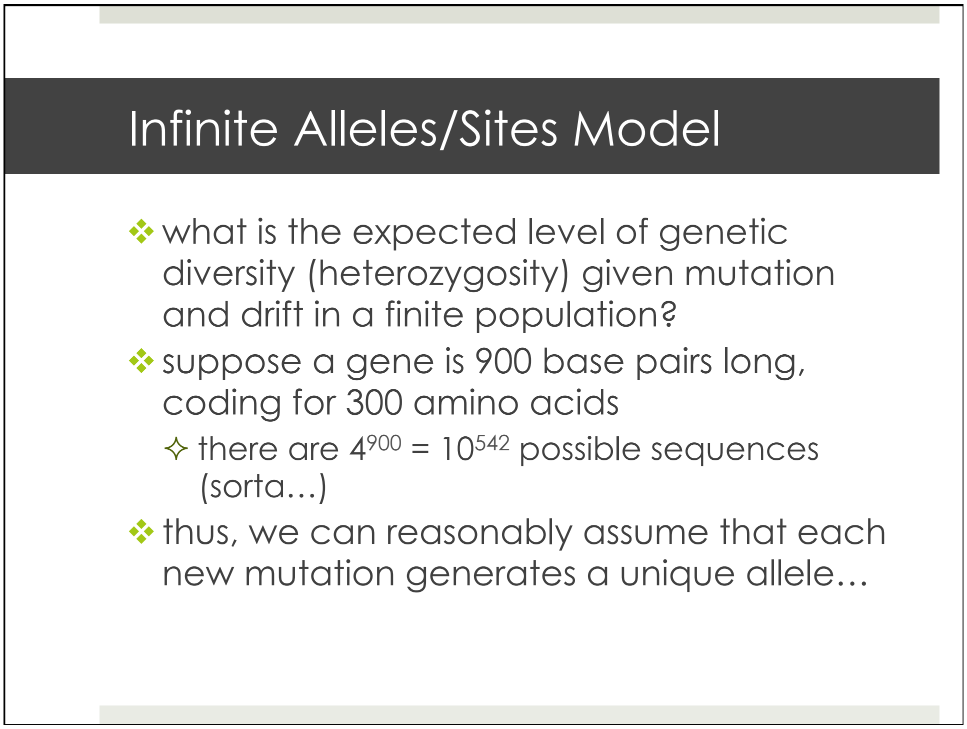
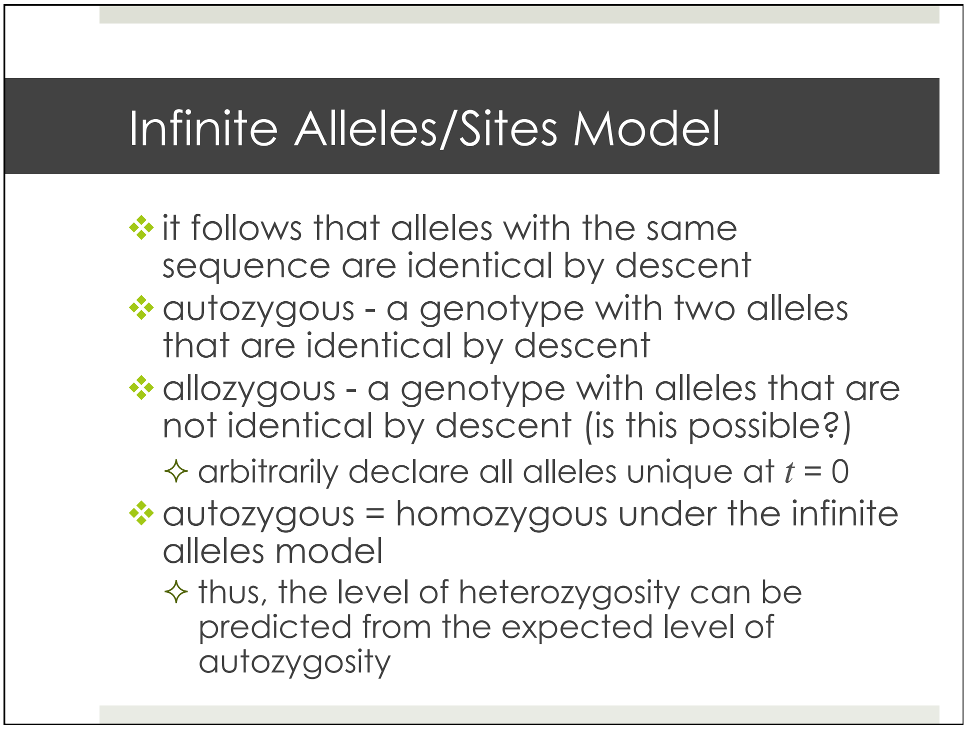


## Infinite Alleles/Sites Model

- what is the expected level of genetic diversity (heterozygosity) given mutation and drift in a finite population?
- suppose a gene is 900 base pairs long, coding for 300 amino acids
  - there are $4^{900} = 10^{542}$ possible sequences (sorta…)
- thus, we can reasonably assume that each new mutation generates a unique allele…



## Infinite Alleles/Sites Model

- it follows that alleles with the same sequence are identical by descent
- autozygous - a genotype with two alleles that are identical by descent
- allozygous - a genotype with alleles that are not identical by descent (is this possible?)
  - arbitrarily declare all alleles unique at $t = 0$
- autozygous = homozygous under the infinite alleles model
  - thus, the level of heterozygosity can be predicted from the expected level of autozygosity

## Infinite Alleles/Sites Model

- $F_t$ = probability that two randomly chosen alleles are *IBD*
  - same as autozygosity if we randomly choose alleles to form genotypes

$$F_t = \left(\frac{1}{2N}\right)(1-\mu)^2 + \left(1-\frac{1}{2N}\right)(1-\mu)^2 F_{t-1}$$

  - in this model, mutations generate new alleles and "erase" *IBD*

## Infinite Alleles Model

- in a random-breeding population of constant size, an equilibrium is reached where the increase in autozygosity (~*IBD*) due to loss of alleles by drift is exactly countered by the increase in heterozygosity produced by new mutations
- solving for $F_t = F_{t-1}$ yields:

$$\hat{F} = \frac{1}{1+4N\mu}$$

## Infinite Alleles Model

$$\hat{F} = \frac{1}{1+4N\mu} = \frac{1}{1+4N_e\mu} = \frac{1}{1+\theta}$$

- given the assumption of infinite alleles, any genotype that is not autozygous is heterozygous, so

$$\hat{H} = 1-\hat{F} = \frac{4N_e\mu}{1+4N_e\mu} = \frac{\theta}{1+\theta}$$

## Neutral Expectations for Genetic Diversity

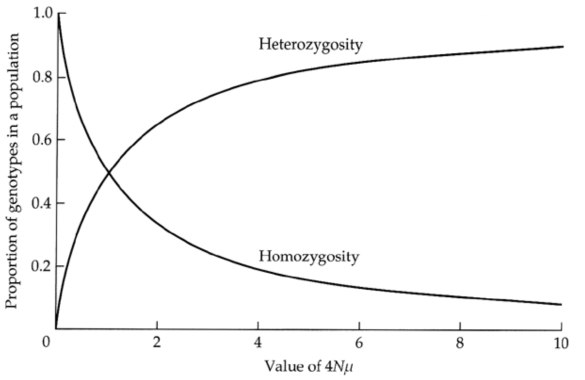


**FIGURE 4.7** Plot of average homozygosity and average heterozygosity for the infinite-alleles model. Intermediate values of heterozygosity are maintained over only a small range of $\theta = 4N_e\mu$.

## Neutral Expectations for Genetic Diversity

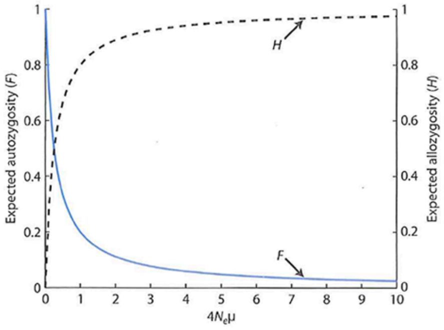


**Figure 5.11** Expected homozygosity (*F* or autozygosity, solid line) and heterozygosity (*H* or allozygosity, dashed line) at equilibrium in a population where the processes of both genetic drift and mutation are operating. The chance that two alleles sampled randomly from the population are identical in state depends on the net balance of genetic drift working toward fixation of a single allele in the population and mutation changing existing alleles in the population to new states. A critical assumption is the infinite alleles model, which guarantees that each mutation results in a unique allele and thereby maximizes the allozygosity due to mutations.

$$\hat{F} = \frac{1}{1+4N\mu}$$

- although an ideal population is expected to reach an "equilibrium" value of *F*, the population is not really *at equilibrium*, but rather in a "dynamic steady state" because there is a continual turnover of alleles
  - the most common allele is periodically replaced by another, other alleles are lost, and new alleles are produced by mutation

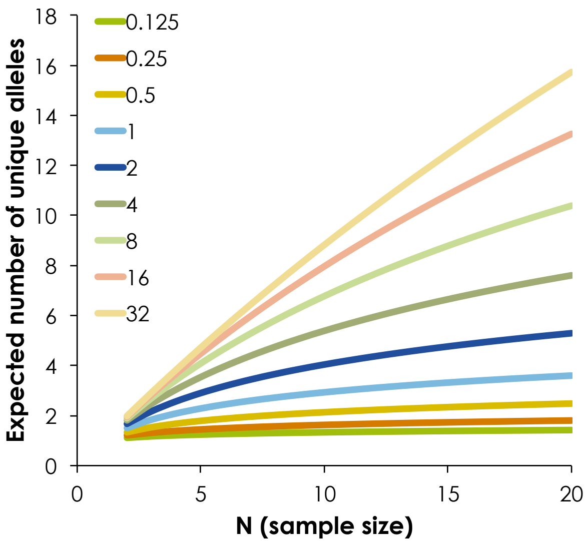
# The Ewens Distribution

- beyond $F$, there is additional "information" available in the number of alleles present and the distribution of allele frequencies
  - "allelic configuration"
  - or "allele-frequency spectrum"
- Ewens (1972) - expected number of alleles $k$ in a sample of size $n$, depends only on $\theta$

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+n-1}$$

Values of $\theta = 4N\mu$

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+n-1}$$



## The Ewens Sampling Formula

- but there's more than just $k$ (number of alleles)
- the Ewens Distribution specifies the probability distribution on the set of all ***partitions*** of the integer $n$
  - a.k.a. the "Chinese Restaurant" problem
  - broad applicability outside of population genetics

## Partitions of 8

- 8
- 7 + 1
- 6 + 2
- 6 + 1 + 1
- 5 + 3
- 5 + 2 + 1
- 5 + 1 + 1 + 1
- 4 + 4
- 4 + 3 + 1
- 4 + 2 + 2
- 4 + 2 + 1 + 1
- 4 + 1 + 1 + 1 + 1
- 3 + 3 + 2
- 3 + 3 + 1 + 1
- 3 + 2 + 2 + 1
- 3 + 2 + 1 + 1 + 1
- 3 + 1 + 1 + 1 + 1 + 1
- 2 + 2 + 2 + 2
- 2 + 2 + 2 + 1 + 1
- 2 + 2 + 1 + 1 + 1 + 1
- 2 + 1 + 1 + 1 + 1 + 1 + 1
- 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1

## Ewens' Sampling Formula (from Wikipedia!)

- Ewens' result provided the basis for a formula (Karlin & McGregor, 1972) giving the probability of a given allele frequency configuration (note: this is just one formulation)…

$$\Pr\{a_1,...,a_n\} = \frac{n!}{\theta(\theta+1)\ ...\ (\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{j^{a_j} a_j!}$$

where $a_1,...,a_n$ are counts of the number of alleles represented one, two,..., $n$ times in the sample. $a_1,...,a_n$ are nonnegative integers that satisfy: $a_1 + 2a_2 + 3a_3 + ... + na_n = n$

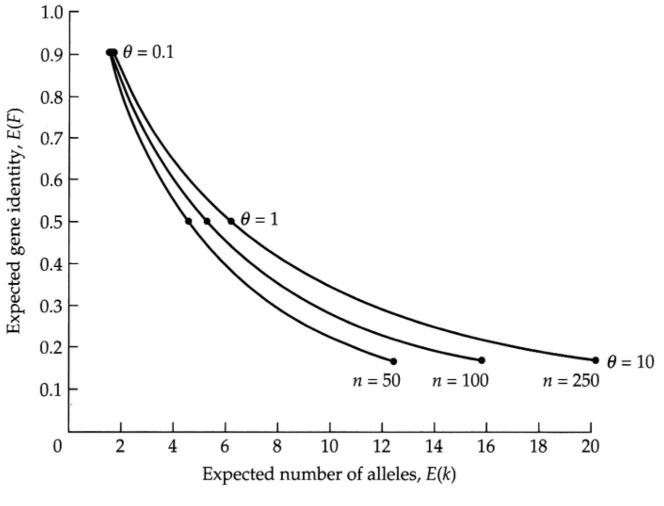


**FIGURE 4.9** The infinite-alleles model prediction of the relation between the expected number of alleles $E(k)$ and the expected gene identity (homozygosity) $F$. The three curves represent a range of values of $\theta = 4N_e\mu$, starting at $\theta = 0.1$ in the upper left, and ending with $\theta = 10$ in the lower right. For the value of $\theta = 1$, the expected $F$, given by the relation $F = 1/(1 + \theta)$, is $\frac{1}{2}$, regardless of the sample size. Larger sample sizes always lead to larger expected numbers of alleles, but the difference is greater in more diverse populations (those with smaller $F$).

observed: 52+9+8+4+4+2+2+1+1+1+1+1+1+1+1+1

## Ewens-Watterson Test

- Ewens-Watterson Test compares the allelic configuration of a sample to neutral expectations (generated with the Ewens sampling formula) using $F$ (homozygosity as a test statistic)

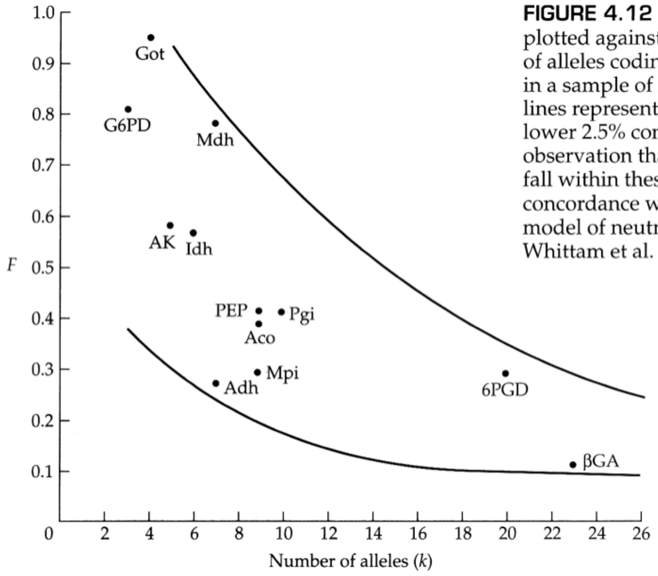


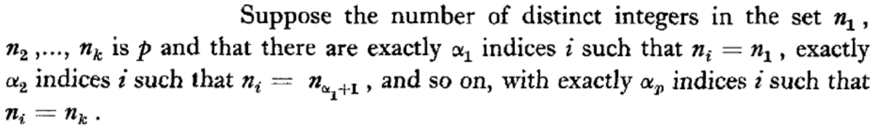
**FIGURE 4.12** Gene identity ($F$) plotted against the observed number of alleles coding for various proteins in a sample of 279 *E. coli*. The solid lines represent the upper 97.5% and lower 2.5% confidence limits, and the observation that all of the tested loci fall within these limits suggests good concordance with the infinite-alleles model of neutral mutation. (From Whittam et al. 1983.)

## Karlin & McGregor 1972

- this equation works too…

$$\Pr(\theta\ ;\ n_1,\ n_2,\ldots,\ n_k) = \frac{r!}{n_1 n_2 \ldots n_k}\,\frac{1}{\alpha_1!\alpha_2!\ldots\alpha_p!}\,\frac{\theta^k}{L_r(\theta)}$$

$$\text{where } L_r(\theta) = \theta(\theta+1)(\theta+1)\ \ldots\ (\theta+r-1)$$



Suppose the number of distinct integers in the set $n_1$, $n_2$,..., $n_k$ is $p$ and that there are exactly $\alpha_1$ indices $i$ such that $n_i = n_1$, exactly $\alpha_2$ indices $i$ such that $n_i = n_{\alpha_1+1}$, and so on, with exactly $\alpha_p$ indices $i$ such that $n_i = n_k$.

## Ewens' Sampling Formula

- *key point:* the Ewens distribution provides a basis for testing observed data against the neutral model

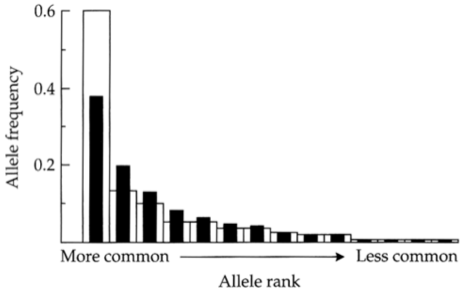
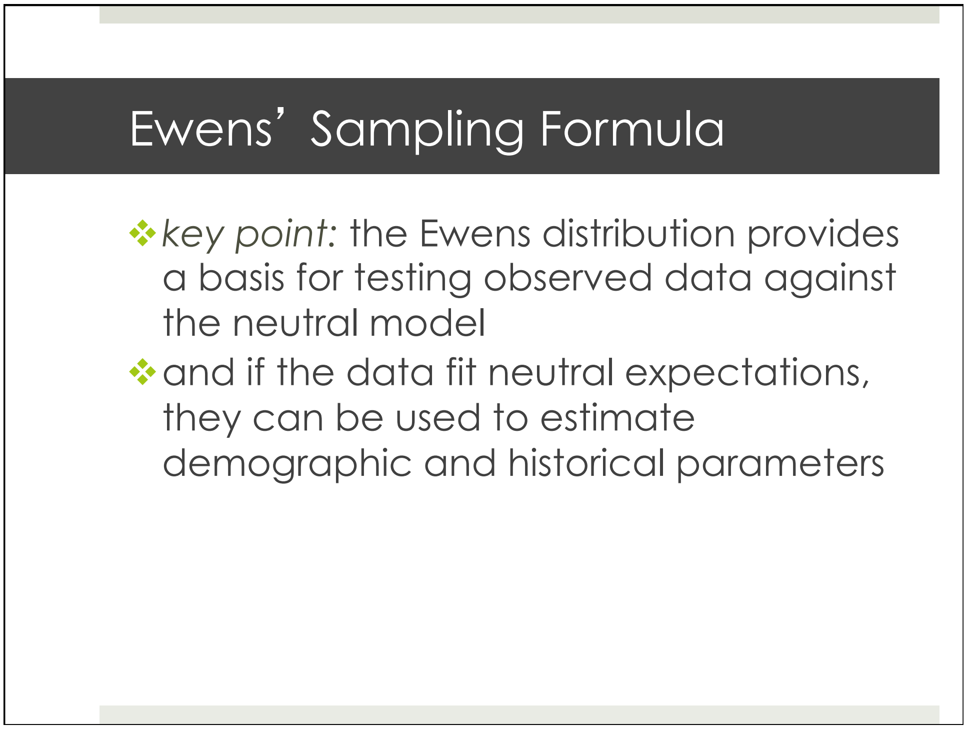
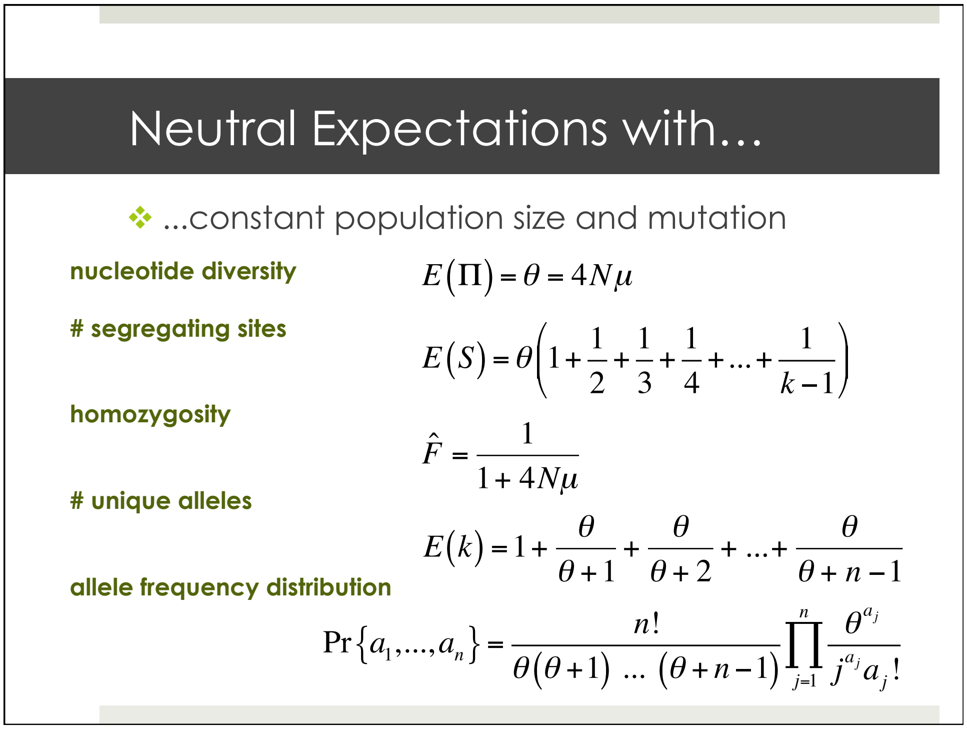


**FIGURE 4.10** Observed (open columns) and expected (black bars) allele frequency spectrum of the *HRAS-1* gene in humans, identified by Southern blotting with the *pLM0.8* probe and *Taq*I digests. Observed data are from Baird et al. (1986). The expected distribution was generated using the Ewens sampling formula. In this sample of 490 genes there were 14 distinct alleles, four of which were present in just one individual. (From Clark 1988.)



## Ewens' Sampling Formula

- *key point:* the Ewens distribution provides a basis for testing observed data against the neutral model
- and if the data fit neutral expectations, they can be used to estimate demographic and historical parameters



## Neutral Expectations with…

- ...constant population size and mutation

**nucleotide diversity**

$$E(\Pi) = \theta = 4N\mu$$

**# segregating sites**

$$E(S) = \theta\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{k-1}\right)$$

**homozygosity**

$$\hat{F} = \frac{1}{1 + 4N\mu}$$

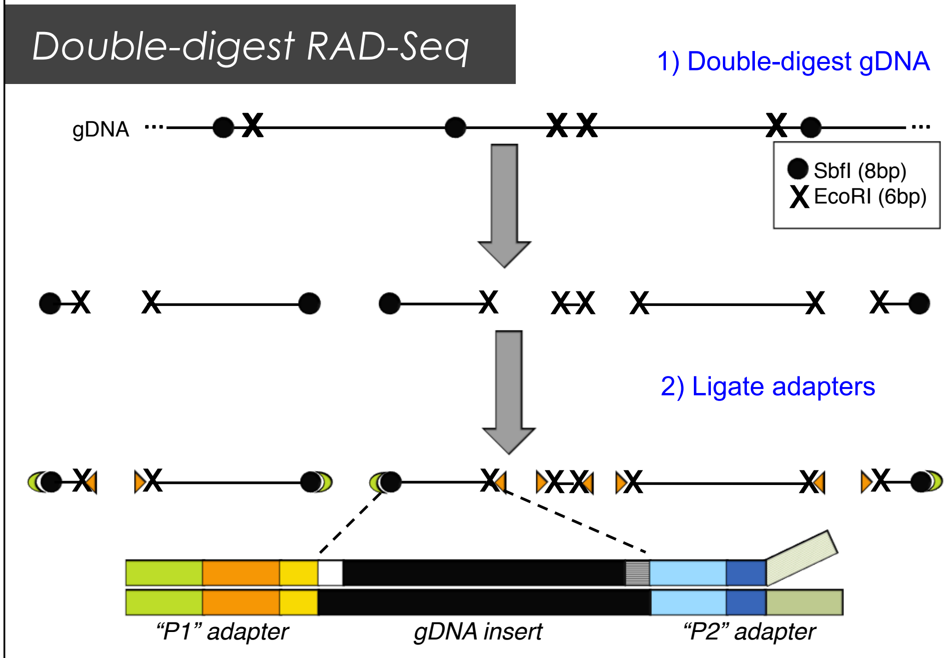
**# unique alleles**

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+n-1}$$

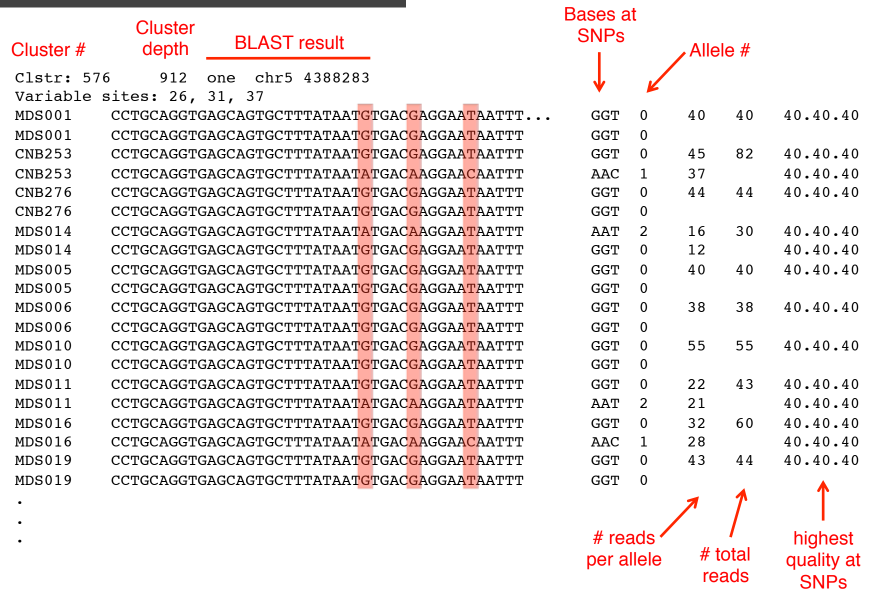
**allele frequency distribution**

$$\Pr\{a_1,\ldots,a_n\} = \frac{n!}{\theta(\theta+1)\ \ldots\ (\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{j^{a_j} a_j!}$$

## Double-digest RAD-Seq



## ddRAD-Seq output


Cluster #
Cluster depth
BLAST result
Bases at SNPs
Allele #
# reads per allele
# total reads
highest quality at SNPs

```
Clstr: 576      912  one  chr5 4388283
Variable sites: 26, 31, 37
MDS001    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT...   GGT  0    40   40   40.40.40
MDS001    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0
CNB253    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    45   82   40.40.40
CNB253    CCTGCAGGTGAGCAGTGCTTTATAATATGACAAGGAACAATTT      AAC  1    37        40.40.40
CNB276    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    44   44   40.40.40
CNB276    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0
MDS014    CCTGCAGGTGAGCAGTGCTTTATAATATGACAAGGAATAATTT      AAT  2    16   30   40.40.40
MDS014    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    12        40.40.40
MDS005    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    40   40   40.40.40
MDS005    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0
MDS006    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    38   38   40.40.40
MDS006    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0
MDS010    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    55   55   40.40.40
MDS010    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0
MDS011    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    22   43   40.40.40
MDS011    CCTGCAGGTGAGCAGTGCTTTATAATATGACAAGGAATAATTT      AAT  2    21        40.40.40
MDS016    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    32   60   40.40.40
MDS016    CCTGCAGGTGAGCAGTGCTTTATAATATGACAAGGAACAATTT      AAC  1    28        40.40.40
MDS019    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0    43   44   40.40.40
MDS019    CCTGCAGGTGAGCAGTGCTTTATAATGTGACGAGGAATAATTT      GGT  0
.
.
.
```