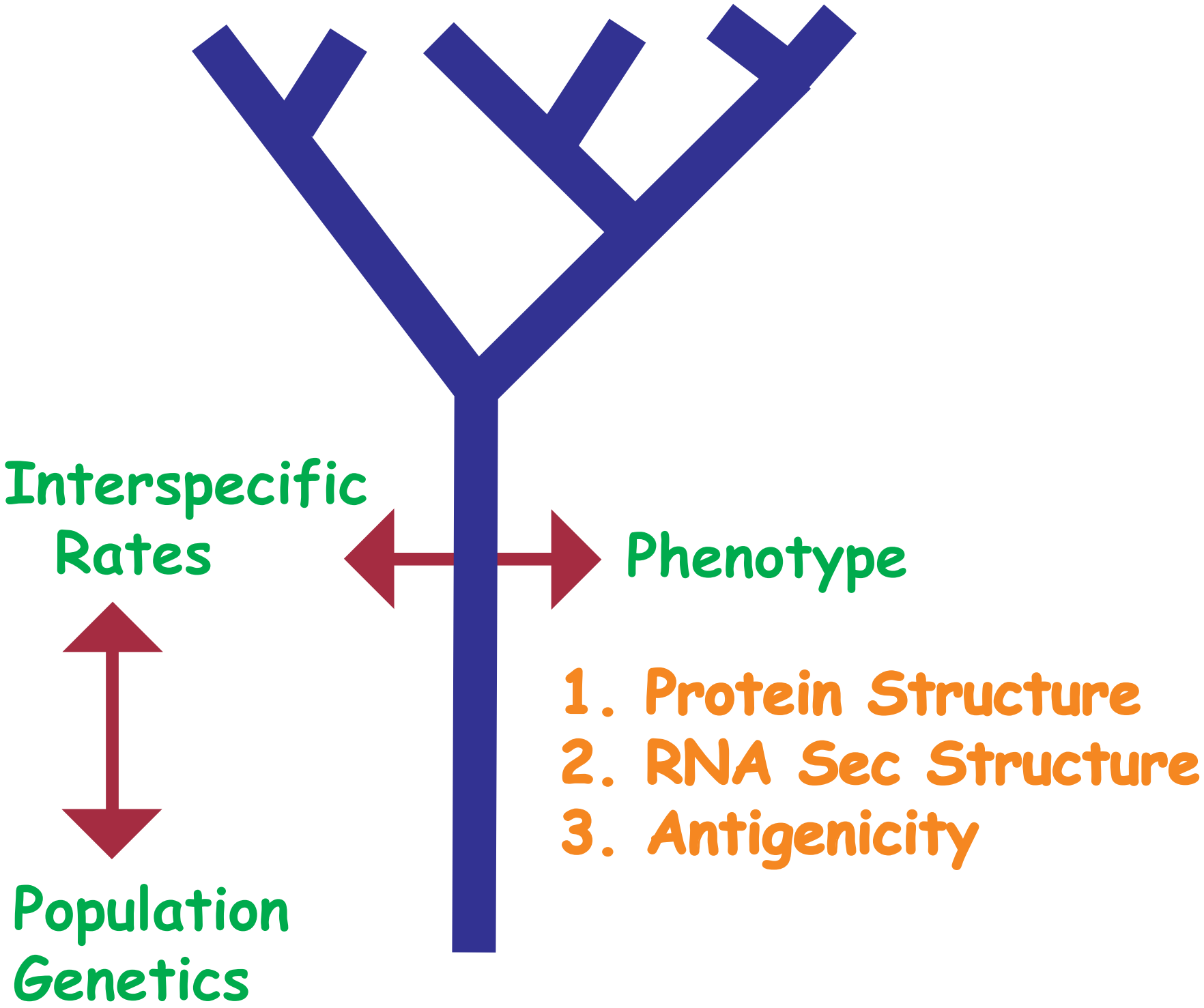
Adding Phenotype and Population Genetics to Interspecific Evolution

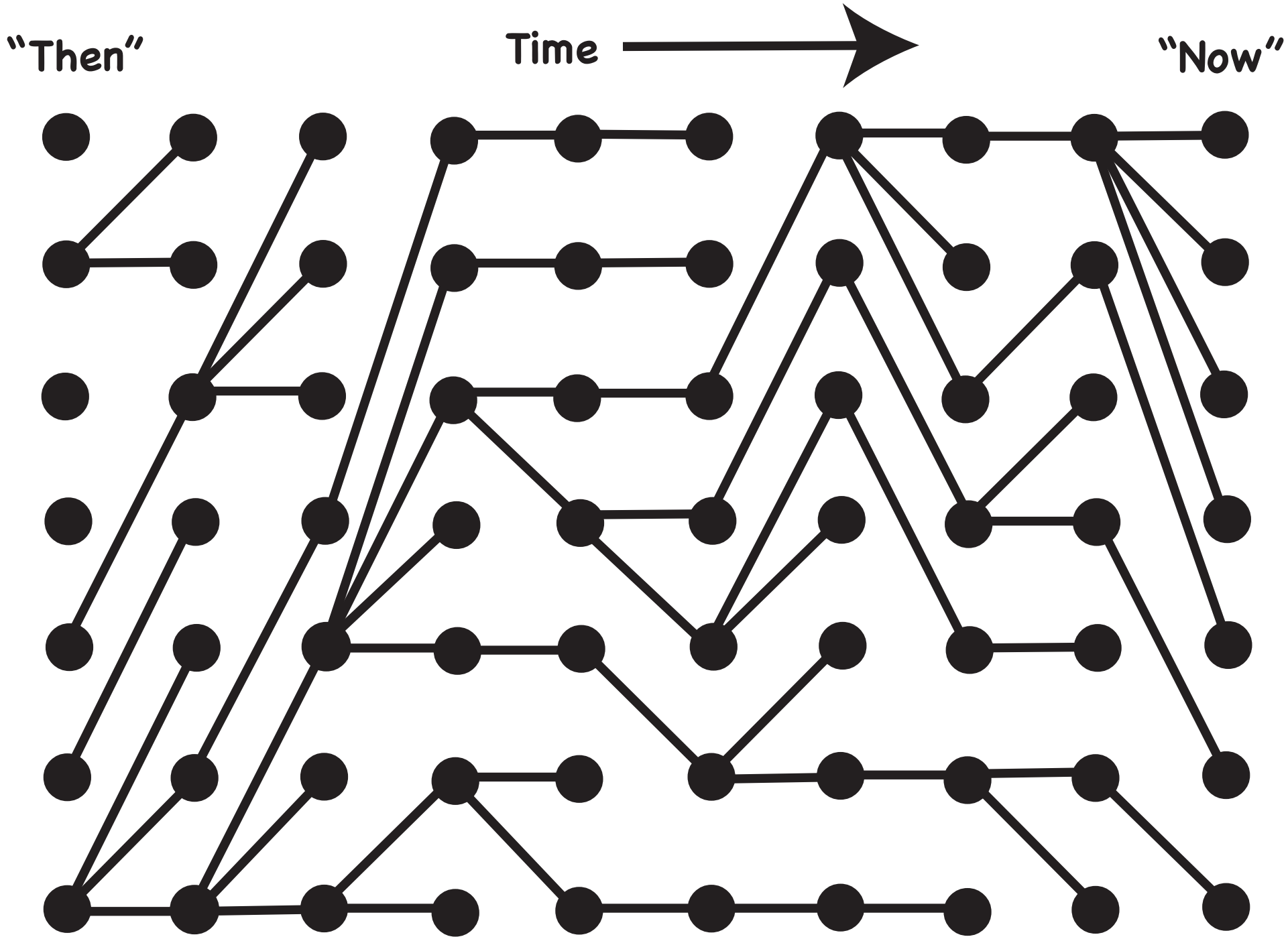
From North Carolina State University:

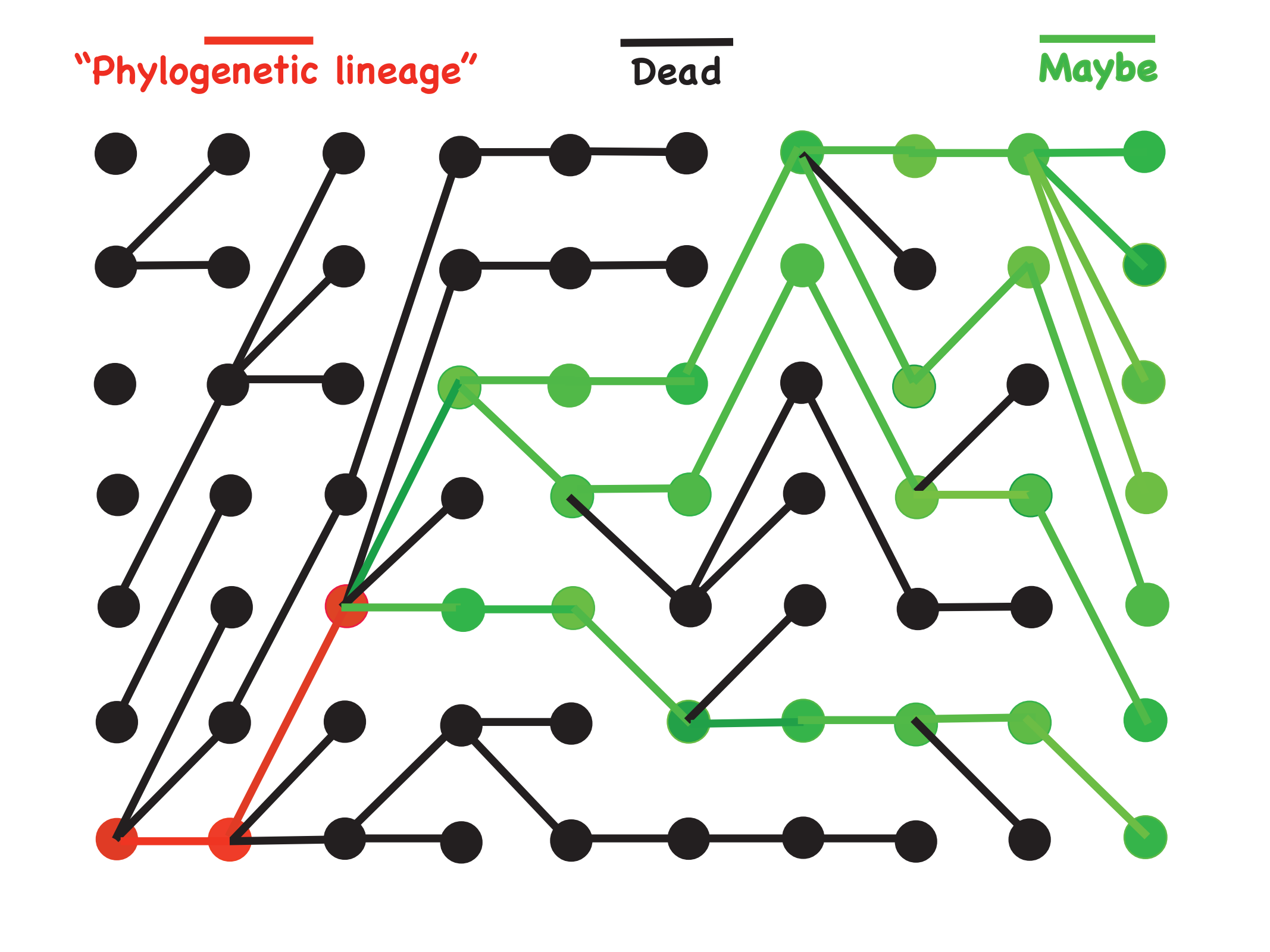
Sang Chul Choi, Benjamin Redelings, Reed Cartwright, Eric Stone, Jeff Thorne

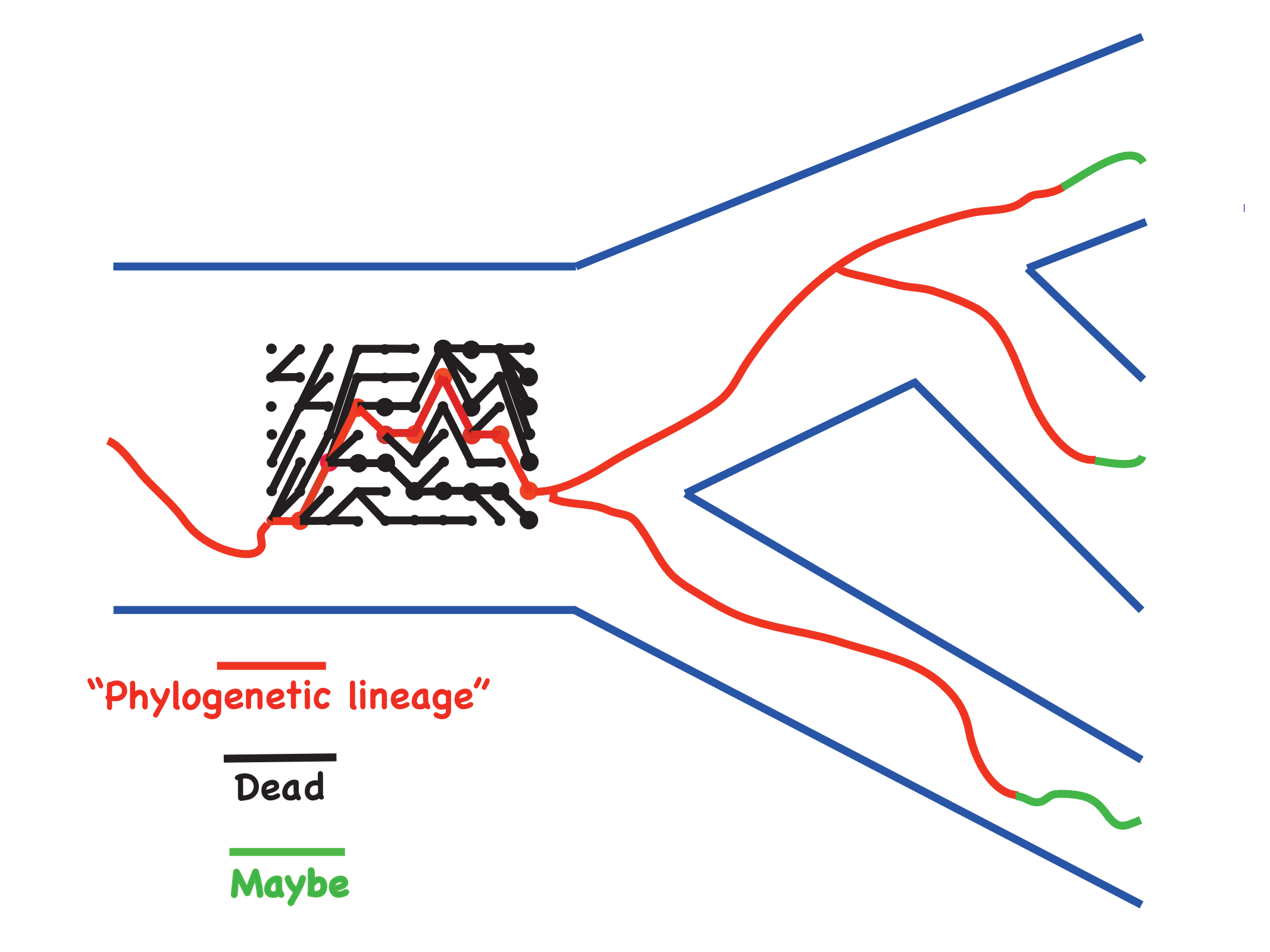
From University of Tokyo:

Hirohisa Kishino


Interspecific Rates
Phenotype
1. Protein Structure
2. RNA Sec Structure
3. Antigenicity
Population Genetics


"Then"
Time
"Now"


"Phylogenetic lineage"
Dead
Maybe

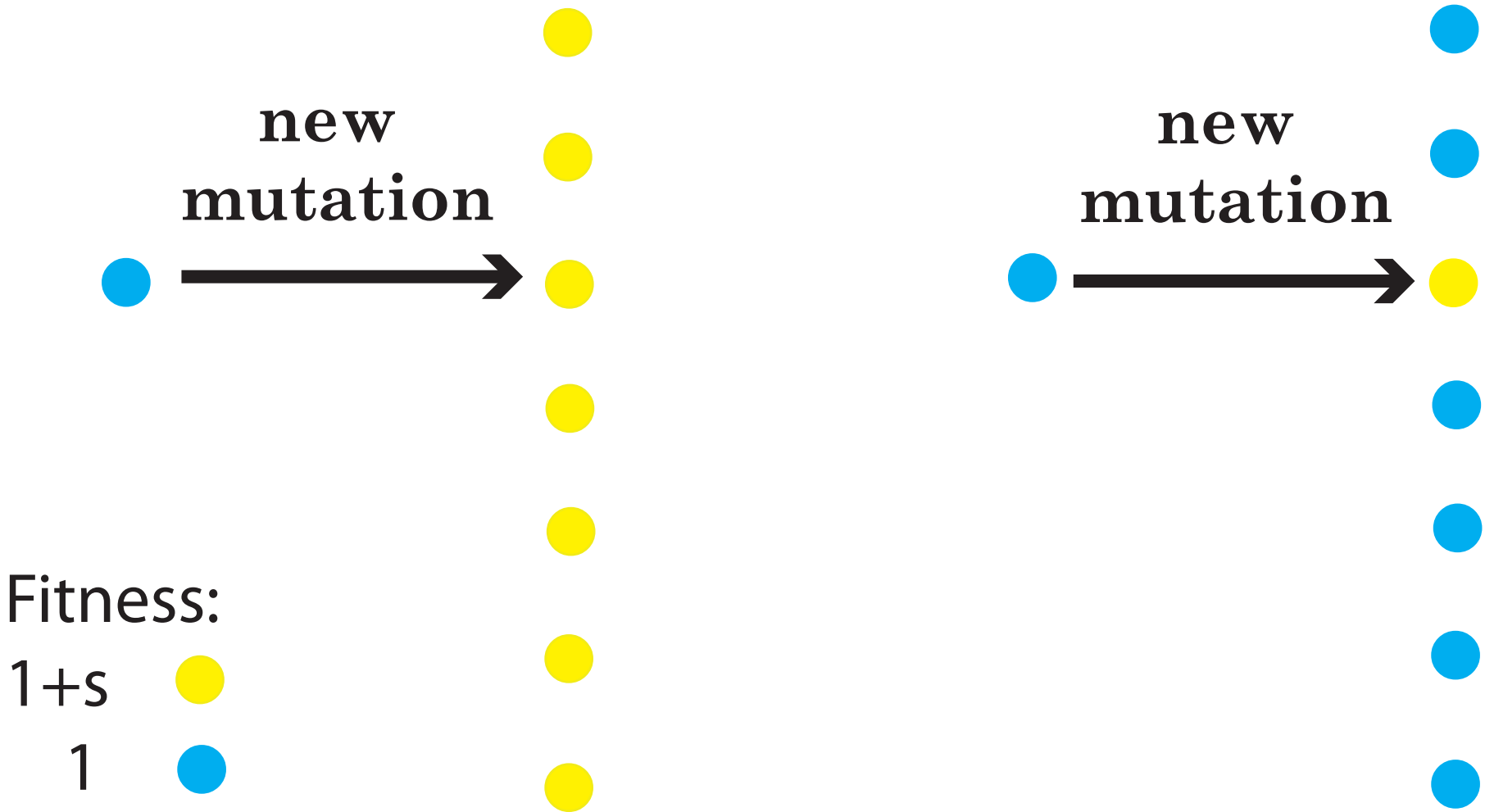

"Phylogenetic lineage"
Dead
Maybe

For change from Sequence i to Sequence j where i & j differ only at one sequence position where j has nucleotide type h, evolutionary rate from i to j is R_{ij} where

R_{ij} = (Mutation Rate) x (Fixation Probability)

(see Halpern & Bruno. 1998. MBE 15:910-917)

Fixation probabilities depend on the other alleles in the population



Wright-Fisher Simulation Conditions:

Ten thousand sites (no recombination)

1 optimal and 3 deleterious nucleotide types per site

No dominance (multiplicative fitness of haplotypes)

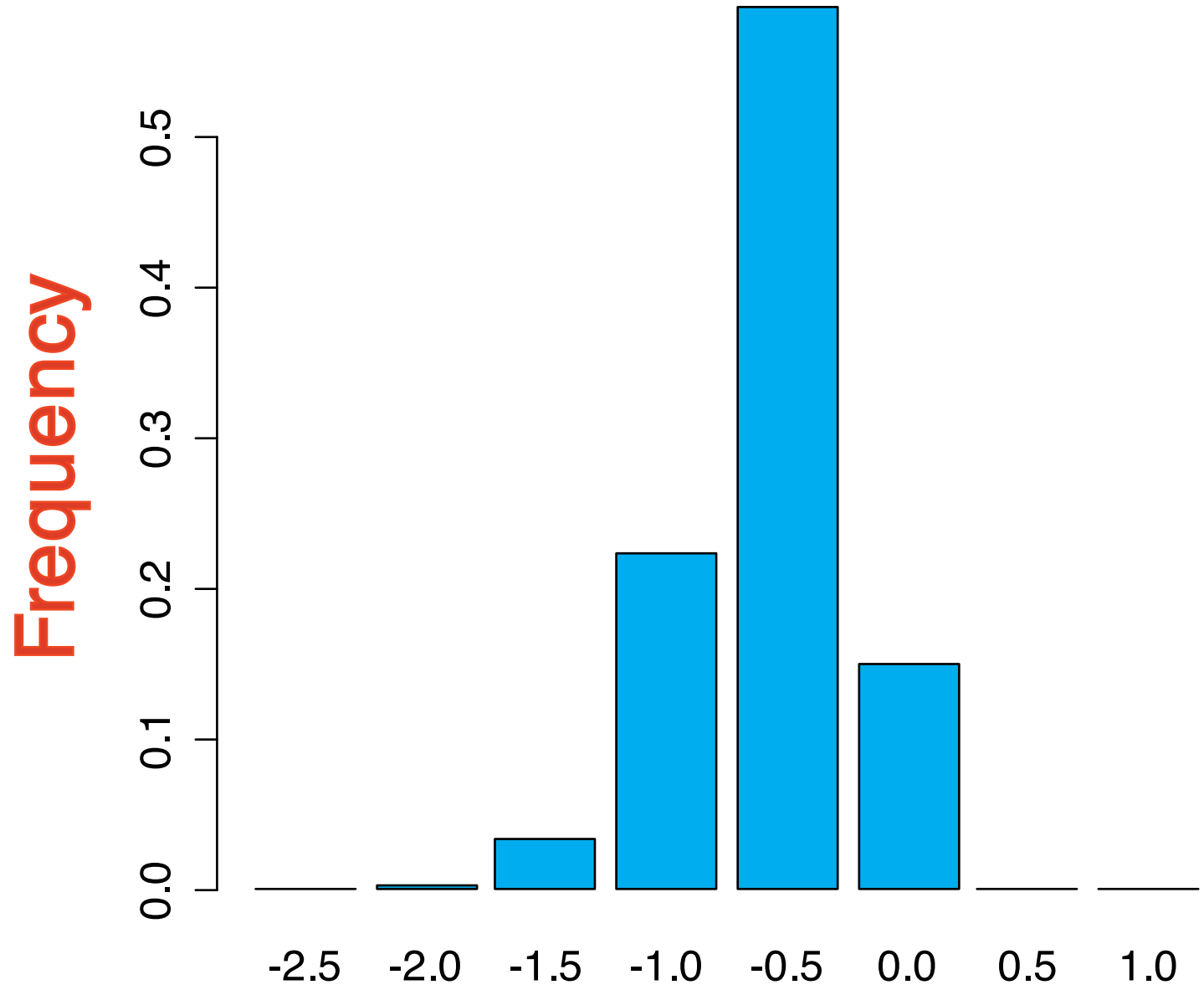
All mutations equally likely (rate per site is 10^{-8})

Relative fitness per sequence = $(1+s)^{m}$

m = # sites with deleterious mutation, s= -0.0001

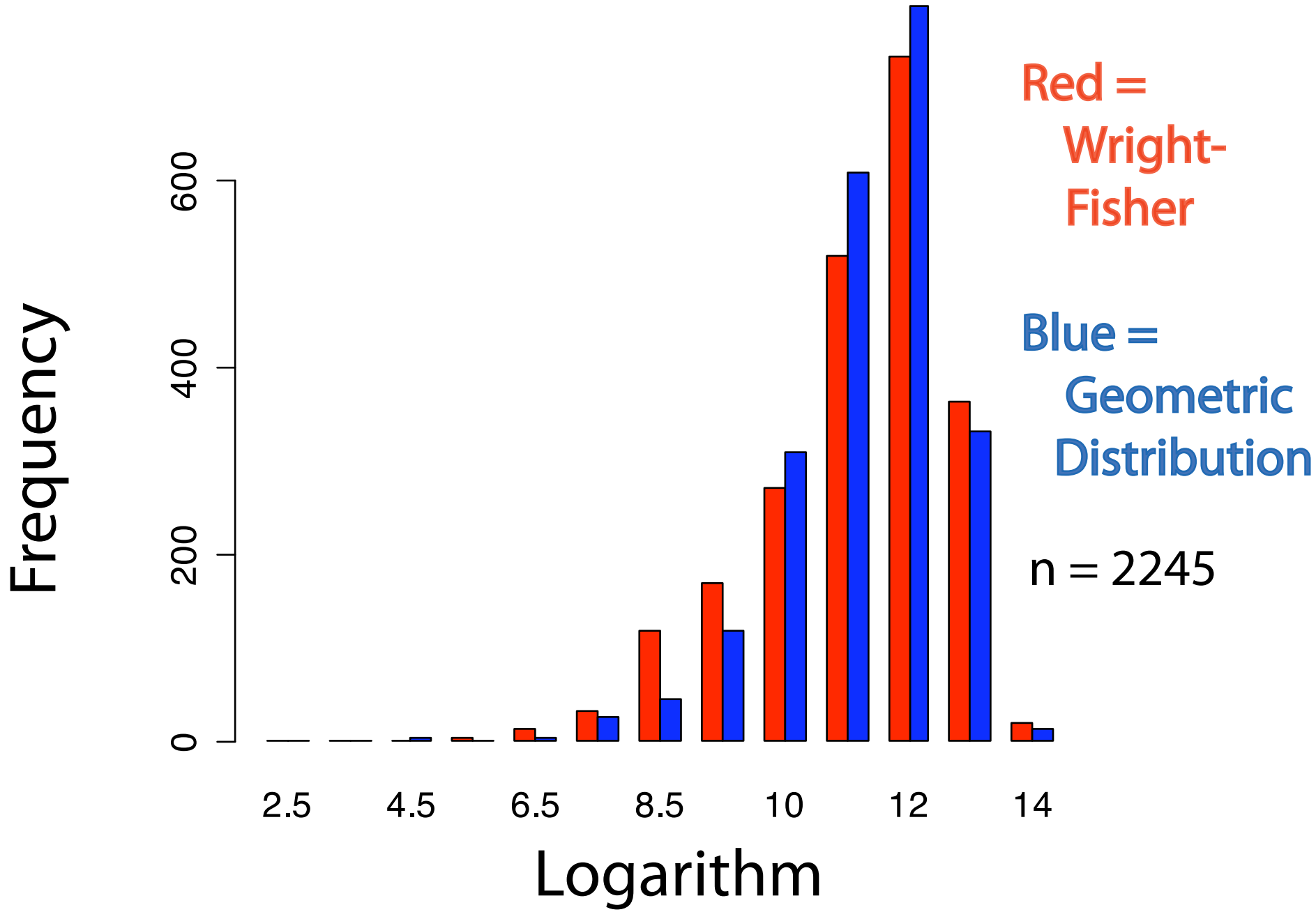
Population Size 2N = One Hundred Thousand

When the phylogenetic lineage has 67 deleterious sites ...

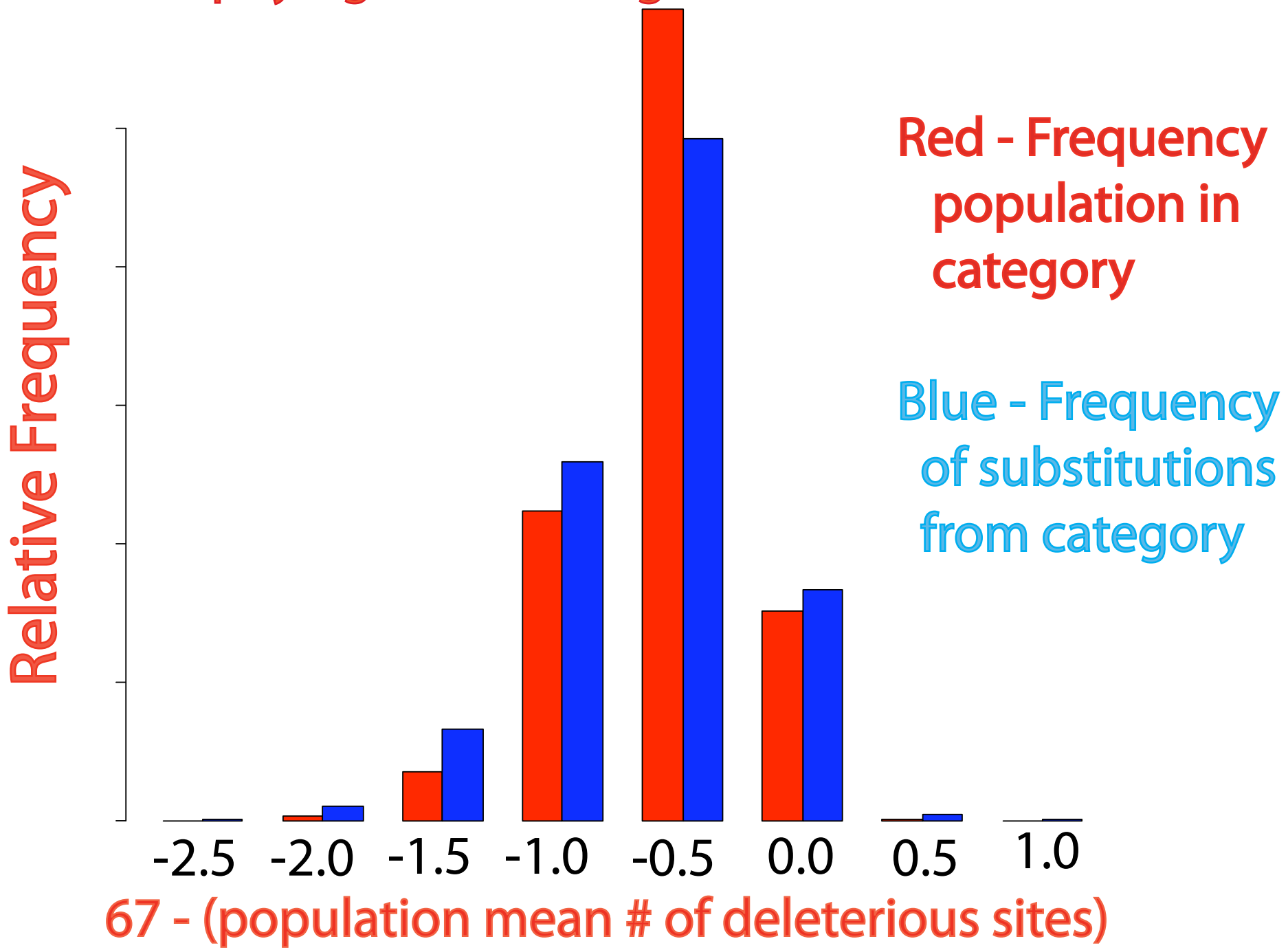


67 - (population mean # of deleterious sites)

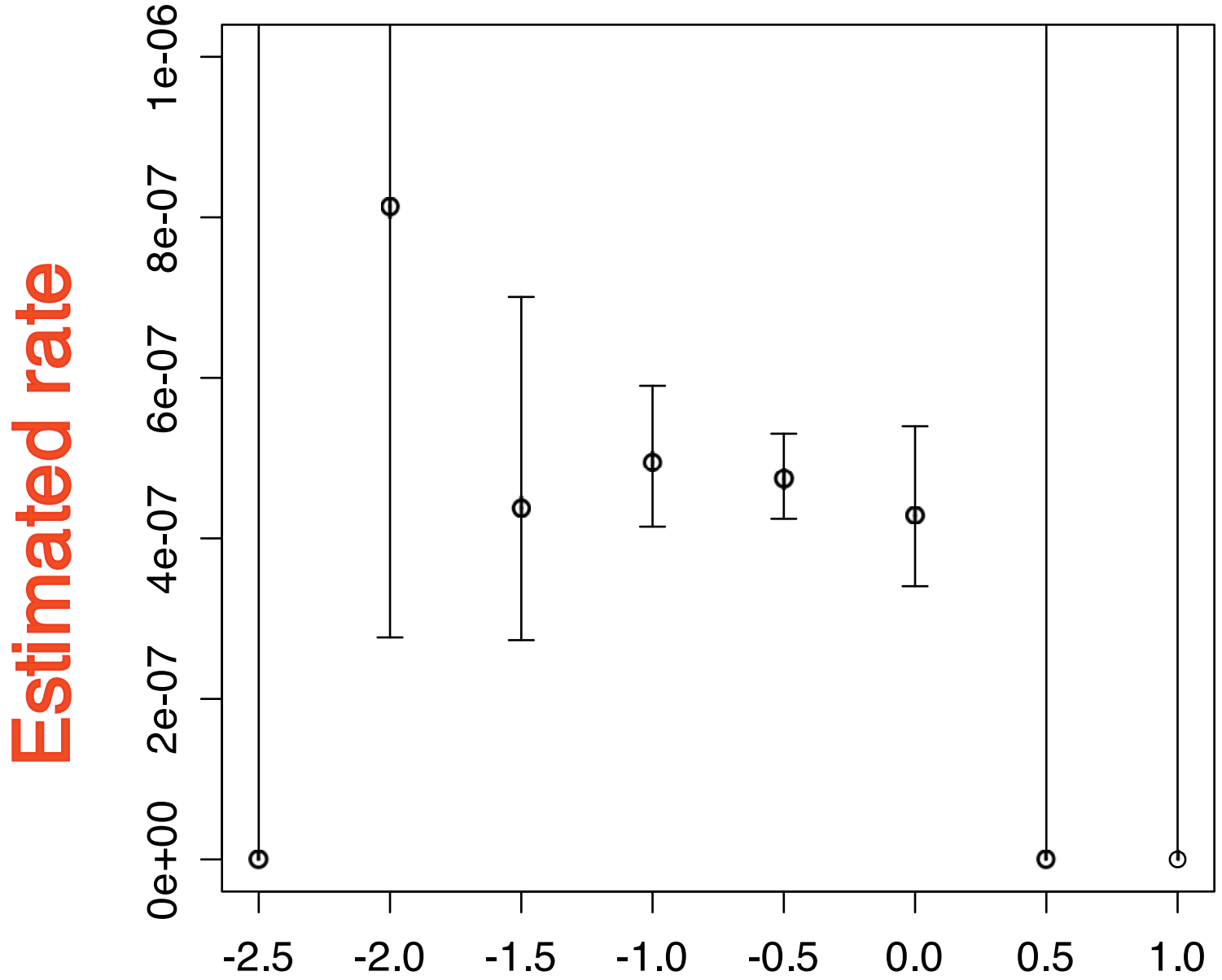
Histogram of log(interevent) times for Wright-Fisher process when phylogenetic lineage has 67 deleterious mutations



When the phylogenetic lineage has 67 deleterious sites ...

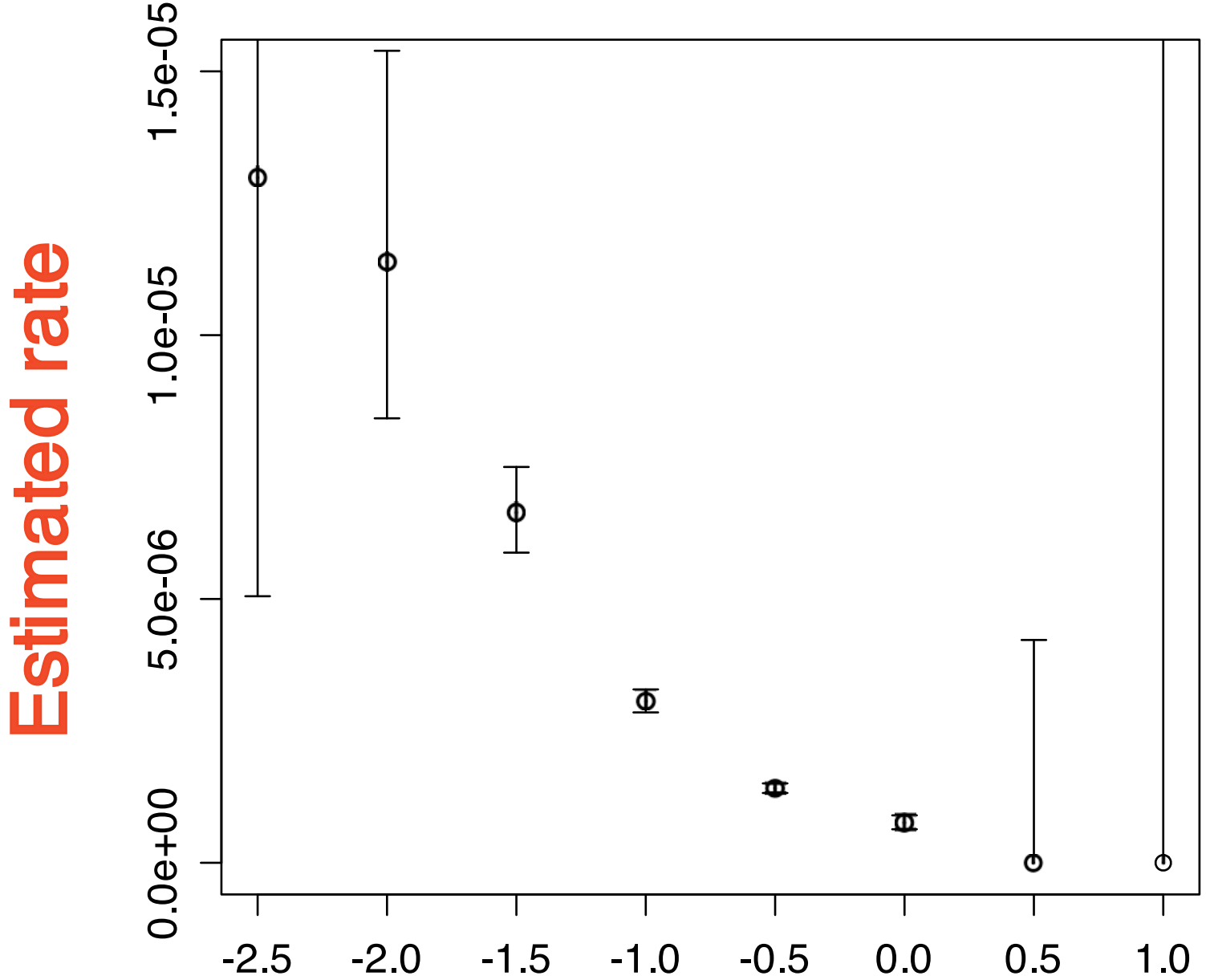


Neutral Changes



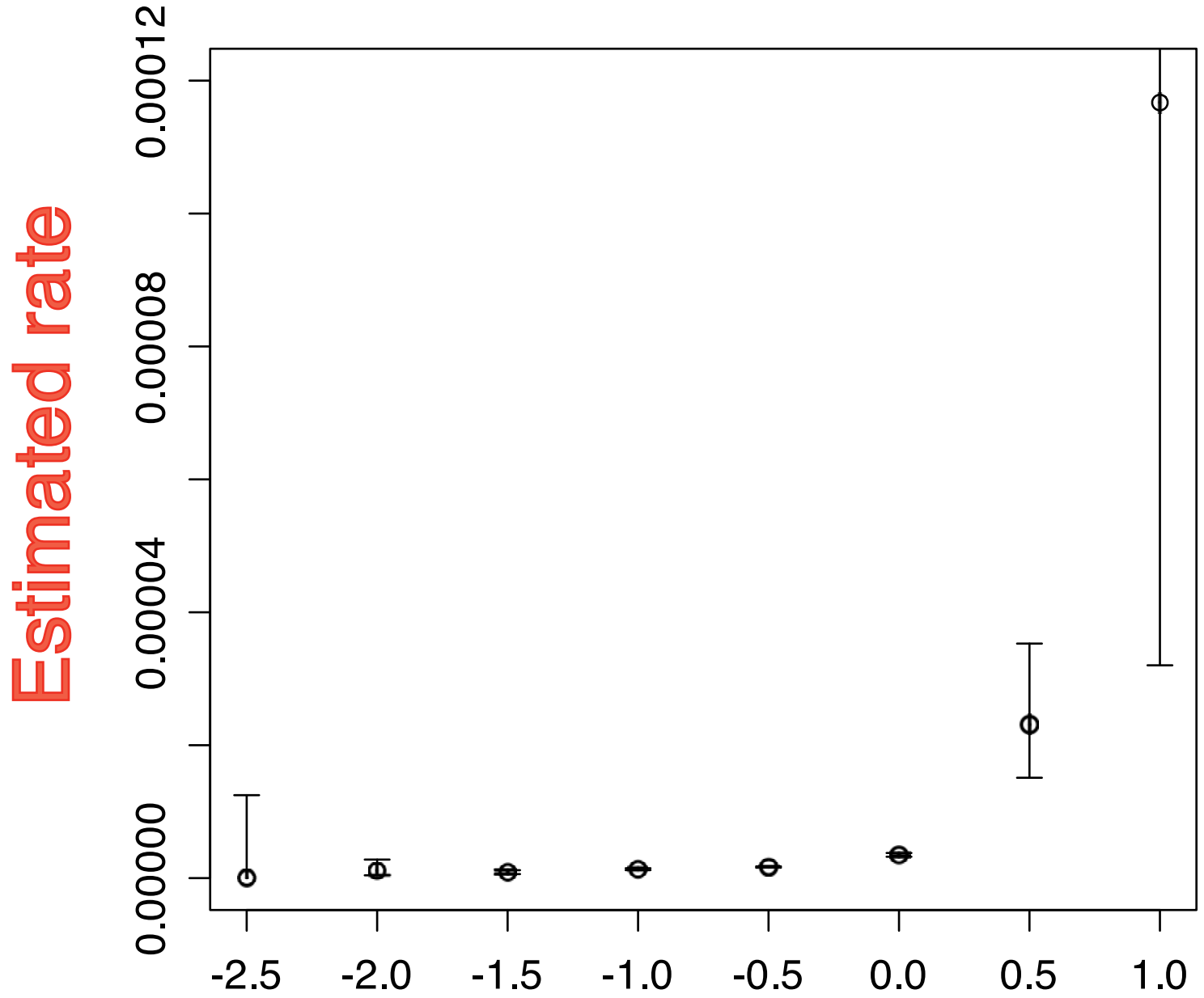
67 - (population mean # of deleterious sites)

Changes that decrease fitness



67 - (population mean # of deleterious sites)

Changes that increase fitness



67 - (population mean # of deleterious sites)

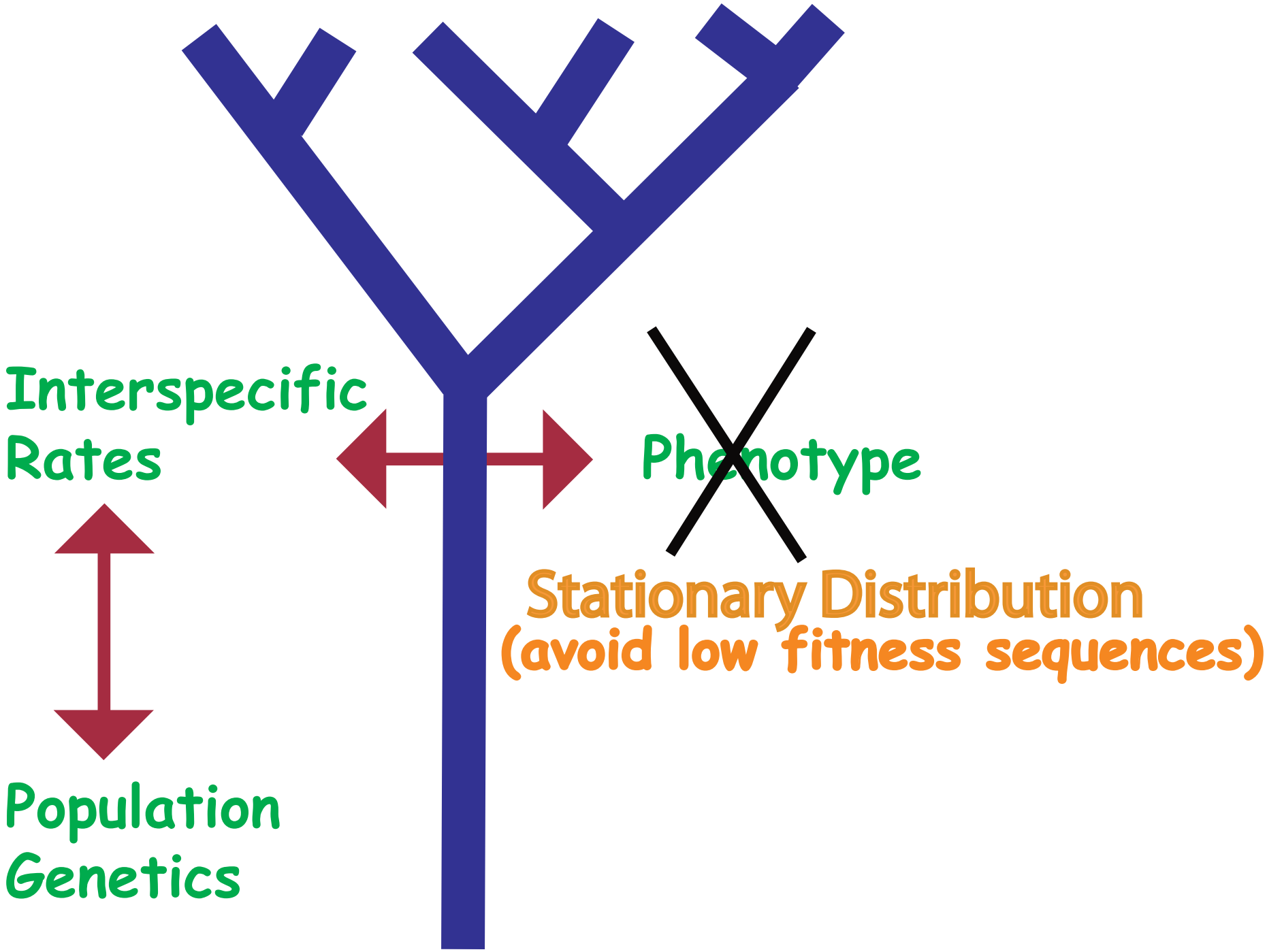
Entering and leaving 67 deleterious sites ...

	From 67 to 68	From 67 to 66
From 66 to 67	416	572
From 68 to 67	581	415

Fisher Exact Test P-Value is about 6×10^{-13}

Is Non-Markovian behavior "important" ?

If it is, should we augment phylogenetic lineage with ancestral selection graph? ... with allele counts at each generation? ... with meaningful summary statistic?



Mutation rates assumed low starting here...

Reversible Model of Insertion-Deletion for "neutral evolution"

+

Desired Probability Distribution of Protein Sequences (e.g. a profile HMM from Pfam)

+

"neutral" nucleotide frequencies

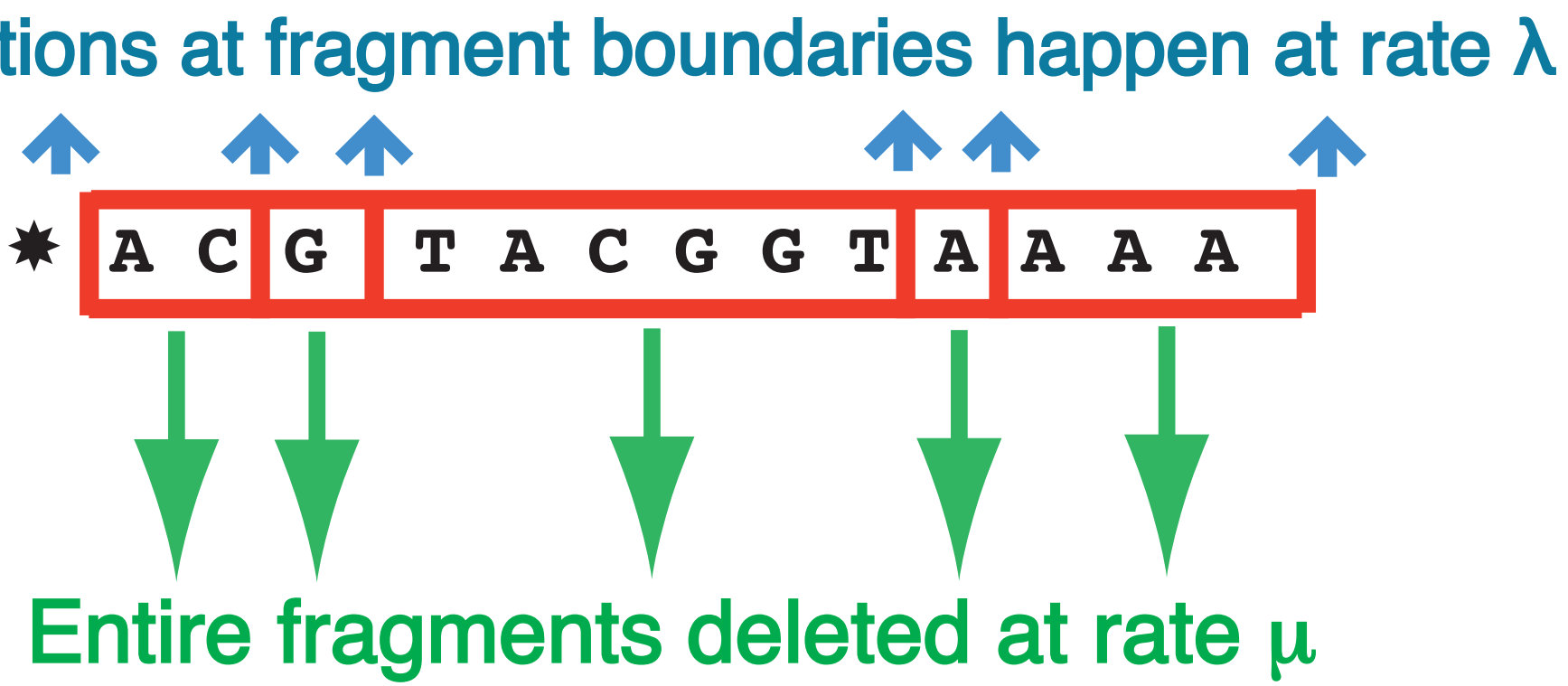
=

Reversible Rate matrix combining insertion-deletion processes and nucleotide substitution processes with desired stationary distribution of sequences

(Inference?)

Fragment model of insertion and deletion (Thorne, Kishino, and Felsenstein. 1992. JME 34:3-16)

Insertion-deletion process independent of nucleotide substitution process and modelled as birth-death process of sequence "fragments"



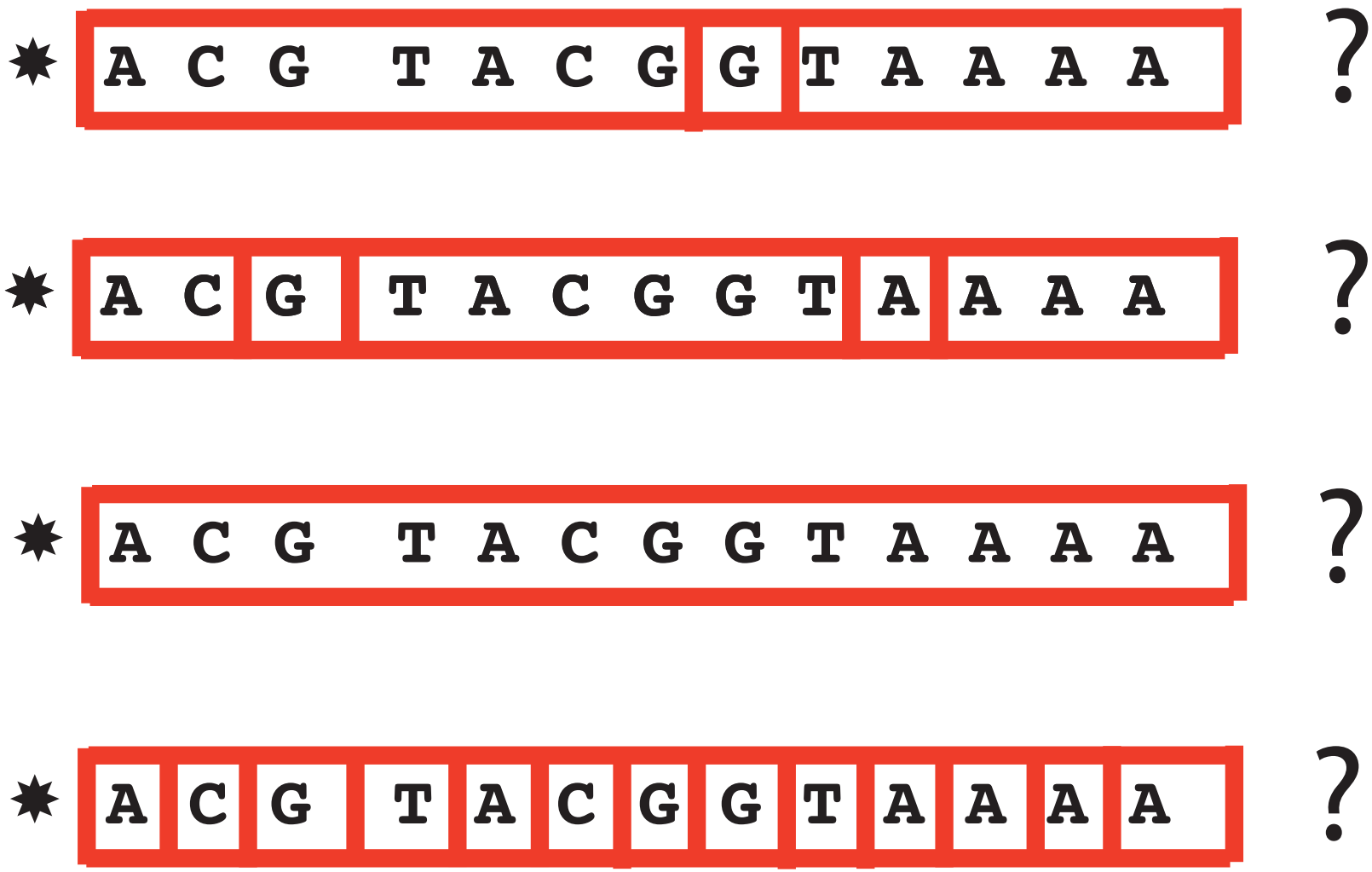
Computationally convenient to use geometric distribution for number of nucleotides per fragment:

Advantages of fragment model:

1. **Explicit transition probabilities**
2. **Known stationary dist. for sequence lengths**

Big disadvantage of fragment model:

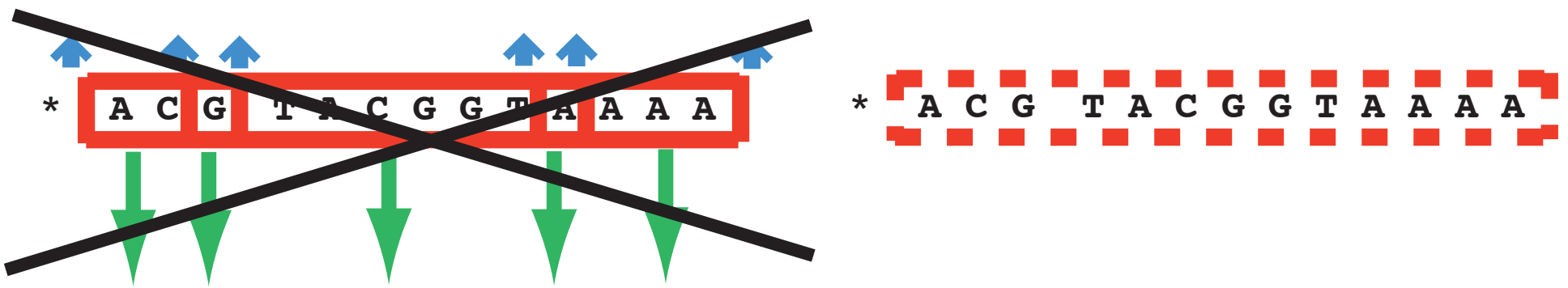
Fragment boundaries are permanent



Modification: Construct process with same stationary length distribution as fragment model but no fixed fragment boundaries:

1. Insertion and deletion lengths still have geometric distribution [insertion-deletion rates are equal to 1992 model's average over all possible fragmentations of a sequence]

2. Lose explicit transition probabilities



Some Notation:

i, j ... Two protein-coding DNA sequences
I,J ... Translated versions of i and j
$1+s_{ij} = w_j / w_i$ (w_i & w_j are fitnesses of i and j)

P(i) ... Stationary Distribution of i
P(I) ... (Desired/Target) Stationary Distribution of I
$P_0(i)$ ... Stationary Distribution of i under neutrality

Goals:

1. Design evolutionary model with $P(i) = P(I)\, P_0(i|I)$
2. Population genetic interpretation of departures between P(i) and $P_0(i)$

Population size 2N is constant over time

Mutation is rare

If i and j differ at more than one site or if j has stop codon, then $R_{ij} = 0$.

Otherwise, if h is nucleotide type at single site in j where i and j differ,

$$R_{ij} = \begin{cases} u\pi_h\kappa \times 2N \times P(Z_{ij}) & \textbf{transition} \\ \\ u\pi_h \times 2N \times P(Z_{ij}) & \textbf{transversion} \end{cases}$$

$P(Z_{ij})$ is event mutation to j from i fixes

Approach 1: Sella-Hirsh (2005) approx. to fixation prob. for mutation to j from i

$$P(Z_{ij}) \doteq \frac{1 - e^{-2\log(w_j/w_i)}}{1 - e^{-4N\log(w_j/w_i)}}$$
$$\doteq \frac{1 - e^{-2\log(1+s_{ij})}}{1 - e^{-4N\log(1+s_{ij})}}$$

and

$$P(j) = \frac{e^{2(2N-1)\log(1+s_{ij})} P_0(j)}{\Sigma_k \, e^{2(2N-1)\log(1+s_{ik})} P_0(k)}.$$

M **defined as** $(P(J)/P_0(J))/(P(I)/P_0(I))$

Combining Sella-Hirsh fixation equation and $\log(M) = 2(2N-1)\log(w_j/w_i)$**,**

$$P(Z_{ij}) \doteq \frac{M^{1/(2N-1)} - 1}{M^{1/(2N-1)} - \frac{1}{M}},$$

and we get model with stationary distribution equal to desired target distribution. For small s_{ij}**,**

$$2Ns_{ij} \doteq \frac{1}{2}\log(M).$$

Approach 2: Can make stationary distribution equal target when

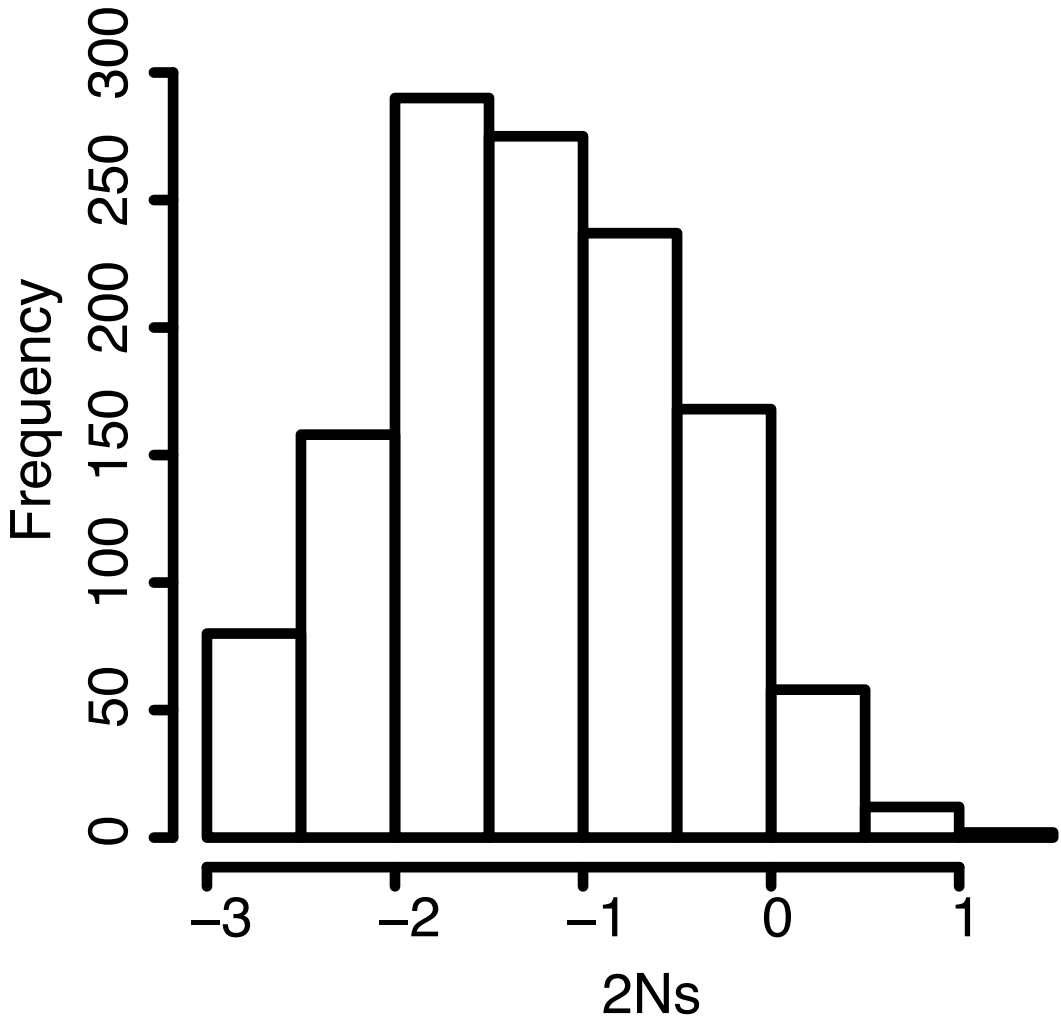
$$R_{ij} = \begin{cases} u\pi_h\kappa \times \sqrt{\dfrac{P(J)/P_0(J)}{P(I)/P_0(I)}} & \textbf{transition} \\ \\ u\pi_h \times \sqrt{\dfrac{P(J)/P_0(J)}{P(I)/P_0(I)}} & \textbf{transversion} \end{cases}$$

If Ns_{ij} is not too big, again have

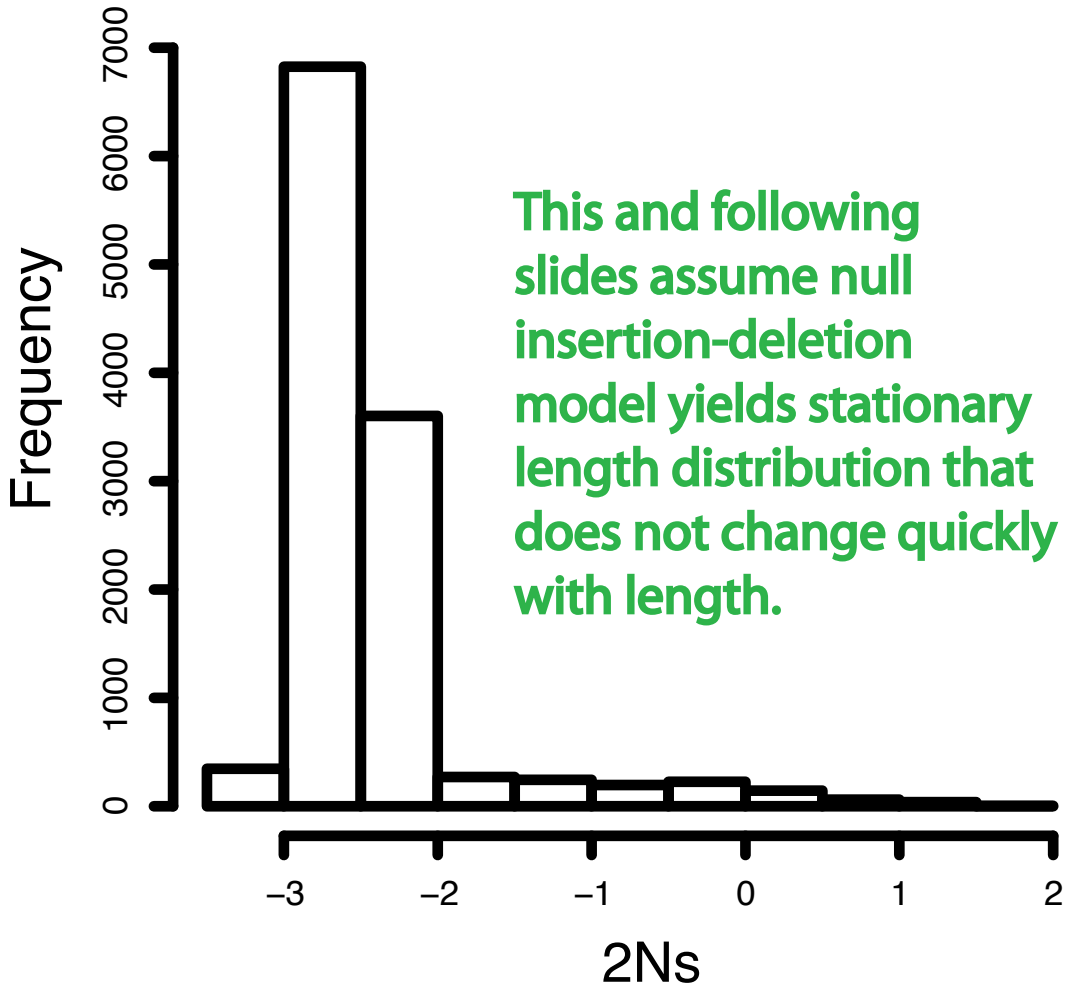
$$2Ns_{ij} \doteq \frac{1}{2}\log\frac{P(J)/P_0(J)}{P(I)/P_0(I)} = \frac{1}{2}\log(M)$$

Parallels Knudsen and Miyamoto (2005) for independent-site models ...

Possible Nonsyn. Changes to human P53

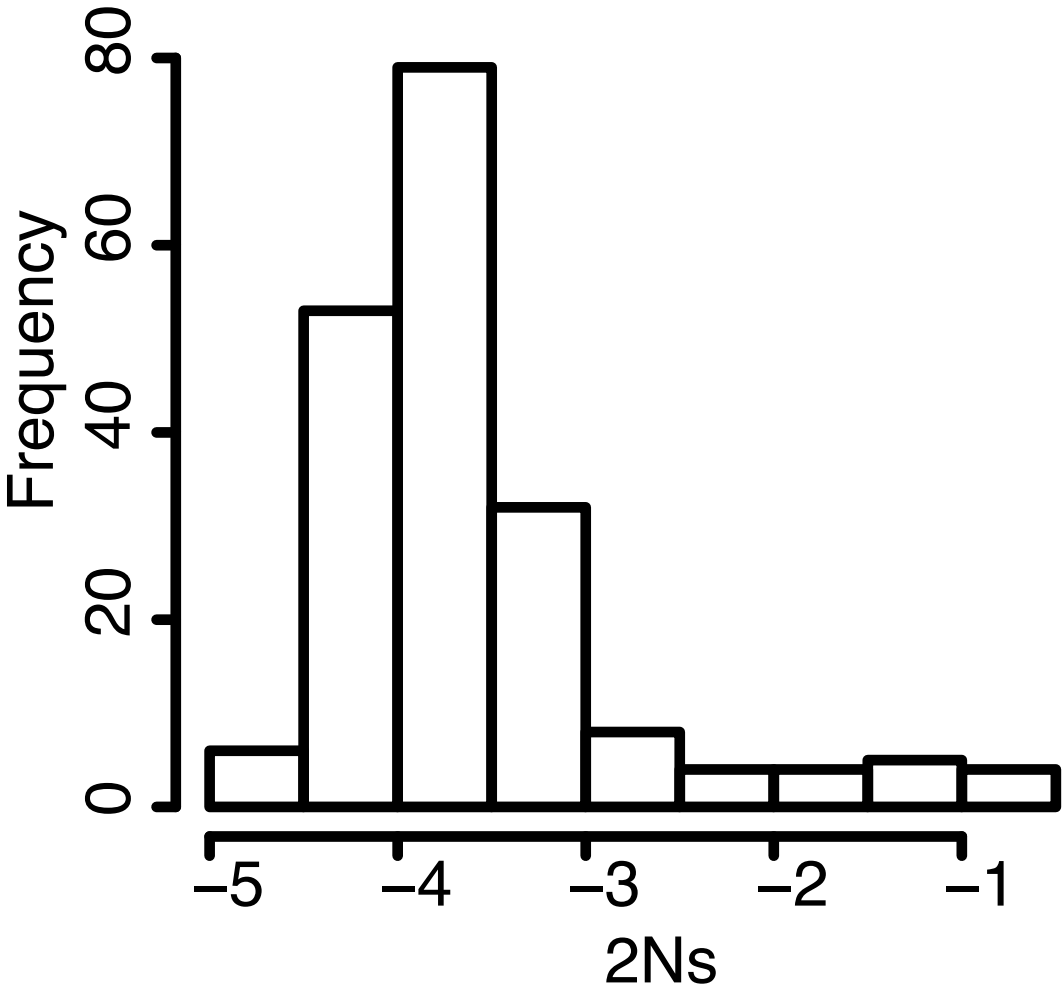


Possible one-codon insertions to human P53

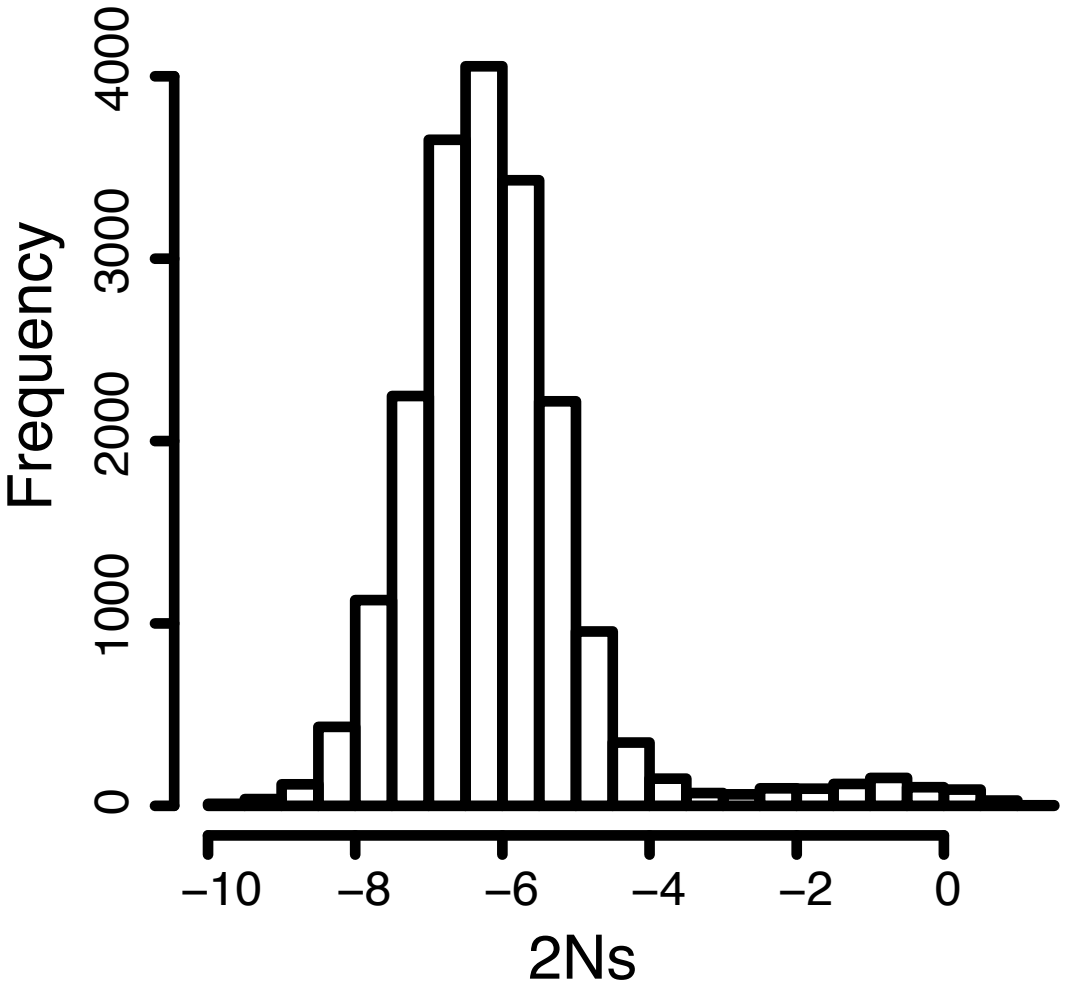


This and following slides assume null insertion-deletion model yields stationary length distribution that does not change quickly with length.

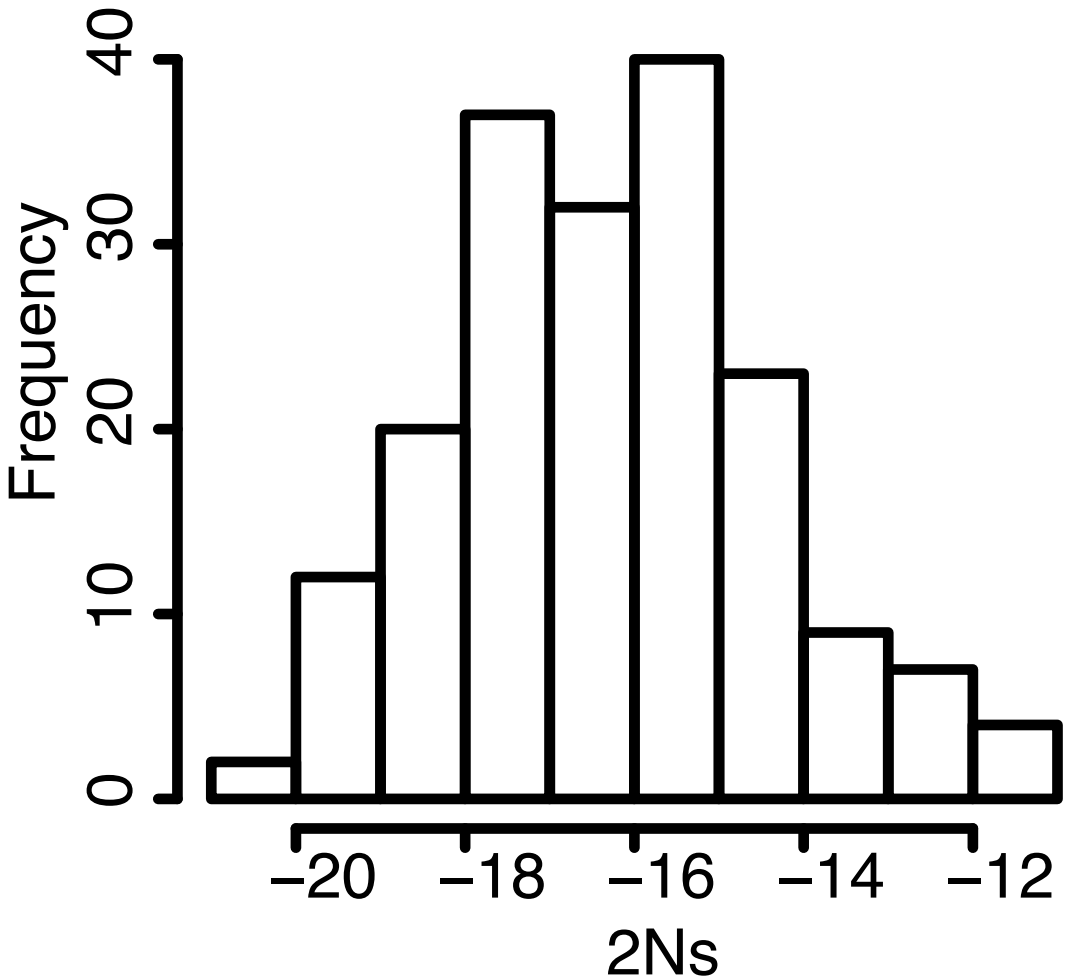
Possible one-codon deletions to human P53



Possible ten-codon insertions to human P53



Possible ten-codon deletions to human P53



Future Directions ...

Other stationary distributions?

Incorporate context-dependent mutation?

Evolutionary inference with realistic models of insertion and deletion?

Less crude reconciliation of population and phylogenetics?

Thanks to NIH and NSF for support !!
(and also thanks to Paul Higgs)