

The frequency spectrum of a mutation

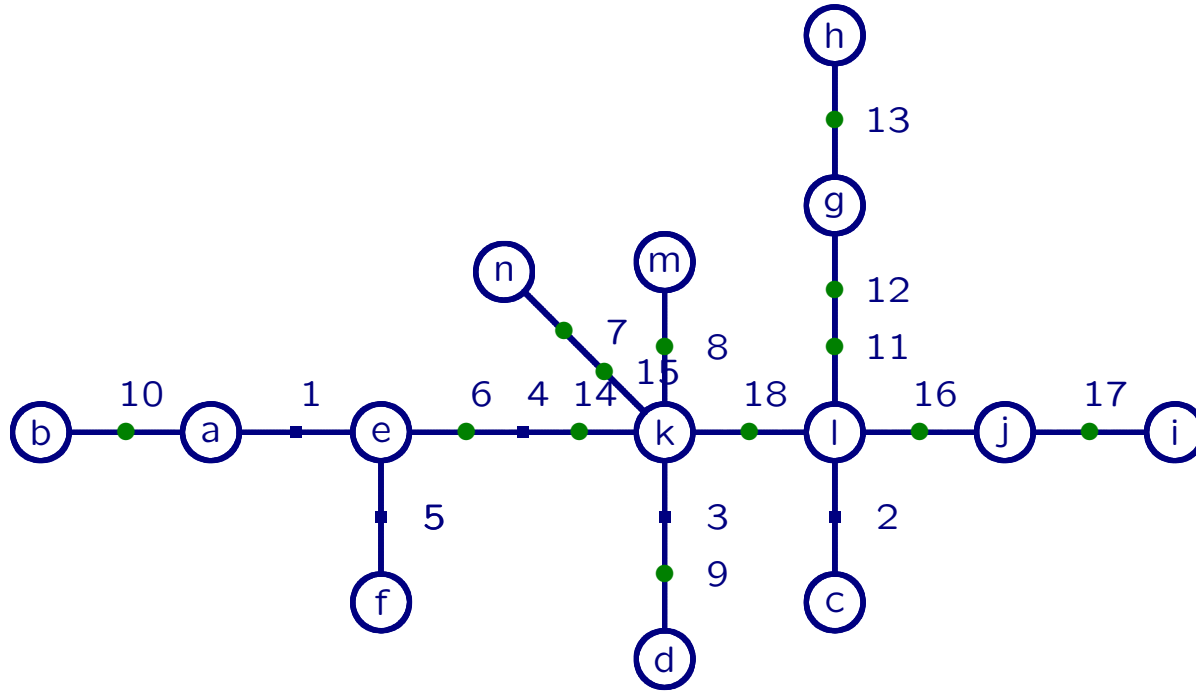
Bob Griffiths

University of Oxford

	MtDNA	Freq
<i>a</i>	AGGAATCCTCTTCTCTTC	2
<i>b</i>	AGGAATCCTTTTCTCTTC	2
<i>c</i>	GAGGACCCTCTTCCCTTT	1
<i>d</i>	GGAGACCCCTTCCCTTC	3
<i>e</i>	GGGAATCCTCTTCTCTTC	19
<i>f</i>	GGGAGTCCTCTTCTCTTC	1
<i>g</i>	GGGGACCCTCCCCCCTTT	1
<i>h</i>	GGGGACCCTCCCTCCTTT	1
<i>i</i>	GGGGACCCTCTTCCCCCT	4
<i>j</i>	GGGGACCCTCTTCCCCTT	8
<i>k</i>	GGGGACCCTCTTCCCTTC	5
<i>l</i>	GGGGACCCTCTTCCCTTT	4
<i>m</i>	GGGGACCTTCTTCCCTTC	3
<i>n</i>	GGGGACTCTTTCCTTTC	1

Ward, R. H. Frazier, B. L., Dew, K. and Paabo, S. (1991)
Proc. Nat. Acad. Sci. USA **88** 8720-8724.

Unrooted Nuu-Chah-Nulth tree



● pyrimidine sites; ■ purine sites

Frequency spectrum of a mutation

Observed frequency spectrum, the distribution of segregating sites with 1, 2, ... mutations.

MtDNA data example

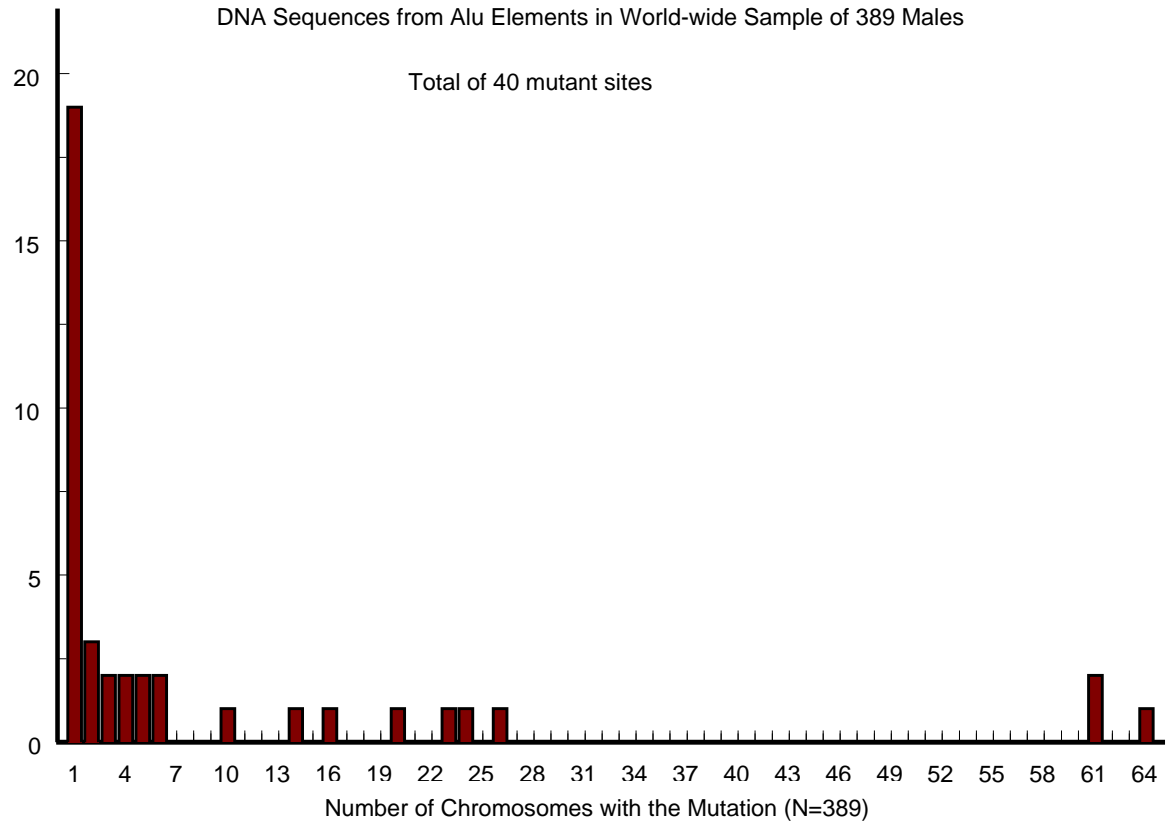
Relative frequencies of sites with 1, 2, ... mutations

Mutations	1	2	3	4	·	12	18	24
Frequency	5/55	3/55	3/55	2/55	·	1/55	1/55	3/55

Frequency Spectrum: Y Chromosome Data of Wilder et. al. 2004

DNA Sequences from Alu Elements in World-wide Sample of 389 Males

Total of 40 mutant sites



General binary coalescent trees

$\{T_j, j = n, \dots, 2\}$ are times while $n, \dots, 2$ ancestors of n individuals.

- T_n, \dots, T_2 are continuous random variables.
- The ancestral tree is binary, and such that when there are k ancestral lines each pair has probability $\binom{k}{2}^{-1}$ of being the next pair to coalesce.

Examples of general binary trees

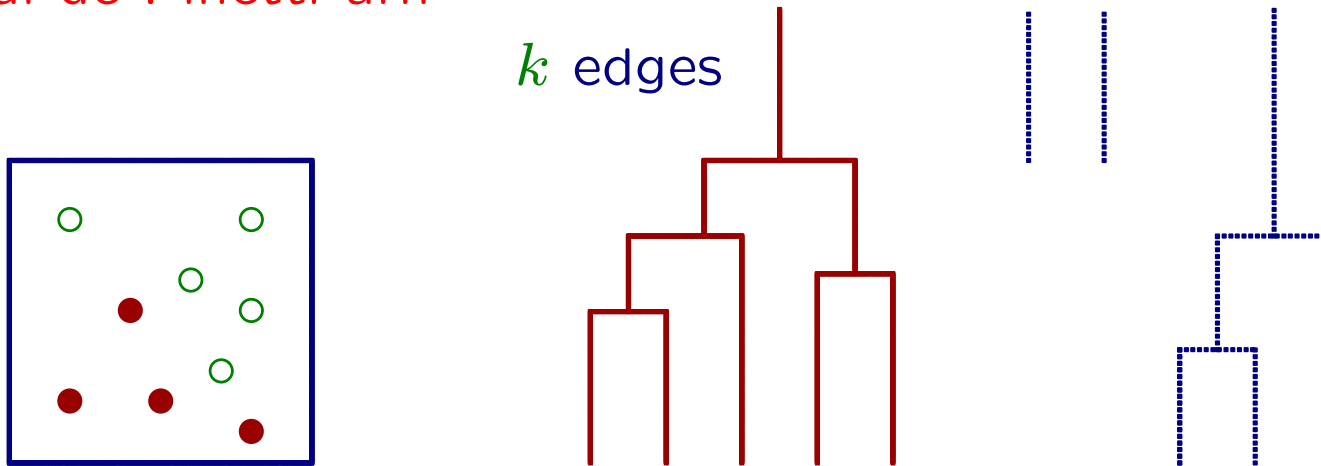
$\{T_j, j = n, \dots, 2\}$ are times while $n, \dots, 2$ ancestors of n individuals.

- Coalescent process where T_n, \dots, T_2 are independent exponential with means $2/n(n-1), \dots, 2/2(2-1)$.
- Coalescent process with a non-homogeneous population size $N(t) = N_0/\nu(t)$ at time t back. $S_j^\nu = T_j^\nu + \dots + T_n^\nu$, with $S_{n+1}^\nu = 0$. $\{S_j^\nu\}$ form a reverse Markov Process with transition density of S_j^ν given $S_{j+1}^\nu = t$ of

$$f(s; t) = \binom{j}{2} \nu(s) \exp\left(-\binom{j}{2} \int_t^s \nu(u) du\right), s > t.$$

- An ancestral tree of a sample of n individuals taken from under a neutral or selected mutation of frequency x in the population.
- An ancestral tree of a sample of n individuals taken at a locus linked to a selectively favoured allele.
- A general death process tree with rates $\mu_k(t)$ at time t back.
- A general binary birth process tree grown forward in time.
- An ancestral tree of a sample of n individuals taken from a birth and death process at a time when $\geq n$ individuals exist.

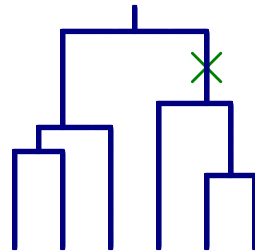
Classical de Finetti urn



- Put one red ball and $k - 1$ white balls in an urn.
- At each trial draw a ball at random and replace with an additional ball of the colour chosen.
- Stop when n balls.

Forward in time branching in the subtree is equivalent to drawing a red ball. The distribution of the number of red balls is the same as the family size distribution of the mutation.

Family size of a mutation



The probability that a particular family has size b is

$$p_{n,k}(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}.$$

In the population limit as $n \rightarrow \infty$, and $b/n \rightarrow x$, $0 < x < 1$, the relative frequency of a particular family has a Beta density

$$(k-1)(1-x)^{k-2}, 0 < x < 1.$$

Frequency spectrum in a general binary tree

A mutation is observed in b out of n genes in a sample, where $0 < b < n$.

What is the probability distribution of the number of copies of b ?

It is helpful in this problem to label mutations as independent uniform random variables on $[0,1]$. Then the probability that a mutation has a label in $(x, x + h)$ is h .

Let $C_h = C(x, b, h)$ denote the event that there is a mutation with label U in the interval $(x, x + h) \subset (0, 1)$ that subtends b copies in the sample.

$$P(C_h \mid T_n, \dots, T_2) = \sum_k p_{nk}(b) \left(k T_k \frac{\theta}{2} h + o(h) \right)$$

Averaging over the distribution of T_n, \dots, T_2

$$P(C_h) \sim \frac{\theta h}{2} \cdot \sum_k k p_{nk}(b) E(T_k), \text{ as } h \rightarrow 0.$$

Summing over b , the probability that there is a mutation with label in $(x, x + h)$ is

$$\sim \frac{\theta h}{2} \cdot \sum_k k E(T_k), \text{ as } h \rightarrow 0.$$

The probability that a particular mutation has b copies is thus

$$q_{nb} = \frac{\sum_{k=2}^n k p_{nk}(b) E(T_k)}{\sum_{k=2}^n k E(T_k)}, \quad 0 < b < n$$

General formula

$$q_{nb} = \frac{E(L_b)}{E(L)}$$

L_b - sum of edge lengths subtending b leaves

L - total sum of edge lengths

If the wild and mutant sites are unidentifiable then the wrapped frequency spectrum is

$$q'_{n,b} = \begin{cases} q_{n,b} + q_{n,n-b}, & 1 \leq b < n/2 \\ q_{n,n/2}, & n \text{ even and } b = n/2, \end{cases}$$

Kingman Coalescent process

$$E(T_k) = \binom{k}{2}^{-1}, \quad k = n, \dots, 2$$

$$q_{nb} = \frac{b^{-1}}{\sum_{j=1}^{n-1} j^{-1}}, \quad 1 \leq b \leq n-1.$$

The mean number of mutations in the distribution is

$$\mu = \frac{n-1}{\sum_{j=1}^{n-1} j^{-1}}$$

As $n \rightarrow \infty$,

$$\mu \sim \frac{n}{\log n}, \quad \sigma^2 \sim \frac{n^2}{2 \log n}$$

Models of population growth

Coalescence times S_n^ν, \dots, S_2^ν can be easily simulated successively from the distribution

$$P(S_j^\nu > s \mid S_{j+1}^\nu = t) = \exp\left(-\binom{j}{2} \int_t^s \nu(u) du\right)$$

by setting the probability equal to a uniform random variable on $[0,1]$, and solving for $S_j = s$.

The frequency spectrum can then be found from the simulated mean times $E(T_n^\nu), \dots, E(T_2^\nu)$.

Multiple Merger Coalescents

r_{nj} ; merger rate to j from n edges.

$r_n = \sum_{j=2}^n r_{nj}$; total merger rate while n edges.

$c_{nj} = r_{nj}/r_n$; probability of a merger from n edges to j edges.

$c_{nj}^{(k)}$; probability of a merger from n edges to j edges, conditional on hitting k edges in the future.

$$p_{nk}(b) = \left\{ \sum_{j=n-b+1}^{n-1} c_{nj}^{(k)} \frac{b - (n - j)}{j} p_{jk}(b - (n - j)) + \sum_{j=b+1}^{n-1} c_{nj}^{(k)} \frac{j - b}{j} p_{jk}(b) \right\}$$

$$p_{kk}(l) = \delta_{l1}$$

Λ -coalescents

$$r_{nj} = \binom{n}{j-1} \int_{[0,1]} x^{n-j-1} (1-x)^{j-1} \Lambda(dx), \quad 1 \leq j < n$$

where Λ is a finite measure on $[0, 1]$.

The total rate is

$$r_n = \int_{[0,1]} \frac{1 - (1-x)^n - nx(1-x)^{n-1}}{x^2} \Lambda(dx)$$

and $c_{nj} = r_{nj}/r_n$.

Cannings, Pitman, Sagitov, Möhle

Beta-Coalescent

$$\Lambda(dx) = B(a, b)^{-1} x^{b-1} (1-x)^{a-1} dx$$

$$r_{nj} = \binom{n}{j-1} \frac{a(n-j-1)b^{(j-1)}}{(a+b)_{(n-2)}}, \quad 1 \leq j < n$$

α - coalescent $a = 2 - \alpha$, $b = \alpha$

Birkner, Blath, Capaldo, Etheridge, Möhle, Schweinsberg, Wakolbinger

Conditioned on hitting k

$u_n^{(k)}$: probability of hitting k from n .

$$u_r^{(k)} = \sum_{j=k}^{r-1} c_{rj} u_j^{(k)}, \quad u_k^{(k)} = 1$$

$$c_{nj}^{(k)} = c_{nj} u_j^{(k)} / u_n^{(k)}$$

Mean coalescence times

$$E_n(T_k) = \sum_{j=k}^{n-1} c_{nj} E_j(T_k), \quad n \geq k, \quad E_k(T_k) = 1/r_k$$

Frequency spectrum

$$q_{nb} = \frac{\sum_{k=2}^n k p_{nk}(b) E_n(T_k)}{\sum_{k=2}^n k E_n(T_k)}, \quad 0 < b < n$$

The frequency spectrum of neutral mutations under a selected mutation of frequency x

If the generator of the process of the selected mutation $\{X(t), t > 0\}$ is

$$L = \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2} + \beta x(1-x)\frac{\partial}{\partial x}$$

then the generator of the reverse process $\{X^*(t), t > 0\}$ back in time starting at x is

$$L = \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2} - \frac{1}{2}\beta x(1-x)\coth\left(\frac{1}{2}\beta(1-x)\right)\frac{\partial}{\partial x}$$

The reverse process partitions the population into selected and non-selected types, and the behaviour of a coalescent taken in the selected group is a stochastic variable-population-size model with

$$N(t) = N_0 X^*(t), \quad X^*(0) = x$$

The mean coalescence times can be calculated by simulation, on each run simulating $\{X^*(t)\}$, then using the variable population size simulation to obtain $E(T_n), \dots, E(T_2)$.

Moran model and diffusion process limit

Let $\{Z(t), t \geq 0\}$ be a birth and death process on $\{0, 1, \dots, N\}$ with rates $\{\lambda_k^{(N)}\}, \{\mu_k^{(N)}\}$.

Denote $X(t) = Z(t)/N$, $\Delta Z(t) = Z(t + \delta t) - Z(t)$, and similarly for $\Delta X(t)$.

Then as $\delta t \rightarrow 0$,

$$E(\Delta X(t) \mid X(t) = x) = \frac{\lambda_z^{(N)} - \mu_z^{(N)}}{N} \delta t + o(\delta t)$$
$$\text{var}(\Delta X(t) \mid X(t) = x) = \frac{\lambda_z^{(N)} + \mu_z^{(N)}}{N^2} \delta t + o(\delta t)$$

To obtain a diffusion process limit with parameters $\sigma^2(x) = x(1-x)$ and $\mu(x)$ suppose that as $N \rightarrow \infty$

$$\begin{aligned}\frac{\lambda_z^{(N)} - \mu_z^{(N)}}{N} &\rightarrow \mu(x) \\ \frac{\lambda_z^{(N)} + \mu_z^{(N)}}{N^2} &\rightarrow \sigma^2(x)\end{aligned}\tag{1}$$

That is

$$\begin{aligned}\lambda_z^{(N)} &\sim \frac{1}{2}N(N\sigma^2(x) + \mu(x)) \\ \mu_z^{(N)} &\sim \frac{1}{2}N(N\sigma^2(x) - \mu(x))\end{aligned}\tag{2}$$

The sets of equations (1) and (2) are of interest, the former as a diffusion limit, and the latter as a birth and death process approximation to a diffusion process.

Diffusion process for the population frequency of a mutation

The relative frequency $\{X(t), t \geq 0\}$ has generator

$$L = \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2} + \mu(x)\frac{\partial}{\partial x}$$

$$\sigma^2(x) = x(1-x), \quad \mu(x) = \beta(x)x(1-x)$$

Transition density reversibility

$$m(x)f(x, p; t) = m(p)f(p, x, ; t)$$

where $m(x)$ is the speed density.

$$m(x) = [\sigma^2(x)s(x)]^{-1}$$

where

$$s(y) = \exp \left\{ - \int_0^y 2\beta(\xi) d\xi \right\}, \quad 0 < y < 1$$

Examples:

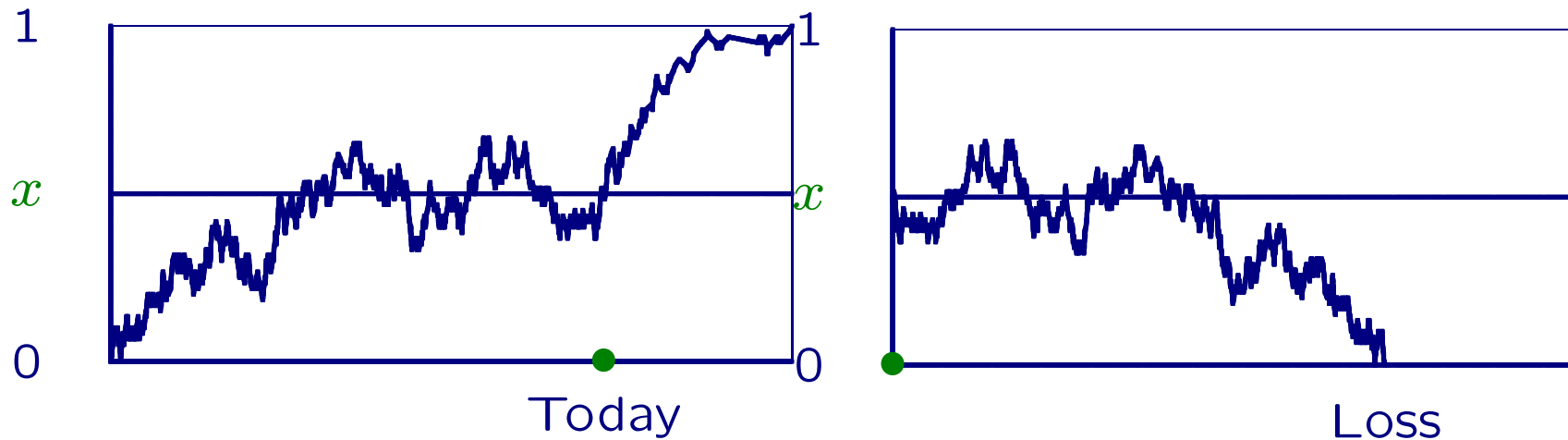
Genic selection

$$\mu(x) = \frac{1}{2}\beta x(1-x), \quad m(x) = [x(1-x)]^{-1}e^{-\beta x}$$

Balancing selection

$$\mu(x) = \frac{1}{2}\beta(2x-1)x(1-x), \quad m(x) = [x(1-x)]^{-1}e^{\beta x(1-x)}$$

Trajectory of a mutation



Reversibility argument. An allele A arises in a pure a population, and is observed to have frequency x at a time later. The distribution of the trajectory of A back in time is the same as the distribution until the time of its eventual loss, from frequency x , conditional on eventual loss.

Frequency spectrum

The probability that there are $0 < b < n$ mutant copies of a gene in a sample of n , conditional on at least one copy is

$$q_{n,b} = \frac{\int_0^1 \binom{n}{b} x^b (1-x)^{n-b} m(x) u_0(x) dx}{\int_0^1 (1-x^n - (1-x)^n) m(x) u_0(x) dx}$$

$$\begin{aligned} u_0(x) &= \text{P(Absorption at 0, starting from } x) \\ &= \frac{\int_x^1 s(y) dy}{\int_0^1 s(y) dy} \end{aligned}$$

In the frequency spectrum as $n, b \rightarrow \infty$ with $b/n \rightarrow x$

$$q_{n,b} \sim m(x) u_0(x)$$

Frequency spectrum

Large sample size $n \rightarrow \infty$, $b/n \rightarrow x$.

$$q_{n,b} \sim m(x)u_0(x)$$

Neutral model

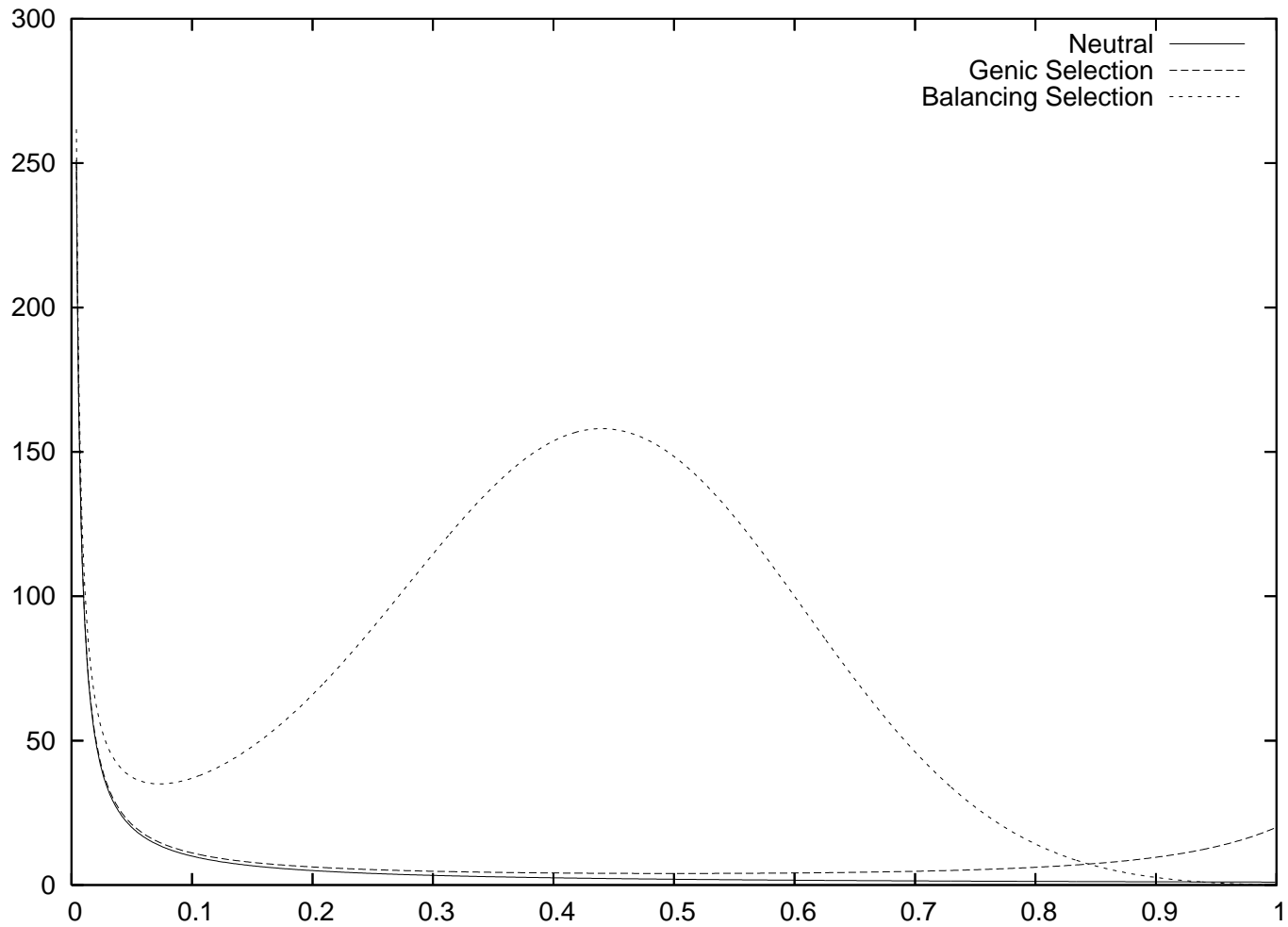
$$q_{n,b} \sim \frac{1}{x}$$

Genic selection

$$q_{n,b} \sim \frac{1}{x} \cdot \frac{1 - e^{-\beta(1-x)}}{(1 - e^{-\beta})(1 - x)}$$

Balancing selection

$$q_{n,b} \sim \frac{e^{\beta x(1-x)}}{x(1-x)} \cdot \frac{\int_x^1 e^{-\beta y(1-y)} dy}{\int_0^1 e^{-\beta y(1-y)} dy}$$



Frequency spectrum plots, $\beta = 20$

Frequency spectrum derivation

$$\begin{aligned} q_{n,b} &= \lim_{p \rightarrow 0} \frac{\int_0^\infty \int_0^1 \binom{n}{b} x^b (1-x)^{n-b} f(p, x; t) dx dt}{\int_0^\infty \int_0^1 (1-x^n - (1-x)^n) f(p, x; t) dx dt} \\ &= \frac{\int_0^1 \binom{n}{b} x^b (1-x)^{n-b} m(x) \int_0^\infty f(x, p; t) dt dx}{\int_0^1 (1-x^n - (1-x)^n) m(x) \int_0^\infty f(x, p; t) dt dx} \\ &= \frac{\int_0^1 \binom{n}{b} x^b (1-x)^{n-b} m(x) u_0(x) dx}{\int_0^1 (1-x^n - (1-x)^n) m(x) u_0(x) dx} \end{aligned}$$

The calculation supposes a uniform prior of time when the frequency is observed between when the mutation arose and was fixed or lost.