

Population Subdivision

When populations are subdivided, movement between subpopulations may be restricted and allele frequencies may be different in the subpopulations (i.e., populations are not panmictic). This is similar to assortative mating with respect to geographical location and/or finite population size (i.e., drift)

- I) Wahlund's Principle – Deficit in number of observed heterozygotes relative to expected due to population subdivision.

Assume one large group of organisms divided into two subpopulations. HWE exists within each population.

OH 5.1.1

$$\text{Separate: } F(aa) = \frac{q_1^2 + q_2^2}{2}$$

$$\text{Overall: } F(aa) = \left(\frac{q_1 + q_2}{2} \right)^2$$

OH 5.1

$$\bar{H}_s = \frac{1}{k} \sum_{i=1}^k \left(1 - \sum_{j=1}^n p_j^2 \right)$$

Where: n = number of alleles

k = number of subpopulations

$$H_T = 1 - \sum_{i=1}^n \left(\frac{1}{k} \sum_{j=1}^k p_i \right)^2$$

That is for all values of p_i :

$$H_T \geq \bar{H}_s$$

In other words:

Where: $\bar{q} = \frac{1}{k} \sum_{i=1}^k q_i$ **Average allele frequency across subpopulations**

And $\bar{Q} = \frac{1}{k} \sum_{i=1}^k q_i^2$ **Averaged observed homozygotes frequency across subpopulations.**

The average, expected homozygotes frequency across subpopulations is:

$$\bar{q}^2 = \left(\frac{1}{k} \sum_{i=1}^k q_i \right)^2$$

Difference between observed and expected is:

$$\begin{aligned} \bar{Q} - \bar{q}^2 &= \frac{1}{k} \sum_{i=1}^k q_i^2 - \left(\frac{1}{k} \sum_{i=1}^k q_i \right)^2 \\ &= \frac{1}{k} \sum_{i=1}^k (q_i - \bar{q})^2 = V_q \text{ (the variance in allele frequency} \\ &\hspace{15em} \text{across subpopulations)} \end{aligned}$$

Then: $\bar{Q} = \bar{q}^2 + V_q$ **and**

$$\bar{P} = \bar{p}^2 + V_p$$

$$\bar{H} = 2\bar{p}\bar{q} - 2V_q$$

Note: $V_q = V_p$ only when two alleles at a locus. **Observed**

Heterozygosity is less than expected (i.e., variance in allele frequency is subtracted from the expected heterozygote frequency to get observed).

II) Wright's F Statistics – Wright, S. 1951. The genetical structure of populations. Ann. Eugen. 15:323-354.

A) Three levels on which heterozygosity can be measured:

- 1) (I)ndividual – effected by breeding behavior or system**
- 2) (S)ubpopulation – effected by migration or subdivision**
- 3) (T)otal Population – effected by both breeding and subdivision.**

H_I – Average heterozygosity of all genes in an individual or the probability of any one gene being heterozygous.

H_S – Average heterozygosity of the subpopulation or the probability of an individual being heterozygous.

H_T – Average heterozygosity of randomly mating pooled subpopulations

Wrights formulation is for one locus with two alleles. A generalized extension is:

$$H_I = \frac{1}{k} \sum_{i=1}^k H_i \quad \text{for } k \text{ subpopulations}$$

$$H_S = 1 - \sum_{i=1}^n p_i^2 \quad \text{in any one subpopulation for } n \text{ alleles}$$

\bar{H}_S = is the average heterozygosity across subpopulations

$$H_T = 1 - \sum_{i=1}^n \bar{p}_i^2 \quad \text{where } \bar{p} = \text{average allele frequency across subpopulations}$$

From these, three statistics can describe the partitioning of genetic variation in populations:

$$F_{IS} = \frac{\bar{H}_s - H_I}{\bar{H}_s}$$

= the reduction of individual heterozygosity due to non-random mating within subpopulations (i.e., inbreeding).

$$F_{ST} = \frac{H_T - \bar{H}_s}{H_T}$$

= the reduction in subpopulation heterozygosity due to non-random mating within the total population (i.e., population subdivision).

$$F_{IT} = \frac{H_T - H_I}{H_T}$$

= the reduction in individual heterozygosity due to non-random mating within the total population and within the subpopulations (i.e., inbreeding and population subdivision).

$$(1 - F_{IT}) = (1 - F_{ST}) (1 - F_{IS})$$

$$F_{ST} = \frac{F_{IT} - F_{IS}}{1 - F_{IS}}$$

$$F_{ST} = \frac{2\text{Var}(q)}{2\bar{q}(1 - \bar{q})} = \frac{\text{Var}(q)}{\bar{q}(1 - \bar{q})}$$

B) Weir and Cockerham – 1984. Estimating F-statistics for the analysis of population substructure. Evolution 38:1358-1370.

As originally formulated, F statistics are effected by sample size, number of subpopulations sampled, equality of sample sizes across subpopulations (single locus). Weir and Cockerham formally corrected for these.

Note: F_{IS} and F_{IT} can be either positive or negative (+ is an excess of heterozygotes and – is a deficit)

$$0.0 \leq F_{ST} \leq 1.0$$

where 0.0 is no difference in allele frequencies among subpopulations and 1.0 is complete difference. This is true because $H_T \geq \bar{H}_S$, always.

OH 5.1

III) Interpretation of F_{ST}

F_{ST} measures the degree that the subpopulations have gone to alternate fixation of alleles NOT the actual degree of gene flow between populations.

OH 5.2

Case 1 versus 6

Remember: $F_{ST} = \frac{\text{Var}(q)}{\bar{q}(1 - \bar{q})}$

Case 6 versus 7

Pair-wise F_{ST} help differentiate between these, however, they are not statistically independent

$\chi^2 = 2NF_{ST}$ where N is the sample size

OH 5.3

Case 1 versus 2

IV) Measures of Genetic Distance and Identity Among Populations

Many available and most equally good and highly correlated.

Some are metric and some are non-metric

A) Triangular equality:

$$\text{Distance from A} \Rightarrow \text{B} + \text{B} \Rightarrow \text{C} = \text{A} \Rightarrow \text{C}$$

B) Nei's Distance

1) does not satisfy the TE (i.e., non-metric)

for a single locus:

$$I = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$$

where x_i and y_i are the frequencies of the i^{th} allele (of n) in the x and y populations, respectively.

$$0.0 \leq I \leq 1.0$$

for multiple loci:

$$I = \frac{J_{xy}}{\sqrt{J_x J_y}}$$

where $J_{xy} = \frac{1}{k} \sum_{i=1}^k \sum_{i=1}^n x_i y_i$, $J_x = \frac{1}{k} \sum_{i=1}^k \sum_{i=1}^n x_i^2$, and $J_y = \frac{1}{k} \sum_{i=1}^k \sum_{i=1}^n y_i^2$ and

k = number of loci and n = number of alleles (i.e., the arithmetic means across loci)

Nei's distance (D) = $-\ln I$

$$0 \leq D \leq \infty$$

C) Roger's Distance

$$D = \sqrt{0.5 \sum_{i=1}^n (x_i - y_i)^2}$$

$$\bar{D} = \frac{1}{k} \sum_{i=1}^k D$$

where k = number of loci

$$0.0 \leq D \leq 1.0$$

within versus between subpopulation variation has the same effect on Roger's D as on F_{ST} .

**D) Cavalli-Sforza and Edwards – overcomes variance problem.
Uses an angular transformation to make variances independent of frequency.**

$$D_{\text{arc}} = \sqrt{\frac{1}{k} \sum_{i=1}^k \left[\frac{2 \left(\cos^{-1} \sum_{i=1}^n x_i y_i \right)}{\pi} \right]^2}$$