

ANALYZING THE GENETIC STRUCTURE OF POPULATIONS

8 populations of *Isotoma petraea* in southwestern Australia surveyed for genotype at *GOT-1* (James et al. *Heredity* **51**:653–663; 1983).

Population	Genotype			p
	A_1A_1	A_1A_2	A_2A_2	
1	14	3	3	0.7750
2	15	2	3	0.8000
3	13	0	0	1.0000
4	23	5	2	0.8500
5	23	3	4	0.8167
6	29	3	1	0.9242
7	5	0	0	1.0000
8	0	1	0	0.5000

$$\bar{p} = 0.8332$$

$$V_p = 0.02250$$

$$F_{ST} = 0.1619$$

$$\begin{aligned} \text{Sample obs. heterozygosity} &= (0.1500 + 0.1000 + 0.0000 + 0.1667 + 0.1000 + 0.0909 + 0.0000 + \\ &\quad 1.0000)/8 \\ &= 0.2009 \end{aligned}$$

$$\begin{aligned} \text{Sample exp. heterozygosity} &= 2(0.8332)(1 - 0.8332) \\ &= 0.2779 \end{aligned}$$

$$\begin{aligned} F_{IT} &= 1 - \frac{\text{sample obs. heterozygosity}}{\text{sample exp. heterozygosity}} \\ &= 1 - \frac{0.2009}{0.2779} \\ &= 0.2769 \end{aligned}$$

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

$$\begin{aligned} F_{IS} &= \frac{F_{IT} - F_{ST}}{1 - F_{ST}} \\ &= \frac{0.2769 - 0.1619}{1 - 0.1619} \\ &= 0.1372 \end{aligned}$$

SUMMARY.

Correlation of gametes due to inbreeding within subpopulations (F_{IS}): 0.1372

Correlation of gametes within subpopulations (F_{ST}): 0.1619

Correlation of gametes in sample (F_{IT}): 0.2769

Note: These calculations are equivalent to calculating Nei's gene diversity statistics without a bias correction introduced below.

STATISTICAL EXPECTATION AND BIASED ESTIMATES

The concept of statistical expectation is actually quite an easy one. It is an arithmetic average, just one calculated from probabilities instead of being calculated from samples. So, for example, if $P(k)$ is the probability that we find k A_1 alleles in our sample, the *expected number* of A_1 alleles in our sample is just

$$\begin{aligned} E(k) &= \sum kP(k) \\ &= np \quad , \end{aligned}$$

where n is the total number of alleles in our sample and p is the frequency of A_1 in our sample.

Now consider the expected value of our sample estimate of the population allele frequency, $\hat{p} = k/n$, where k now refers to the number of A_1 alleles we actually found.

$$\begin{aligned} E(\hat{p}) &= E\left(\sum(k/n)\right) \\ &= \sum(k/n)P(k) \\ &= (1/n)\left(\sum kP(k)\right) \\ &= (1/n)E(k) \\ &= (1/n)(np) \\ &= p \quad . \end{aligned}$$

Because $E(\hat{p}) = p$, \hat{p} is said to be an unbiased estimator of p .

What about estimating the frequency of heterozygotes within a population? The obvious estimator is $H = 2\hat{p}(1 - \hat{p})$. Well,

$$\begin{aligned} E(H) &= E(2\hat{p}(1 - \hat{p})) \\ &= 2(E(\hat{p}) - E(\hat{p}^2)) \\ &= ((n - 1)/n)2p(1 - p) \quad . \end{aligned}$$

Because $E(H) \neq 2p(1 - p)$, H is said to be a biased estimator of $2p(1 - p)$. If we set $\hat{H} = (n/(n - 1))H$, however, \hat{H} is an unbiased estimator of $2p(1 - p)$.

THE GORY DETAILS¹

Starting where we left off above:

$$\begin{aligned} E(H) &= 2((E\hat{p}) - E(\hat{p}^2)) \\ &= 2(p - E((k/n)^2)) \quad , \end{aligned}$$

¹ Skip these unless you are *really*, *really* interested in how I got from the second equation to the third equation in the last paragraph. This is more likely to confuse you than help unless you know that the variance of a binomial sample is $np(1 - p)$ and that $E(k^2) = \text{Var}(p) + p^2$.

where k is the number of A_1 alleles in our sample and n is the sample size.

$$\begin{aligned}
 E\left(\left(\frac{k}{n}\right)^2\right) &= \sum (k/n)^2 P(k) \\
 &= (1/n)^2 \sum k^2 P(k) \\
 &= (1/n)^2 (\text{Var}(p) + p^2) \\
 &= (1/n)^2 (np(1-p) + p^2) \quad .
 \end{aligned}$$

Substituting this back into the equation above yields the following:

$$\begin{aligned}
 E(H) &= 2 \left(p - (1/n)^2 (np(1-p) + p^2) \right) \\
 &= 2 (p(1-p) - p(1-p)/n) \\
 &= (1 - 1/n) 2p(1-p) \\
 &= ((n-1)/n) 2p(1-p) \quad .
 \end{aligned}$$

CORRECTIONS FOR SAMPLING ERROR.

There are two sources of allele frequency difference among populations in our sample: (1) real differences in the allele frequencies among our sampled populations and (2) differences that arise because allele frequencies in our population samples differ from those in the population.

BIAS-CORRECTION METHOD DUE TO NEI AND CHESSER (*Ann. Hum. Genet.* **47**:253-259; 1983. implemented in **Genestat-PC**²).

$$\begin{aligned}
 H_I &= 1 - \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^m X_{kii} \\
 H_S &= \frac{\tilde{n}}{\tilde{n}-1} \left[1 - \sum_{i=1}^m \hat{x}_i^2 - \frac{H_I}{2\tilde{n}} \right] \\
 H_T &= 1 - \sum_{i=1}^m \bar{x}_i^2 + \frac{H_S}{\tilde{n}} - \frac{H_I}{2\tilde{n}N}
 \end{aligned}$$

where we have N subpopulations, $\hat{x}_i^2 = \sum_{k=1}^N x_{ki}^2/N$, $\bar{x}_i = \sum_{k=1}^N x_{ki}/N$, \tilde{n} is the harmonic mean of the population sample sizes, i.e., $\frac{\tilde{n} = (1/((1/N) \sum_{k=1}^N (1/n_k)))}$, X_{kii} is the frequency of genotype $A_i A_i$ in population k , x_{ki} is the frequency of allele A_i in population k , and n_k is the sample size from population k . Recall that

$$\begin{aligned}
 F_{IS} &= 1 - \frac{H_I}{H_S} = 0.2166 \\
 F_{ST} &= 1 - \frac{H_S}{H_T} = 0.0866 \\
 F_{IT} &= 1 - \frac{H_I}{H_T} = 0.2846
 \end{aligned}$$

² Actually, you have to calculate H_I by hand, but that's only tedious, not hard.

METHOD DUE TO WEIR AND COCKERHAM (*Evolution* **38**:1358-1370; 1984. implemented in **Genetic Data Analysis**)

Often the populations we have sampled are only a subset of all populations in which we might be interested. We may then be interested in making inferences about *all* populations, not only those that we happened to sample. Weir and Cockerham's method allows us to make this additional inference (see also pp. 161–190 in Weir, *Genetic Data Analysis II*).³ The formulas are pretty complicated, so I won't present them here. You can look them up in *GDAll*, if you're interested.

$$\begin{aligned}F_{IS} &= 0.5356 \\F_{ST} &= 0.0160 \\F_{IT} &= 0.5430\end{aligned}$$

AN EXAMPLE FROM WRIGHT.

Hierarchical analysis of variation in the frequency of the Standard chromosome arrangement of *Drosophila pseudoobscura* in the western United States (data from Dobzhansky and Epling, *Carnegie Inst. Wash. Publ.* 554; 1944. analysis by Wright, *op. cit.*).

66 populations (demes) studied. Demes are grouped into eight regions. The regions are grouped into four primary subdivisions.

RESULTS.

Correlation of gametes within demes (F_{ID}):	0.0444
Correlation of gametes within regions (F_{RS}):	0.0373
Correlation of gametes within subdivisions (F_{ST}):	0.1478
Correlation of gametes in sample (F_{IT}):	0.2160

$$1 - F_{IT} = (1 - F_{IR})(1 - F_{RS})(1 - F_{ST})$$

INTERPRETATION.

There is great geographical differentiation among the populations in the frequency of the Standard chromosome arrangement ($F_{IT} = 0.2160$, but $F_{ID} = 0.0444$).

Most of this geographical differentiation is a result of differences among the four primary subdivisions ($F_{ST} = 0.1478$). Relatively minor differentiation is found among regions within a given subdivision ($F_{RS} = 0.0373$) and among demes within a region ($F_{DR} = 0.0444$).

Thus, an explanation for the chromosomal diversity that predicted great local differentiation and little or no differentiation at a large scale would be inconsistent with these observations.

³ For those of you who have had some statistics, Nei and Chesser's approach corresponds, roughly, to a fixed-effect ANOVA. Weir and Cockerham's approach corresponds to a random-effect ANOVA.