Resolving the ASG: A general framework for computing likelihoods under selection

Vivaswat Shastry[1] and Jeremy Berg[2]
[1]Committee on Genetics, Genomics & Systems Biology,
[2]Department of Human Genetics,
University of Chicago, Chicago IL
vivaswat@uchicago.edu

Motivation

- The **Ancestral Selection Graph (ASG)** is a branching-coalescing random graph that contains within it all the possible genealogies of a sample under selection. As a result, this graph **explodes in size with larger samples and stronger selection**.
- But, with the development of methods for **fast & accurate estimation of ARGs** and **numerical Wright-Fisher diffusion**, we can quickly compute likelihoods under a model based on the ASG.

Methods

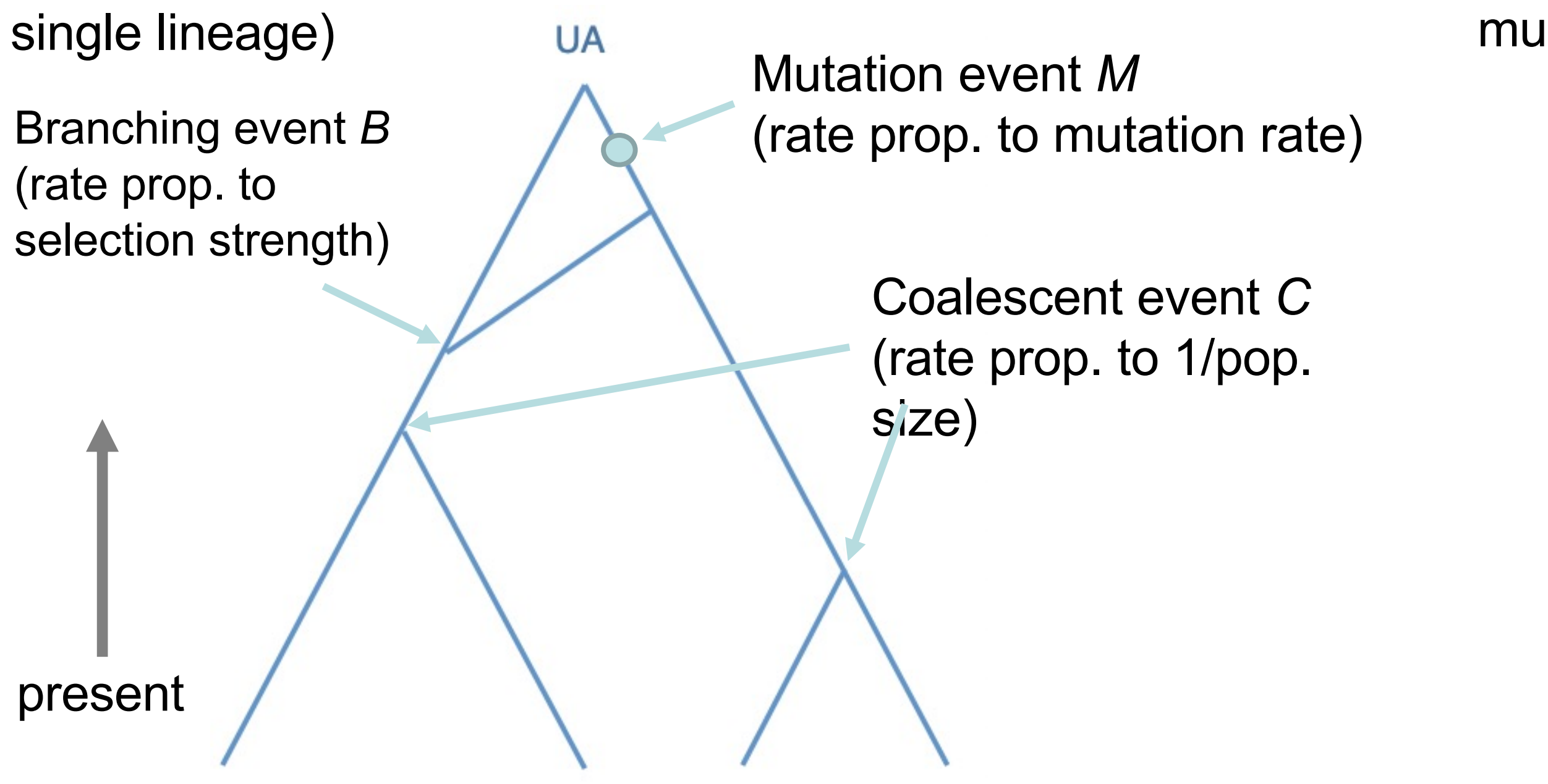
- **Stephens & Donnelley, 2003** present a general method to approximate the posterior distributions of genealogies using **importance sampling** to compute stationary distributions under the ASG.
- Here, we sidestep this expensive scheme by **assuming an infinite sites model** and **replacing their stationary transition probabilities by transitional probabilities conditional on the age of the mutation**.
- Then, we apply this **general maximum-likelihood framework** to estimate selection coefficients given the age of mutation under any demographic history.

Results

- In **simulations with true trees**, our method produces estimates with low error across a range of selection coefficients, on-par with current methods like **CLUES2** (Vaughn *et al*, 2023).

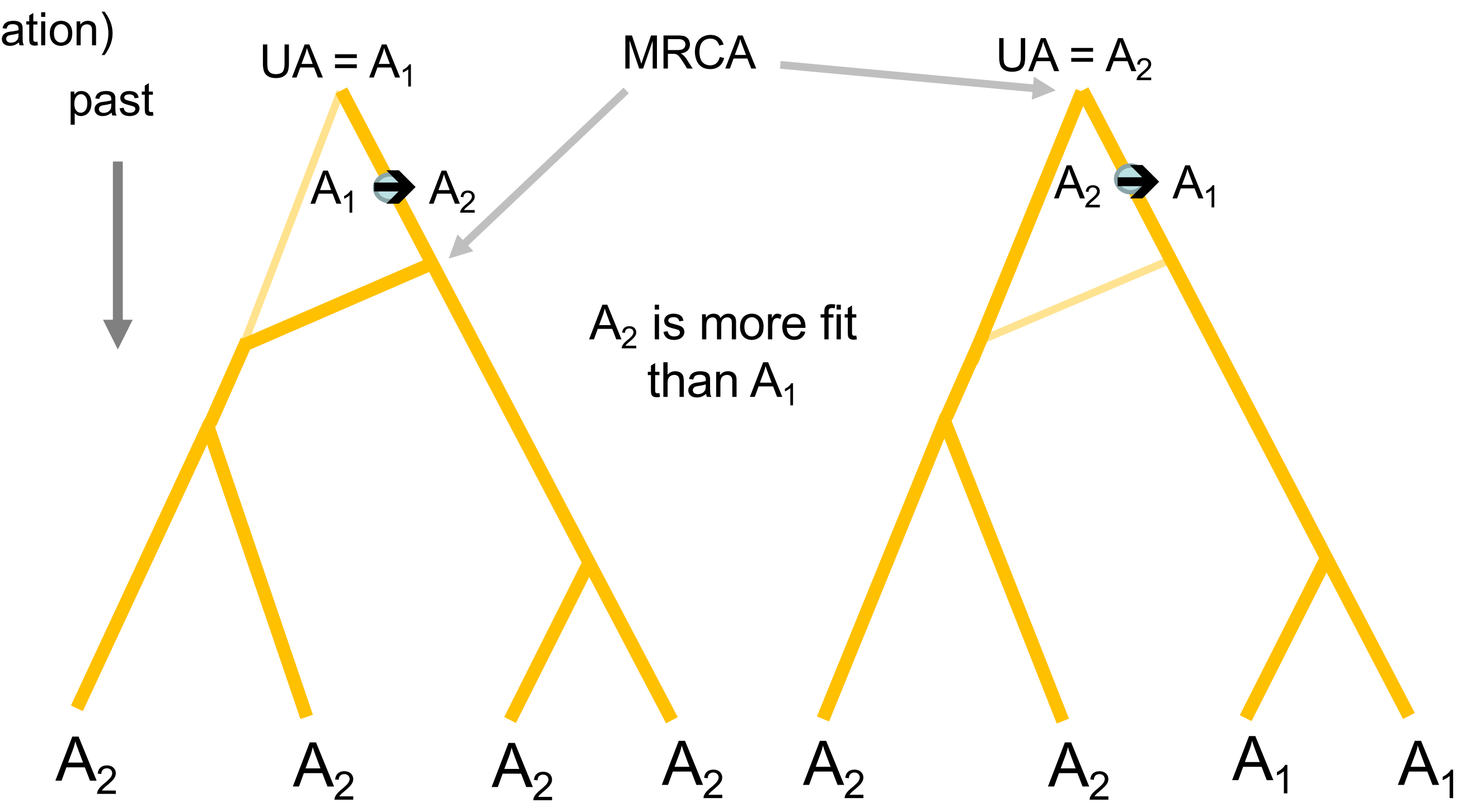
Ancestral Selection Graph (ASG) (Neuhauser & Krone, 1997)

1. Go back in time from *n*=4 *present-day* samples and place branching, coalescent & mutation events (stop when left with single lineage)



Neutral case reduces to the *n*-coalescent with no branching events

2. Follow the multiple paths down from "**U**ltimate **A**ncestor" (UA) to determine the true genealogy of the sample (assume additive fitness with symmetric mutation)



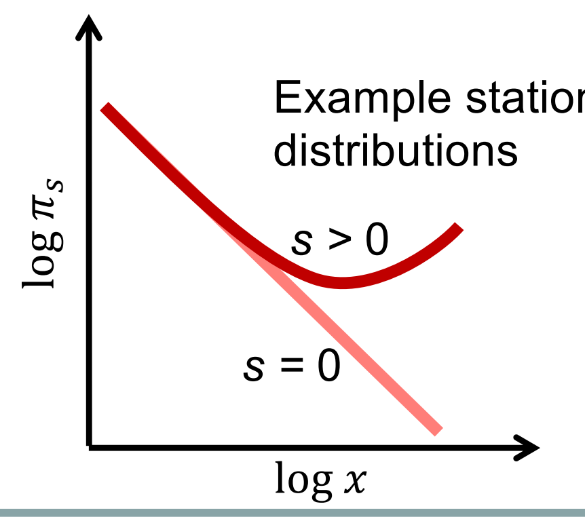
Stephens & Donnelly, 2003 model

Derive the rates of these events as ratios of sampling frequency distributions at stationarity $\pi_s^n(.)$ of a Markovian process.

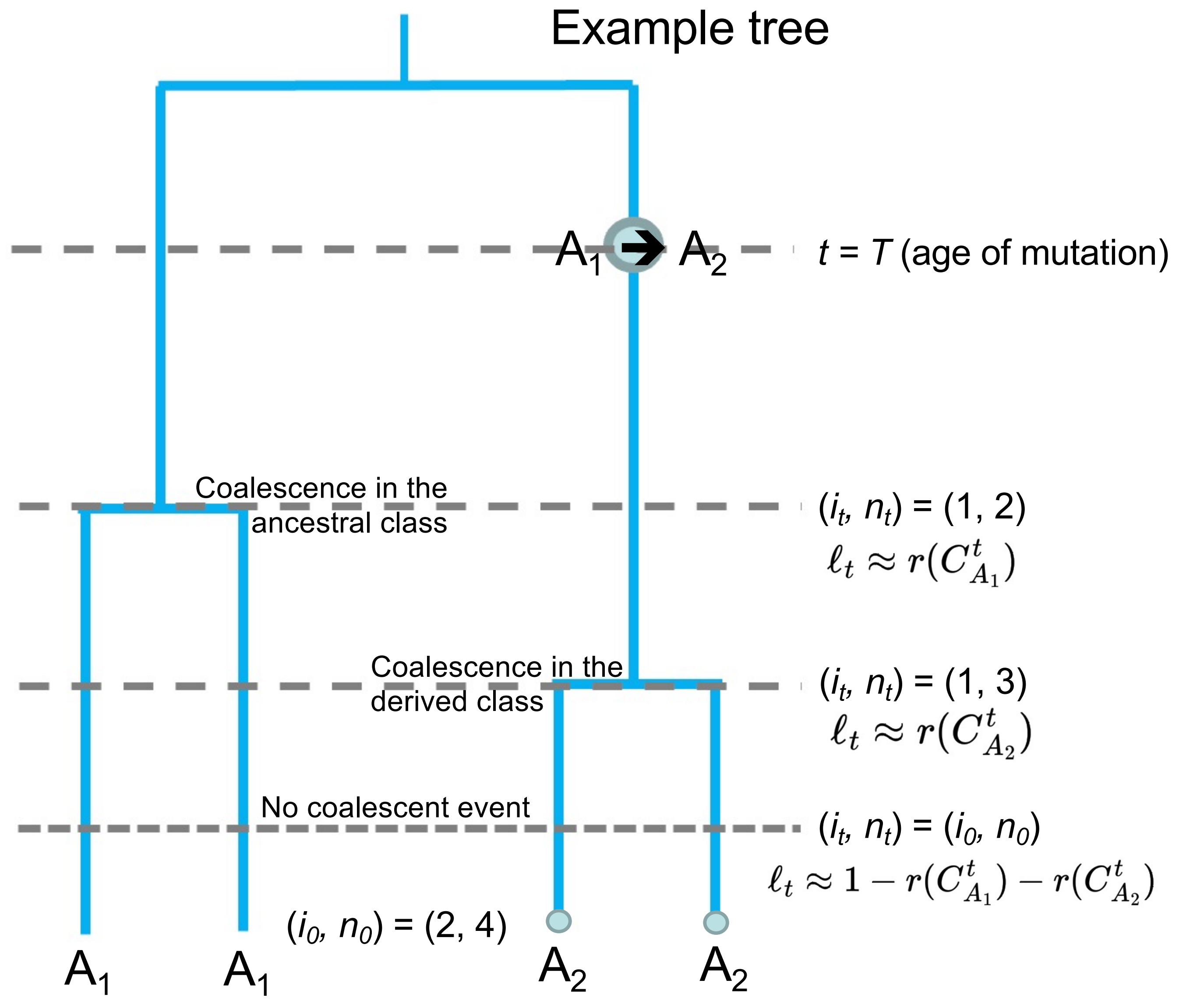
Coalescent rates in the derived and ancestral alleles: $r(C_{A_2}) = \frac{\pi_s^{n-1}(i-1)}{\pi_s^n(i)}$ $r(C_{A_1}) = \frac{\pi_s^{n-1}(n-i-1)}{\pi_s^n(n-i)}$

Two types of branching events: $r(B_1) = Ns\frac{\pi_s^{n+1}(i+1)}{\pi_s^n(i)}$ $r(B_C) = Ns\frac{\pi_s^{n+1}(i)}{\pi_s^n(i)}$

Rate of mutation along a branch:
$r(M) = 2N\mu\frac{\pi_s^n(i-1)}{\pi_s^n(i)}$



Our model


Example tree
A$_1$ → A$_2$
$t = T$ (age of mutation)
Coalescence in the ancestral class
$(i_t, n_t) = (1, 2)$
$\ell_t \approx r(C^t_{A_1})$
Coalescence in the derived class
$(i_t, n_t) = (1, 3)$
$\ell_t \approx r(C^t_{A_2})$
No coalescent event
$(i_t, n_t) = (i_0, n_0)$
$\ell_t \approx 1 - r(C^t_{A_1}) - r(C^t_{A_2})$
$(i_0, n_0) = (2, 4)$
A$_1$
A$_1$
A$_2$
A$_2$

In general, rates are multiplied by the number of opportunities, so if we have i_t derived lineages at generation *t*,

$$\ell_t \approx \binom{i_t}{2} r(C^t_{A_2})$$

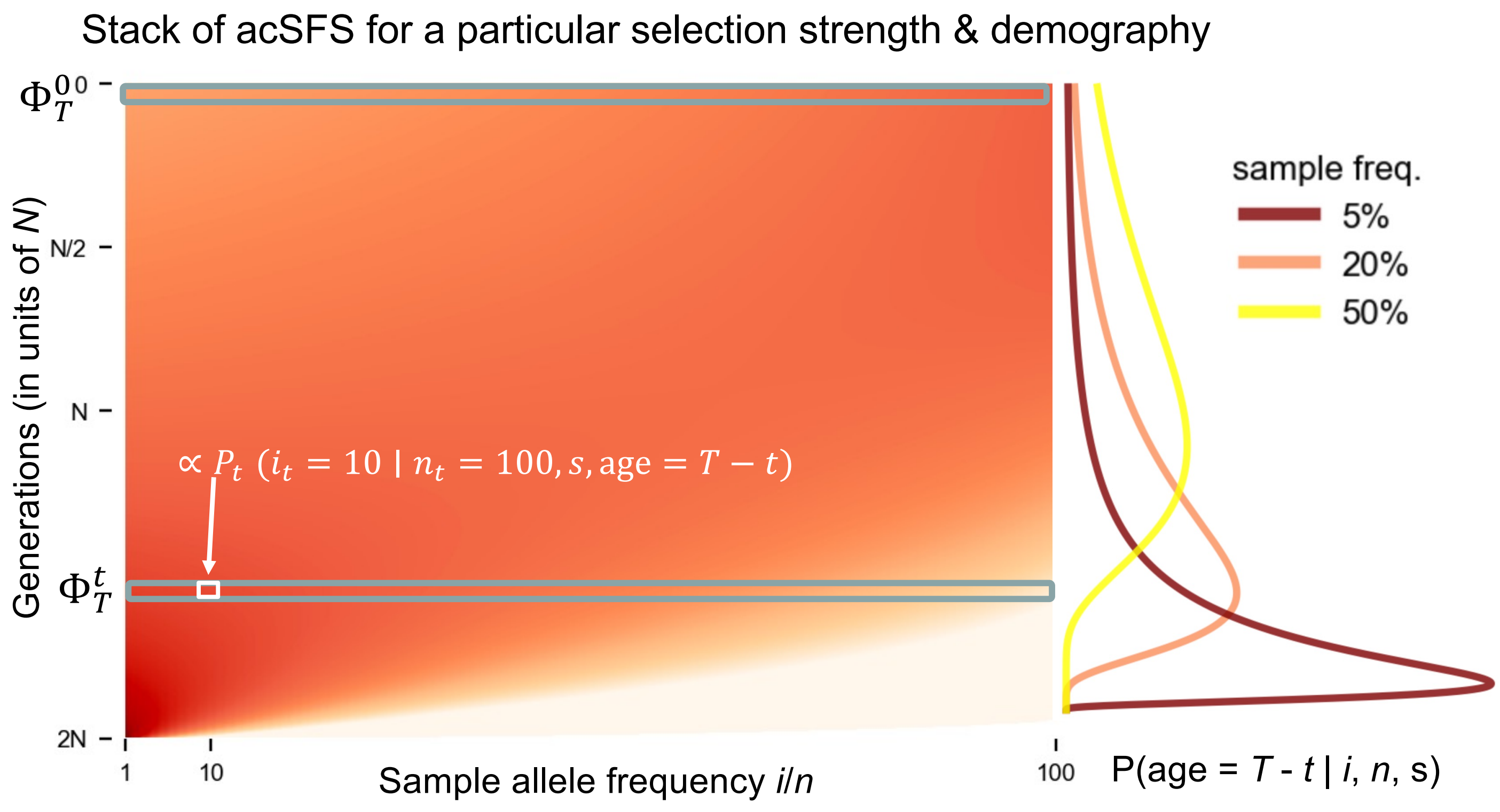
Final likelihood:

$$\boxed{\mathcal{L}(s; \{(i_t, n_t)\}, \text{age} = T) = \prod_{t=0}^{T} \ell_t \times P(\text{age} = T \mid s)}$$

Conditioning on the age of the mutation allows us to calculate per-generation transition probabilities in the sample using forward-in-time Wright-Fisher diffusion (and this also means we can set *r(M) = 0*). We are now left to calculate the coalescent and branching rates. To do this, we use an approach in which we construct an **age-conditioned SFS (acSFS)**.

If Φ_T^T represents the expected SFS of *de-novo* mutations in generation *T* (mass in the singleton bin of $\frac{n\theta}{4N}$, zeros everywhere else), then we can evolve this acSFS forward up to the present day using a probability transition matrix Ξ derived from *moments* (Jouganous *et al*, 2017) that captures the effects of drift and selection.

$$\Phi_T^0 = \Xi^{(1)} \times \ldots \times \Xi^{(T-1)} \times \Phi_T^T$$
$$= \prod_{t=1}^{T-1} \Xi^{(t)} \times \Phi_T^T.$$


Stack of acSFS for a particular selection strength & demography

We can now approximate the stationary distributions in Stephens and Donnelly, 2003 with **appropriate entries in the normalized acSFS** above.

$$\pi_s^{n_t}(i_t)^{(t)} = \frac{\Phi_T^t(i_t)}{\sum_{i'=1}^{n_t-1} \Phi_T^t(i')} = P_t(i_t \mid n_t, s, T)$$

This normalized acSFS is just the **probability of seeing i_t copies out of n_t at time *t* given a selection coefficient *s***. Then, the per-generation rate is:

$$r(C^t_{A_2}) = \frac{P_{t+1}(i_t - 1 \mid n_t - 1, s, T)}{P_t(i_t \mid n_t, s, T)} \times 1/2N_t$$

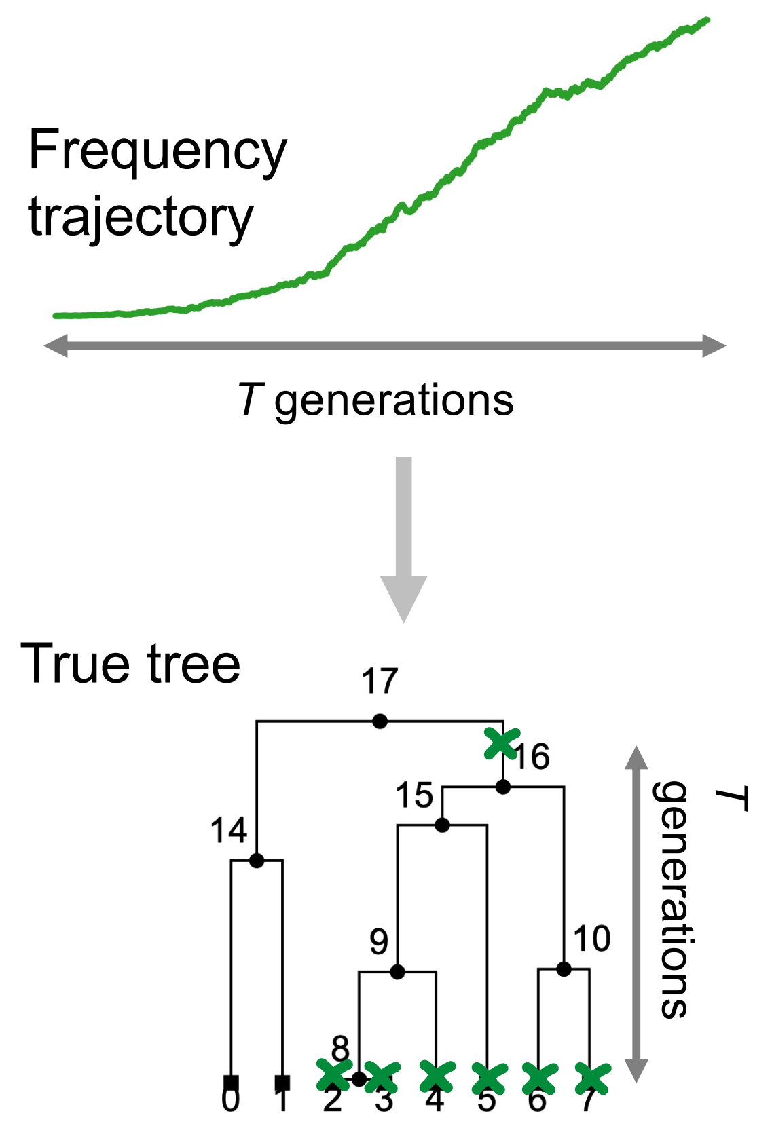
Now, if we have a tree that records the coalescent event between pairs of samples, we show that we do not need to know the branching rate to compute the likelihood for a given value of selection. This calculation essentially **averages over all the possible virtual lineages** at each generation.

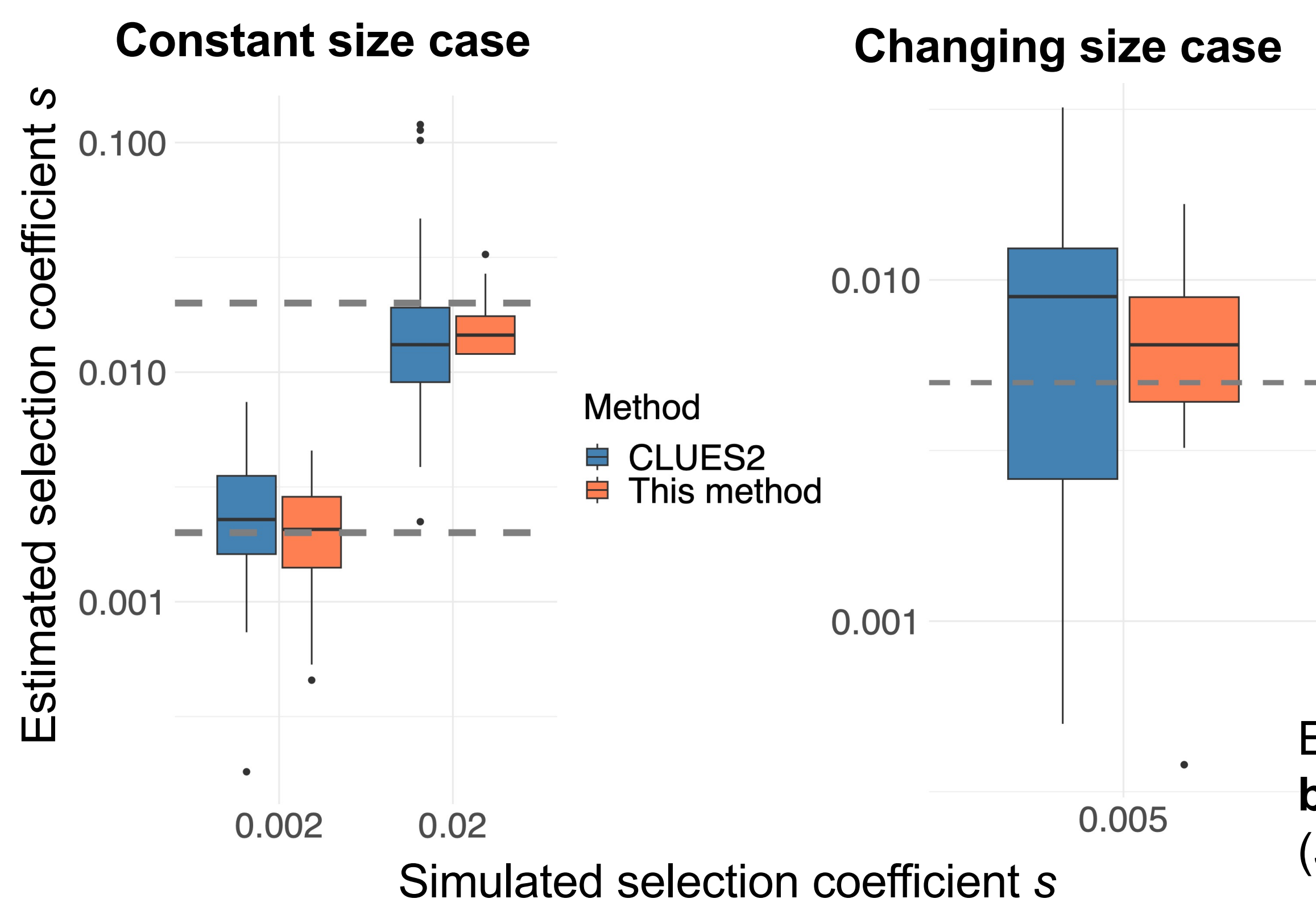
$$\begin{aligned}\text{Prob. of no change} &= \text{prob. of no coalescence} \times \text{prob. of no branching} + \text{prob. of branching}\\ &\approx (1 - (r(C_{A_1}) + r(C_{A_2}))) \times (1 - (r(B_1) + r(B_C))) + (r(B_1) + r(B_C))\\ &= 1 - (r(C_{A_1}) + r(C_{A_2})) - \overset{\approx 0}{\cancel{(r(C_{A_1}) + r(C_{A_2})) \times (r(B_1) + r(B_C))}} - \cancel{(r(B_1) + r(B_C))} + \cancel{(r(B_1) + r(B_C))}\\ &\approx 1 - (r(C_{A_1}) + r(C_{A_2}))\end{aligned}$$

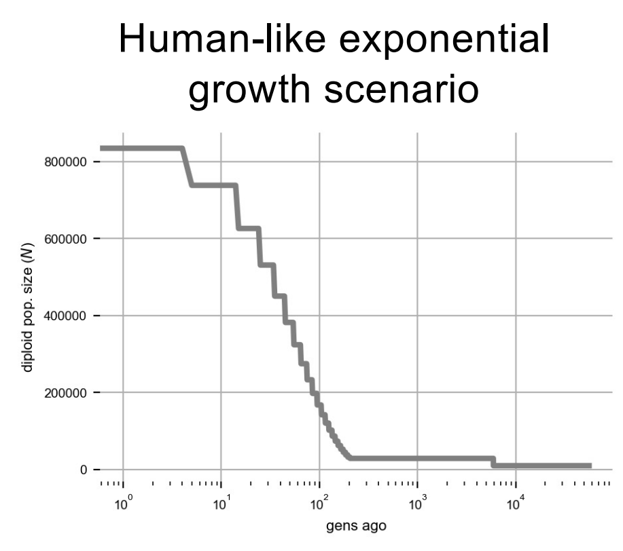
(If these rates $\ll 1$, which is a valid approximation since they're per-generation, we can approximate the probability of an exponentially distributed event with this small rate to be equal to the rate of the event)

Results

Simulations in *mssel* (Hudson, 2002) across 50 replicates conditioned on segregation in a sample of 40 haploids after *T* generations:







Estimates are **downward biased for strong selection** (*s* >= 0.01) with both methods.

Future directions

1. Incorporate importance sampling scheme to account for tree estimation being performed under neutral prior (using ARG-based methods)
   - Selected alleles tend to be **younger** than their neutral counterparts
2. Perform estimation in a complex demographic history (for example, two-population model with migration)

References

- Neuhauser, C., & Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics*, *145*(2), 519-534.
- Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*, *206*(3), 1549-1567.
- Stephens, M., & Donnelly, P. (2003). Ancestral inference in population genetics models with selection (with discussion). *Australian & New Zealand Journal of Statistics*, *45*(4), 395-430.
- Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321-1329.
- Vaughn, A., & Nielsen, R. (2023). Fast and accurate estimation of selection coefficients and allele histories from ancient and modern DNA. *bioRxiv*, 2023-12.
- Hudson, R. R. (2002). ms a program for generating samples under neutral models. *Bioinformatics*, *18*(2), 337-338.