Measuring Genotypic Variation

“There must be some other principle regulating the infinitely varied forms of ... life”

What do we mean by variation?

Biologists often talk about variation and diversity. Recognition of the “varied forms of life” is explicit from the 19th century, “Biodiversity” is a catch-phrase from the late 20th century, and as plant breeders we are interested in “genetic diversity”. In the last several lectures we have been discussing techniques to measure and describe this diversity.

In our discussion of heritability we have been trying to define and quantify the genetic contribution to traits. The statistical definition of “variance” \( \frac{\text{SUM}(X - \bar{X})^2}{n-1} \) provides a numerical description of the phenotypic diversity. Heritability is a numerical description of the genetic contribution to this diversity. Numerical descriptions are often based on assumptions of the estimation procedure, and we should keep in mind the limitations of techniques. Thus, some familiarity with the derivation and assumptions is helpful.

We now turn our attention to the measurement of genotypic variation. Genotypic variation refers to variation in DNA or genes (rather than phenotype). It differs from heritability in that we are trying to measure diversity at “neutral” sites in the genome, and then relate this information to the genetic structure of populations and individuals.

Experimental approaches to measuring genetic variation

Plant breeders and geneticists must strive to balance the need for new traits (new genetic variation) with the need to preserve adapted genes and gene complexes. We must make decisions on where to sample new variation (from the wild or from collections) and how to use it in the context of plant improvement.

Nearly a century has been spent collecting and preserving genetic diversity in plants. Germplasm banks—living seed collections that serve as repositories of genetic variation—have been established as a source of genes for improving agricultural crops. Genetic linkage maps have made it possible to study the chromosomal locations of genes for improving yield and other complex traits important to agriculture. The tools of genome research may finally unleash the genetic potential of our wild and cultivated germplasm resources for the benefit of society. From Tanksley SD, McCouch SR. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science 1997 Aug 22;277(5329):1063-6

A narrow genetic base has resulted in severe disease epidemics in the past, and the need to monitor genetic diversity in breeding programs is widely recognized. The use of wild relatives of cultivated plants to increase genetic diversity and introduce new traits is accepted by academic researchers as a valid practice. However based on a recent survey of tomato breeders, 60% felt that there was no lack of genetic variability in adapted germplasm (see Heisey et al, 2002, Proceedings of the 2001 Tomato Breeders Roundtable, Hort and Crop Science Series No. 724. http://www.oardc.ohio-state.edu/tomato/prgm01.htm). Some of the controversy behind use of wild relatives relates to the theoretical arguments alluded to in the first lecture: Allard’s
proposal that breeders should seek to maintain favorable combinations by choosing breeding strategies that minimize recombination and conserve combinations of unlinked genes (linkage disequilibrium) is contrasted with the dramatic and often negative effects of introducing wild germplasm into breeding stock. Use of wild germplasm may be essential for introducing new alleles not found in the adapted germplasm, but at the same time we may introduce many “unfavorable” alleles that have a negative effect on yield or quality. We also may break up favorable combinations of adapted alleles if we are not careful with our introgression strategy.

A fundamental question with applied ramifications is “how do we measure genetic variation?” The methodology we will discuss is part of the tool chest that will let you answer questions such as:

Which germplasm collection has the most genetic variation?

Which wild relative has the most genetic variation?

How does the mating system of related plants compare to that of the crop species?

If a number of accessions have the same phenotypic trait, is there any way we can decide whether it is conditioned by the same gene(s) or is genetically distinct?

If a collection has 2,000 accessions and we can only screen 200/year is there a way of concentrating on the important ones first?

Do in situ conservation methods preserve genetic diversity?

While we cover the topic of “experimental approaches to measuring genetic variation” I want you to keep in mind the following:

Genetic variation may be precisely defined by statistics such as the coefficient of coancestry, inbreeding coefficients, etc…

Genetic variation can be measured in more than one way, but the definitions do not change.

Review of Plant Genetic Resources.

Germlasm collections
http://tgrc.ucdavis.edu
http://www.ars-grin.gov/npgs/

Tomato Genetics Resource Center
USDA Plant Germplasm System

Genetic Data Bases
http://tgrc.ucdavis.edu
http://soldb.cit.cornell.edu

Solanaceae genome network data base

Genomic Data Bases
http://www.tigr.org/

Nat. Center for Biotechnology Information
The Institute for Genomics Research
How do we measure and quantify genetic variation? e.g. from pedigree analysis:

Breeders often ask questions about how closely related varieties are. The relationship between varieties can be expressed in terms of Coancestry coefficients ($F_{ST}$), gene diversities ($G_{ST}$), Inbreeding coefficients ($F_{IT}$), and gene identities ($D_{ST}$). This information may be useful in determining the probability that a gene is shared (for example a resistance gene) and is helpful for guiding breeding efforts. The following is a pedigree for two varieties, U and W.

\[
\begin{array}{cccc}
A \times B & C \times D & A \times B & C \times E \\
\mid & \mid & \mid & \\
P & X & Q & R \\
\mid & \mid & \\
U & W & S & \\
\end{array}
\]

The relationship between variety U and W can be described in terms of the coancestry coefficient $\theta$ (for most discussions we can assume that $\theta = f_{po} = F_{ST} = G_{ST}$, and we will come back to these).

**Note: Review Chapter 7, Resemblance Between Relatives.**

\[
\begin{array}{cccc}
A \times B & C \times D & A \times B & C \times E \\
\mid & \mid & \mid & \\
P & X & Q & R \\
\mid & \mid & \mid & \\
U & W & S & (CE) \\
\end{array}
\]


If A, B, C, D, & E are inbred and unrelated, U and W share 18.75% (3/16) of their genes by descent ($\theta = .1875$).

**From Lecture 5 Definitions (note that we will see these again and again and …)**

$\theta = f_{po} = (1 + F_p)/2$ (coefficient of ancestry also called coancestry coefficient): the probability that a random allele from X is identical by descent to a random allele from Y.

$F_p = (1-(1/2)^{n-1})$ (inbreeding coefficient): the probability that two alleles are identical by descent i.e. homozygous. For selfing $F_p = (1-(1/2)^{n-1})$ where $n = \text{generation number}$.

Also remember that this information can be used to estimate heritability through parent offspring regression (or some other family structure..).
Covpq = r(σ²(a)) + u(σ²(d))

Covpq = 2fpq(σ²(a)) + (fac*fbd + fad*fbc)(σ²(d)) where a & b are parents of p and c & d are parents of q.

Cockerham developed the following more general equation that can be used for inbreeding (for two alleles of equal frequency):

\[\text{Cov}_{gg'} = (1 + F_t)(\sigma^2(a)) + [(1 + F_t)(1 - F_g)(1 - F_{g'}))/(1 - F_t)] \sigma^2(d)\]

\[\text{CovF2:F3} \; F_t = 0 \text{ where } F_g = (1-(1/2)^2-1) \text{ and } g' = (1-(1/2)^3-1)\]

\[\text{CovF3:F4} \; F_t = (1-(1/2)^2-1) \text{ where } g = (1-(1/2)^3-1) \text{ and } g' = (1-(1/2)^4-1)\]

\[\text{etc…}\]

**Table 1. Covariance among relatives.**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>D</th>
<th>VarA</th>
<th>VarD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CovF2:F3</td>
<td>1/2</td>
<td>1/8</td>
<td>1</td>
<td>1/2</td>
</tr>
<tr>
<td>CovF3:F4 (F3 fam:F4 fam)</td>
<td>1/2</td>
<td>1/32</td>
<td>1</td>
<td>1/8</td>
</tr>
<tr>
<td>CovF3:F4 (F3 ind:F4 fam)</td>
<td>1/4</td>
<td>1/16</td>
<td>1/2</td>
<td>1/4</td>
</tr>
</tbody>
</table>

Let’s now turn our attention to a breeding population derived from line crosses.

Consider the following F₂ population for an organism with N = 2:
We can describe the whole population and individuals in terms of relationships:

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected segregation of alleles</td>
<td>1:2:1</td>
</tr>
<tr>
<td>Number of alleles at each loci</td>
<td>At most, 2</td>
</tr>
<tr>
<td>Alleles shared between prog. and parents</td>
<td>50% of alleles are shared by decent</td>
</tr>
<tr>
<td>Alleles shared between F2 progeny</td>
<td>50% of alleles are shared by decent</td>
</tr>
<tr>
<td>Alleles within individuals</td>
<td>50% homozygous 50% heterozygous</td>
</tr>
</tbody>
</table>

If we take this population to the next generation

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alleles shared between prog. and parents</td>
<td>50% (of the Parental genotype)</td>
</tr>
<tr>
<td>Alleles shared within each F3 family</td>
<td>75%</td>
</tr>
<tr>
<td>Alleles shared between F3 family</td>
<td>50%</td>
</tr>
<tr>
<td>Alleles within individuals</td>
<td>75% homozygous 25% heterozygous</td>
</tr>
</tbody>
</table>

As the population is selfed toward homozygosity (Recombinant inbred lines)

- Each line will share 50% of alleles with each parent
- Each line will share 100% of alleles within each line
- Each line will share 50% of alleles with each progeny line
- Each line will be 100% homozygous

These relationships are straight-forward for line crosses, but can be more complicated as pedigrees and crossing schemes become more complicated. In the context of a breeding program, how many alleles progeny share with parents is the co-ancestry coefficient. The inbreeding coefficient describes the amount of homozygosity in the genome.

The Definitions from lecture 5 can help us determine these relationships:

- \( \theta = f_{po} = (1 + F_p)/2 \) (coefficient of ancestry also called coancestry coefficient): the probability that a random allele from X is identical by descent to a random allele from Y.

- \( F_p = \text{inbreeding coefficient} \): the probability that two alleles are identical by descent i.e. homozygous. For selfing \( F_p = (1-(1/2)^{n-1}) \) where \( n = \text{generation number} \).
Application of coefficients to inbred population derived from inbreeding:

Table 2. Inbreeding coefficients for selfing.

<table>
<thead>
<tr>
<th>Generation</th>
<th>Fp = (1-(1/2)^(n-1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>0.875</td>
</tr>
<tr>
<td>5</td>
<td>0.9375</td>
</tr>
<tr>
<td>6</td>
<td>0.96875</td>
</tr>
<tr>
<td>7</td>
<td>0.984375</td>
</tr>
<tr>
<td>8</td>
<td>0.992188</td>
</tr>
<tr>
<td>9</td>
<td>0.996094</td>
</tr>
<tr>
<td>10</td>
<td>0.998047</td>
</tr>
</tbody>
</table>

Table 3. Coancestry coefficient for selfing (within and between families).

<table>
<thead>
<tr>
<th></th>
<th>Within families (θ = (1 + Fp)/2)</th>
<th>Between Families (θ = (1 + 0)/2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2 share</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>F3 share</td>
<td>0.75</td>
<td>0.5</td>
</tr>
<tr>
<td>F4 share</td>
<td>0.875</td>
<td>0.5</td>
</tr>
<tr>
<td>F6 share</td>
<td>0.9375</td>
<td>0.5</td>
</tr>
<tr>
<td>F7 share</td>
<td>0.96875</td>
<td>0.5</td>
</tr>
<tr>
<td>F8 share</td>
<td>0.984375</td>
<td>0.5</td>
</tr>
<tr>
<td>F9 share</td>
<td>0.992188</td>
<td>0.5</td>
</tr>
<tr>
<td>F10 share</td>
<td>0.996094</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The coefficient Fp is the inbreeding coefficient from the previous generation. Shared alleles are between F2 individuals or within families (F3 – on). Can you make similar calculations for BC families? What about inbred backcross families? For the relationship between families it is assumed that P1 and P2 are unrelated and Fp = 0.

Discussion point: with pedigrees we need to make assumptions about the relationship between A, B, C, D, & E. With line crosses we must make assumptions about the relationship of the two parents, P1 and P2 (we have assumed Fp = 0). Markers may provide more accurate estimates.
Experimental approaches to measuring genetic variation

The genetic structure of populations includes numbers, composition by age and sex, and state of subdivision. Discussions of population structure usually begin by considering large random-mating populations in which structure (or subdivision) is absent. In contrast, for most crop plants, discussions of population structure reduce down to questions of numbers and subdivision. Our discussion will focus on measuring the number of alleles, allele frequencies, and subdivision within populations. Over the next three class periods we will consider:

Tools
- Genetic markers

Sampling genetic variation
- phenotype
- genotype (neutral genetic markers)

Measurements of genetic structure
- The amount of variation among individuals in a population
- The ways in which variation is partitioned in time and space
- The relationships among individuals within and between sub-populations

Use of marker data to characterize populations or germplasm

Because crosses are necessary to demonstrate that a band on a gel corresponds to an allelic form of a single gene it is assumed that these crosses have been carried out prior to population level work. Always keep in mind the expected genetic behavior of your germplasm (if known), your crosses, and the limitations of your marker system. A second assumption of most analyses is that the sampled loci will be independent. Crosses and linkage analysis can be used to verify this assumption.

Marker Review

Techniques for detecting loci include Isozymes, RFLPs, RAPDs, AFLPs, SSRs, and more. Many of these have been reviewed in Chapter 3 of Liu. Students should review Section 3.4 on pp. 62-80 and come prepared with questions next class. It will be important to keep in mind the experimental strengths and limitations of marker systems (e.g., Table 2). More importantly it will be important to understand the expected genetic properties of the loci detected and the effect of these genetic properties on analysis. Confirmation of expected segregation and assortment and allelic relationships is an essential control. Always keep in mind that a band on a gel is a phenotype from which we infer genotype.
### Table 2. Marker systems, genetic properties, strengths, and limitations.

<table>
<thead>
<tr>
<th>System</th>
<th>Genetic Properties</th>
<th>Strengths &amp; Limitations</th>
</tr>
</thead>
</table>
| Isozymes | Co-dominant | Enzyme activity based  
Limited number of loci  
Limited alleles per locus  
Protein is measured (therefore not exact measure of genotype)  
Tissue specificity/environmental regulation  
Fast |
| RFLPs | Co-dominant | “Restriction Fragment Length Polymorphism”  
Pre-screen for single copy sequences to be used as probes  
Large number of loci  
Slower than isozymes  
Assumption that when samples share a fragment, they share flanking cleavage sites |
| RAPD | Dominant | PCR based “Random amplified polymorphic DNA”  
Fast  
Sensitive to conditions (reproducibility issues)  
Assumption that when two samples share a fragment, it is the same locus  
Measures phenotype in outcrossing species  
Multiple loci can be scored in single reaction |
| AFLP | Dominant | PCR based “Amplified Fragment Length Polymorphism”  
Detects large number of bands and therefore polymorphism  
Multi-step, therefore high technical requirements |
| SCAR | Co-dominant | PCR based “sequence characterized region”  
Fast  
Requires sequence data, therefore expensive to develop primers |
| CAPs | Co-dominant | PCR based “cut amplified polymorphism”  
Requires restriction enzyme digestion of PCR product  
Requires sequence data, therefore expensive to develop primers |
| SSRs | Co-dominant | PCR based “Simple Sequence Repeats” also called “Micro-satellites”  
Fast  
Requires sequence data, therefore expensive to develop primers  
Commercially available for some crops  
Detect multiple alleles |
| SNPs | Co-dominant | PCR or sequencing based “Single Nucleotide Polymorphism”  
Fast  
Requires sequence data, therefore expensive to develop primers  
The next wave of marker technology will focus on methods to detect SNPs |

*Note: Review Chapter 14 “Principles of Marker-Based Analysis” in Lynch and Walsh*
**Review of Genetic Behavior**

Because crosses are necessary to demonstrate that a band on a gel corresponds to an allelic form of a single gene it is assumed that these crosses have been carried out prior to population level work. Always keep in mind the expected genetic behavior of your germplasm (if known), your crosses, and the limitations of your marker system. A second assumption of most analyses is that the sampled loci will be independent. Crosses and linkage analysis can be used to verify this assumption. *This is the second place where I’ve noted the importance of understanding the genetic behavior of markers and of testing assumptions in crosses. Either I’m getting senile or I must think this is an important point…*

What are the possible expected segregation patterns for the following banding pattern:

Parent A  X  Parent B

___  ___
___  ___
___  ___

If both parents are homozygous, we expect no segregation. If both parents are heterozygous and each band represents an allele, we expect 1:2:1 segregation:

___  ___
___  ___
___  ___

Complex banding patterns can be more difficult to sort out. Consider the following RFLP pattern:

Parent A  X  Parent B

___  ___  ___  ___  ___  ___  ___  ___
___  ___  ___  ___  ___  ___  ___  ___
___  ___  ___  ___  ___  ___  ___  ___

Two bands could indicate a single allele or independent loci. Only segregation data will help distinguish the possibilities:

___  ___  ___  ___  ___  ___  ___  ___
___  ___  ___  ___  ___  ___  ___  ___
___  ___  ___  ___  ___  ___  ___  ___

The allelic relationships of bands are often easier to understand if parents are known to be homozygous and banding patterns are simple. A band cannot be interpreted as an allele without segregation data.
The first order of analysis or summary statistics characterize genetic states.

Indices for Phenotypic data
For data from some molecular markers (e.g. RAPDs and AFLPs) bands are generally scored as + (designated 1) or – (designated 0). For inbreeding species these character states may often correspond to alternate alleles. However for outcrossing species dominance of the bands will complicate the analysis. For any two individuals we can define four classes based on a two way contingency table:

\[
\begin{align*}
&\ a = 1,1 \\
&\ b = 0,1 \\
&\ c = 1, 0 \\
&\ d = 0,0
\end{align*}
\]

and \( n = a + b + c + d \)

When two bands of the same size are detected, the assumption of a high degree of sequence similarity is often made. Note that this assumption can be tested experimentally (Discuss). For classes b and c the assumption is that sequence divergence due to insertions, deletions, and base substitution has occurred. One of the problems with markers such as RAPDs and AFLPs is that multiple genetic changes can lead to the same phenotype (0). The “d” class is considered the least informative regarding the relationship between any two individuals.

Several similarity coefficients are available for plus/minus data. Similarity coefficients differ in how they weigh missing bands.

Jaccard coefficient \( \{a/(n-d)\} \)

Simple Matching coefficient \( \{(a + d)/n\} \)

Composite genotypes (haplotypes)
If segregation analysis is lacking and the genetic control of band phenotypes cannot be deduced it may be better to calculate frequency of a composite genotype (haplotype). This approach will generally underestimate genetic differences because many differences are weighted the same as a single difference. In the example below the three haplotypes are weighted equally despite a lower band sharing index between A & C and B & C relative to A & B.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shannon Index : \( SI = \sum (Pi)(\ln Pi) \) where \( Pi \) is the frequency of the \( i \) th phenotype.
Other summary statistics:

Polymorphsim (P = Nd/(Na + Nb) where Nd is the number of polymorphic bands, Na and Nb are total number of bands in lanes a & b). P is 1-BS (band sharing).

Number of alleles per locus (requires that allelic relationships have been demonstrated or deduced)

Summary statistics do not allow us to compare levels of variation. For purposes of comparison, allelic frequencies offer a suitable starting point. Frequency A = (2AA + ½ Aa)/N

Sampling genetic variation

Sampling genetic variation based on phenotype

   Strength: In a breeding context we will be selecting on the basis of phenotype.
   Weakness: Under convergent evolution we might infer a close relationship in samples that differ widely in genetic background. In a breeding context we may not be able to distinguish different sources of a phenotype (e.g. disease resistance).

Sampling genetic variation based on neutral genetic markers

   Marker systems
   Genetic properties of marker systems
   Genotype vs Phenotype of marker banding pattern

Sampling strategies

In a natural population the way variation is arranged in space and time is important to the design of a sampling strategy. Because we rarely have information about this distribution a priori, sampling is often iterative (i.e. the scale of sampling is adjusted after the analysis of preliminary results).

The objective is to seek the most efficient way to deduce the genetic relationships among individuals. If nothing is known it will be best to sample broadly and aim to determine how variation is partitioned. There are two important considerations. First, what is the question that we seek to answer through our sampling? And second, how many samples and markers do we need to answer that question. If we randomly select individuals from the three examples (Below) we will conclude that the frequency of A = ¼ and the frequency of B = ¾ in all three cases. However a random sample will miss the fact that in the first example the population consists of two sub-populations. To avoid the inability of random samples to detect underlying structure, a hierarchical sampling strategy is often recommended. Note that a problem of scale still exists (i.e. on what scale might we detect genetic structure…). Hierarchical sampling is also valid for germplasm stored in collections especially if there is sufficient passport data to allow for systematic structuring of the hierarchy.
The sample size chosen will determine the detection limit of the analysis and a realistic sample size should be chosen to yield adequate information without expending excessive effort for a small incremental gain (Table 2). Answering a question such as “do samples from region X and region Y represent samples from different populations” will require some knowledge of allele frequencies in the two populations in order to design an appropriate sampling strategy.

Patchy

\[
\begin{array}{cccc}
A & B & B & B & B & B & B & B \\
B & B & B & B & B & B & B & B \\
\end{array}
\]

Frequency of A = \( \frac{1}{4} \)

Frequency of B = \( \frac{3}{4} \)

Gradient

\[
\begin{array}{cccc}
A & A & B & B & A \\
A & B & B & A & BBB & AB & BBB & B & B \\
B & B & B & B & B & B & B & B & B \\
\end{array}
\]

Frequency of A = \( \frac{1}{4} \)

Frequency of B = \( \frac{3}{4} \)

Random

\[
\begin{array}{cccc}
AB & A & AB & BB & A & BB & AB \\
AB & B & A & BB & A & B & B & B & B \\
B & B & A & AB & BB & A & B & B & B \\
B & B & A & AB & BB & A & B & B & B \\
\end{array}
\]

Frequency of A = \( \frac{1}{4} \)

Frequency of B = \( \frac{3}{4} \)

Example of a hierarchical sample (scale neutral).

\[
\begin{array}{ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]
Table 3. The number of individuals from each of two populations needed to detect differences (Source: Hillis and Moritz, 1990. Molecular Systematics. Sinauer Associates, Inc).

<table>
<thead>
<tr>
<th>Power</th>
<th>allele f diff.</th>
<th>frequency of allele</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.55</td>
</tr>
<tr>
<td>50%</td>
<td>0.05</td>
<td>760</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>6</td>
</tr>
<tr>
<td>80%</td>
<td>0.05</td>
<td>1554</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>389</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>16</td>
</tr>
<tr>
<td>90%</td>
<td>0.05</td>
<td>2081</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>520</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>22</td>
</tr>
</tbody>
</table>

The concept of genetic structure.

Our discussion of genotypic variation is related to our previous discussion of heritability through coefficients of coancestry and inbreeding. We can answer questions about the relationships between and within populations and individuals through pedigree analysis or a priori knowledge of population structure. We can also deduce these relationships by measuring allele frequencies and distributions within sampled populations.

Under Hardy-Weinberg equilibrium (HWE) genetic structure is absent and the distribution of alleles within and between individuals and populations can be described by $p^2 + 2pq + q^2 = 1$. However isolation of small populations, selection, and mating system (to name a few factors) may result in a distribution of alleles within and between individuals and populations that departs from Hardy-Weinberg equilibrium. Most breeding programs have some genetic structure by design. All genetic mapping relies on genetic structure.

Wright’s F statistics

Wright first proposed the concept of using “F” statistics to describe the degree of relatedness of various pairs of alleles between populations, in individuals within populations, and within individuals.

F is also sometimes called the fixation index and is a measure of genetic differentiation of a sub-population relative to the total population (presumably due to non-random mating).

$$F = 1 - \frac{\text{observed heterozygosity}}{\text{heterozygosity expected under random mating in an infinite population}}$$

Note that all F statistics are “fixation indexes” and are related to the panmitic index $P = (1 - F)$.
F statistics can be formulated to express fixation at different levels of subdivision in a population (within the total population, within sub-populations, and within individuals).

- Coancestry coefficient \( F_{ST} \) is the correlation of genes from individuals in the same population.
- Inbreeding coefficient \( F_{IT} \) is the correlation of genes within individuals.
- Within pop. inbreeding coefficient \( F_{IS} \) is the correlation of genes within individuals within populations.

**An area of Active Research in Population Genetics:**

The formulas and statistical approaches used to calculate these parameters (or statistics) differ and make different assumptions about population sizes, balanced data, alleles per locus, and more. Different statistics (parameters) and notation have been developed to address problems and assumptions of Wright’s methodology. These problems include 1) high sampling variance for \( F \) in the range of \((0<F<0.3)\) occur when gene frequencies are in the most useful range \((0.2 < p < 0.5)\); 2) the inability to test the hypothesis \( F_{ST} = 0 \) due to a lack of normal distribution for alleles of a gene; disagreement over whether fixation indexes are parameters with precise definitions or statistics; and 4) confusion (and disagreement) over the appropriate sampling variance. We will briefly review three approaches for measuring population substructure. The approaches are equivalent for large diploid populations with two alleles per locus.

**Assignment:**

Read Chapter 7 and Chapter 14 in Lynch and Walsh.

Go to the World Wide Web and Identify sites relevant to Plant Genetic Resources for
- Wheat
- Corn
- Soybean
- Cotton
- Arabidopsis
- Rye (and other grasses)

Identify sites relevant to
- Markers and genetic mapping
- Germplasm resources