Species Tree Inference

Bruce Rannala @ UC Davis

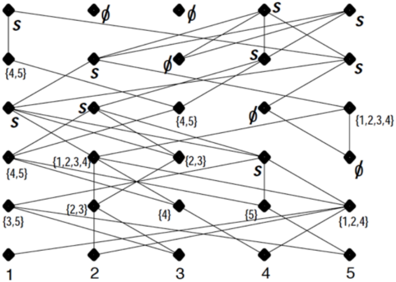
Outline

Coalescent Theory

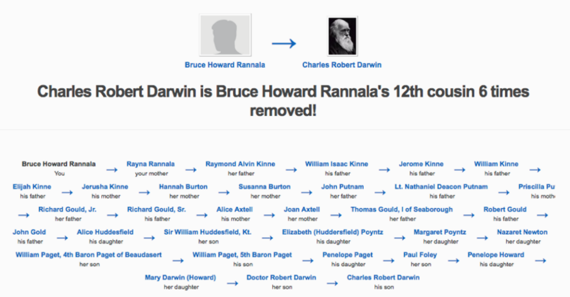
Multispecies Coalescent

Phylodemographic Inference

Pedigree



Pedigree

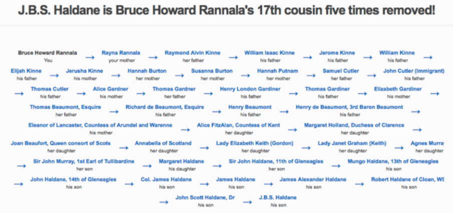


Bruce Howard Rannala → Charles Robert Darwin

Charles Robert Darwin is Bruce Howard Rannala's 12th cousin 6 times removed!

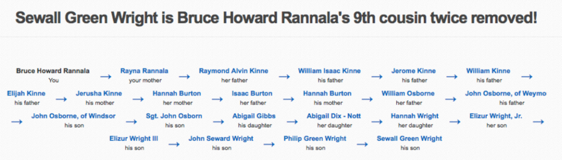
Bruce Howard Rannala You → Rayna Rannala your mother → Raymond Alvin Kinne her father → William Isaac Kinne his father → Jerome Kinne his father → William Kinne his father →
Elijah Kinne his father → Jerusha Kinne his mother → Hannah Burton her mother → Susanna Burton her mother → John Putnam her father → Lt. Nathaniel Deacon Putnam his father → Priscilla Pu his moth
→ Richard Gould, Jr. her father → Richard Gould, Sr. his father → Alice Axtell his mother → Joan Axtell her mother → Thomas Gould, I of Seaborough her father → Robert Gould his father →
John Gold his father → Alice Huddesfield his daughter → Sir William Huddesfield, Kt. her son → Elizabeth (Huddersfield) Poyntz his daughter → Margaret Poyntz her daughter → Nazaret Newton her daughter
William Paget, 4th Baron Paget of Beaudasert her son → William Paget, 5th Baron Paget his son → Penelope Paget his daughter → Paul Foley her son → Penelope Howard his daughter →
Mary Darwin (Howard) her daughter → Doctor Robert Darwin her son → Charles Robert Darwin his son

Pedigree



J.B.S. Haldane is Bruce Howard Rannala's 17th cousin five times removed!

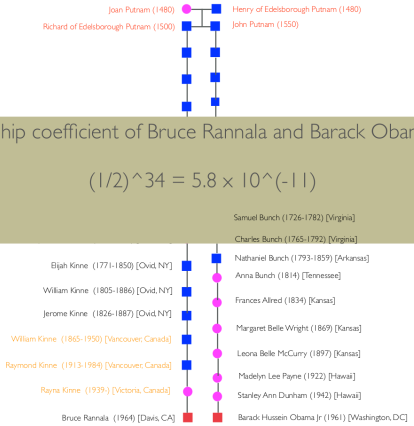
Bruce Howard Rannala You → Rayna Rannala your mother → Raymond Alvin Kinne her father → William Isaac Kinne his father → Jerome Kinne his father → William Kinne his father →
Elijah Kinne his father → Jerusha Kinne his mother → Hannah Burton her mother → Susanna Burton her mother → Hannah Putnam her mother → Samuel Cutler her father → John Cutler (Immigrant) his father
→ Thomas Cutler his father → Alice Gardner his mother → Thomas Gardner her father → Henry London Gardiner his father → Thomas Gardiner his father → Elizabeth Gardiner his mother →
Thomas Beaumont, Esquire her father → Richard de Beaumont, Esquire his father → Henry Beaumont his father → Henry de Beaumont, 3rd Baron Beaumont his father →
Eleanor of Lancaster, Countess of Arundel and Warenne his mother → Alice FitzAlan, Countess of Kent her daughter → Margaret Holland, Duchess of Clarence her daughter →
Joan Beaufort, Queen consort of Scots her daughter → Annabella of Scotland her daughter → Lady Elizabeth Keith (Gordon) her daughter → Lady Janet Graham (Keith) her daughter → Agnes Murra her daughter
→ Sir John Murray, 1st Earl of Tullibardine her son → Margaret Haldane his daughter → Sir John Haldane, 11th of Gleneagles her son → Mungo Haldane, 13th of Gleneagles his son
→ John Haldane, 14th of Gleneagles his son → Col. James Haldane his son → James Haldane his son → James Alexander Haldane his son → Robert Haldane of Cloan, WS his son
→ John Scott Haldane, Dr his son → J.B.S. Haldane his son



Sewall Green Wright is Bruce Howard Rannala's 9th cousin twice removed!

Bruce Howard Rannala You → Rayna Rannala your mother → Raymond Alvin Kinne her father → William Isaac Kinne his father → Jerome Kinne his father → William Kinne his father →
Elijah Kinne his father → Jerusha Kinne his mother → Hannah Burton her mother → Isaac Burton her father → Hannah Burton his mother → William Osborne her father → John Osborne, of Weymo his father
→ John Osborne, of Windsor his son → Sgt. John Osborn his son → Abigail Gibbs his daughter → Abigail Dix - Nott her daughter → Hannah Wright her daughter → Elizur Wright, Jr. her son →
Elizur Wright III his son → John Seward Wright his son → Philip Green Wright his son → Sewall Green Wright his son

Pedigree



Joan Putnam (1480)

Henry of Edelsborough Putnam (1480)

Richard of Edelsborough Putnam (1500)

John Putnam (1550)

Kinship coefficient of Bruce Rannala and Barack Obama:

$$(1/2)^{\wedge}34 = 5.8 \times 10^{\wedge}(-11)$$

Samuel Bunch (1726-1782) [Virginia]

Charles Bunch (1765-1792) [Virginia]

Elijah Kinne (1771-1850) [Ovid, NY]

Nathaniel Bunch (1793-1859) [Arkansas]

Anna Bunch (1814) [Tennessee]

William Kinne (1805-1886) [Ovid, NY]

Frances Allred (1834) [Kansas]

Jerome Kinne (1826-1887) [Ovid, NY]

Margaret Belle Wright (1869) [Kansas]

William Kinne (1865-1950) [Vancouver, Canada]

Leona Belle McCurry (1897) [Kansas]

Raymond Kinne (1913-1984) [Vancouver, Canada]

Madelyn Lee Payne (1922) [Hawaii]

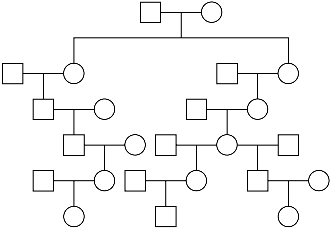
Rayna Kinne (1939-) [Victoria, Canada]

Stanley Ann Dunham (1942) [Hawaii]

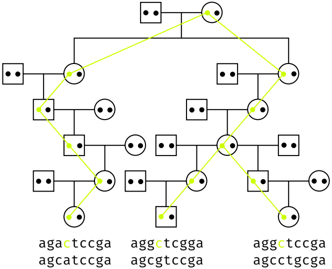
Bruce Rannala (1964) [Davis, CA]

Barack Hussein Obama Jr (1961) [Washington, DC]

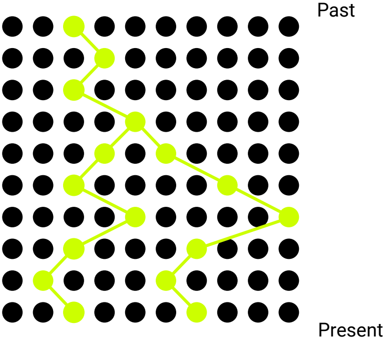
Pedigree



Gene Tree Within a Pedigree

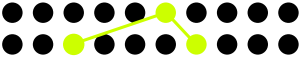


Population Coalescent ($n = 2$)



Population Coalescent ($n = 2$)

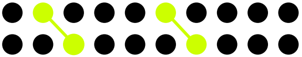
Probability of common parent ($P_1 = \frac{1}{2N}$):



Past

Present

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



Past

Present

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 ($n = 2$)

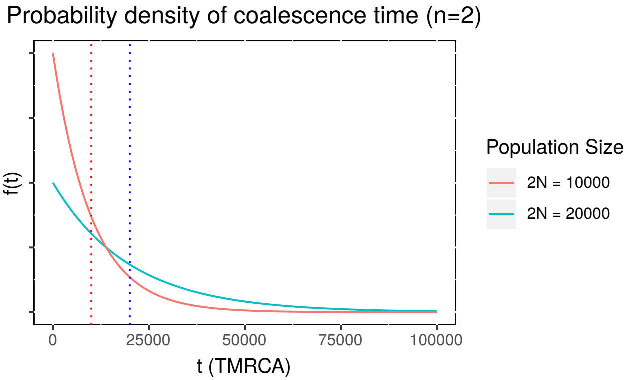
The mean and variance of the coalescence time for two sequences on the original timescale (generations) are

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

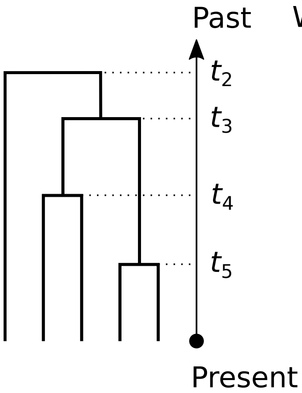
and

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

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

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

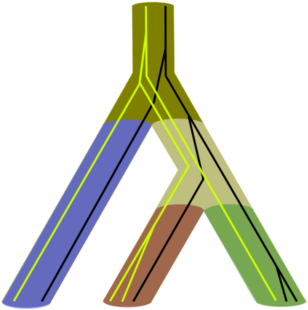
This is an exponential distribution with rate parameter

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

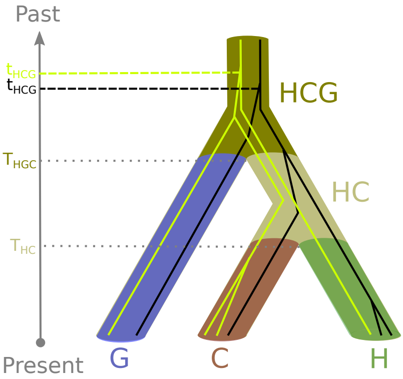
The mean and variance are therefore

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

Gene Trees Within Species Trees

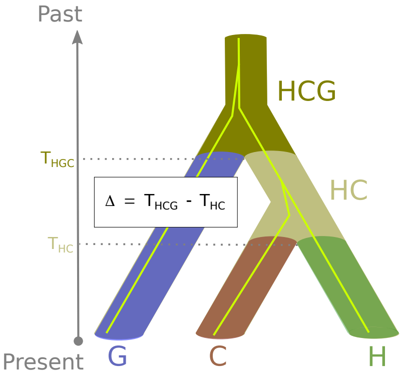


Gene Trees Within Species Trees



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

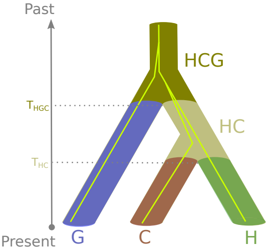
Species Tree = Gene Tree

Pr = 1/3

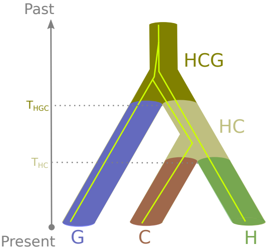
Species Tree != Gene Tree

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

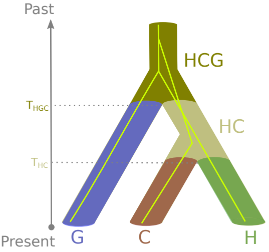
1/3



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

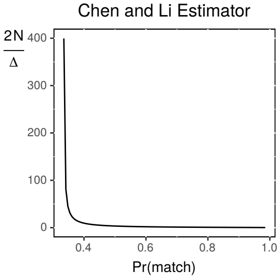
$$P = 1 - \frac{2}{3} \exp\left(\frac{-\Delta}{2N}\right)$$

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)

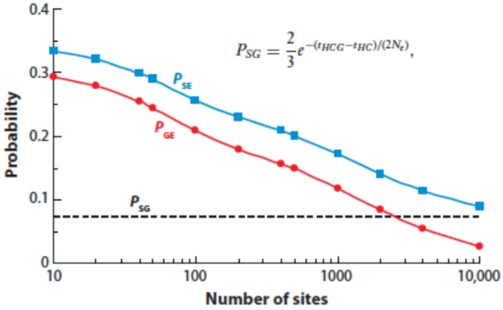
Population Size Versus Match Probability



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

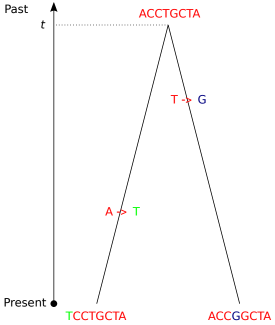
$$\begin{aligned} P &= 1 - \frac{2}{3}\exp\left(-\frac{200000}{2 \times 100000}\right) \\ &= 1 - \frac{2}{3}\exp(-1) \\ &= 1 - \frac{2}{3} \times 0.368 = 0.755 \end{aligned}$$

Estimator of Ancestral N (Chen and Li, 2001)



What is θ?

two sequences



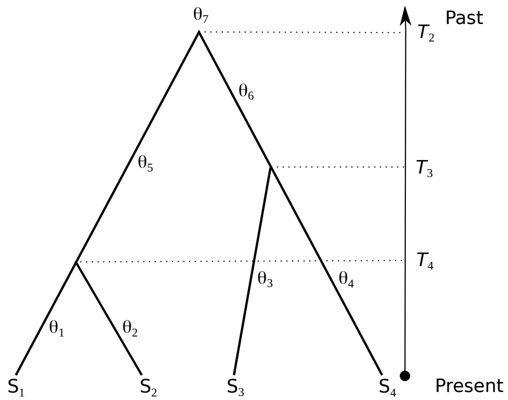
Number of segregating sites under infinite-sites model (Watterson, 1981)

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

We define

$$\theta = 4N\mu$$

Parameters of Phylodemographic Model



Bayesian Phylodemographic Inference

Let $\Theta = \{\theta, \tau\}$. For s species, θ contains at most $2s - 3$ and at least $s - 1$ parameters. τ contains $s - 1$ parameters. The posterior distribution of demographic parameters given sequence data D is

$$f(\Theta|D) = \int \frac{f(D|G)f(G|\Theta)f(\Theta)}{f(D)} dG$$

where $f(D|G)$ is the “Felsenstein Likelihood” and $f(G|\Theta)$ is the “Multispecies Coalescent” prior on gene trees.

Bayesian Phylodemographic Inference

Assumptions

Unlinked genes

$$f(G|\Theta) = \prod_{i=1}^{L} f(G_i|\Theta)$$

Independent sites

$$f(D|G_i) = \prod_{i=1}^{n} f(D_i|G_i)$$

No gene flow between populations.

Bayesian Phylodemographic Inference

Markov chain Monte Carlo

1. Simulate a proposed value for a paraneter

$$\theta^* \approx g(\theta^*|\theta)$$

2. Accept proposed value with probability

$$\alpha = \min\left(\frac{f(D|\theta^*)f(\theta^*)g(\theta|\theta^*)}{f(D|\theta)f(\theta)g(\theta^*|\theta)}, 1\right).$$

Metropolis et al. (1953) + Hastings (1970)

Bayesian Phylodemographic Inference

Metropolis-Hastings Algorithm

Edward Teller (father of the H-bomb)

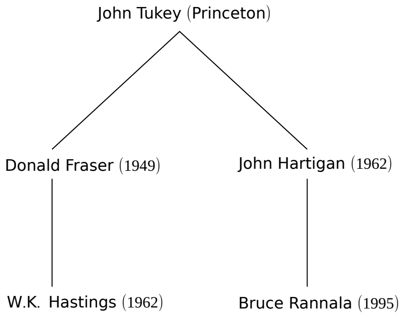
Bayesian Phylodemographic Inference

Metropolis-Hastings Algorithm

Keith Hastings (statistician)

Bayesian Phylodemographic Inference

Metropolis-Hastings Algorithm



Bayesian Phylodemographic Inference

MCMC Proposal Moves

- Propose changes to coalescent times in gene trees that respect constraints of species tree
- Propose changes to gene trees by subtree pruning and re-grafting (respect species tree constraints)
- Propose changes to effective population size parameters
- Propose new speciation times in the species tree and transforming gene trees to respect constraints
- Jointly propose proportional changes to all effective population sizes, divergence, and coalescence times

Bayesian Phylodemographic Inference

Making Sense of BPP Parameters

$$\theta = 4N\mu$$

units are expected DNA substitutions.
To obtain N we specify a mutation rate and generation time

$$N = \frac{\theta}{4 \times \mu \times g}$$

Example: if $\theta_H = 0.00057$, $g = 20$ years/generation and $\mu = 10^{-9}$ mutations/year

$$N_H = \frac{0.00057}{4 \times 10^{-9} \times 20} = 7125$$