Lecture Handout 5-Genetic diversity

Exam1 will cover class lectures, chapters 1 and 2 in textbook, and handouts (including problem sets) on Compass.

Exam1 from two previous years will be posted in Compass for your review

More info will be provided Friday and Monday.

For further reference, several additional textbooks on population genetics are on reserve at the ACES library.

A commonly used measure of genetic variation in a population is the amount of **heterozygosity**. For a population at Hardy-Weinberg equilibrium:

$$H_E = 1 - \sum_{i=1}^{n} p_i^2$$

Or one minus the HW homozygosities

Estimated heterozygosity should be corrected for small sample sizes:

$$\hat{H}_E = \frac{2N}{2N-1}\left(1 - \sum_{i=1}^{n} \hat{p}_i^2\right)$$

For sets of loci (e.g., microsatellites), the mean or overall heterozygosity across all tested loci can be estimated as:

$$\hat{H} = \frac{1}{Nm} \sum_{i=1}^{N} \sum_{j=1}^{m} H_{ij}$$

Genetic polymorphism: the occurrence of two or more alleles at one locus "each with appreciable frequency" "within the same population".

Proportion of polymorphic loci, P, for a population, is estimated as:

$$\hat{P} = \frac{x}{m}$$

Where x is number of polymorphic loci and m is the number of loci examined (not useful if most loci have many polymorphism)

Allele diversity or allele richness, A, is a count of the number of alleles (n) observed at a locus. But: very dependent on sample size, so should be subject to rarefaction.

Effective number of alleles, n_e, is calculated as the inverse of the expected homozygosity:

$$\hat{n}_e = \frac{1}{1 - \hat{H}}$$

Measures of DNA diversity

segregating sites, S: number of nucleotide sites that are polymorphic in a sequence when the locus is compared across individuals. S has not been adjusted for the sample size.

Proportion of nucleotide sites at which 2 sequences differ, also called the **p distance** is estimated as:

$$\hat{p}_S = \frac{S}{N}$$

where N is the sequence length

Nucleotide diversity, π: Over all possible pairs of sequences, determine the proportion of nucleotide differences between pairs of sequences, weighed by the frequencies of the sequences:

$$\pi = \sum_{ij} p_i p_j \pi_{ij}$$

p is frequency of sequences i or j; πij is the proportion of nucleotides that differ between sequence i and sequence j.

Protein diversity: Similar formulae are used, comparing amino acid sites rather than nucleotides, in order to determine amino acid sites that are segregating, or amino acid diversity.

Measures of genetic distance

- consolidate the data on variation

- can help in visualizing general relationships

- but can also obscure details

- distance measures are analogous to geometric distances

Are two populations significantly different in allele frequencies?

- use a chi-square test:

$$\chi^2 = 2N\sum_{i=1}^{n}\frac{V(\hat{p}_i)}{\bar{p}_i}$$

ie, twice the sample size times (the variance times the frequency of allele i, divided by the average frequency of allele i).

Nei's standard genetic distance, D

$$D = -\ln (I)$$

To first calculate **genetic identity, I:**

$$I = \frac{J_{xy}}{(J_x J_y)^{1/2}}$$

$$J_{xy} = \sum_{i=1}^{n} p_{i\cdot x} p_{i\cdot y}, \qquad J_x = \sum_{i=1}^{n} p_{i\cdot x}^2, \qquad J_y = \sum_{i=1}^{n} p_{i\cdot y}^2$$

Note: I (or I' for multiple loci) goes from zero (no alleles shared) to one (all alleles shared)

Nei's genetic distance for multiple loci, D'

J'xy, J'x and J'y values are calculated by summing over alleles at all loci in study; average value is calculated (divide by number of loci)

Average J'xy, J'x and J'y values are used to calculate I'

$$D' = -\ln(I')$$

Note: I (or I' for multiple loci) goes from zero (no alleles shared) to one (all alleles shared)