Lecture 05:
The population structure of neutral variation

UNE course:

The search for selection

3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

Marker-based tests

- We now move to test based on molecular markers, either in a candidate region (or regions) or in a genomic scan
- Trait-independent, use allele and/or gamete frequencies, not additive variation.
  - Hence, no information of effect sizes
- Need to have an understanding of the patterns expected for strictly neutral alleles

Neutral equilibrium model

- The standard neutral model, or the neutral equilibrium model, assumes
  - Strictly neutral alleles
  - The population is in mutation-drift equilibrium
  - Hence, the population size has been constant for a sufficiently long amount of time to reach mutation drift equilibrium
  - No population structure

Behavior under drift alone

- Results in one lineage ultimately becoming fixed
  - Coalescent theory
- Loss of variation (in the absence of new mutation)
- Neutral allele frequency as a function of age

Loss of heterozygosity under drift

$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t$$

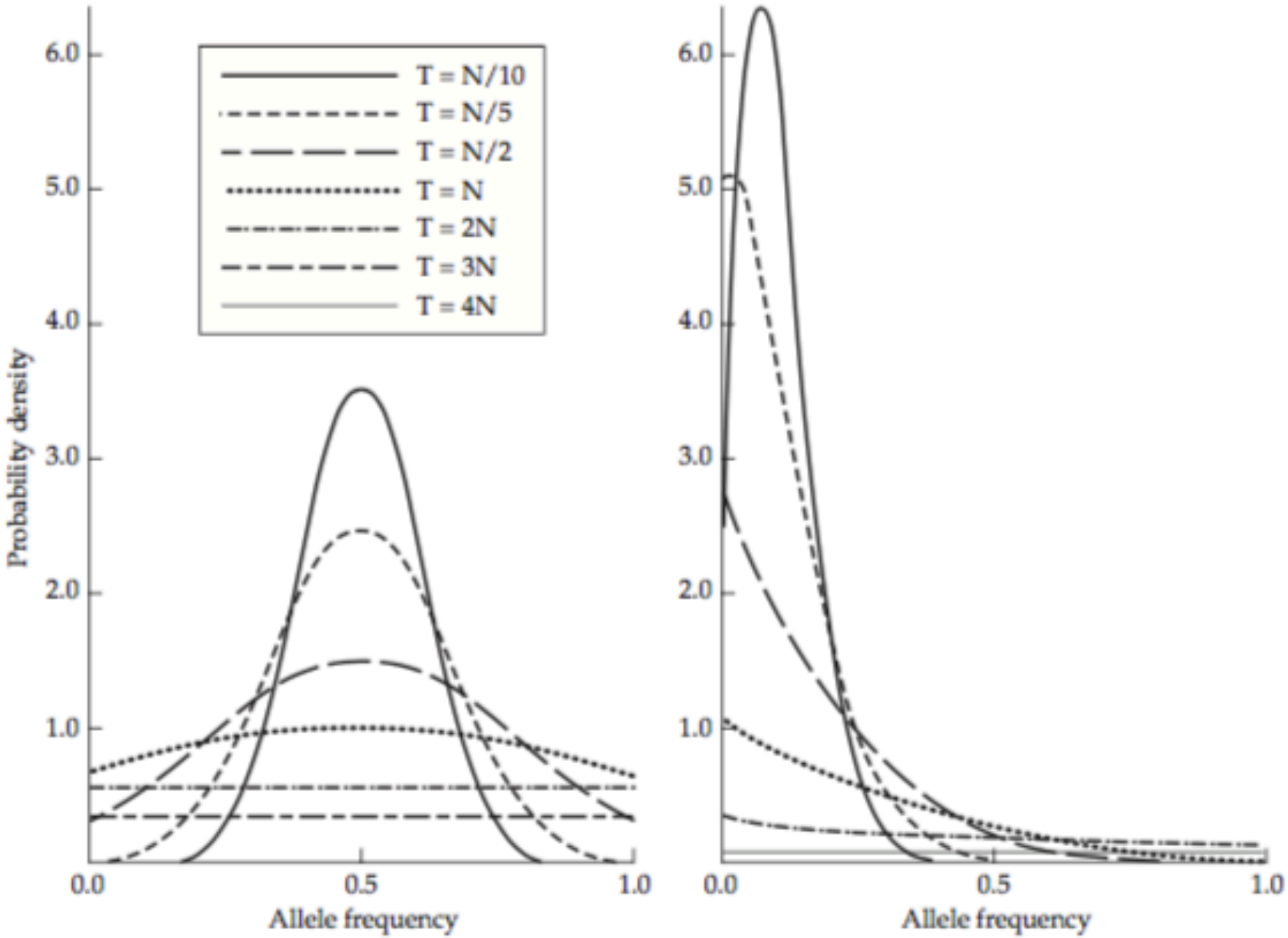
$$H_t \simeq H_0 e^{-t/(2N)}$$

$$t = -2N \ln(H_t/H_0)$$

Pure drift distribution

- Can solve using Diffusion theory (WL Appendix 1)

$$p_f(p_0,t) = p_0 + p_0(1-p_0)\sum_{i=1}^{\infty}(2i+1)(-1)^i \cdot F(1-i,i+2,2,p_0)\cdot e^{-i(i+1)t/4N}$$


T = N/10
T = N/5
T = N/2
T = N
T = 2N
T = 3N
T = 4N
Probability density
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0.5
1.0
Allele frequency
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0.5
1.0
Allele frequency

Mean time to loss or fixation

Mean sojourn time

$$\bar{t}_a(p_0) \simeq -4N[p_0 \ln(p_0) + (1-p_0)\ln(1-p_0)]$$

Mean sojourn time, conditioned on fixation

$$\bar{t}_f(p_0) \simeq -\frac{4N(1-p_0)\ln(1-p_0)}{p_0}$$

Mean sojourn time, conditioned on loss

$$\bar{t}_l(p_0) \simeq -\frac{4Np_0\ln(p_0)}{1-p_0}$$

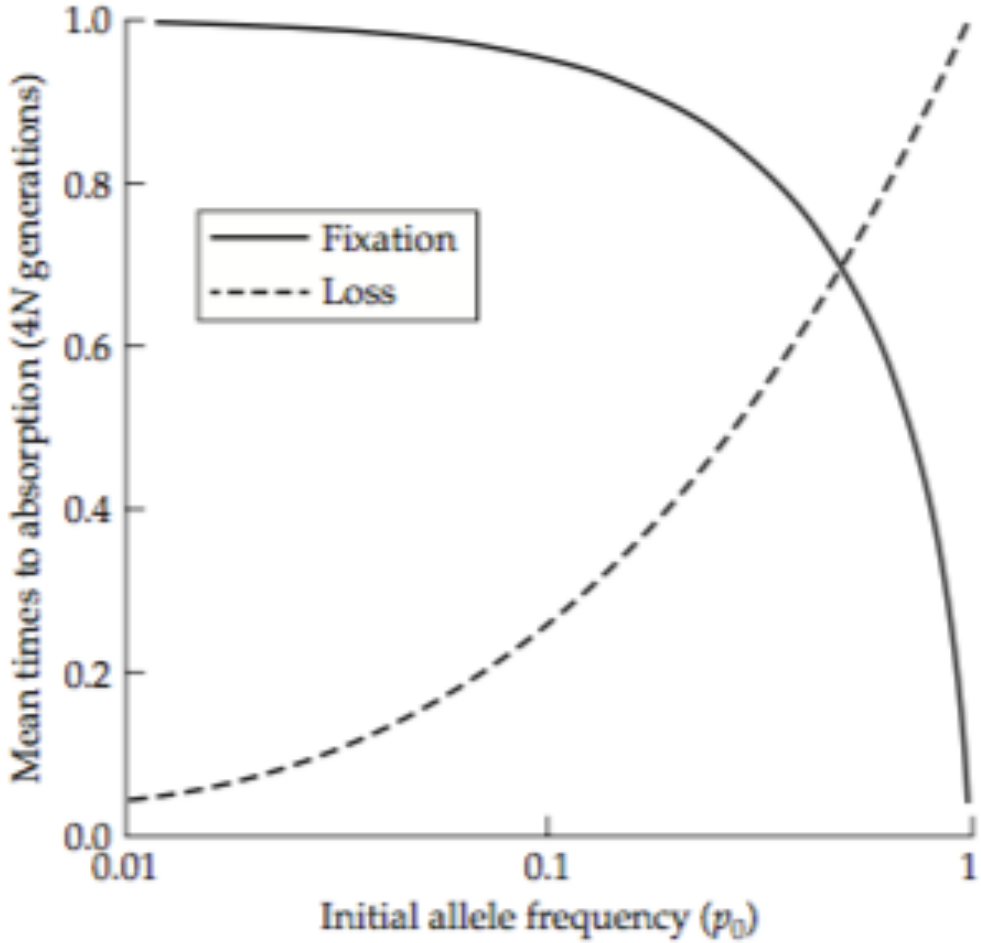


Figure 2.2 Mean times to fixation and loss of neutral alleles with starting frequency p_0 (from Equations 2.11b and 2.11c). The times are scaled in units of $4N$ generations, and thus need to be multiplied by $4N$ to obtain absolute numbers of generations.

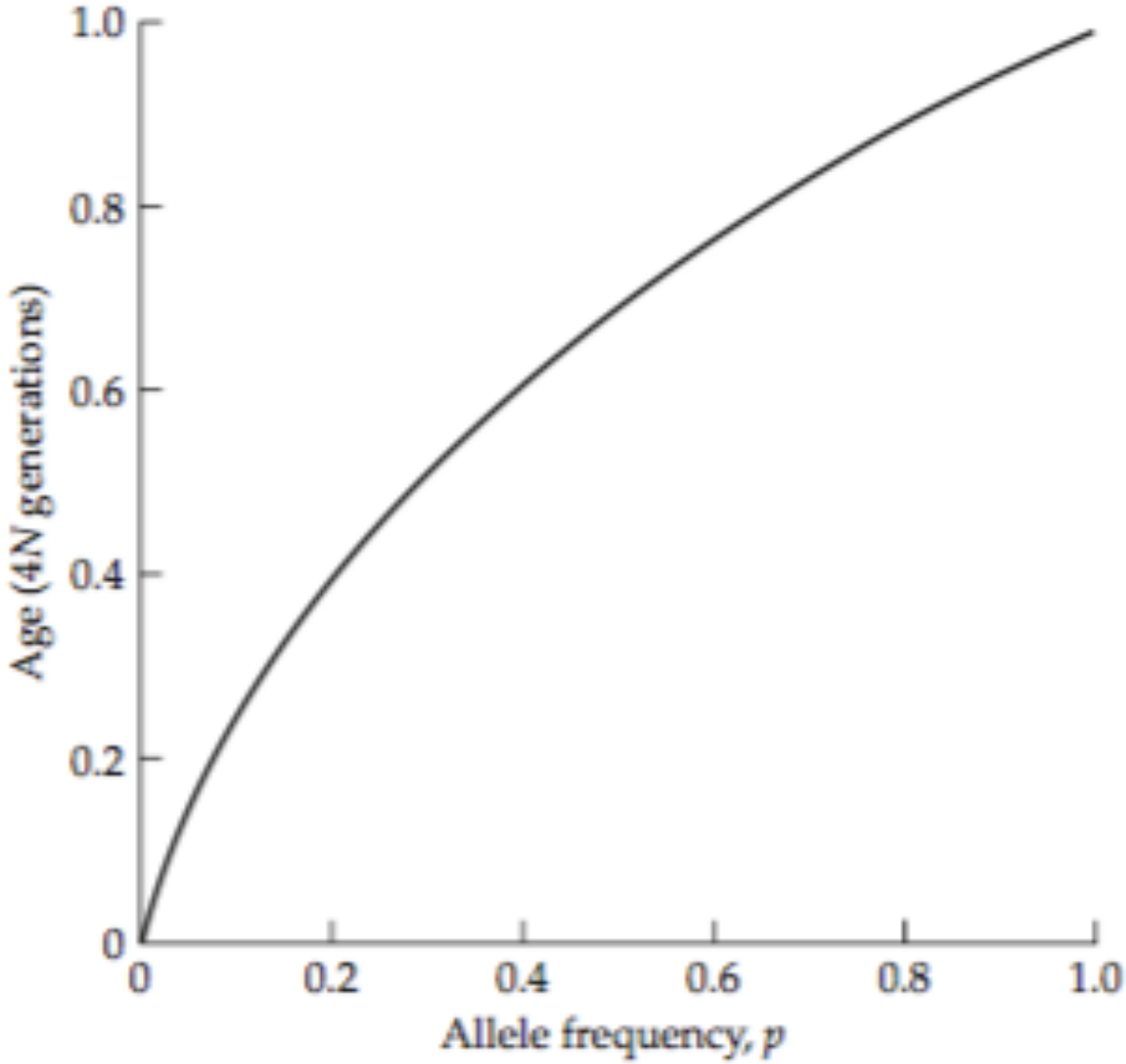
Age of a neutral allele

- A common allele is an old allele

$$E(t) = -\frac{4Np\ \ln(p)}{1-p}$$

Example 2.3. The mutation *CCR5-δ32* destroys the human *CCR5* receptor, which is used by the HIV virus to enter the cell, leading to significant resistance against HIV infection. This deletion occurs at frequencies up to 14% in Eurasians, but is absent in Africans, Native Americans, and East Asians. Assuming a frequency of $p = 0.10$ and an effective population size $N = 5000$ for Caucasians, Stephens et al. (1998) used Equation 2.12 to estimate the age of this allele (under the assumption of neutrality) to be

$$\widehat{t} = -\frac{4 \cdot 5000 \cdot 0.1\ \log(0.1)}{0.9} = 5116 \text{ generations}$$


1.0
0.8
0.6
0.4
0.2
0
Age (4N generations)
0
0.2
0.4
0.6
0.8
1.0
Allele frequency, p

Coalescent time for 2 alleles

Consider a random sample of n alleles drawn from a current population, assumed to obey all the properties of the idealized Wright-Fisher model, and with no recombination within alleles. Focusing initially on just two of the sampled alleles, we first evaluate the probability that both members of the pair are direct copies of a single allele in the preceding generation. Assuming that each individual produces a large number of gametes, because there are $2N$ gene copies in the population in each generation, this probability is simply $1/(2N)$, whereas $\lambda_1 = 1-(1/2N)$ is the probability that coalescence occurred at some earlier generation. Conditional on coalescence not having occurred in generation one, the probability of coalescence one further generation in the past is again equal to $1/(2N)$, yielding $\lambda_1(1/2N)$ as the unconditional probability of coalescence two generations back. This simple rule can be generalized to give the probability of coalescence exactly t generations in the past,

$$P_c(t) = \lambda_1^{t-1}(1/2N) \tag{2.37}$$

which defines a **geometric distribution**, with the sum of $P_c(t)$ over the interval $t = 1$ to ∞ being equal to one. One simple related point is that the probability that the **most recent common ancestor** (**MRCA**) between two sampled alleles occurred within the last t generations is $1-\lambda_1^t \simeq 1-e^{-t/2N}$, namely one minus the probability of no common ancestor over the first t generations into the past.

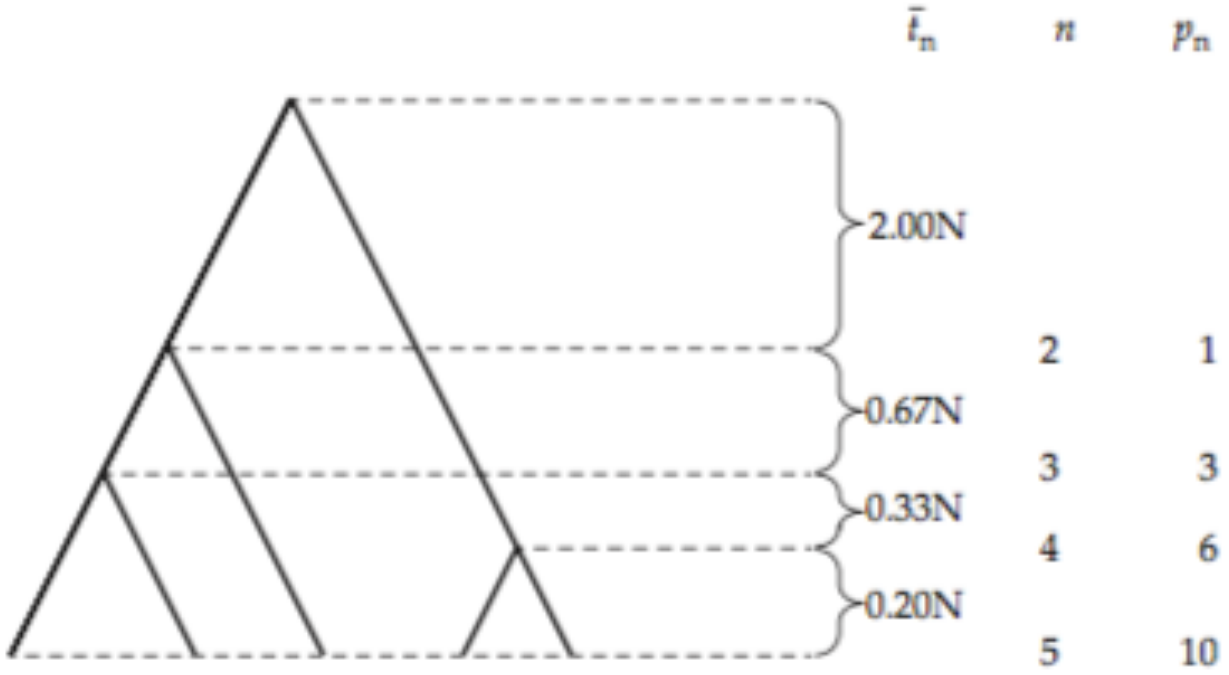
The logic used to derive this result is easily extended to the entire sample of n gene copies. There are $p_n = n(n-1)/2$ possible pairs of n copies, each of which will or will not coalesce in the preceding generation with respective probabilities $1/(2N)$ and $[1-(1/2N)]$. If the sample size is much smaller than the population size, the probability of coalescence for any pair in the sample in the preceding generation is simply the product $p_n/(2N)$. Thus, the probability distribution for the coalescence time of one pair within a set of n sequences is

$$P_c(p_n, t) = [1-(p_n/2N)]^{t-1}[p_n/(2N)] \tag{2.39}$$

Namely, a geometric random variable with success parameter $(p_n/2N)$. The mean time to coalescence of the first pair is then $2N/p_n$ generations (as opposed to $2N$ generations with a single pair). Because at this point two copies have coalesced into one, the sample size has been reduced by one, and the mean time to coalescence of the next pair is found by resetting p_n to $p_{n-1} = (n-1)(n-2)/2$. This procedure can be followed recursively down to the final pair ($p_n = 1$), which again has an expected coalescence time of $2N$ generations (Figure 2.10). The implication of these results is that the expected time for merging n random lineages into $n-1$ lineages,

$$\bar{t}_n = 2N/p_n = \frac{4N}{n(n-1)} \tag{2.40}$$

increases with decreasing sample size.



$$\bar{t}_n = 2N/p_n = \frac{4N}{n(n-1)}$$

$$\bar{t}_c(n) = \sum_{i=2}^{n} \frac{4N}{i(i-1)} = 4N\left(1 - \frac{1}{n}\right)$$

Drift and mutation

At equilibrium, key parameter is
$\theta = 4N_e u$.

- k-alleles $$E(H) = \frac{\theta}{1 + [\theta k/(k-1)]}$$
- Infinite-alleles $$E(H) \simeq \frac{\theta}{1+\theta}$$
- SNP (k=4) $$E(H) = \frac{\theta}{1 + (4/3)\theta}$$

Descriptors of neutral variation

- Single summary statistics
  - Heterozygosity
  - Nucleotide diversity
  - Number of singletons
  - Allele frequency
- Frequency spectrum (full distribution of the number of alleles and their frequencies)
  - Site-frequency spectrum (SFS): single SNPs
  - Allele-frequency spectrum (AFS): single haplotypes (SNPs + LD)

Frequency spectra

- For a sample of m sequences, this is given by $(n_1, n_2, , \ldots , n_m)$ where n_k is the number of alleles in the sample present as exactly k copies
  - Monomorphic sample, $n_1 = \cdots = n_{m-1} = 0$, $n_m = 1$
  - All singletons, $n_1 = m$, $n_2 = \cdots = n_m = 0$
  - 10 sequences: 4 singletons, 1 present as two copies, 1 present as four copies $n_1 = 4$, $n_2 = 1$, $n_4 = 1$, all others 0
  - The constraint on the n_k and m: $\sum k^* n_k = m$

Infinite alleles vs. infinite sites

- Infinitely many alleles (infinite alleles)
  - Consider a block of DNA that has no recombinants in your sample
  - Each different DNA sequence (haplotype) is a different allele
  - Requires phased data
- Infinitely many sites (infinite sites)
  - Each nucleotide is considered a different site
  - Again, no recombinants in your sample
  - Does not require phased data, but may used polarized data (ancestral vs. derived alleles)

Infinite alleles vs infinite sites

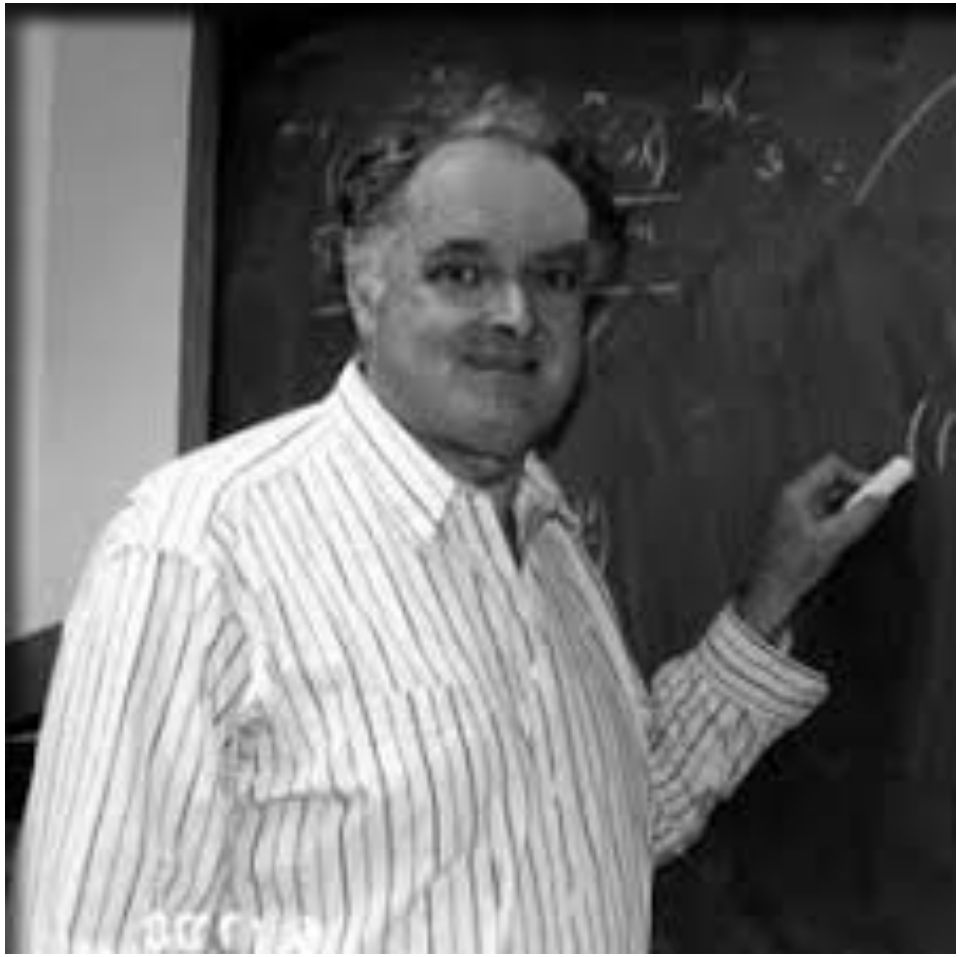
Infinite alleles: Ewen's sampling formula

- Number of alleles, k, in a sample of size n

$$\Pr(k \mid \theta_L, n) = \frac{S_n^k \, \theta_L^k}{S_n(\theta_L)}$$

$$S_n(\theta_L) = \theta_L(\theta_L + 1)(\theta_L + 2) \cdots (\theta_L + n - 1)$$

Warren Ewens



Ewen's (cont)

- Prob. Monomorphic

$$\Pr(k=1) = \frac{(n-1)!}{(\theta_L+1)(\theta_L+2)\cdots(\theta_L+n-1)}$$

- Mean and variance in k

$$E(k) = 1 + \theta_L \cdot \sum_{j=2}^{n} \frac{1}{\theta_L + j - 1}, \qquad \sigma^2(k) = \theta_L \cdot \sum_{j=1}^{n-1} \frac{j}{(\theta_L + j)^2}$$

Allele-frequency spectrum

- Let n_i = number of alleles with exactly i copies in sample (size n)

$$\sum_{i=1}^{n} i \cdot n_i = n$$

$$\Pr(n_1, n_2, \cdots, n_n \,|\, n,\, k) = \frac{n!}{S_n^k \, (1^{n_1} 2^{n_2} \cdots n^{n_n}) n_1! n_2! \cdots n_n!}$$

$$\Pr(m_1, \cdots, m_k, k \,|\, n, \theta_L) = \frac{n! \, \theta_L^k}{k! \, (m_1 \, m_2 \, \cdots \, m_k) S_n(\theta_L)}$$

Infinite sites

- Ancestral (original) vs. derived (new mutation)
- Nucleotide diversity, π
- Number of segregating sites, S
- Site-frequency spectrum
  - Number, s_j, of sites with exactly j derived alleles in the sample

Nucleotide Diversity

Suppose a population sample of n random sequences has been obtained for a particular genomic region. In principle, such a stretch of DNA might consist of intronic or intergenic sequence or of the subset of **silent (synonymous)** sites in one or more coding regions. Letting k_{ij} be the number of site-specific differences between observed sequences i and j, and L be the number of sites per sequence, the average fraction of pairwise differences between the sampled sites,

$$\widehat{\theta}_{\pi} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i}^{n} k_{ij}/L \tag{4.1}$$

yields a heterozygosity-based estimate of $\theta = 4N_e u$ (Tajima 1983). This formulation, frequently called the **Tajima estimator**, is often denoted by π in the literature.

Number of Segregating Sites

Although nucleotide diversity is the most transparent means of estimating θ, it is by no means the only, or even the most efficient, approach. Watterson (1975) pointed out an alternative statistical measure of allelic diversity—the total number of segregating sites (S) in the region analyzed over the full set of n sequences. Because a segregating site is any nucleotide position that harbors two or more variants, S clearly increases with the length L of the sequence and the number of individuals assayed. Watterson (1975) showed that under the assumptions of neutrality and drift-mutation equilibrium, an unbiased estimator of the per-site parameter $\theta = 4N_e u$ is

$$\widehat{\theta}_S = S/(L\,a_n) \tag{4.3a}$$

where

$$a_n = \sum_{j=1}^{n-1} 1/j \tag{4.3b}$$

By rearranging, it can be seen that Equation 4.3a relates directly to the expected site-frequency spectrum for a sample under drift-mutation equilibrium with a known value of θ (Equation 2.35a). A central point here is that when the nucleotide sites surveyed are neutral and in drift-mutation equilibrium, like the Tajima estimator (Equation 4.1), the **Watterson estimator** provides a separate estimate of θ. In Chapter 9, we will see that when the assumptions of neutrality and/or equilibrium are violated, the values of $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ deviate from each other in ways that yield insight into past population-genetic processes.

Site frequency spectrum

- Can be unfolded or folded.
- Unfolded SFS assumes the polarity of the alleles are known
- Folded SFS simply uses the minor allele frequency
- Can express the SFS as either
  - the fraction, x, of sites in a particular allele frequency in the population,
  - Or the number, n_k, of sites with k derived alleles in a sample

Watterson distribution

- Let x = population frequency of all sites with a fraction of x derived alleles

$$\phi(x) = \frac{\theta}{x} \quad \text{for} \quad \frac{1}{2N} \leq x \leq 1 - \frac{1}{2N}$$

Folded Watterson distribution, x = freq of minor allele ($x \leq 0.5$)

$$\phi(x) = \frac{\theta}{x} + \frac{\theta}{1-x} = \frac{\theta}{x(1-x)} \quad \text{for} \quad \frac{1}{2N} \leq x \leq 1/2$$

Expected number of sites in a sample

unfolded

$$E(s_i) = \frac{\theta_L}{i}, \quad \text{for} \quad 1 \leq i \leq n-1$$

folded

$$E(s_i) = \frac{\theta_L}{i} + \frac{\theta_L}{n-i} = \frac{\theta_L n}{i(n-i)}, \quad \text{for} \quad 1 \leq i \leq [n/2]$$