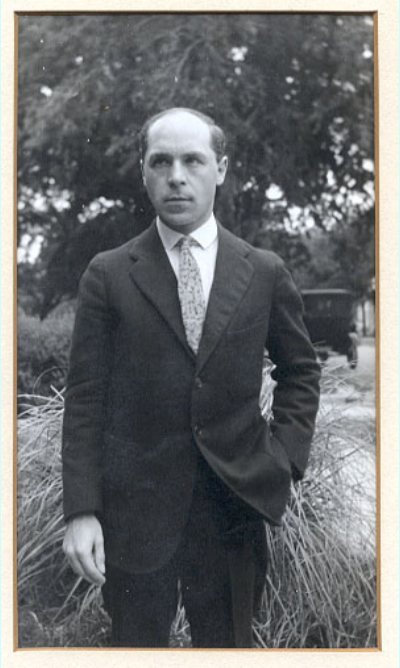
Molecular variation

Joe Felsenstein

GENOME 453, Autumn 2013

Views of genetic variation before 1966

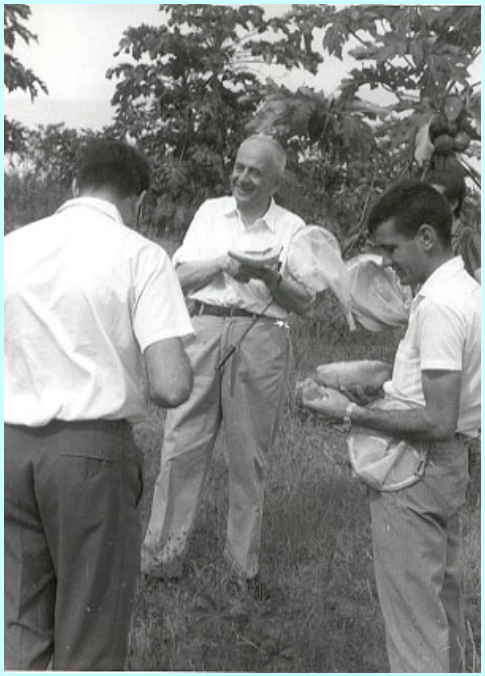
The Classical view



Hermann Joseph Muller

Most loci will be homozygous
for the “wild-type allele"
but a few mutants will exist

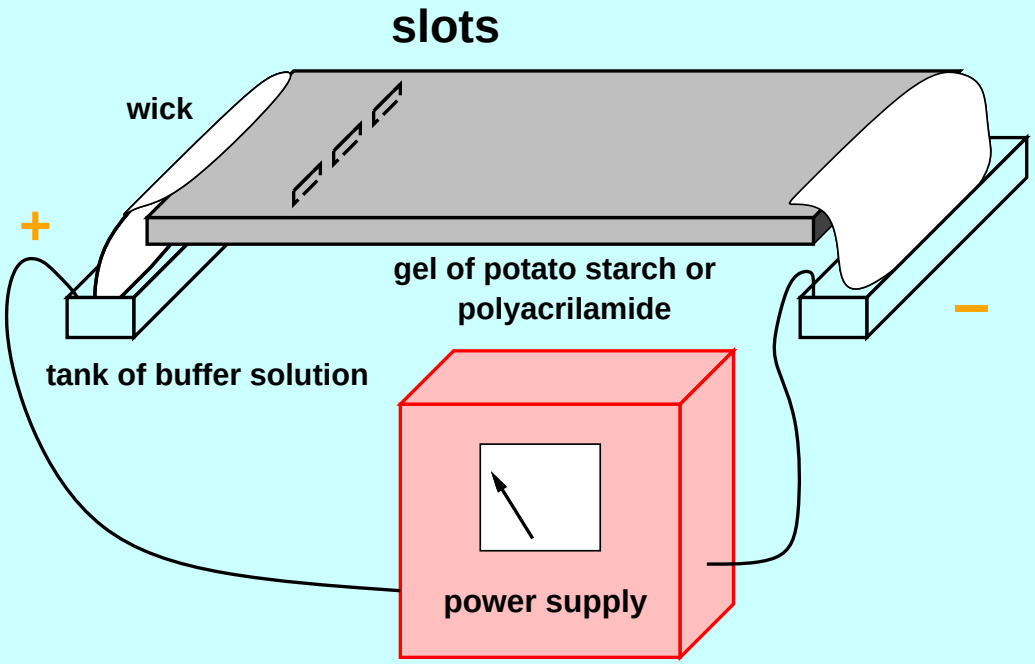
The Balancing Selection view



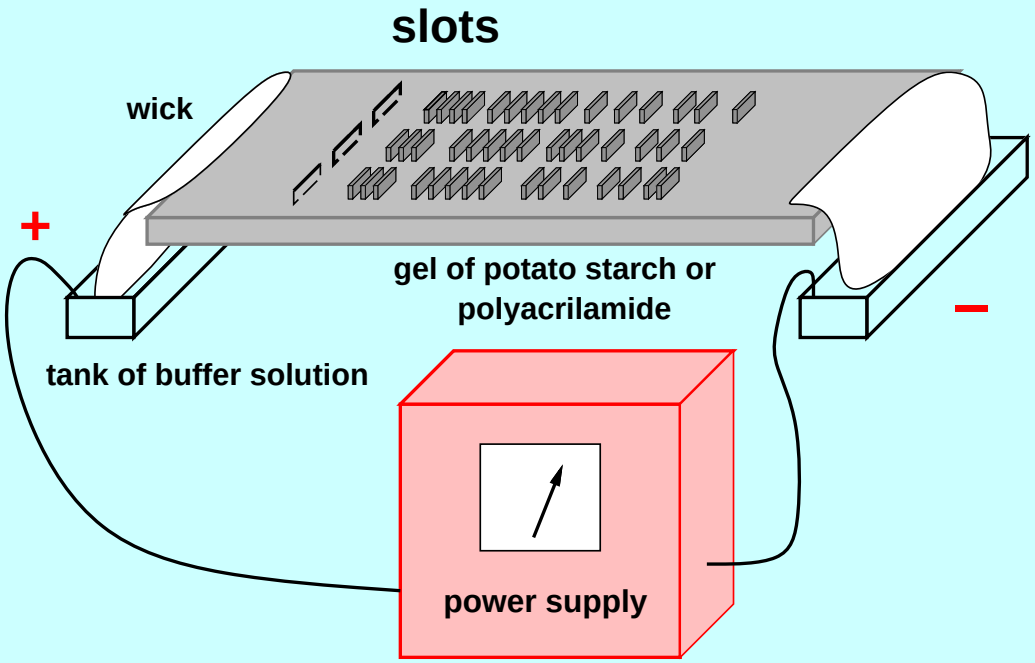
Theodosius Dobzhansky

Most loci will be polymorphic
due to balancing selection
with strong selection

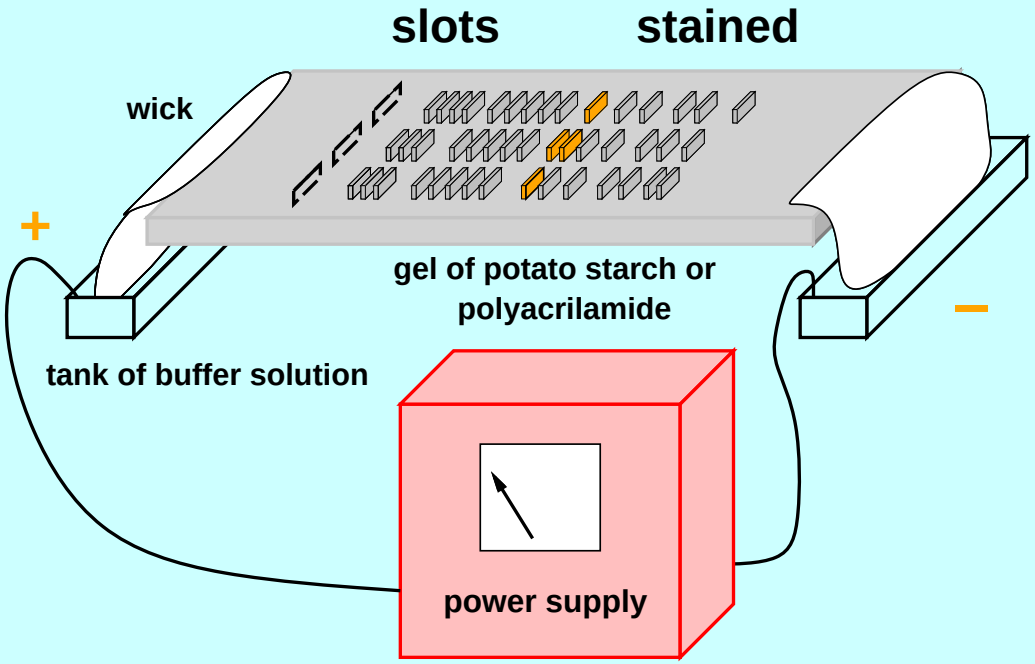
Gel electrophoresis

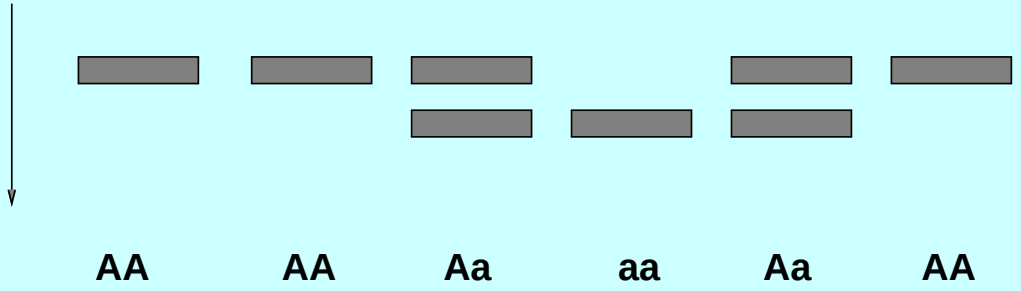


After running the current

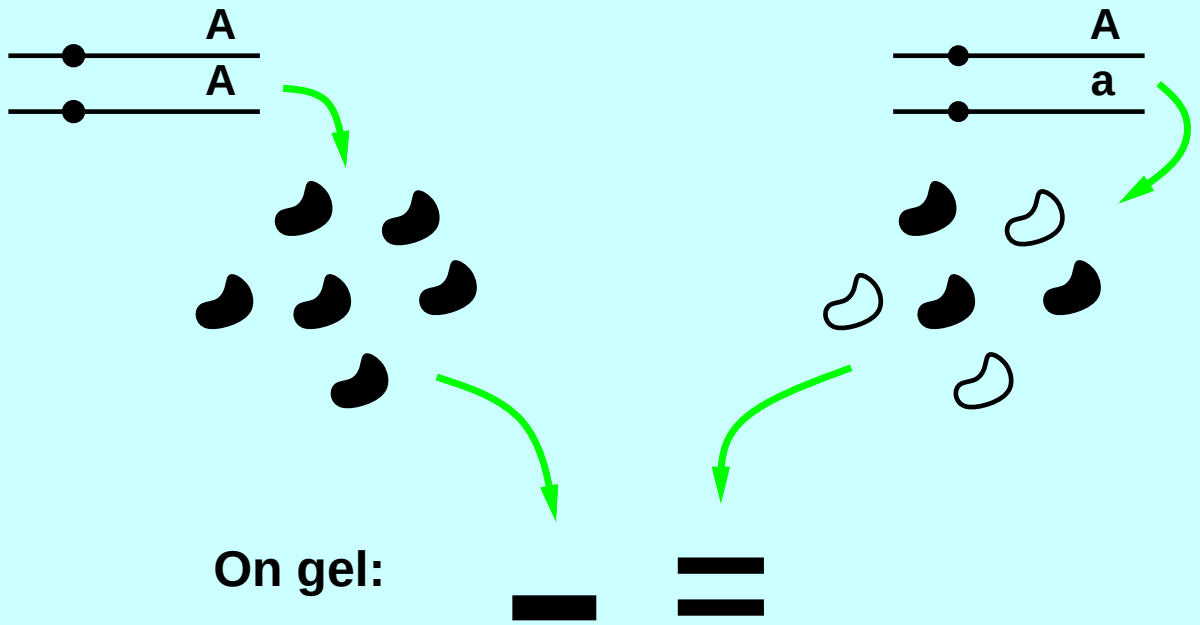


Making one locus visible by staining

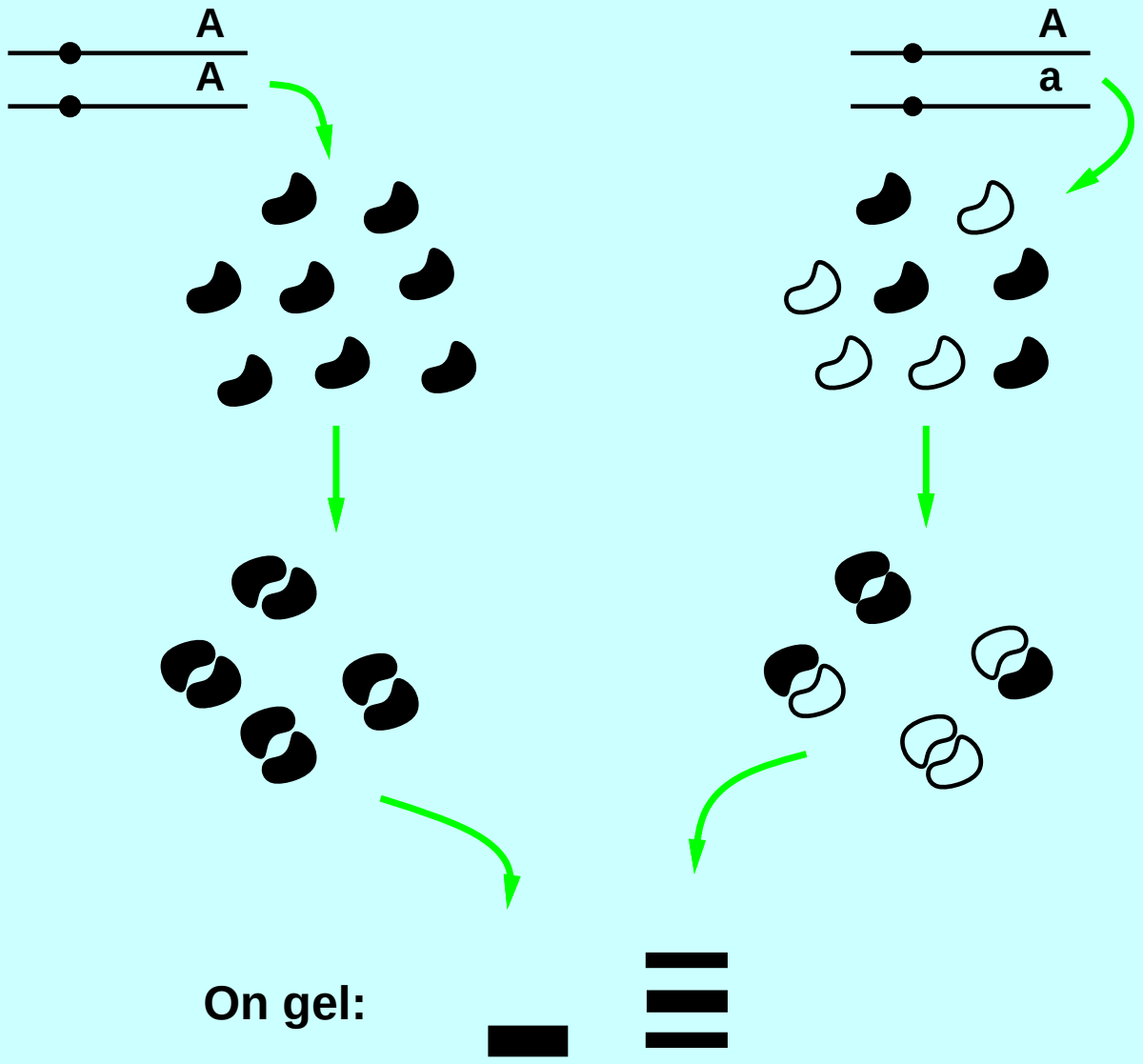




A monomeric enzyme



A dimeric enzyme



Polymorphism on a gel

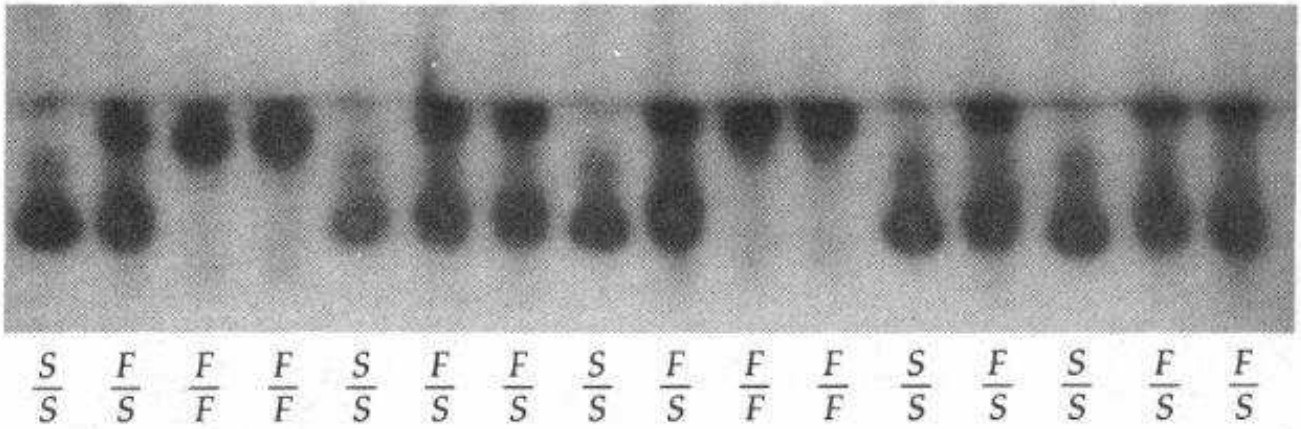


Figure 4. Results of electrophoresis of the enzyme glucose phosphate isomerase-1 from 16 cultured cell lines originating from individuals of the mouse, *Mus musculus*. The gene that codes for the enzyme is *Gpi-1*. In this sample, some individuals are homozygous for an allele (*S*) corresponding to a slow-migrating enzyme, some are homozygous for an allele (*F*) corresponding to a fast-migrating enzyme, and the rest are heterozygous (*F*/*S*). The inferred genotypes of the cell lines are indicated beneath the enzyme bands. This enzyme is a monomer, so the heterozygotes exhibit two enzyme bands of differing mobility. (Courtesy of S. E. Lewis and F. M. Johnson.)

Lewontin and Hubby's 1966 work

Richard Lewontin, about 1980

Lewontin, R. C. and J. L. Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* **54:** 595-609.

Measures of variability with multiple loci

Lewontin and Hubby (Genetics, 1966) suggested two measures of variability: polymorphism and heterozygosity.

Polymorphism is the fraction of all loci that have the most common allele less than 0.95 in frequency (i.e. all the rarer alleles together add up to less than 0.05.

Heterozygosity is the estimated fraction of all individuals who are heterozygous at a random locus.

Computing the average heterozygosity

If p_i is the frequency in the sample of allele i at a locus, then the heterozygosity for that locus is estimated by taking the sum of squares of the gene frequencies (thus estimating the homozygosity) and subtracting from 1:

$$H = 1 - \sum_{i}^{\text{alleles}} p_i^2$$

An example:

locus 1	1
locus 2	1
locus 3	0.8 0.2
locus 4	0.94 0.04 0.02

The heterozygosities are calculated as:

locus 1	$1 - 1^2 = 0$
locus 2	$1 - 1^2 = 0$
locus 3	$1 - \left(0.8^2 + 0.2^2\right) = 0.32$
locus 4	$1 - \left(0.94^2 + 0.04^2 + 0.02^2\right) = 0.1144$

The average heterozygosity in this example is

$$\mathrm{H} \;=\; (0 + 0 + 0.32 + 0.1144) \;/\; 4 \;=\; 0.1086$$

Amounts of heterozygosity

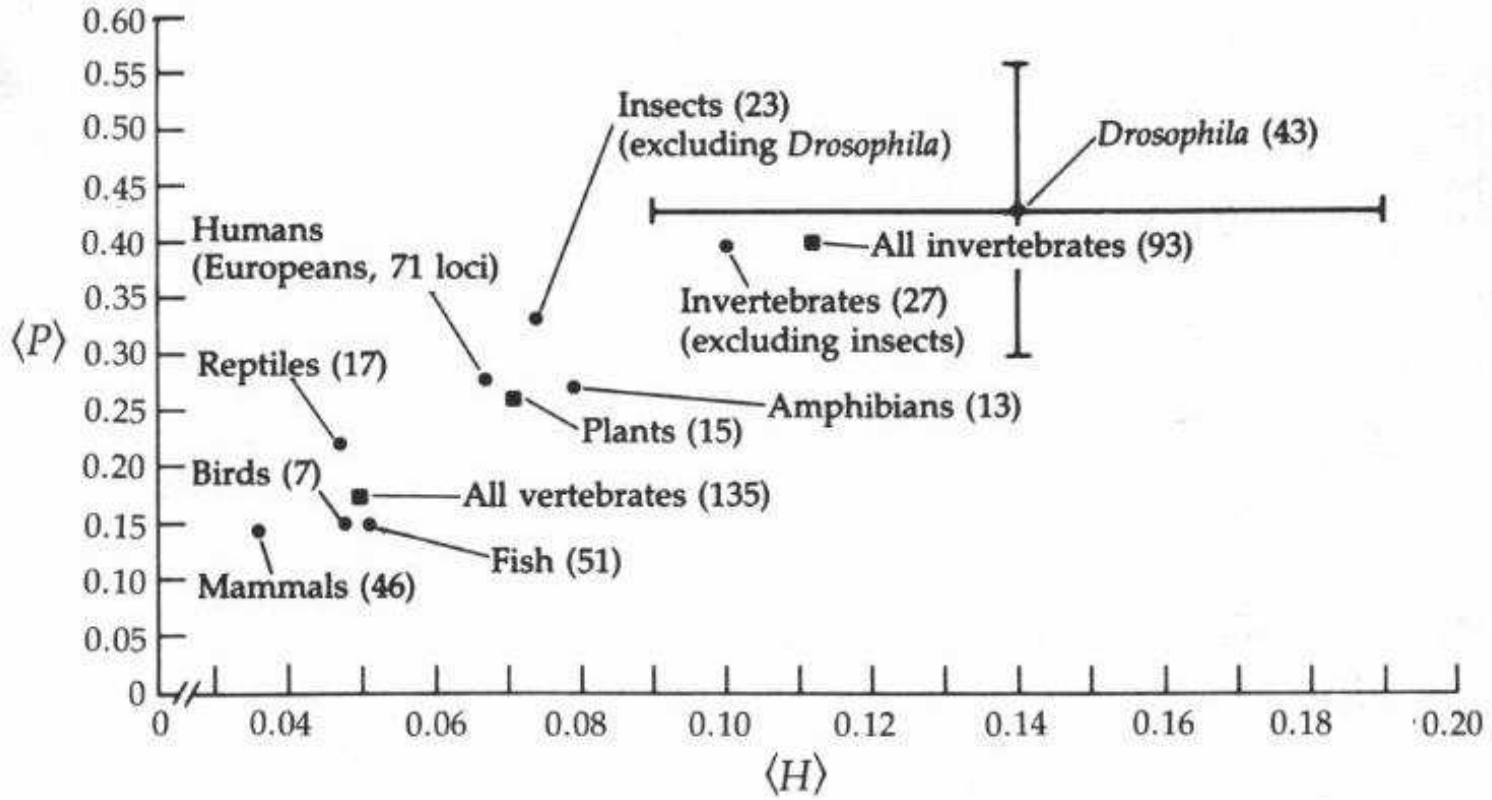


Figure 10. Estimated levels of heterozygosity ($\langle H \rangle$) and proportion of polymorphic genes ($\langle P \rangle$) derived from allozyme studies of various groups of plants and animals. The number of species studied is shown in parentheses beside each point. Squares denote averages for plants, invertebrates, and vertebrates. The bars across the *Drosophila* point show the range of H and P within which 68% of the *Drosophila* species fall. Other groups would have similarly large bars. (Data from Nevo 1978.)

Kimura's neutral mutation theory

Motoo Kimura and family, 1966

Tomoko Ohta, recently

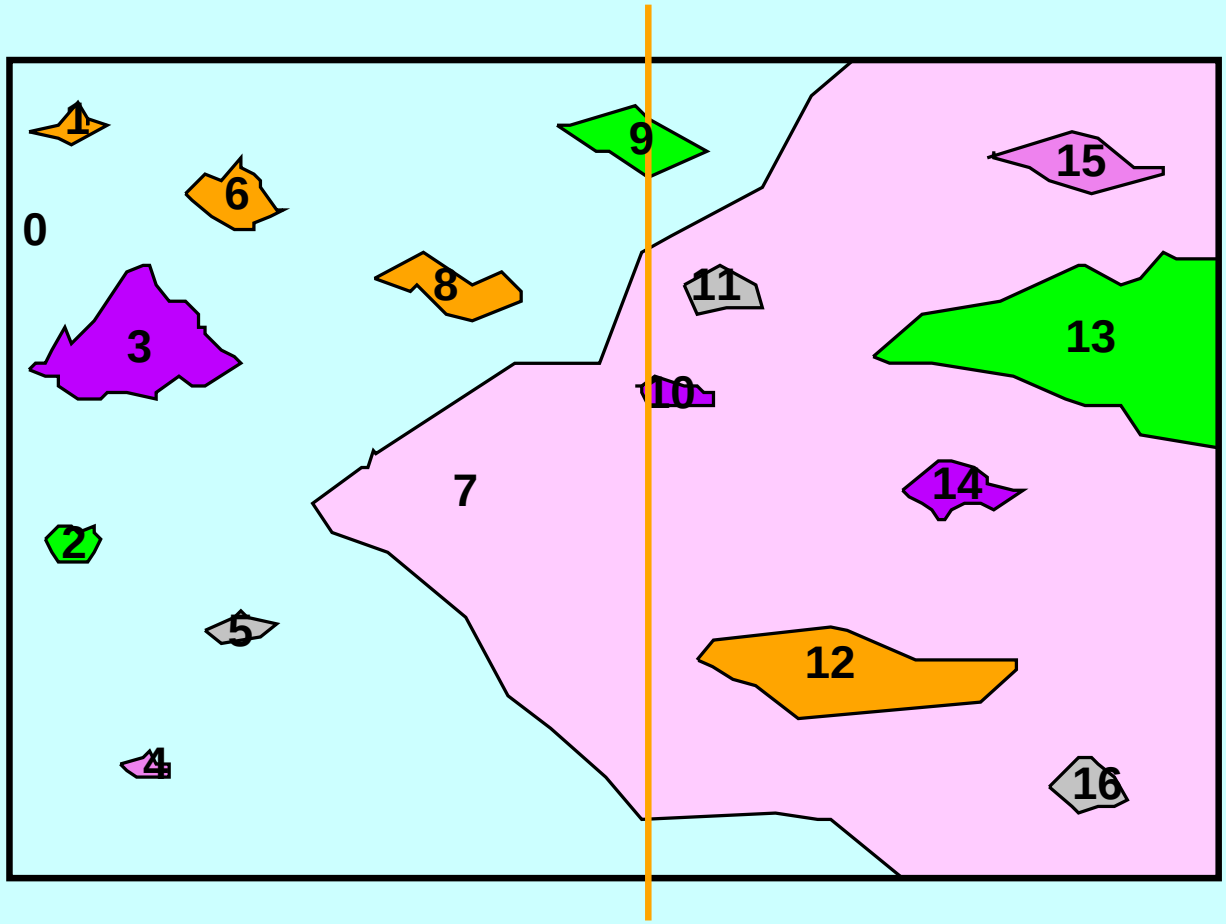
Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217:** 624-626.

Kimura, M., and T. Ohta. 1971. Protein polymorphism as a phase of molecular evolution. *Nature* **229:** 467-469.

Neutral mutation theory

Crow and Kimura, 1964; Lewontin and Hubby, 1966;
Kimura, 1968; King and Jukes, 1969; Kimura and Ohta, 1971

assume: population size N, rate u of neutral mutations, all different



Heterozygosity at any point is expected to be

$$\frac{4Nu}{4Nu + 1}$$

Crow and Kimura's theoretical calculation

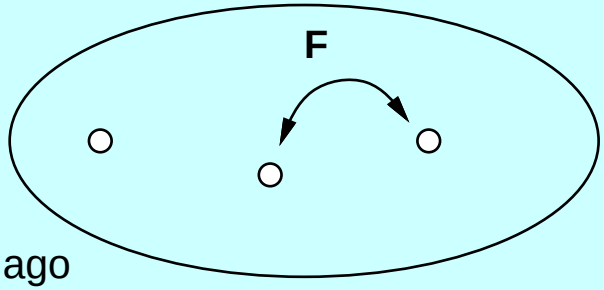
James F. Crow, about 1990

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* **49:** 725-738.

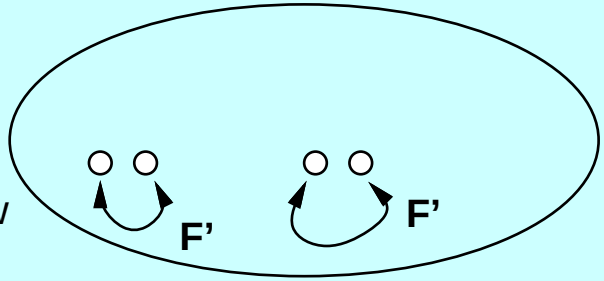
Expected heterozygosity with neutral mutation

In a random−mating population with neutral mutation,
a fraction **F** of the pairs of copies will be homozygous.
Suppose all mutations create completely new alleles,
and the rate of these neutral mutations is **u**

diploid population of size **N**



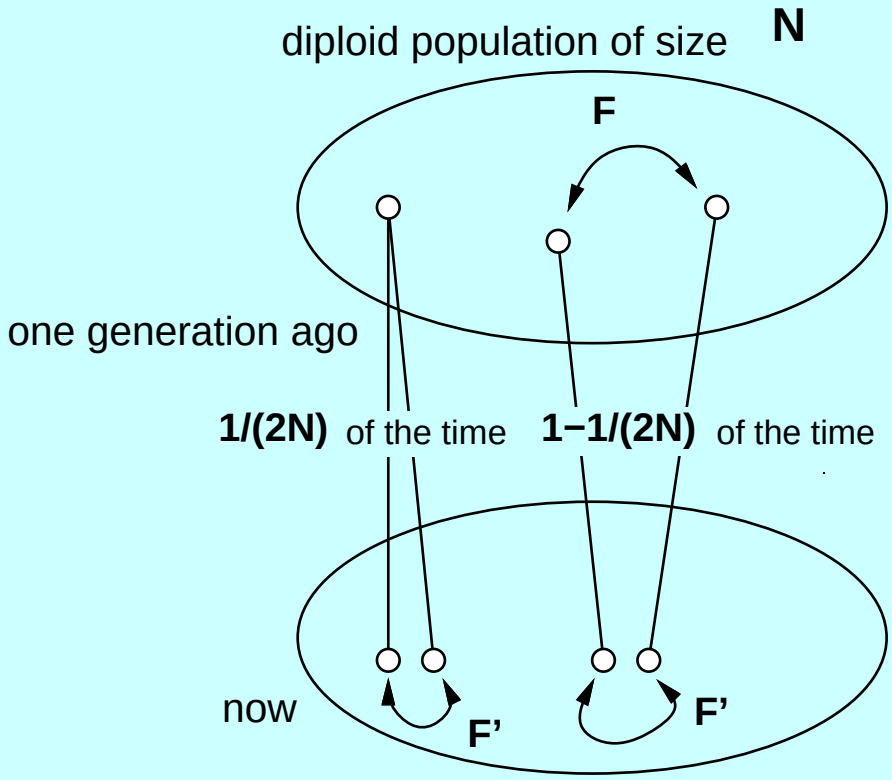
one generation ago



now

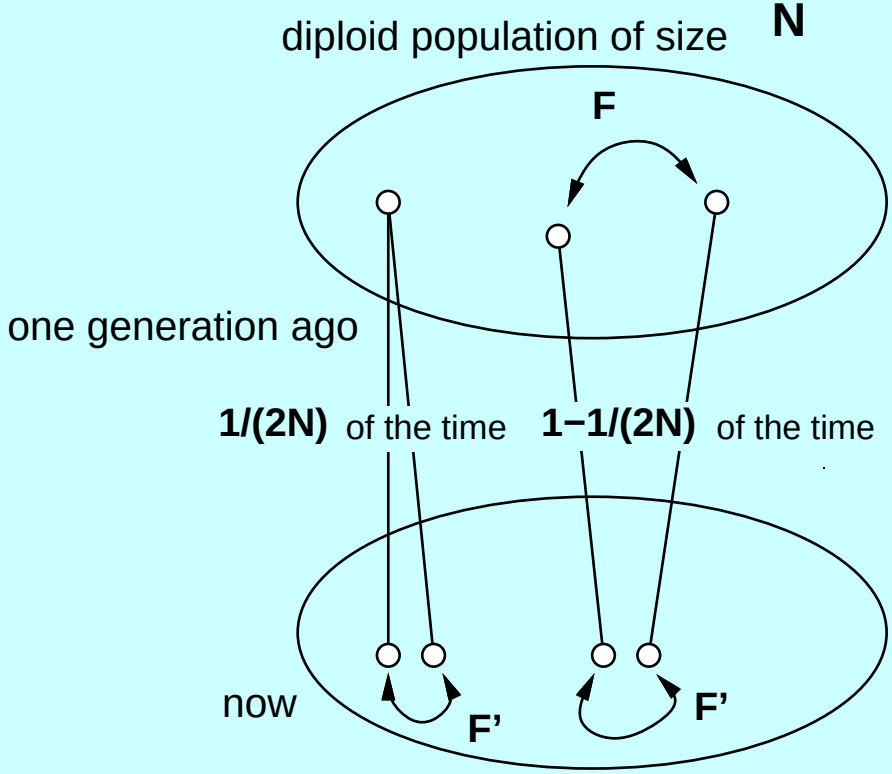
Expected heterozygosity with neutral mutation

In a random−mating population with neutral **mutation, a fraction F of the pairs of copies** will be homozygous. Suppose all mutations create completely new alleles, and the rate of these neutral mutations is **u**



Expected heterozygosity with neutral mutation

In a random−mating population with neutral **mutation, a fraction F of the pairs of copies** will be homozygous. Suppose all mutations create completely new alleles, and the rate of these neutral mutations is **u**

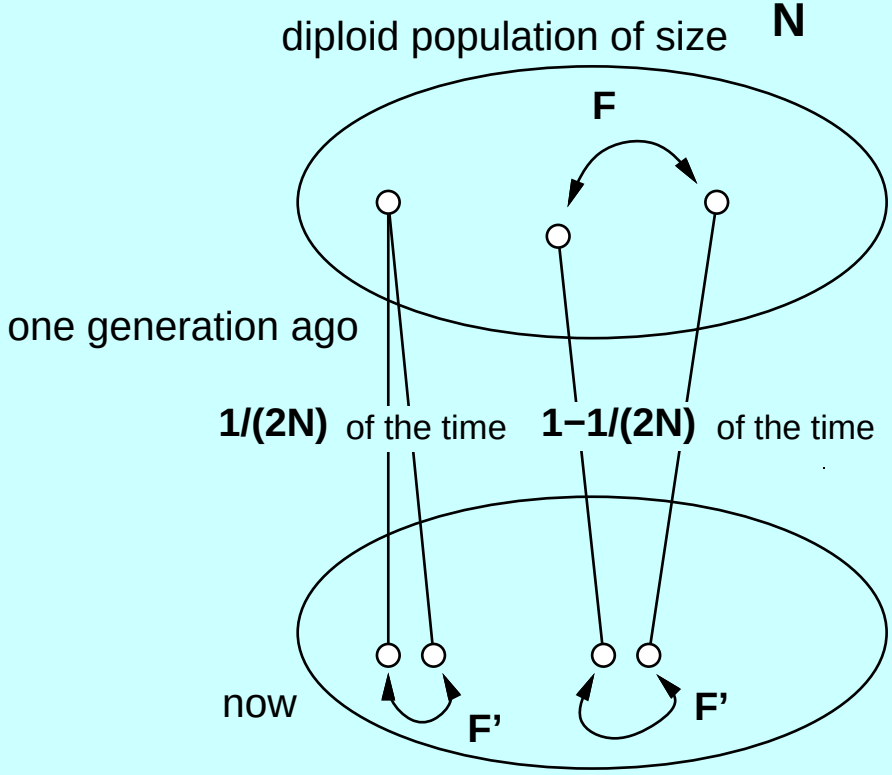


To be identical, both copies must not be new mutants, and the probability of this is $(1-u)^2$

$$F' = (1-u)^2 \left[(1/(2N)) \times 1 + (1-1/(2N))\, F \right]$$

Expected heterozygosity with neutral mutation

In a random−mating population with neutral **mutation, a fraction F of the pairs of copies** will be homozygous. Suppose all mutations create completely new alleles, and the rate of these neutral mutations is **u**



To be identical, both copies must not be new mutants, and the probability of this is $(1-u)^2$

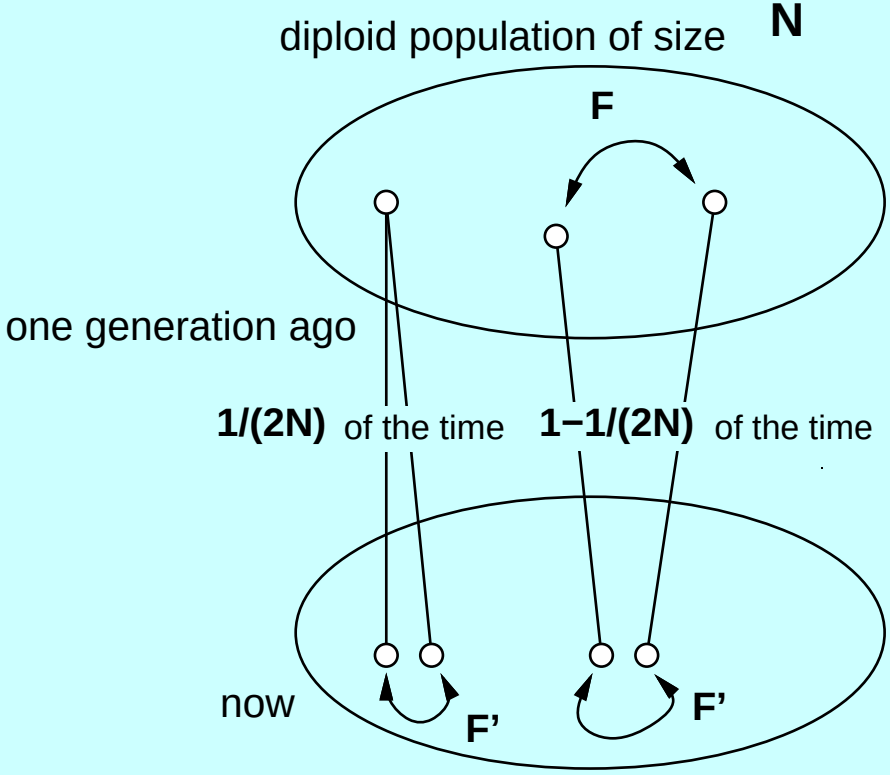
$$F' = (1-u)^2 \left[(1/(2N)) \times 1 + (1-1/(2N))\, F \right]$$

So if we have settled down to an equilibrium level of heterozygosity, $F' = F$, so that

$$F = (1-u)^2 \left[(1/(2N)) \times 1 + (1-1/(2N))\, F \right]$$

Expected heterozygosity with neutral mutation

In a random−mating population with neutral **mutation, a fraction F of the pairs of copies** will be homozygous. Suppose all mutations create completely new alleles, and the rate of these neutral mutations is **u**



To be identical, both copies must not be new mutants, and the probability of this is $(1-u)^2$

$$F' = (1-u)^2 \left[(1/(2N)) \times 1 + (1-1/(2N))\, F \right]$$

So if we have settled down to an equilibrium level of heterozygosity, $F' = F$, so that

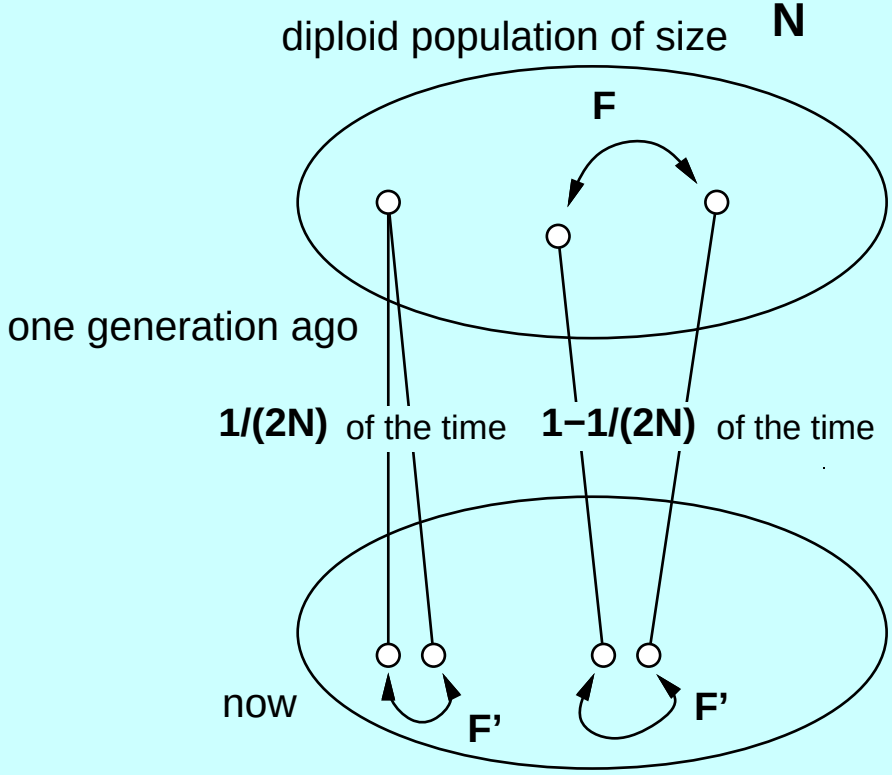
$$F = (1-u)^2 \left[(1/(2N)) \times 1 + (1-1/(2N))\, F \right]$$

which is easily solved to give

$$F = \frac{(1-u)^2 (1/(2N))}{1 - (1-u)^2 (1 - 1/(2N))}$$

Expected heterozygosity with neutral mutation

In a random−mating population with neutral **mutation, a fraction F of the pairs of copies** will be homozygous. Suppose all mutations create completely new alleles, and the rate of these neutral mutations is **u**



To be identical, both copies must not be new mutants, and the probability of this is $(1-u)^2$

$$F' = (1-u)^2 \left[(1/(2N)) \times 1 + (1-1/(2N))\, F \right]$$

So if we have settled down to an equilibrium level of heterozygosity, $F' = F$, so that

$$F = (1-u)^2 \left[(1/(2N)) \times 1 + (1-1/(2N))\, F \right]$$

which is easily solved to give

$$F = \frac{(1-u)^2 (1/(2N))}{1 - (1-u)^2 (1 - 1/(2N))}$$

or to good approximation:

$$F = \frac{1}{1 + 4Nu}$$

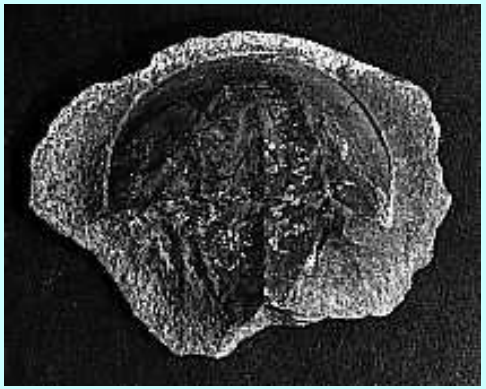
heterozygosity is:

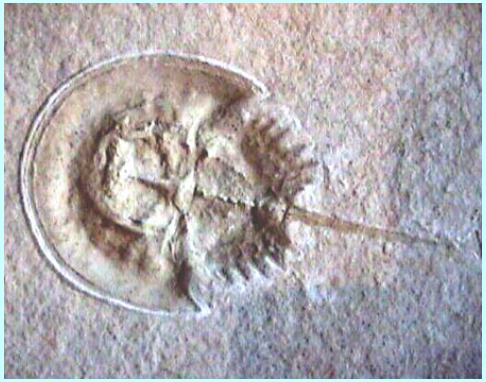
$$1-F = \frac{4Nu}{1 + 4Nu}$$

Heterozygosity in marine invertebrates (Valentine, 1975)

species	which is a	samples/locus	no. loci	Het.
Asterias vulgaris	Northern sea star	19-27	26	1.1
Cancer magister	Dungeness crab	54	29	1.4
Asterias forbesi	Common sea star	19-72	27	2.1
Lyothyrella notorcadensis	brachiopod	78	34	3.9
Homarus americanus	lobster	290	37	3.9
Crangon negricata	shrimp	30	30	4.9
Limulus polyphemus	horseshoe crab	64	25	5.7
Euphausia superba	Antarctic krill	124	36	5.7
Upogebia pugettensis	blue mud shrimp	40	34	6.5
Callianassa californiensis	ghost shrimp	35	38	8.2
Phoronopsis viridis	horseshoe worm	120	39	9.4
Crassostrea virginica	Eastern oyster	200	32	12.0
Euphausia mucronata	small krill	50	28	14.1
Asteriodea (4 spp.)	deep sea stars	31	24	16.4
Frielea halli	brachiopod	45	18	16.9
Ophiomusium lymani	large brittlestar	257	15	17.0
Euphausia distinguenda	tropical krill	110	30	21.5
Tridacna maxima	giant clam	120	37	21.6

An interesting case: Limulus polyphemus





Carboniferous (300 mya) Jurassic (155 mya) today

Selander, R.K., S.Y. Yang, R.C. Lewontin, W.E. Johnson. 1970. Genetic variation in the horseshoe crab (*Limulus polyphemus*), a phylogenetic "relic." *Evolution* 24:402-414.

An interesting case

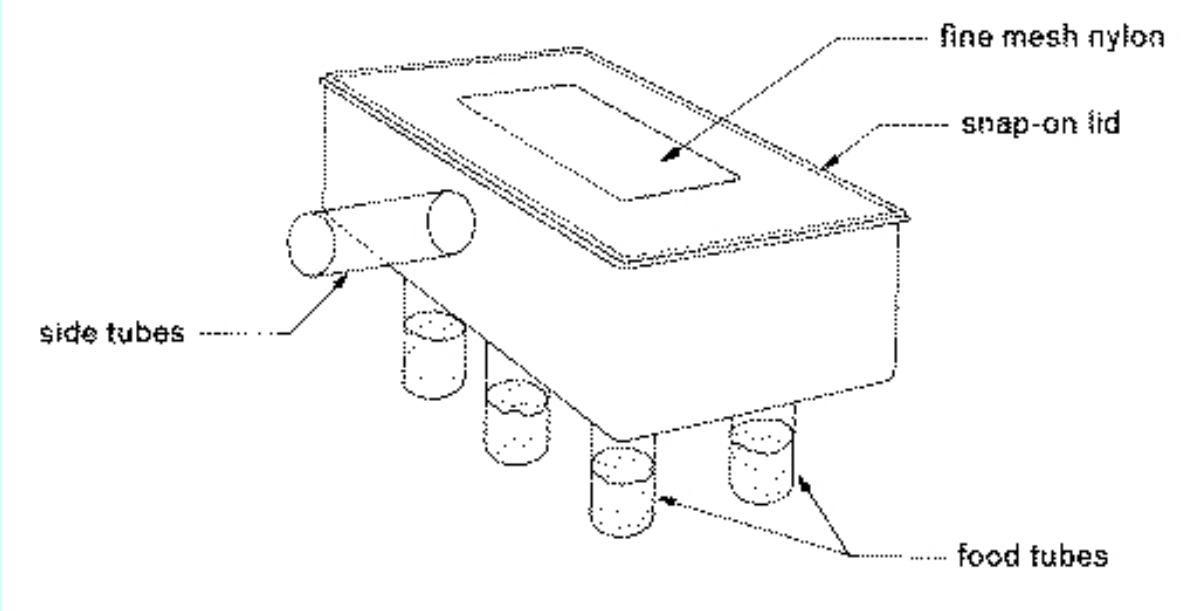
Northern elephant seal
Mirounga angustirostris

Southern elephant seal
Mirounga leonina

Northern elephant seal: Population in 1890's: 2-10 ?
Population today: 150,000 or so ("911? help! there's a monster dying on my beach")

Bonnell, M.L., and R.K. Selander. 1974. Elephant seals: genetic variation and near extinction. *Science* **184:** 908-909.

A “population cage" for *Drosophila*



Yamazaki's population cage experiment

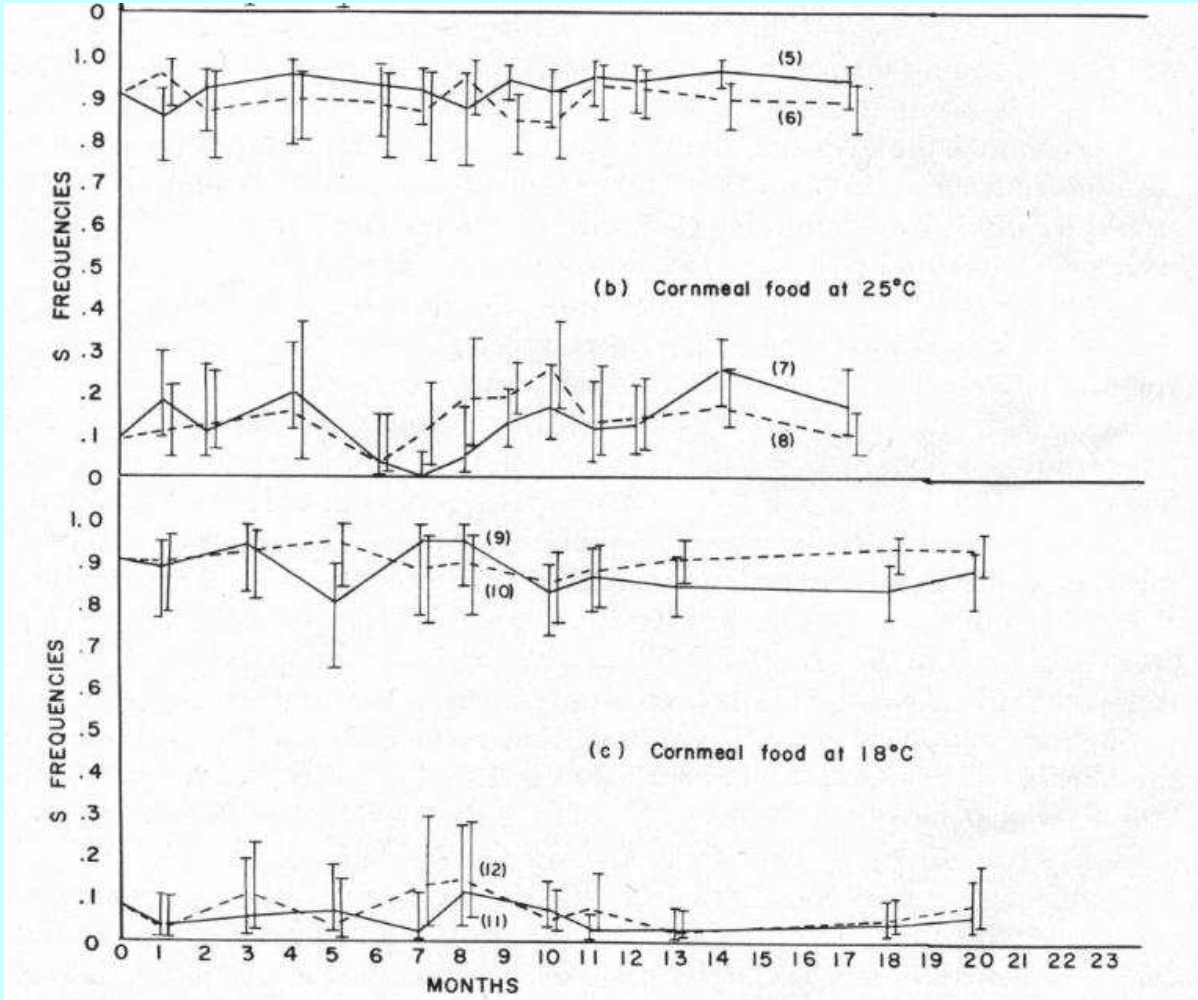


FIGURE 1.—Gene frequency changes in population cages at two different temperatures (18°C and 25°C) with two different foods (SPASSKY's and cornmeal). Vertical lines represent 95% confidence intervals.

Yamazaki, T. 1971. Measurement of fitness at the esterase-5 locus in *Drosophila melanogaster*. *Genetics* **67:** 579-603.

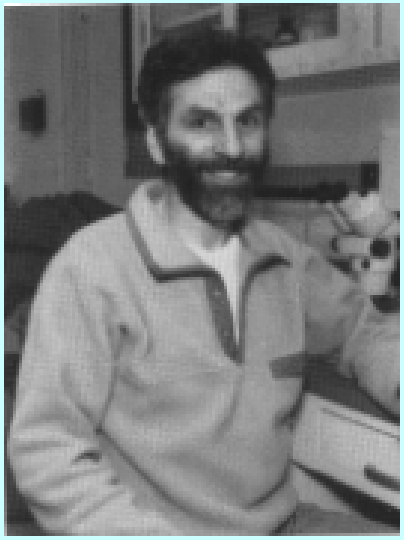
Explaining Electrophoretic Polymorphisms

Can do it either way:

By neutral mutation: If $H = 0.15$ then if $N_e = 1,000,000$ we need $4N_e\mu = 0.176$ to predict this, so that implies $\mu = 4.4 \times 10^{-8}$. So we can explain the level of variation by a neutral mechanism.

By selection: To be effective in a population with $N_e = 1,000,000$ selection would need to be big enough that $4N_e s > 1$ so $s > 1/4,000,000$ which is quite small, and impossible to detect in laboratory settings.

DNA sequencing reveals a similar picture



Marty Kreitman

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol-dehydrogenase Locus of *Drosophila melanogaster*. *Nature* **304:** 412-417.

Kreitman's sample of 11 ADH gene sequences, front end

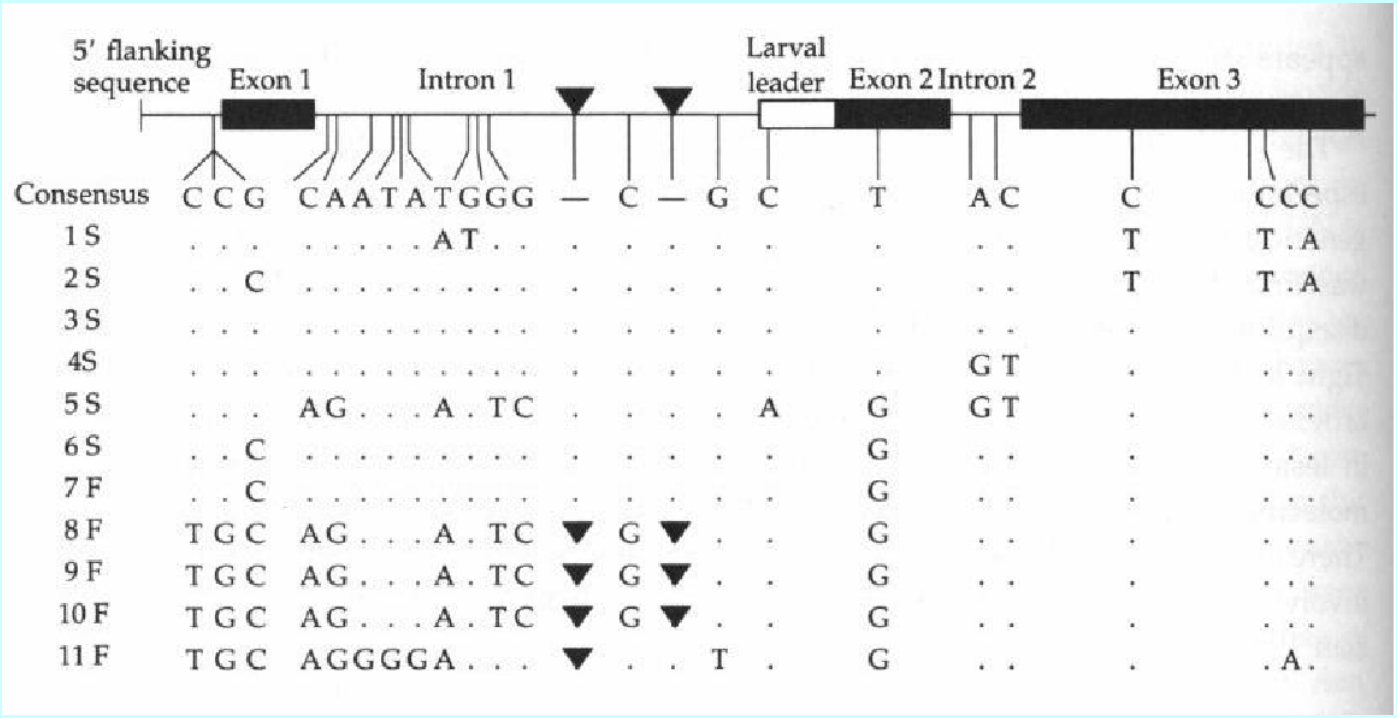


Figure 4. Polymorphic nucleotide sites among 11 alleles of the alcohol dehydrogenase gene of *Drosophila melanogaster*. The first line gives a consensus sequence for *Adh* at sites that vary; subsequent lines give the nucleotides from each copy for the polymorphic sites. A dot indicates that the site is identical to the consensus sequence. The triangles indicate sites of insertion or deletion relative to the consensus sequence. The star in exon 4 indicates the site of the amino acid replacement (lysine for threonine) responsible for the *fast-slow* mobility difference in the *Adh* protein. (After Kreitman 1983.)

Kreitman's sample of 11 ADH gene sequences, tail end

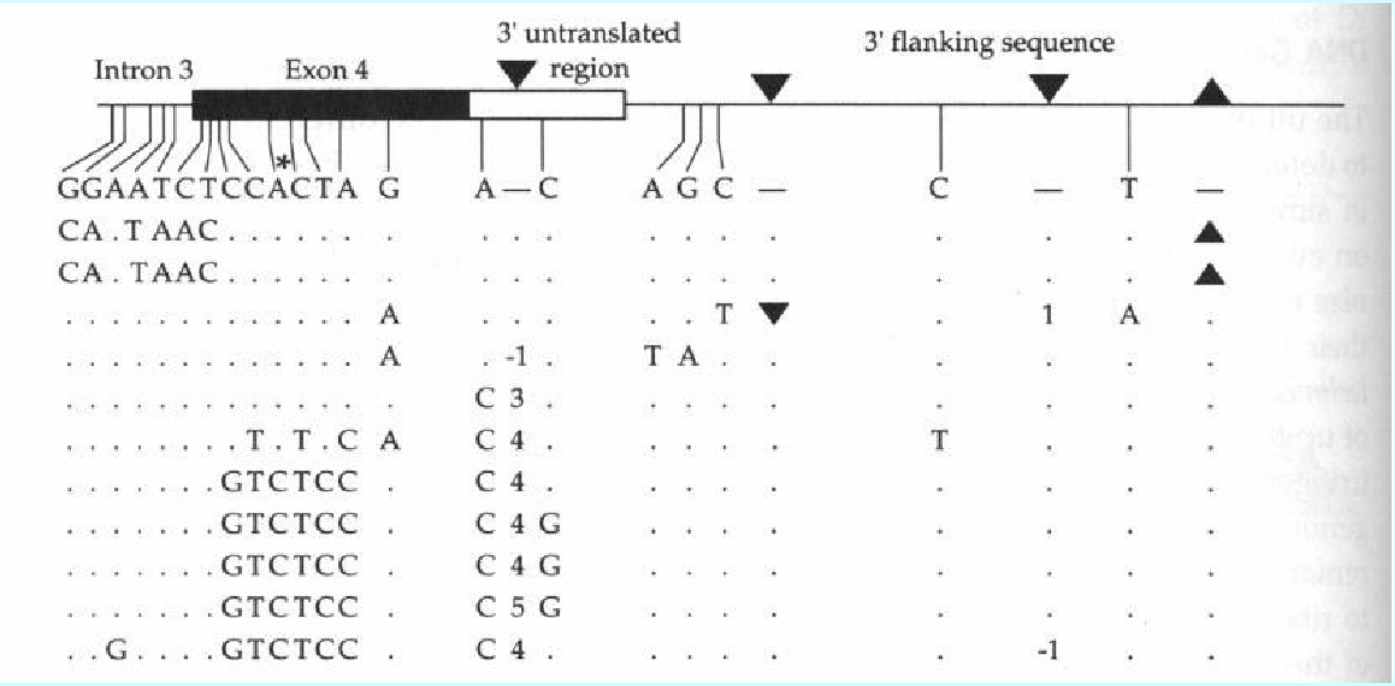
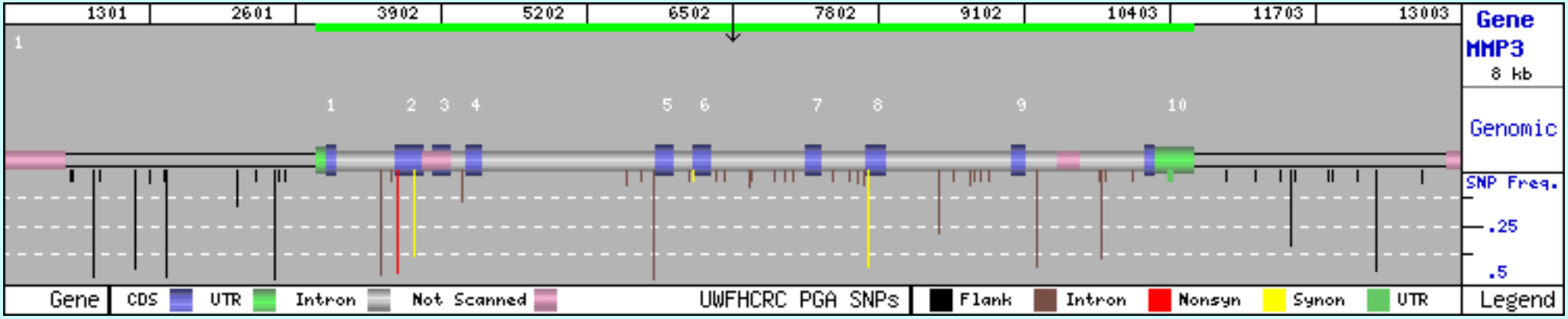


Figure 4. Polymorphic nucleotide sites among 11 alleles of the alcohol dehydrogenase gene of *Drosophila melanogaster*. The first line gives a consensus sequence for *Adh* at sites that vary; subsequent lines give the nucleotides from each copy for the polymorphic sites. A dot indicates that the site is identical to the consensus sequence. The triangles indicate sites of insertion or deletion relative to the consensus sequence. The star in exon 4 indicates the site of the amino acid replacement (lysine for threonine) responsible for the *fast-slow* mobility difference in the *Adh* protein. (After Kreitman 1983.)

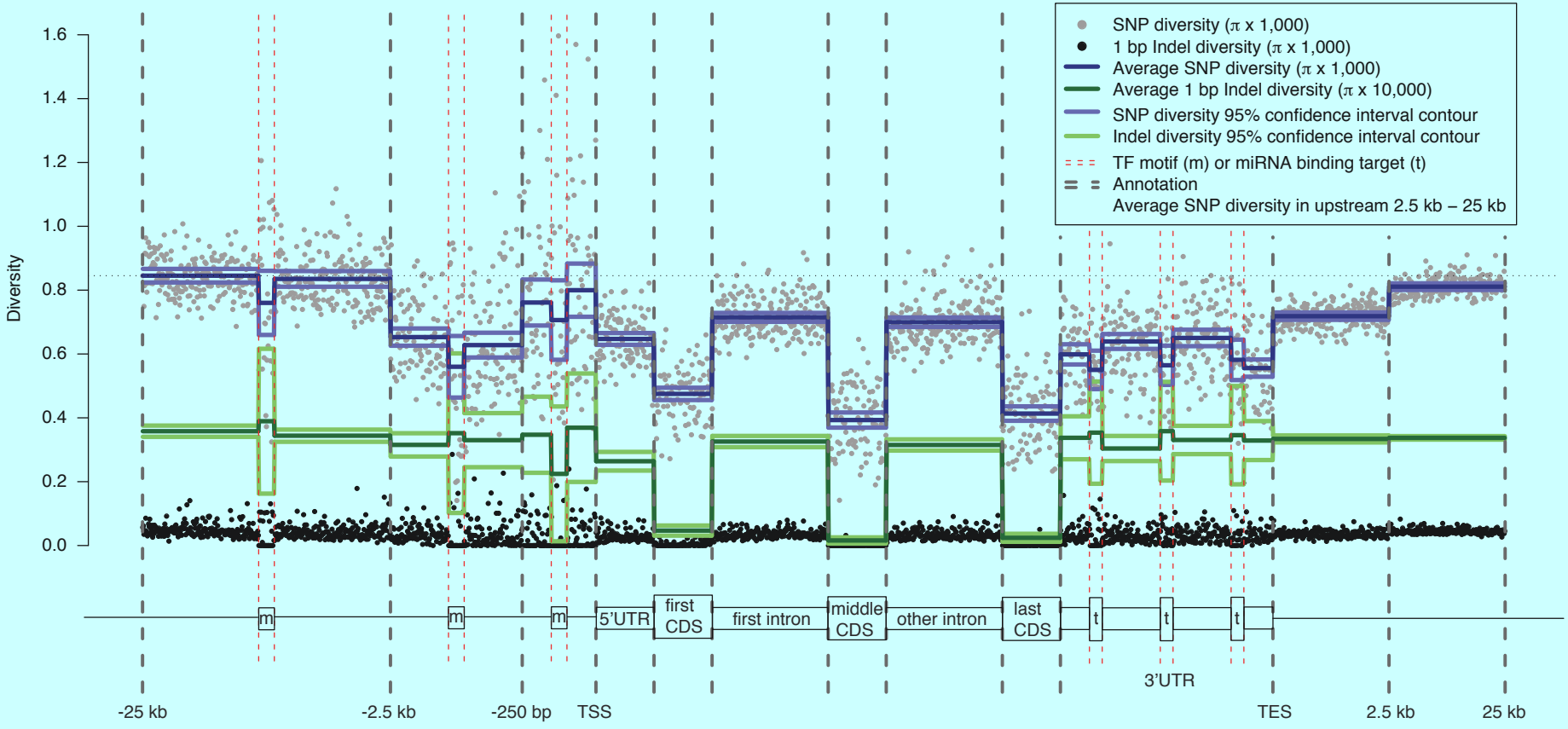
SeattleSNPs data (Nickerson lab)



Matrix Metalloproteinase 3 SNP data

Variation from the 1000 Genomes project

Supplementary Figure S7



From paper:

Mu, X.J., Z.J. Lu, Y. Kong, H.Y. Lam, M.B. Gerstein. 2011. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Research* **39(16):** 7058-7076.