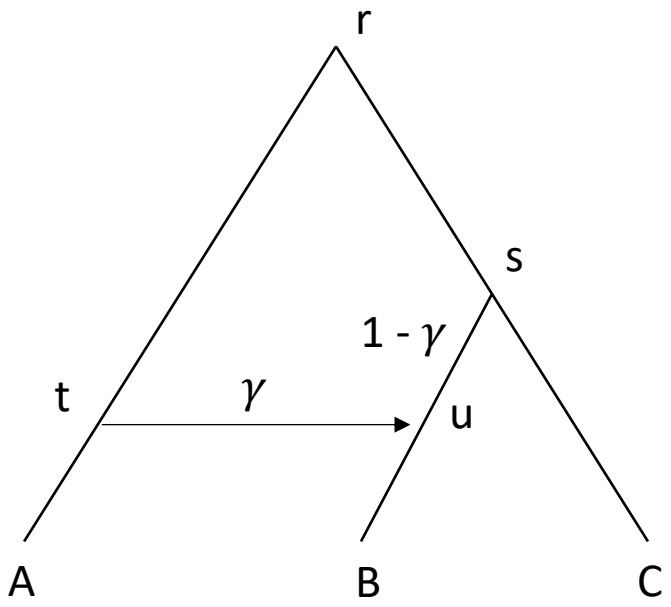
$$D = \frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}}$$

Multiple simplistic assumptions will render the test inappropriate for deeper divergences. Generating the null distribution can take a lot of data.

Malinsky et al. 2020. Mol Ecol Res. 21:584-595
Green et al. 2010. Science, 328: 710-722.

Beyond counting – probabilistic modeling

Introgression



This is modeled as an episodic process

Meng and Kubatko (2009)

Yu et al. (2012)

γ is the inheritance probability

The distribution of gene trees is not a simple mixture with proportions γ and 1-γ, it is a polynomial function of γ

Solís-Lemus and Ané (2016)

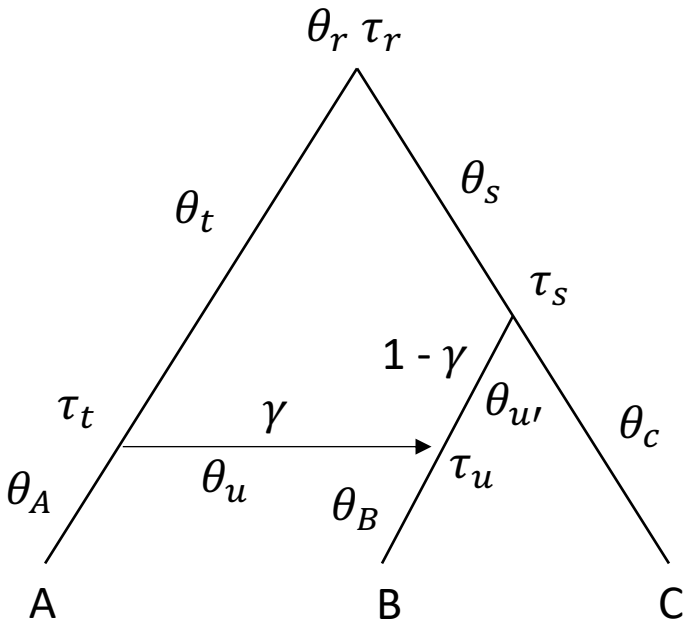
Meng C, Kubatko LS. 2009. Theor Popul Biol. 75:35-45

Yu Y, Degnan JH, Nakhleh. 2012 PLoS Genet. 8:e1002660

Solís-Lemus C, Ané C. 2016. PLoS Genet. 12:e1005896

Beyond counting – probabilistic modeling

Introgression



Optimizing the likelihood function for this MSC model with introgression (MSci, MSNC, NMSC) is very difficult except for very small examples

Estimation of the likelihood through Bayesian MCMC can incur a heavy time and computing burden

gene tree density from Yu et al. 2014

$$f(\tau,\theta,\gamma \mid X) \propto f(\tau,\theta,\gamma) \prod_{i=1}^{L} \int_{G_i} f(G_i|\tau,\theta,\gamma) \times f(X_i|G_i) \partial G_i$$

Felsenstein Likelihood

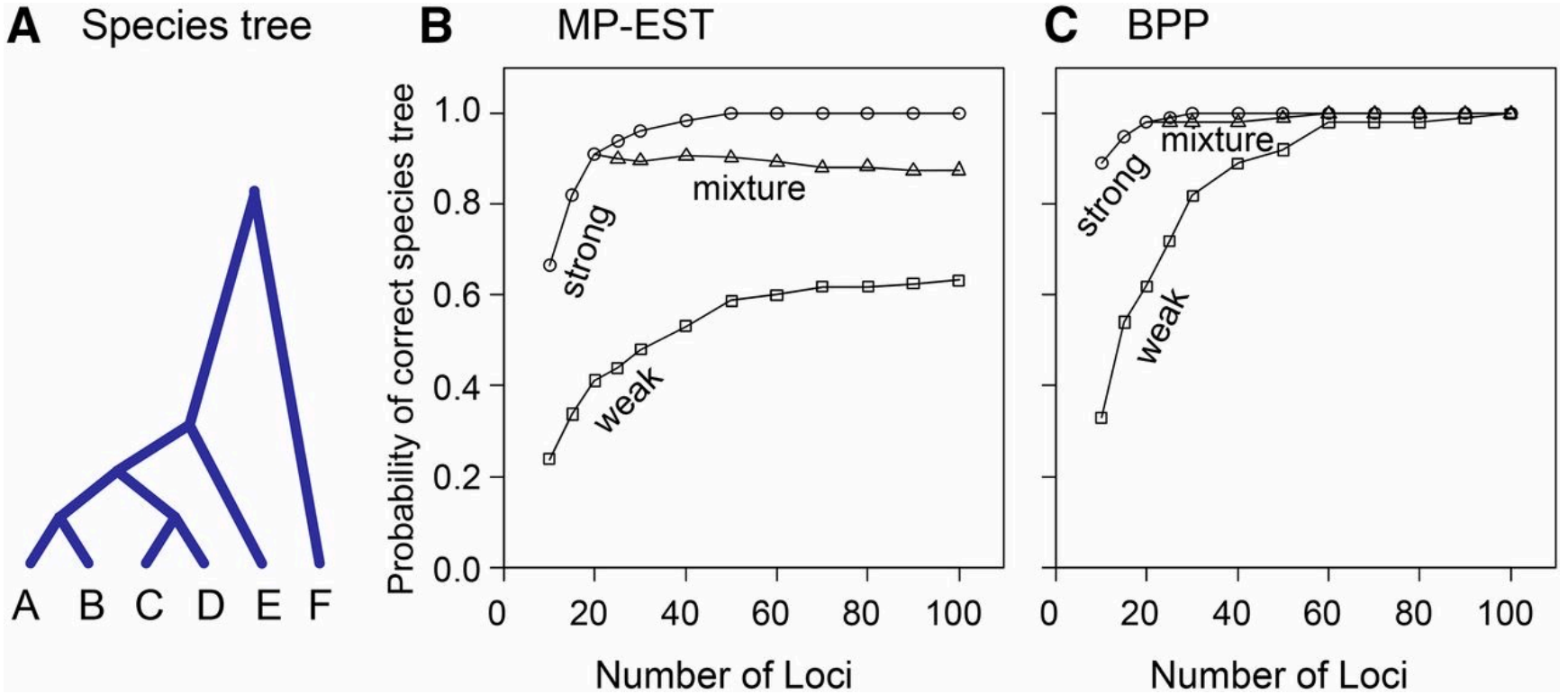
Eq. 1 from Flouri et al. (2020)

Most strategies employed will fix some aspects of the model and use smaller pieces of information (**unrooted quartets** or **rooted triples**) for searches

Meng C, Kubatko LS. 2009. Theor Popul Biol. 75:35-45
Yu Y, Degnan JH, Nakhleh. 2012 PLoS Genet. 8:e1002660
Solís-Lemus C, Ané C. 2016. PLoS Genet. 12:e1005896
Yu Y, et al. 2014. Proc Natl Acad Sci USA.111:16448-16453
Flouri T, et al. 2020. Mol Biol Evol. 37: 1211-1223

Beyond counting – probabilistic modeling

A benefit of joint modeling of species trees (probably networks too) and gene trees is the damage from low-information tress is mitigated.



Xu and Yang. 2016. Genetics. 204:1353-1368

Marginal Likelihoods and Bayes Factors

Some model selection approaches with likelihoods do not work for Bayesian estimators

For multiple competing hypotheses

$$AIC = 2k - 2\ln L$$

For nested cases

$$\mathrm{LRT} = -2 \times (lnL_1 - lnL_0)$$

Marginal Likelihoods and Bayes Factors

Some model selection approaches with likelihoods do not work for Bayesian estimators

Likelihood

Prior Probability Density

Posterior Probability Density

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

Marginal Probability of the data

Marginal Likelihoods and Bayes Factors

Some model selection approaches with likelihoods do not work for Bayesian estimators

Likelihood

Prior Probability Density

Posterior Probability Density

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

Marginal Probability of the data "Marginal Likelihood"

If we were to, for example, take the mean lnL from the posterior, that includes the effect of the prior. We want the denominator.

Marginal Likelihoods and Bayes Factors

$$BF_{10} = \frac{f(D|M_1)}{f(D|M_0)}$$

The ratio of marginal likelihoods in favor of model 1

$f(D|M_1)$ = Marginal Likelihood of **Model 1**

$f(D|M_0)$ = Marginal Likelihood of **Model 0**

Jeffreys. 1935. Proc Camb Phil Soc. 31:203-222.
Kass and Raftery. 1995. J Am Stat Assoc. 90:773-795.

Marginal Likelihoods and Bayes Factors

$$BF_{10} = \frac{f(D|M_1)}{f(D|M_0)}$$

$f(D|M_1)$ = Marginal Probability of the **D**ata under **M**odel **1**

$$f(D|M_1) = \int f(D|\theta)\, f(\theta)\, d\theta$$

Posterior

Prior

Marginal Likelihoods and Bayes Factors

$$BF_{10} = \frac{f(D|M_1)}{f(D|M_0)}$$

$$BF_{10} > 1 = evidence\ in\ favor\ of\ model\ 1$$

$$BF_{10} < 1 = evidence\ in\ favor\ of\ model\ 0$$

Marginal Likelihoods and Bayes Factors

Marginal Likelihoods can be used to calculate model probabilities when there are multiple competing hypotheses

$$prob\ of\ model\ w\ of\ i = \frac{exp(lnL_w - lnL_{max})}{\sum_i [exp(lnL_j - lnL_{max})]}$$

Marginal Likelihoods and Bayes Factors

An example

lnL	lnLi - lnLmax	exp(lnLi - lnLmax)	exp(lnLi - lnLmax)/sum
-452715.5			
-452713.7			
-452704.6			
-452705.2			
-452704.7			

Marginal Likelihoods and Bayes Factors

An example

$$log\left(\frac{m}{z}\right) = \log(\mathrm{m}) - \log(\mathrm{z})$$

lnL	lnLi - lnLmax	exp(lnLi - lnLmax)	exp(lnLi - lnLmax)/sum
-452715.5	-10.9		
-452713.7	-9.1		
-452704.6	0		
-452705.2	-0.6		
-452704.7	-0.1		

Marginal Likelihoods and Bayes Factors

The Bayes factors that say how less likely a model is wrt the least worst one

An example

$$log\left(\frac{m}{z}\right) = \log(\mathrm{m}) - \log(\mathrm{z}) \qquad exp^{log(m)} = m$$

lnL	lnLi - lnLmax	exp(lnLi - lnLmax)	exp(lnLi - lnLmax)/sum
-452715.5	-10.9	1.84582E-05	7.52237E-06
-452713.7	-9.1	0.000111666	4.55077E-05
-452704.6	0	1	0.407534634
-452705.2	-0.6	0.548811636	0.223659749
-452704.7	-0.1	0.904837418	0.368752586

Marginal Likelihoods and Bayes Factors

We will focus on methods that use *power posteriors*

19 Dubious Ways to Compute the Marginal Likelihood of a Phylogenetic Tree Topology

MATHIEU FOURMENT[1], ANDREW F. MAGEE[2], CHRIS WHIDDEN[3], ARMAN BILGE[3], FREDERICK A. MATSEN IV[3], AND VLADIMIR N. MININ[4,*]

TABLE 1. Names, abbreviations, and number of required MCMC chains involved in applying the 19 methods

Abbreviation	Full name	# MCMC chains
ELBO	Evidence lower bound	0
GLIS	Gamma Laplus importance sampling	0*
VBIS	Varational Bayes importance sampling	0*
BL	Beta′ Laplus	0
GL	Gamma Laplus	0
LL	Lognormal Laplus	0
MAP	Maximum un-normalized posterior probability	0
ML	Maximum likelihood	0
NMC	Naïve Monte Carlo	0*
BS	Bridge sampling	1
CPO	Conditional predictive ordinates	1
HM	Harmonic mean	1
SHM	Stabilized harmonic mean	1
NS	Nested sampling	Multiple short chains
PPD	Pointwise predictive density	1
PS	Path sampling	50
MPS	Modified path sampling	50
SS	Stepping stone	50
GSS	Generalized stepping stone	50

Note: GLIS, VBIS, and NMC (*) do not require MCMC samples but perform importance sampling. Stepping stone and path sampling methods employ an unspecified number of steps; we found 50 to be sufficient.

Fourment et al. 2020 Syst Biol. 69:209-220

Marginal Likelihoods and Bayes Factors

Goal: Create a path from the prior to the posterior given some weights and expectation of the log likelihood to create an integral that factors out the effect of the prior.

A weighted average of power posteriors.

$$f(D|M) = \int f(D|\theta)f(\theta)d\theta$$

Gelman and Meng. 1998. Stat Sci. 13:163-185.
Lartillot and Philippe 2006. Syst Biol. 55:195-207.

Marginal Likelihoods and Bayes Factors

Goal: Create a path from the prior to the posterior given some weights and expectation of the log likelihood to create an integral that factors out the effect of the prior.

A weighted average of power posteriors.

$$logf(D|M) = \int_0^1 E_\beta[logf(D|\theta)]d\beta$$

For computers, an integral means to sum

Gelman and Meng. 1998. Stat Sci. 13:163-185.
Lartillot and Philippe 2006. Syst Biol. 55:195-207.

Marginal Likelihoods and Bayes Factors

Goal: Create a path from the prior to the posterior given some weights and expectation of the log likelihood to create an integral that factors out the effect of the prior.

A weighted average of power posteriors.

How many weights are enough?

$$logf(D|M) \approx \frac{1}{2} \sum_{k=1}^{K} w_k \times E_\beta[logf(D|\theta)]$$

Approximated by the mean!

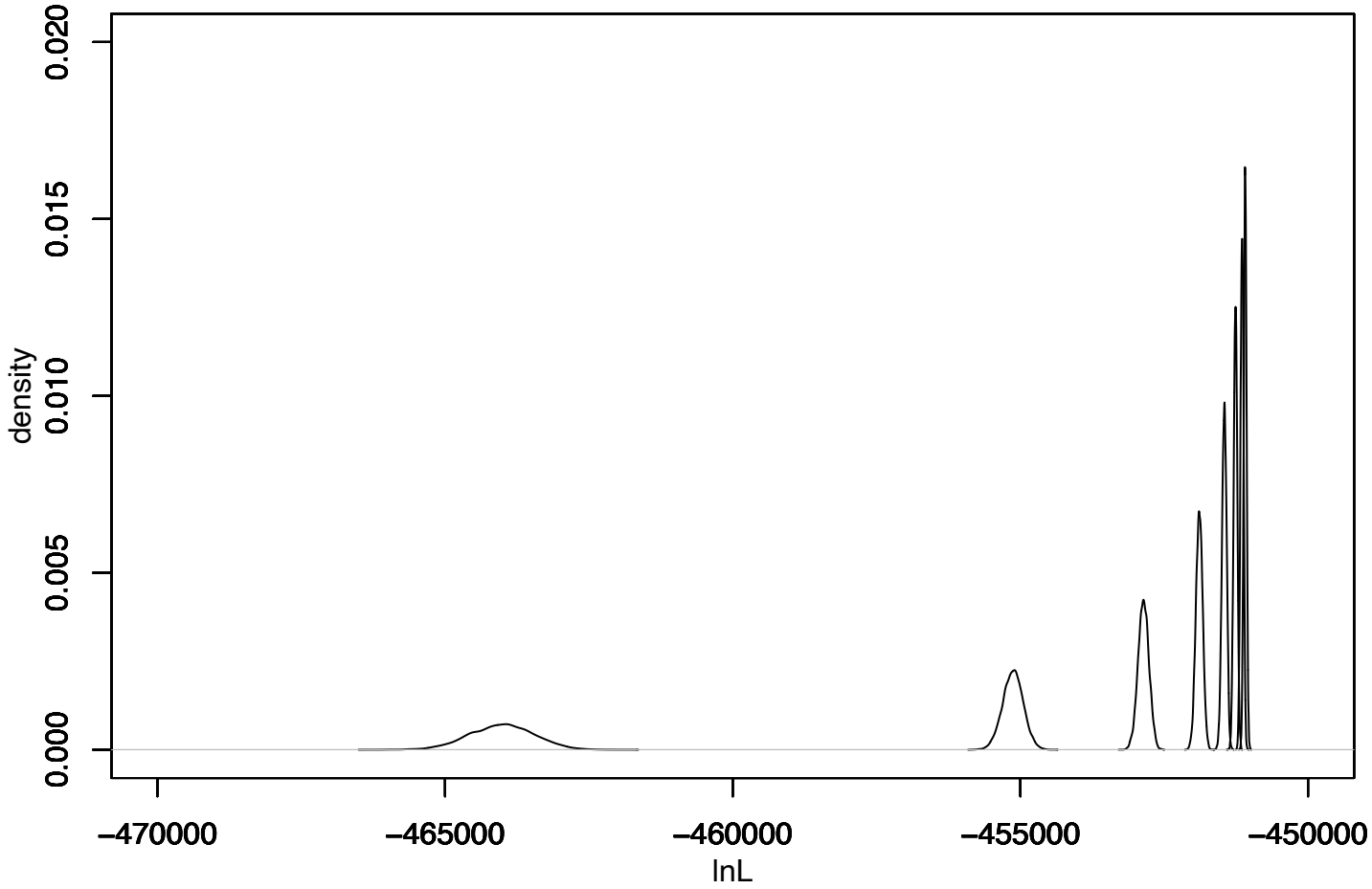
The ½ sneaks in because of a change of interval limits for Gauss-Legendre Quadrature

Gelman and Meng. 1998. Stat Sci. 13:163-185.
Lartillot and Philippe 2006. Syst Biol. 55:195-207.
Rannala and Yang. 2016. Syst Biol. 66:823-842

-452715.5

Marginal Likelihoods and Bayes Factors

$$logf(D|M) \approx \frac{1}{2}\sum_{k=1}^{K} w_k \times E_\beta[logf(D|\theta)]$$



The log marginal likelihood

-452715.5

The posterior mean

-451080.5

Gelman and Meng. 1998. Stat Sci. 13:163-185.
Lartillot and Philippe 2006. Syst Biol. 55:195-207.
Rannala and Yang. 2016. Syst Biol. 66:823-842

Marginal Likelihoods and Bayes Factors

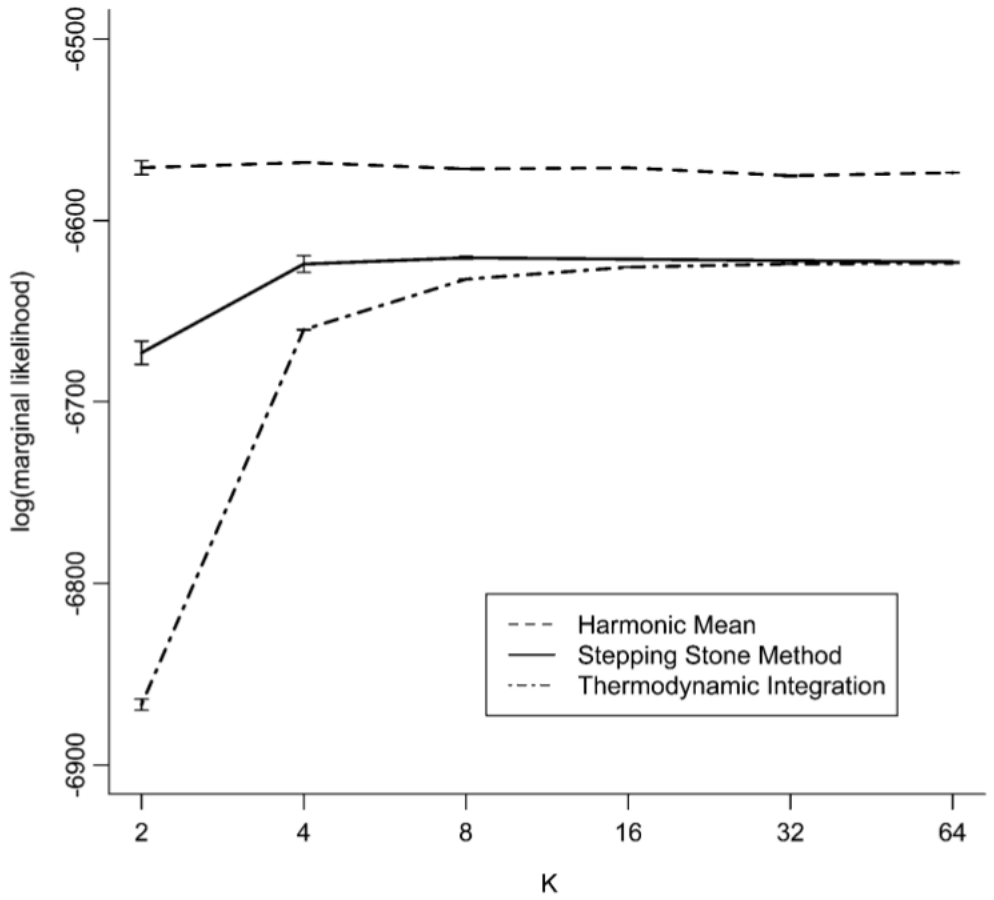


FIGURE 5. Log marginal likelihood for three estimation methods as a function of the number of β intervals, *K*, for the green plant Ribulose Bisphosphate Carboxylase/Oxygenase large subunit (*rbc*L) example. β values are evenly spaced quantiles from a Beta(0.3,1.0) distribution. Error bars represent ±1 standard error based on 30 independent MCMC analyses.

Stepping-Stone Sampling

- potentially more efficient (i.e. less steps needed) by sampling a skewed distribution of beta

- Attempts to create better numerical stability in the expectation by factoring out the largest likelihood in the sample

- more complicated to calculate expectations

Xie et al. 2011. Syst Biol. 60:150—160.

Marginal Likelihoods and Bayes Factors

Estimating marginal likelihoods is an area of active research

Syst. Biol. 72(3):639–648, 2023

https://doi.org/10.1093/sysbio/syad007
Advance Access Publication February 28, 2023

LoRaD: Marginal likelihood estimation with haste (but no waste)

YU-BO WANG[1], ANALISA MILKEY[2], AOLAN LI[3], MING-HUI CHEN[3], LYNN KUO[3], AND PAUL O. LEWIS[2,*]

[1]*School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634, USA*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA*
[3]*Department of Statistics, University of Connecticut, Storrs, CT 06269, USA*
**Correspondence to be sent to: Paul O. Lewis, Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269, USA; E-mail: paul.lewis@uconn.edu*

Approximating Bayes Factors

We established that marginal likelihood estimation, though useful, can incur a computational burden. When the null and alternative hypothesis is well-defined and nested, the Bayes factor can be approximated.

Approximating Bayes Factors

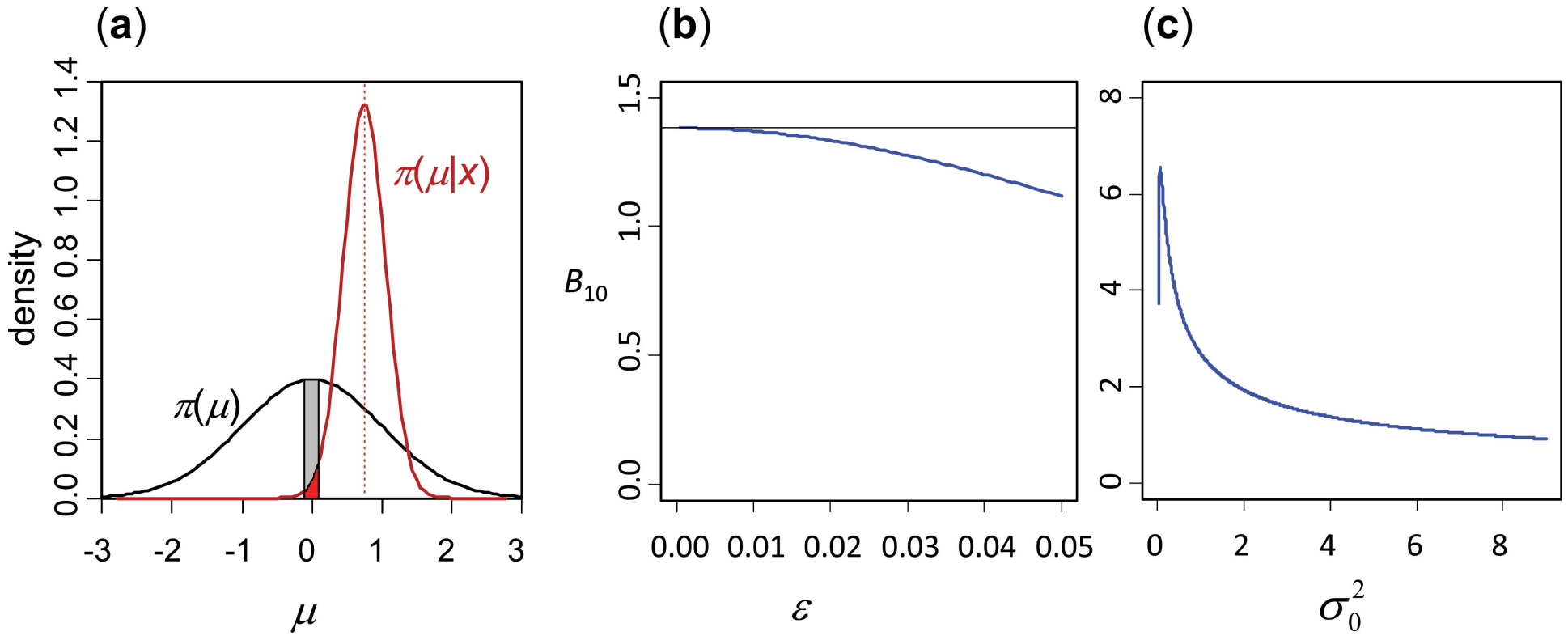
$$BF_{10} = \frac{f(D|M_1)}{f(D|M_0)} = \frac{q(\varphi = \varphi_0|M_1)}{p(\varphi = \varphi_0|D, M_1)}$$

Let the parameter of interest have some region of null effects. This is phi in the case of testing introgression. The ratio of the prior probability to the posterior probability at the null region should yield the Bayes factor with respect to the more complex model.

Verdinelli and Wasserman. 1995. J Am Stat Assoc. 90:614-618.
Suchard et al. 2001. Mol Biol Evol. 18:1001-1013.

Approximating Bayes Factors

However, this only holds as the null region approaches 0.

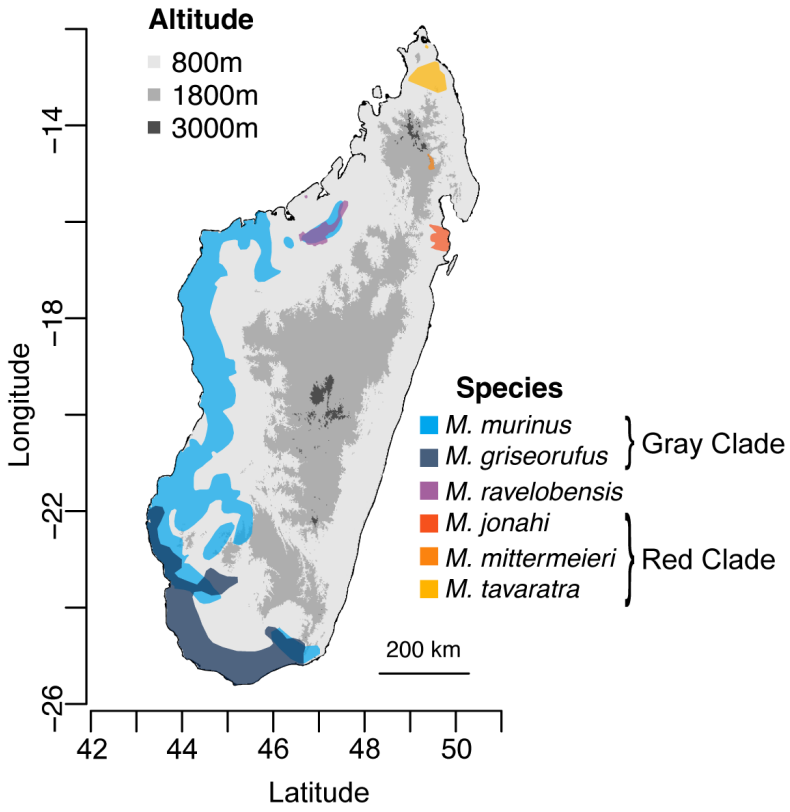


And can be affected by other prior choices too.

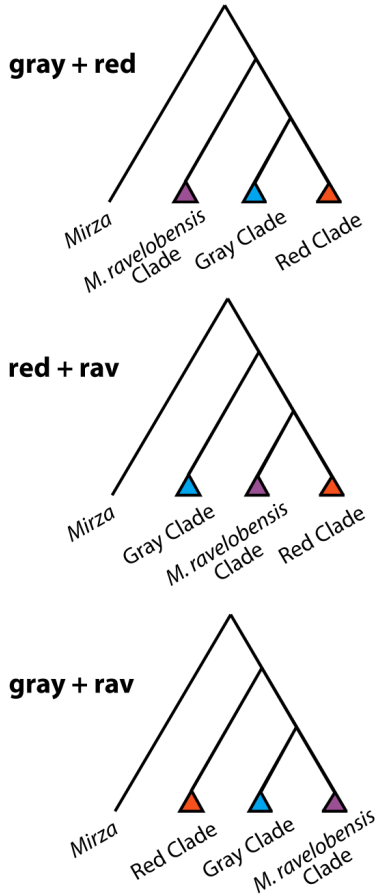
Ji et al. 2023 Syst Biol. 72:446-465

Putting it all together

Species distributions



Phylogenetic hypotheses



Citations

Louis *et al*. 2006

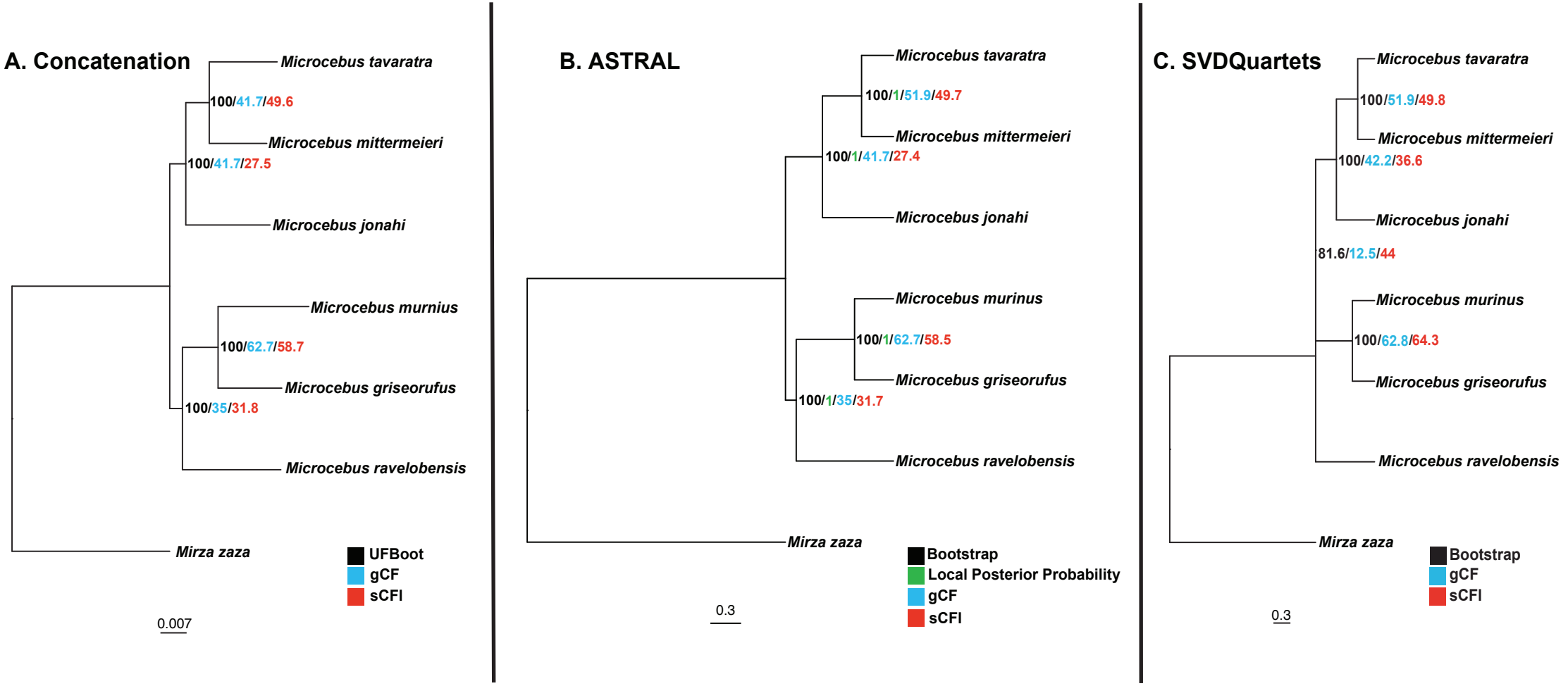
Yoder *et al*. 2000
Weisrock *et al*. 2010
Herrera *et al*. 2016
Hotaling *et al*. 2016
dos Reis *et al*. 2018

Weisrock *et al*. 2012
Everson *et al*. 2023

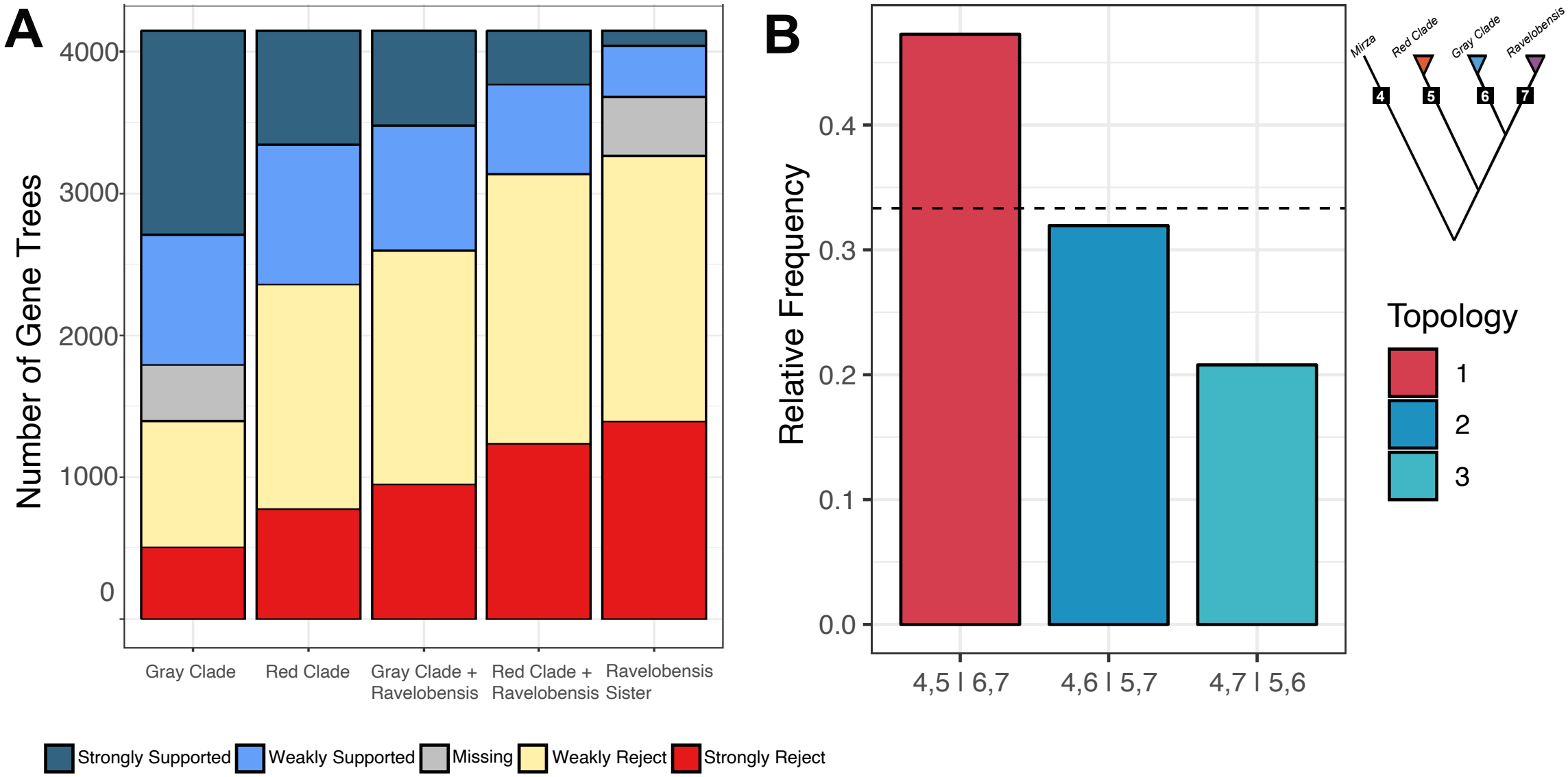
David Haring, ©
Duke Lemur Center

Blake Fauskee
PhD Candidate
Duke University

Putting it all together

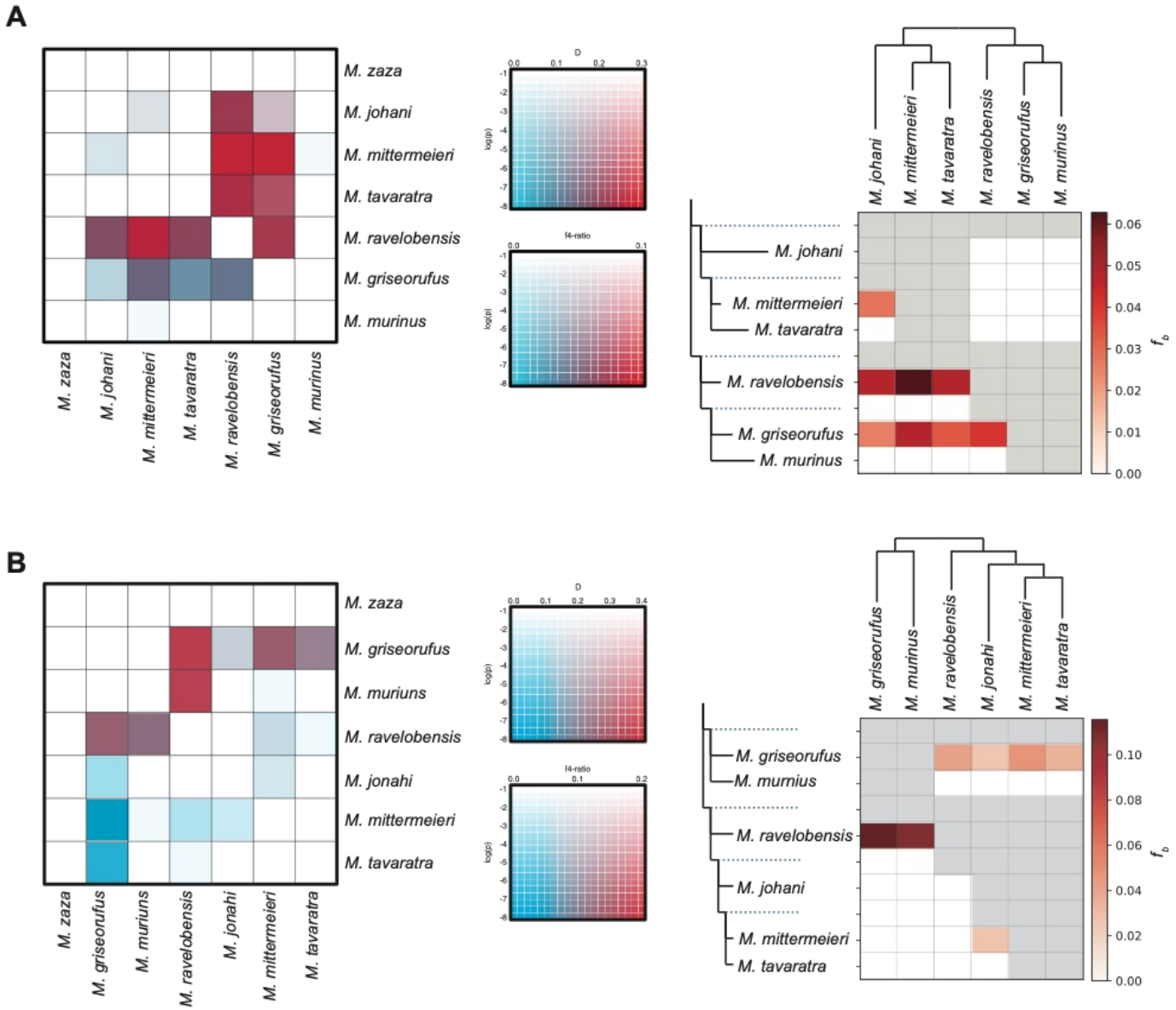


Putting it all together

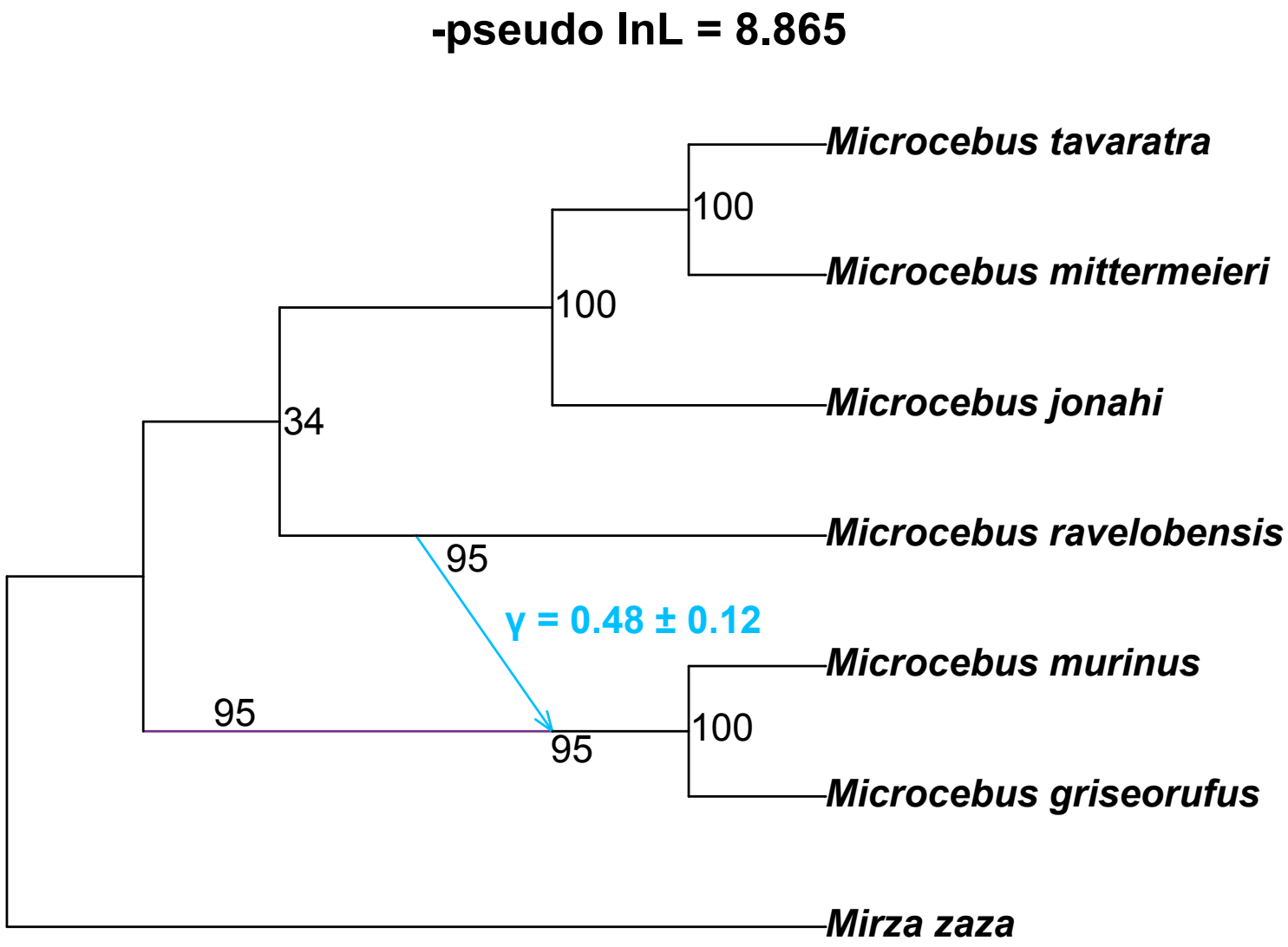


Putting it all together

D-statistics and related site-pattern-based methods are a great way to explore data. It is nice when they support *a priori* hypotheses or corroborate additional evidence.



Putting it all together



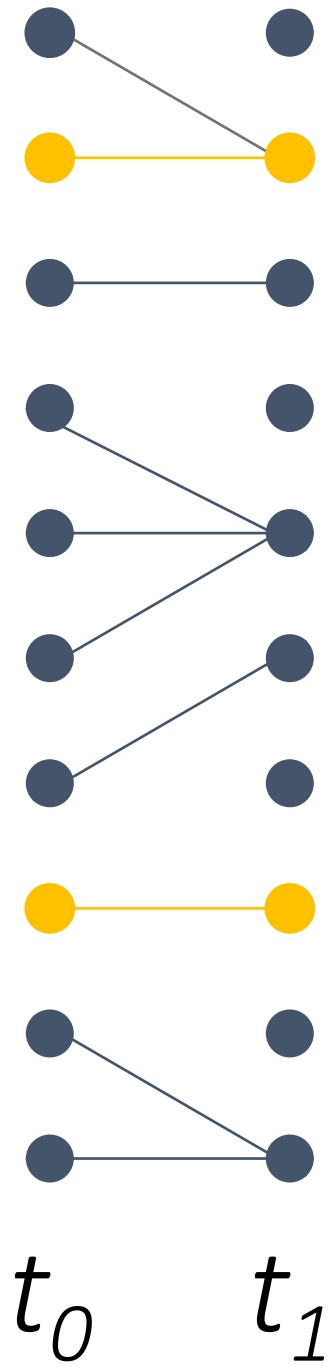
Putting it all together

Topology	marginal lnL	SE	2lnBF†
(Mrav,(Grey,Red))	-11255520	4.164485	-1478
(Red,(Grey,Mrav))	-11255150	4.190757	-738
(Grey,(Mrav,Red))	-11255557	4.088052	-1552
Network Hypothesis	-11254781	4.148756	0
Two-rate Model A	-11255439	4.377716	-658
Two-rate Model B	-11255465	4.22489	-684

†log Bayes Factors calculated with respect to the best network model

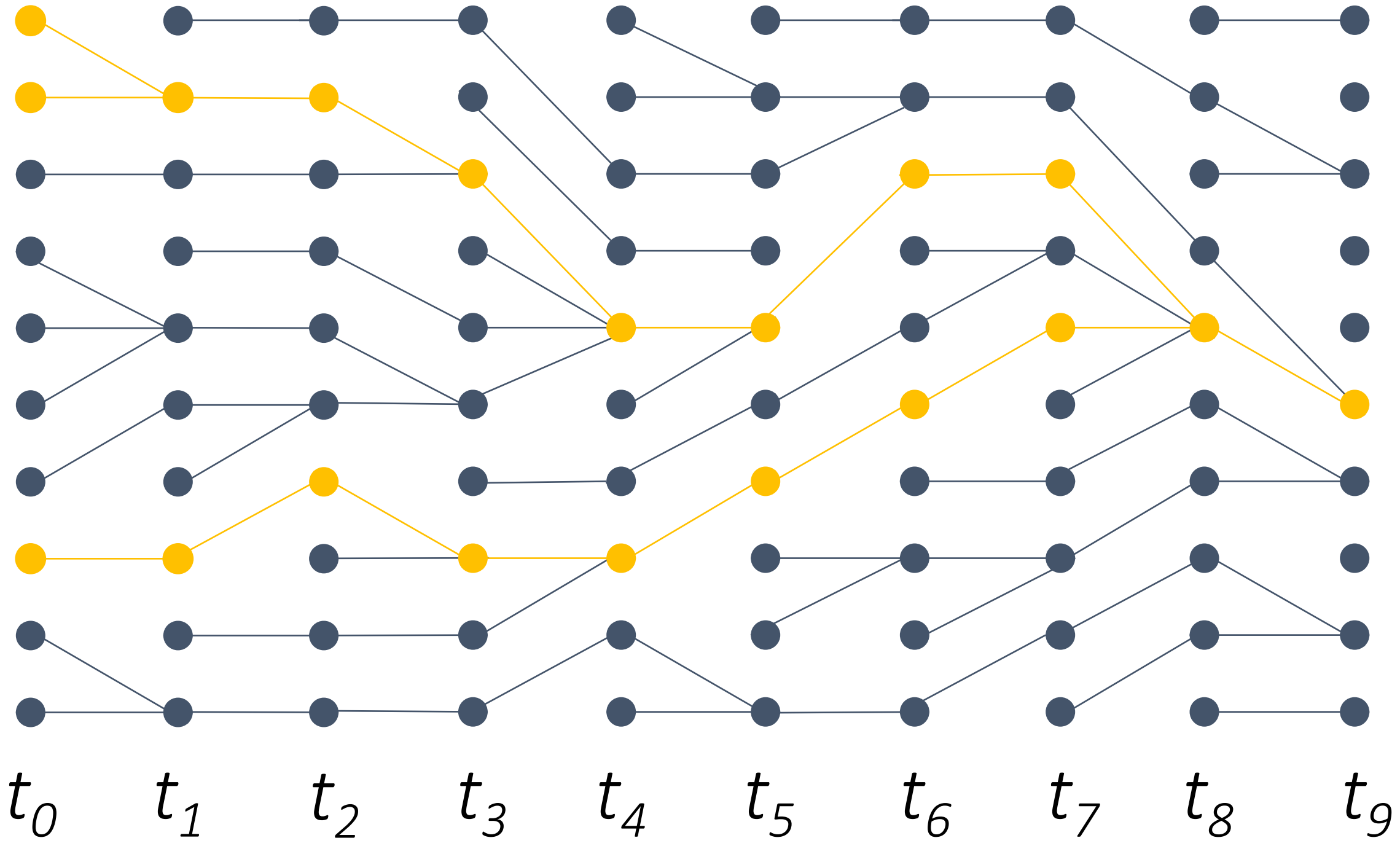
End

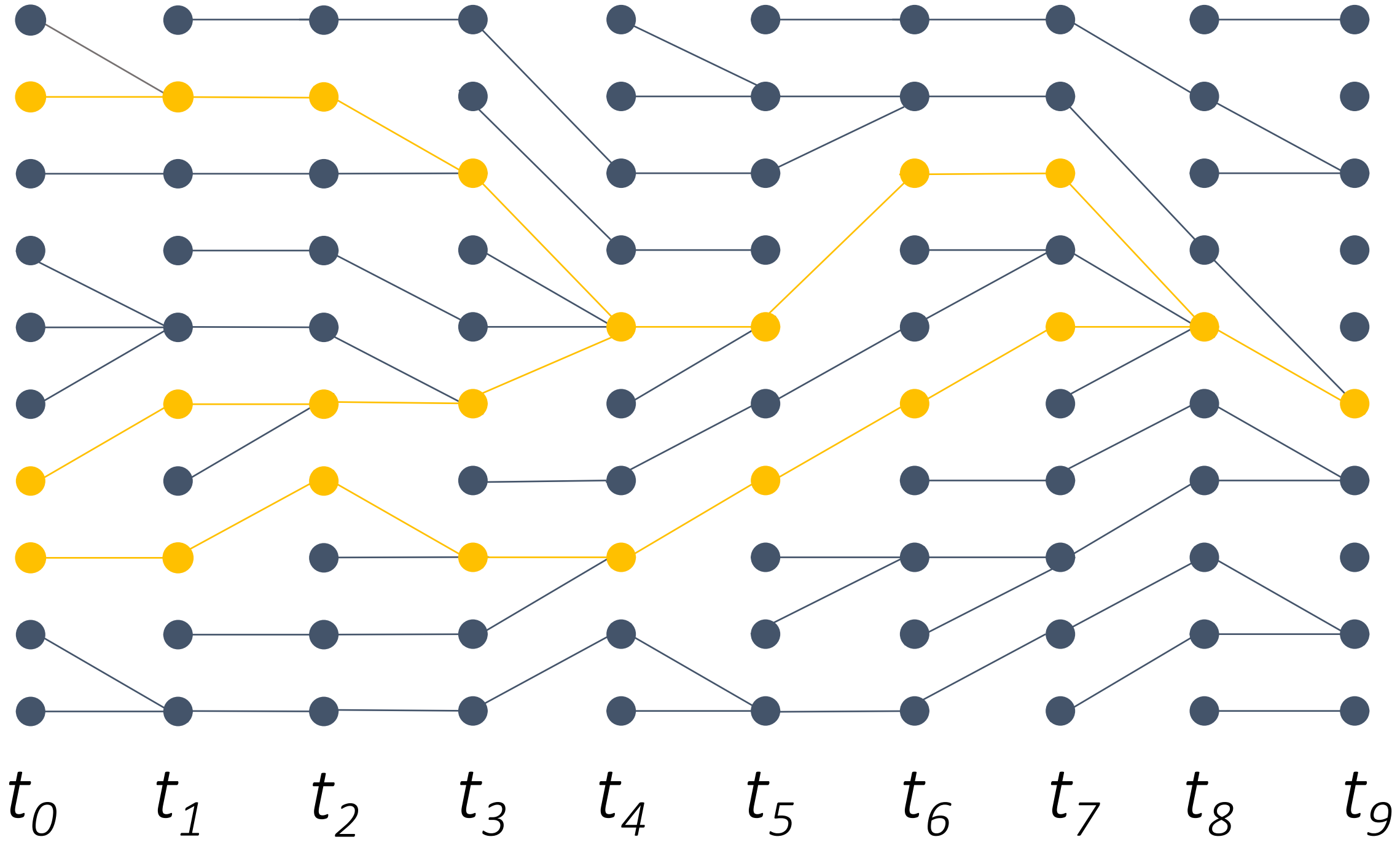
Time to look at the baobab data a little more!

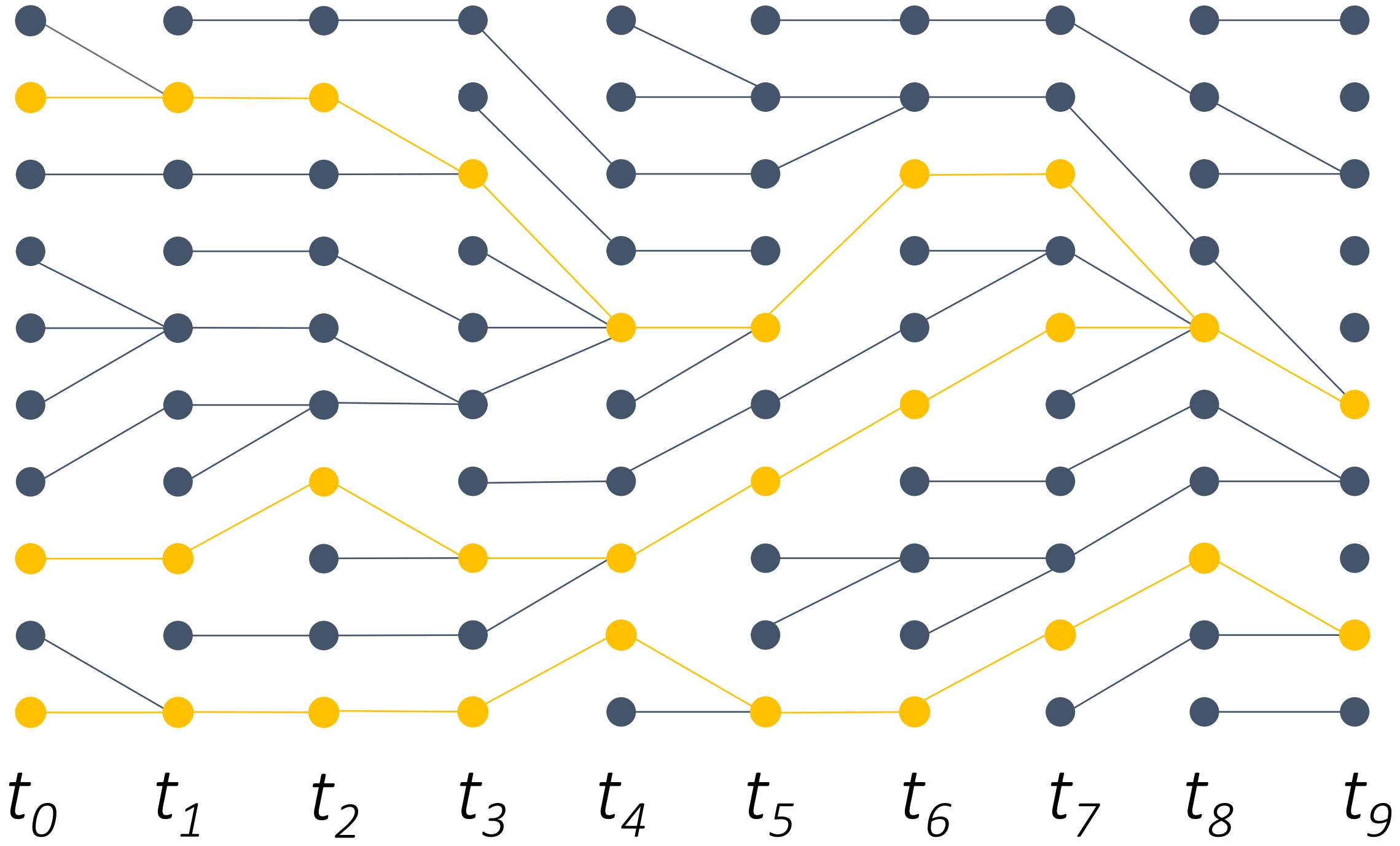


Secret slides about 4N and T_{MRCA}

What about more than 2 alleles?


t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9

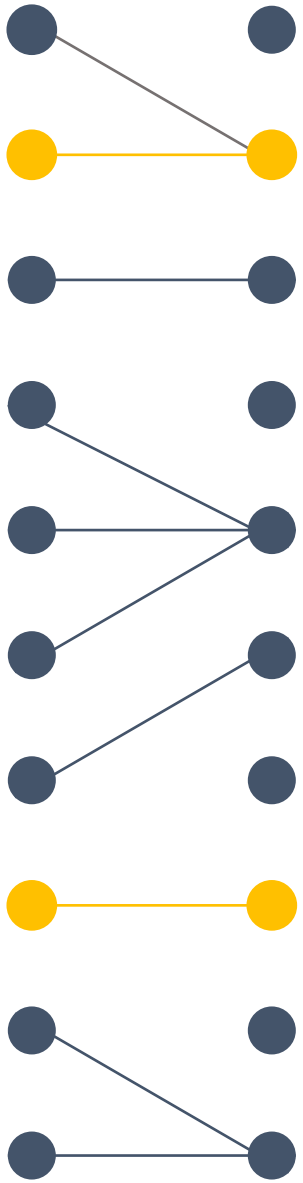

t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9


t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9

P(3 alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{2}{2N}\right)$$

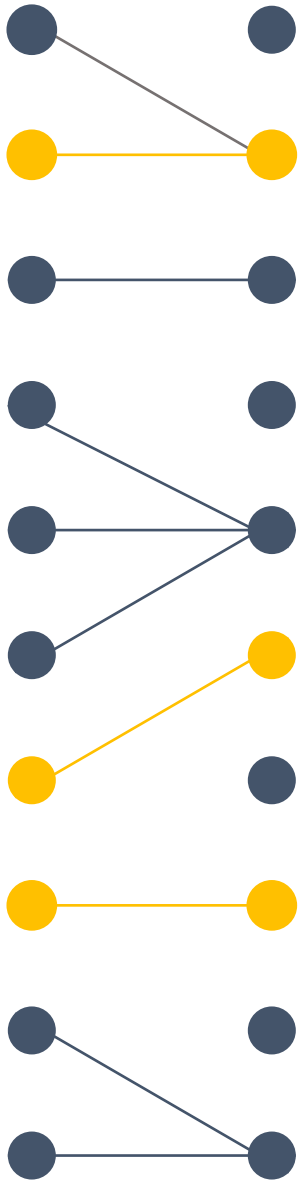
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(2 alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right)$$

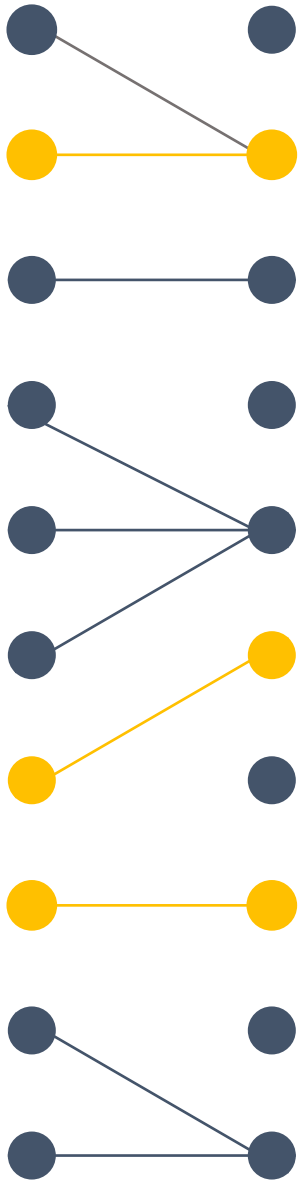
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(3 alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{2}{2N}\right)$$

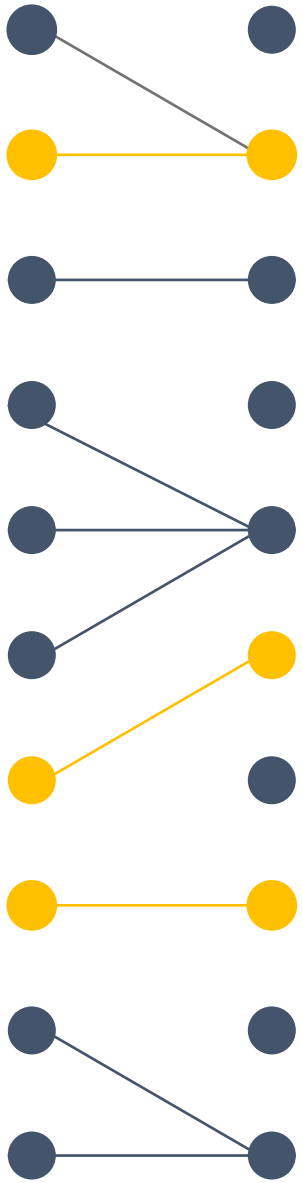
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(*n* alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{2}{2N}\right) \times \cdots \times \left(1 - \frac{n-1}{2N}\right)$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

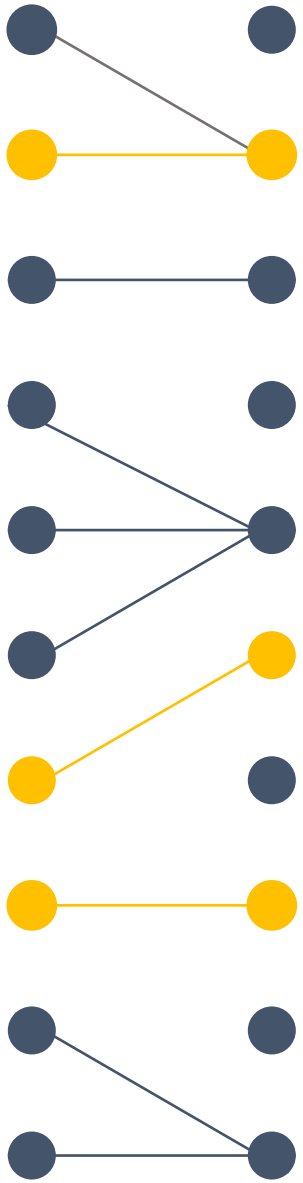


P(*n* alleles do not coalesce in 1 generation)

$$\approx \left(1 - \frac{1 + 2 + \cdots + (n-1)}{2N}\right)$$

$$= 1 - \binom{n}{2}\left(\frac{1}{2N}\right)$$

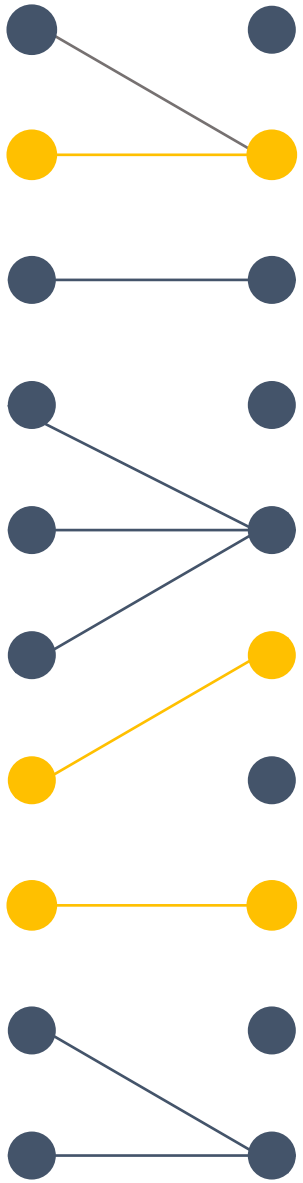
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(*n* alleles coalesce in the i^{th} generation)

$$= \left[1 - \binom{n}{2}\left(\frac{1}{2N}\right)\right]^{i-1} \binom{n}{2}\frac{1}{2N}$$

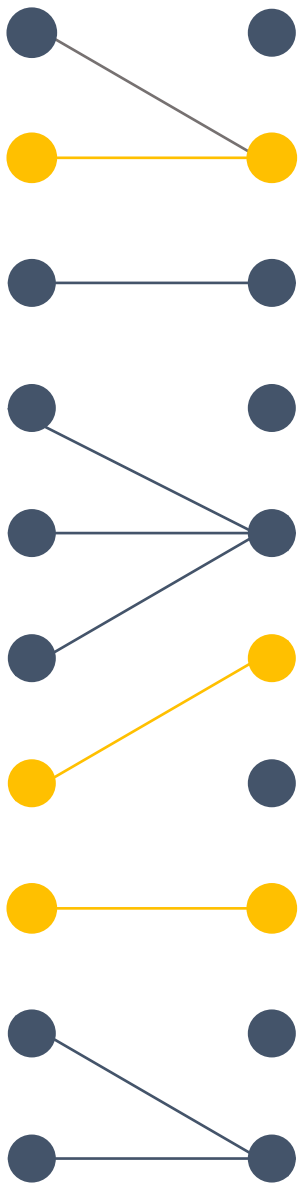
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



This gets things into a similar geometric distribution, thus

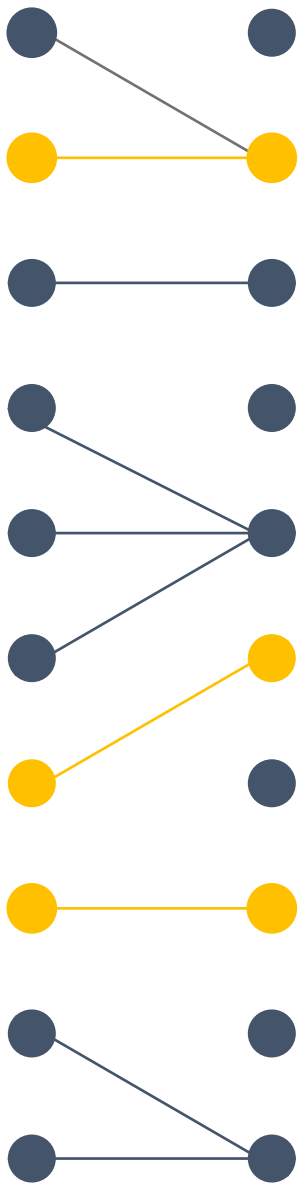
$$\mathrm{E}[t] = 1/p = \frac{2N}{\binom{n}{2}}$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



And we can approximate the probability of coalescent waiting times with an exponential

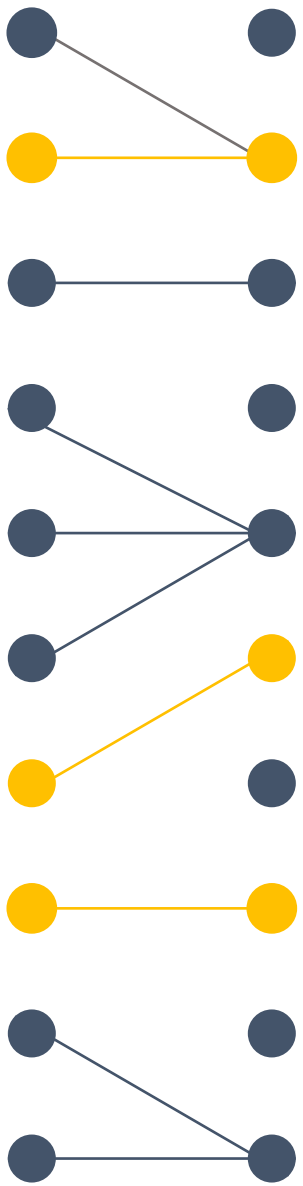
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



And we can approximate the probability of coalescent waiting times with an exponential

$$T = \frac{1}{2N}$$

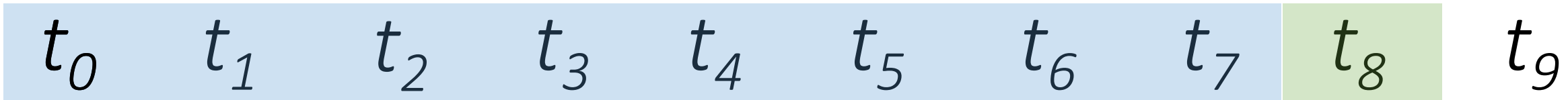
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

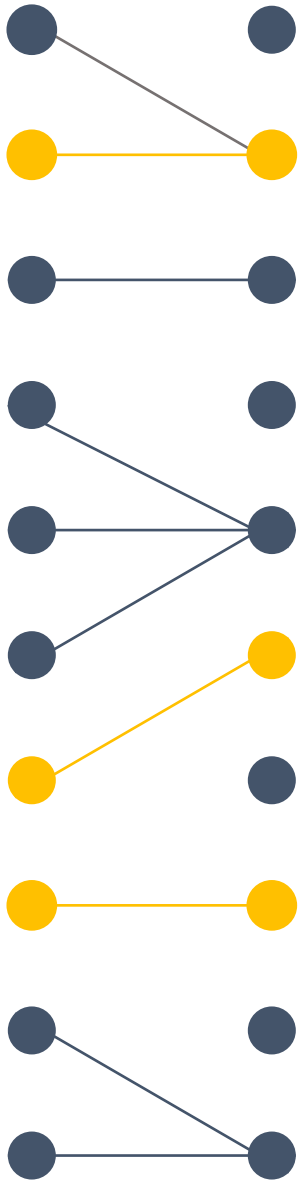


And we can approximate the probability of coalescent waiting times with an exponential

$$T = \frac{1}{2N}$$

But there are *n-1* coalescent events to happen and $\binom{n}{2}$ ways to get there





$$\binom{n}{2} = \frac{n(n-1)}{2}$$

So for the j^{th} coalescence

$$f(T_j) = \frac{j(j-1)}{2} \exp\left\{-\frac{j(j-1)}{2} T_j\right\}$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

And all coalescences for a given genealogy *G*

$$f(T|G) = \prod_{j=2}^{n} \frac{j(j-1)}{2} \exp\left\{-\frac{j(j-1)}{2} T_j\right\}$$

We can derive expectations for the T_{MRCA}

$$f(T|G) = \prod_{j=2}^{n} \frac{j(j-1)}{2} \exp\left\{-\frac{j(j-1)}{2} T_j\right\}$$

$$E[T_j] = \frac{2}{j(j-1)}$$

$$E[T_{MRCA}] = E(T_n + T_{n-1} + \cdots + T_2)$$

We can derive expectations for the T_{MRCA}

$$E[T_{MRCA}] = E(T_n + T_{n-1} + \cdots + T_2)$$

$$E[T_{MRCA}] = \sum_{j=2}^{n} \frac{2}{j(j-1)} = 2\sum_{j=2}^{n}\left(\frac{1}{j-1} - \frac{1}{j}\right)$$

$$E[T_{MRCA}] = 2\left(1 - \frac{1}{n}\right) \approx 2 = 4N$$