Species Tree Inference

Bruce Rannala @ UC Davis

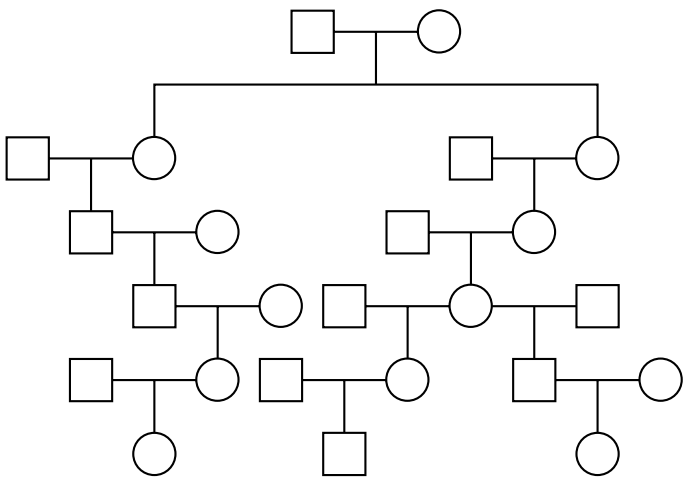
Outline

Coalescent Theory

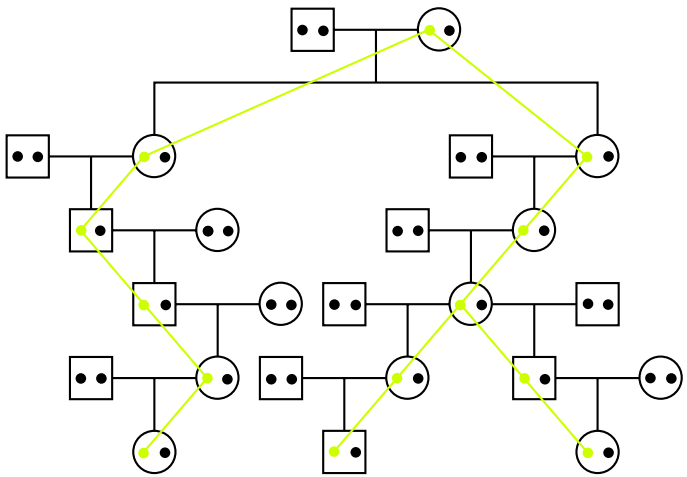
Multispecies Coalescent

Phylodemographic Inference

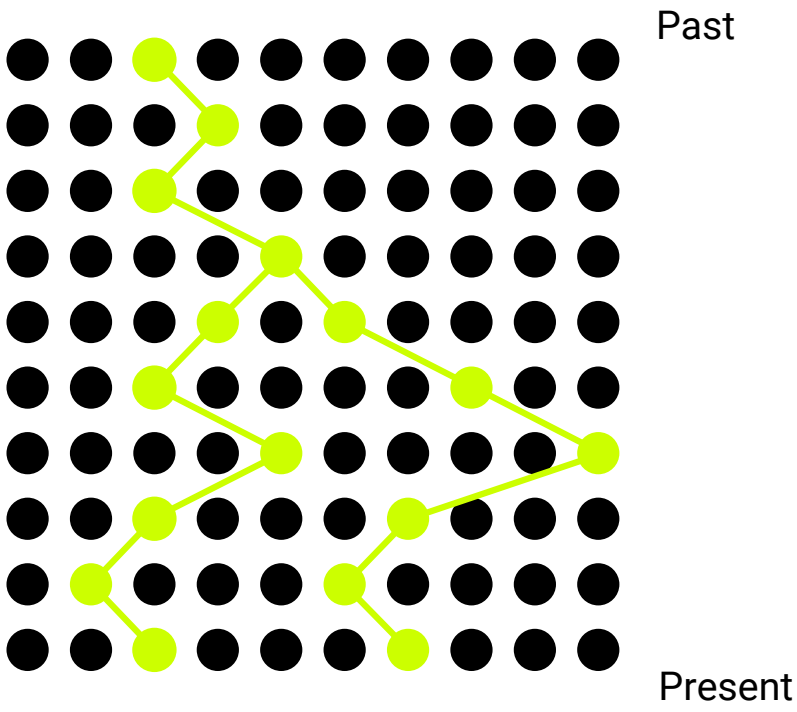
Pedigree



Gene Tree Within a Pedigree

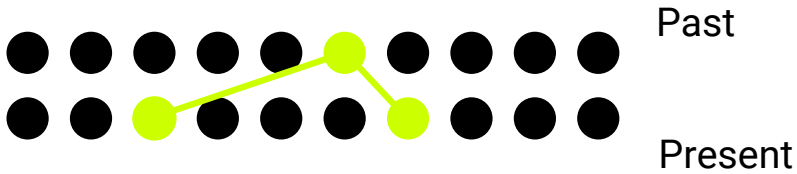


Population Coalescent ($n = 2$)

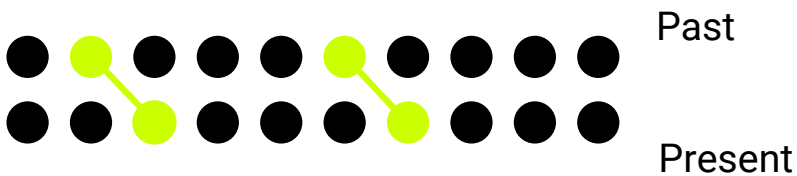


Population Coalescent ($n = 2$)

Probability of common parent ($P_1 = \frac{1}{2N}$):



Probability of distinct parents ($P_2 = 1 - \frac{1}{2N}$):



Population Coalescent ($n = 2$)

Probability no coalescence by generation t

$$P_2^{(t)} = \left(1 - \frac{1}{2N}\right)^t$$

Transform to "diffusion" timescale

$$t = (2N)\tau$$

Unit of time is now $2N$ generations. One generation on this timescale is

$$d\tau = \frac{1}{2N}$$

so as $N \to \infty$ time appears continuous.

Population Coalescent ($n = 2$)

Probability of no coalescence (on the transformed timescale) is

$$P_2^{(\tau)} = \left(1 - \frac{1}{2N}\right)^{\tau(2N)}$$

and for large population size this converges to

$$\lim_{N \to \infty} \left(1 - \frac{1}{2N}\right)^{\tau(2N)} = \mathrm{e}^{-\tau}$$

Expectation of Coalescence Time ($n = 2$)

On the original discrete generation timescale the expected time for 2 sequences to coalesce is

$$\begin{aligned}\mathbb{E}(t) &= \sum_{t=0}^{\infty} t P_2^{(t)} \\ &= \sum_{t=1}^{\infty} \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} = 2N.\end{aligned}$$

On the transformed continuous timescale the expected time to coalescence is

$$\mathbb{E}(\tau) = \int P_2^{(\tau)} d\tau = \int_0^{\infty} \tau \mathrm{e}^{-\tau} d\tau = 1.$$

Distribution of TMRCA ($n = 2$)

Recall: an exponentially distributed random variable x with rate λ has probability density function

$$f(x) = \lambda \mathrm{e}^{-\lambda x},$$

with mean (expectation)

$$\mathbb{E}(x) = 1/\lambda,$$

and variance

$$\mathrm{Var}(x) = 1/(\lambda^2).$$

Distribution of TMRCA ($n = 2$)

The probability density function of the coalescence time for two sequences on the transformed timescale ($2N$ generations) is

$$f(\tau) = \mathrm{e}^{-\tau},$$

which is an exponential distribution with $\lambda = 1$. The mean and variance are

$$\mathbb{E}(\tau) = 1,$$

and

$$\mathrm{Var}(\tau) = 1.$$

Distribution of TMRCA ($n = 2$)

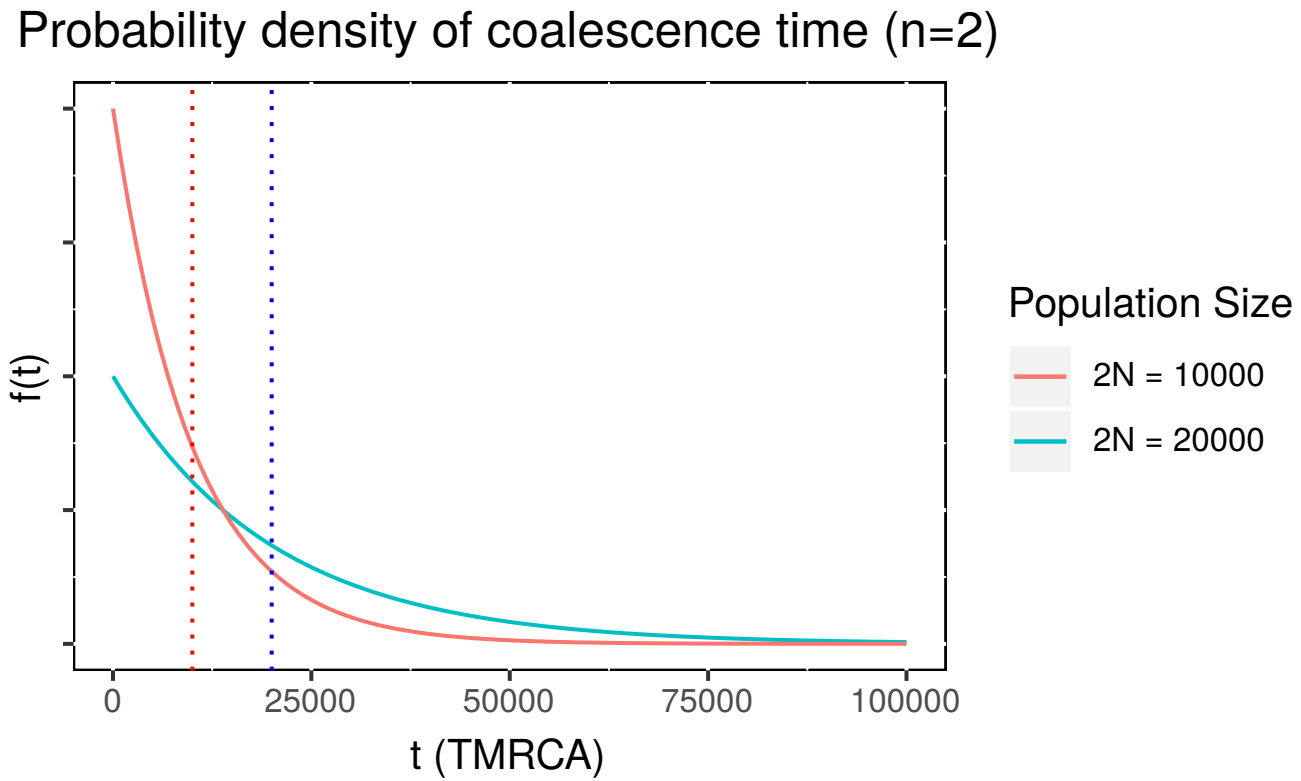
The mean and variance of the coalescence time for two sequences on the original timescale (generations) are

$$\begin{aligned} \mathbb{E}(t) &= \mathbb{E}[(2N)\tau] \\ &= (2N)\mathbb{E}(\tau) \\ &= 2N, \end{aligned}$$

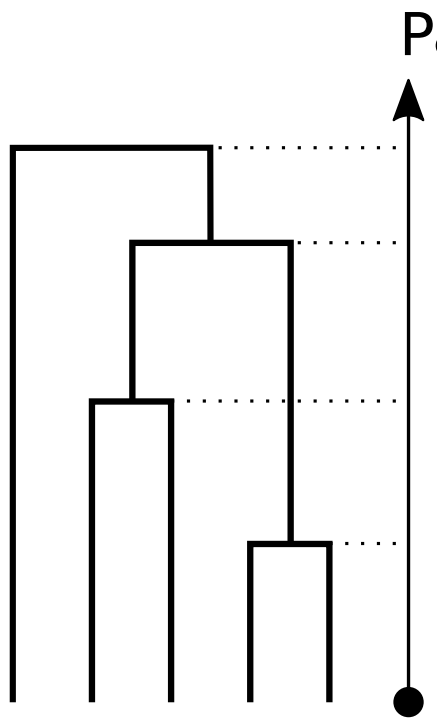
and

$$\begin{aligned} \mathrm{Var}(t) &= \mathrm{Var}[(2N)\tau] \\ &= (2N)^2\mathrm{Var}(\tau) \\ &= 4N^2. \end{aligned}$$

Distribution of TMRCA ($n = 2$)



Population Coalescent ($n \geq 2$)



Waiting times

$T_2 = t_2 - t_3$

$T_3 = t_3 - t_4$

$T_4 = t_4 - t_5$

$T_5 = t_5$

Population Coalescent ($n \geq 2$)

Waiting time for i lineages to coalesce to $i-1$ lineages

$$f(\tau_i) = \frac{i(i-1)}{2} \mathrm{e}^{-\frac{i(i-1)}{2}\tau_i}$$

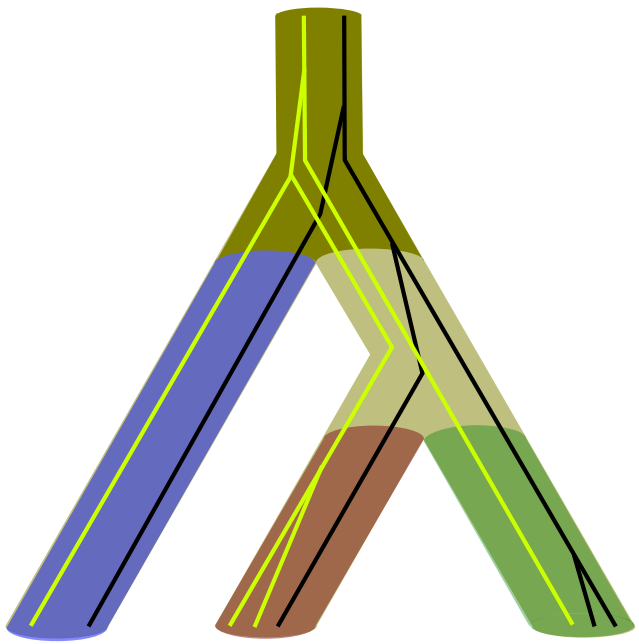
This is an exponential distribution with rate parameter

$$\frac{i(i-1)}{2}$$

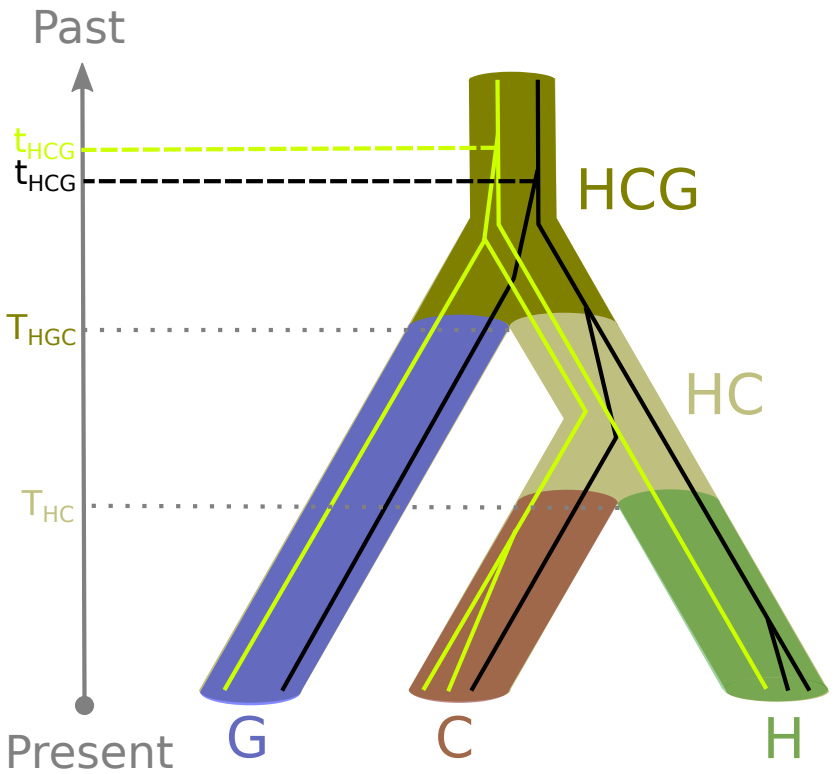
The mean and variance are therefore

$$\mathbb{E}(\tau_i) = \frac{2}{i(i-1)}, \ \mathrm{Var}(\tau_i) = \frac{4}{i^2(i-1)^2}.$$

Gene Trees Within Species Trees

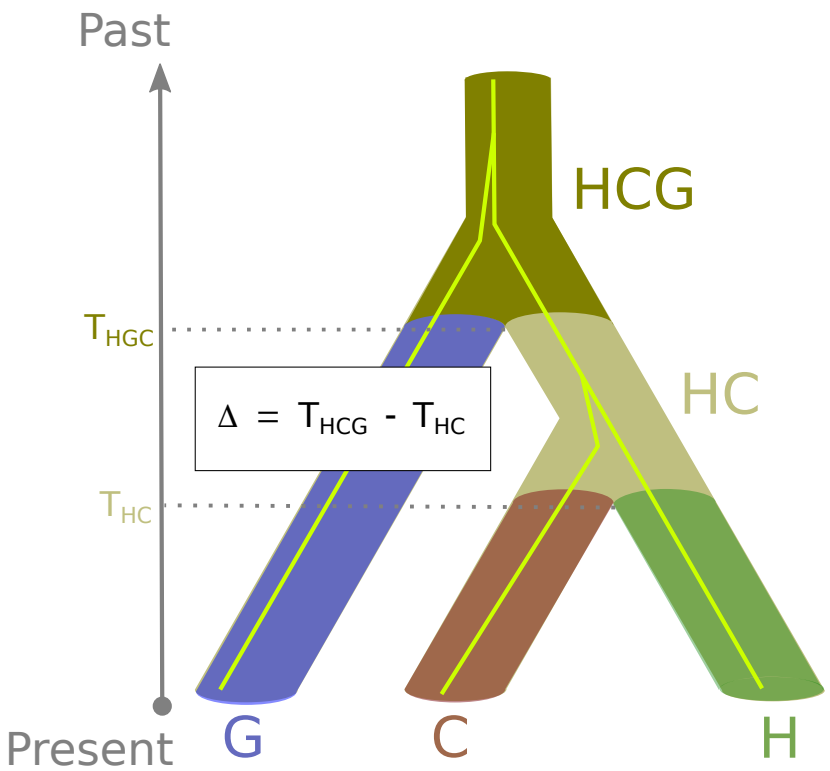


Gene Trees Within Species Trees



Gene Tree Probabilities Within Species Trees

3 species: 1 sequence each



Probability H and C coalesce in HC

$$\int_0^{\Delta} \frac{e^{-\frac{x}{2N_{HC}}}}{2N_{HC}} dx = 1 - e^{-\frac{\Delta}{2N_{HC}}}.$$

Probability H and C do not coalesce in HC

$$e^{-\frac{\Delta}{2N_{HC}}}.$$

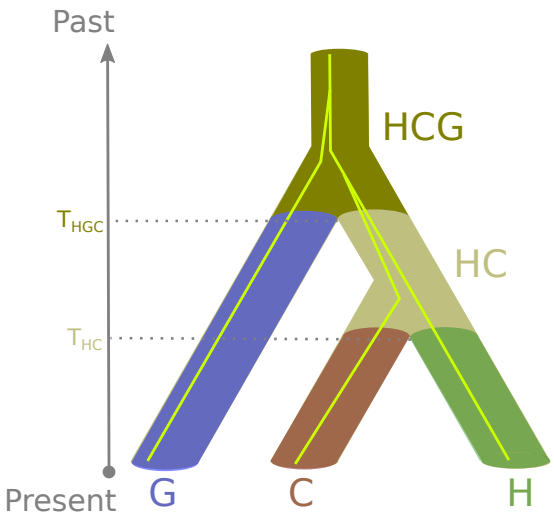
Gene Tree Probabilities Within Species Trees

3 species: 1 sequence each

Species Tree = Gene Tree

Pr = 1/3

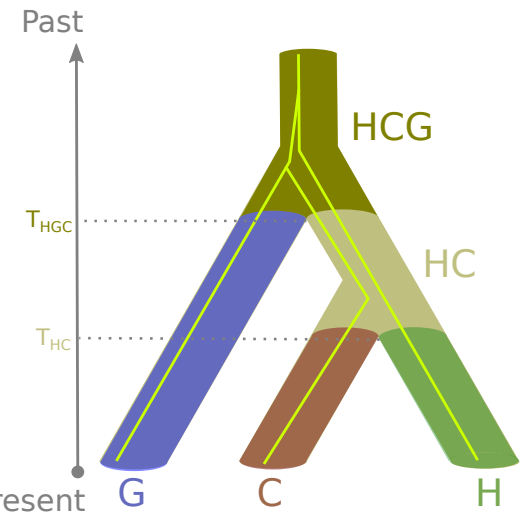
1/3



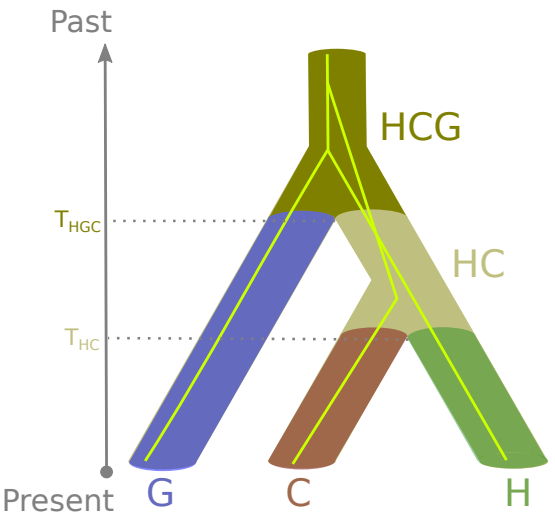
Species Tree != Gene Tree

Pr = 1/3 + 1/3 = 2/3

1/3



1/3



Gene Tree Probabilities Within Species Trees

3 species: 1 sequence each

$$\begin{aligned}\Pr(G = S) &= 1 - \exp\left(\frac{-\Delta}{2N_{HC}}\right) + \frac{1}{3}\exp\left(\frac{-\Delta}{2N_{HC}}\right) \\ &= 1 - \frac{2}{3}\exp\left(\frac{-\Delta}{2N_{HC}}\right)\end{aligned}$$

$$\Pr(G \neq S) = \frac{2}{3}\exp\left(\frac{-\Delta}{2N_{HC}}\right)$$

Estimator of Ancestral N (Chen and Li, 2001)

Procedure: estimate gene trees from sequence data and check match with a known species tree. The expected proportion of matches is

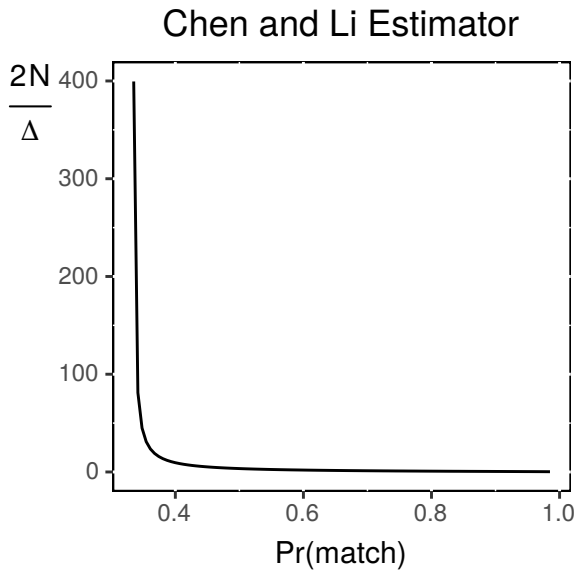
$$P = 1 - \frac{2}{3}\exp\left(\frac{-\Delta}{2N}\right)$$

Solving for $2N/\Delta$ gives the estimator:

$$\frac{2N}{\Delta} = \frac{1}{-\log(3/2) - \log(1-P)}.$$

Estimator of Ancestral N (Chen and Li, 2001)

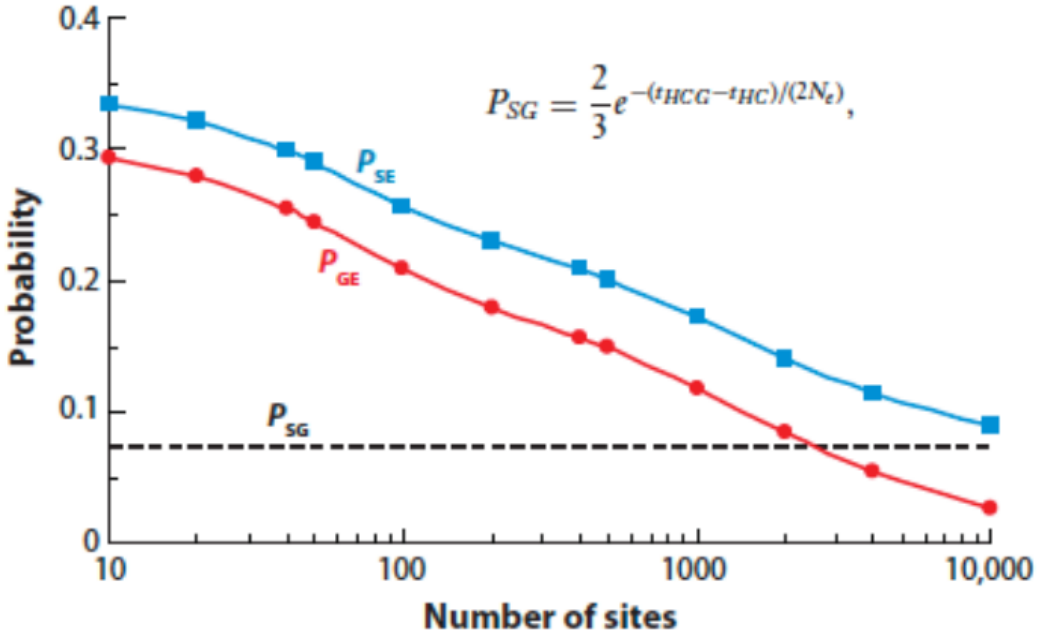
Population Size Versus Match Probability



Example: Let $\Delta_{HC} = 4 \times 10^6$ years (2×10^5 generations if $g = 20$) and $N_{HC} = 10^5$.

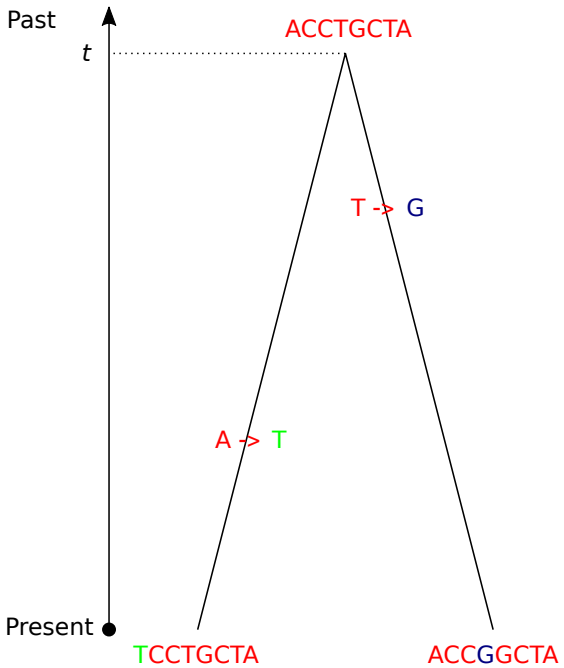
$$\begin{aligned} P &= 1 - \frac{2}{3}\exp\left(-\frac{200000}{2 \times 100000}\right) \\ &= 1 - \frac{2}{3}\exp(-1) \\ &= 1 - \frac{2}{3} \times 0.368 = 0.755 \end{aligned}$$

Estimator of Ancestral N (Chen and Li, 2001)



What is θ?

two sequences



Number of segregating sites under infinite-sites model (Watterson, 1981)

$$\begin{aligned}\mathbb{E}(S) &= \mathbb{E}(2\mu t) \\ &= 2\mu\mathbb{E}(t) \\ &= 2\mu(2N) \\ &= 4N\mu\end{aligned}$$

We define

$$\theta = 4N\mu$$