Molecular variation

Joe Felsenstein

GENOME 453, Autumn 2011
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

slots

stained

gel of potato starch or polyacrilamide

tank of buffer solution

wick

power supply

AA AA Aa aa Aa AA

AA AA Aa aa Aa AA
A monomeric enzyme
A dimeric enzyme
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

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:

$$H = 1 - \sum_{i} p_i^2$$
### An example:

<table>
<thead>
<tr>
<th>locus</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.94</td>
<td>0.04</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The heterozygosities are calculated as:

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

The average heterozygosity in this example is

\[
H = \frac{(0 + 0 + 0.32 + 0.1144)}{4} = 0.1086
\]
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

Neutral mutation theory


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

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.

$F'$
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

$\frac{1}{2N}$ of the time

$1-\frac{1}{2N}$ of the time

now

$F'$

$F'$
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 $(1-u)^2$

$$F' = (1-u)^2 \left[ \frac{1}{2N} \times 1 + (1-1/(2N))F \right]$$
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 $u$ per generation. To be identical, both copies must not be new mutants, and the probability of this is $(1-u)^2$

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

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

$$F = (1-u)^2 \left[ \frac{1}{2N} \times 1 + \frac{1-1/(2N)}{F} \right]$$
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 $N$.

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

$$F' = (1-u)^2 \left[ \frac{1}{2N} \times 1 + \left(1 - \frac{1}{2N}\right) F \right]$$

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

$$F = (1-u)^2 \left[ \frac{1}{2N} \times 1 + \left(1 - \frac{1}{2N}\right) F \right]$$

which is easily solved to give

$$F = \frac{(1-u)^2 \left(1 \div 2N\right)}{1 - (1-u)^2 \left(1 - 1 \div 2N\right)}$$
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 \( (1-u)^2 \)

\[
F' = (1-u)^2 \left[ \frac{1}{(2N)} \times 1 + (1-1/(2N)) F \right]
\]

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

\[
F = (1-u)^2 \left[ \frac{1}{(2N)} \times 1 + (1-1/(2N)) F \right]
\]

which is easily solved to give

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

or to good approximation: heterozygosity is:

\[
F = \frac{1}{1 + 4Nu}
\]

\[
1-F = \frac{4Nu}{1 + 4Nu}
\]
### Heterozygosity in marine invertebrates (Valentine, 1975)

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

Carboniferous (300 mya)  Jurassic (155 mya)  today

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”)

A “population cage" for *Drosophila*
Yamazaki’s population cage experiment

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 $4N_es > 1$ 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

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

From paper: