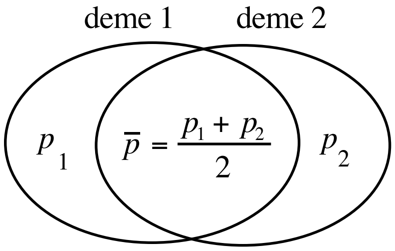
GENETIC POPULATION STRUCTURE

DEFINITIONS

Panmictic Index: $P = \dfrac{H_{\text{obs}}}{H_{\text{exp}}} = 1 - F$, where H is heterozygosity, or gene diversity

Fixation Index: $F = 1 - P = 1 - \left(\dfrac{H_{\text{obs}}}{H_{\text{exp}}}\right)$.

THE WAHLUND EFFECT



If you sampled from two reproductively isolated demes, allelic frequencies would equal the average of those in the two demes.

Wahlund (1928) determined the heterozygosity in the pooled (T) population as

$$H_T = 2\overline{p}(1 - \overline{p}) - 2\sigma_p^2.$$

You would see deficiency of heterozygotes and a corresponding excess of homozygotes:

	AA	Aa	aa
Population 1:	p_1^2	$2p_1q_1$	q_1^2
Population 2:	p_2^2	$2p_2q_2$	q_2^2
Average:	$\overline{p_i^2} = (p_1^2 + p_2^2)/2$	$\overline{2pq} = (2p_1q_1 + 2p_2q_2)/2$	$\overline{q}^2 = (q_1^2 + q_2^2)/2$
Wahlund's equation:	$\overline{p}^2 + \sigma_p^2$	$2\overline{p}(1 - \overline{p}) - 2\sigma_p^2$	$\overline{q}^2 + \sigma_p^2$
Wright's equation:	$\overline{p}^2(1 - F) + \overline{p}F$	$2\overline{p}\overline{q}(1 - F)$	$\overline{q}^2(1 - F) + \overline{q}F$
Pooled HW	$\overline{p}^2$	$2\overline{p}\overline{q}$	$\overline{q}^2$

If you solve for F from the Wahlund and Wright equations,

$$\overline{p}^2 + \sigma_p^2 = \overline{p}^2(1 - F) + \overline{p}F = \overline{p}^2 - \overline{p}^2F + \overline{p}F$$

$$\sigma_p^2 = \overline{p}F + -\overline{p}^2F = F(\overline{p} - \overline{p}^2) = F\overline{p}(1 - \overline{p}) \therefore$$

$$F = \frac{\sigma_p^2}{\overline{p}(1 - \overline{p})}$$

Define F as the **Standardized Variance**, or F_{ST} of Wright (1931) the denominator, $\overline{p}(1 - \overline{p})$, is the **limiting** (maximal) **variance**. If the binomial variance for a frequency is $p(1 - p)/N$, then you could imagine the limiting variance as the maximum variance in a situation where $N = 1$.

The variance in p is easily estimated from

$$\overline{p^2} = \overline{p}^2 + \sigma_p^2 \ .$$

Solving for the variance

$$\sigma_p^2 = \overline{p_i^2} - \overline{p}^2 \ .$$

In a genetically subdivided population, the frequency of homozygotes is greater than the Hardy-Weinberg expectation for a pooled population**. Looks like inbreeding doesn't it, but it is not caused by consanguinous matings. This could fool you if you did not know the population was subdivided.**

The checkerboard example; variance in p *is maximum*

	Genotypic frequencies			Allelic frequencies	
	AA	*Aa*	*aa*	*pA*	*qa*
Population 1	1	0	0	1	0
Population 2	0	0	1	0	1
Pooled HW	0.25	0.50	0.25	0.50	0.50
Wahlund	0.50	0.00	0.50		
Wright	0.50	0.00	0.50		

$s_p^2 = \overline{p_i^2} - \overline{p}^2 = \dfrac{1^2 + 0^2}{2} - 0.5^2 = 0.25.$

$H_T = 2\overline{p}\overline{q} - 2s_p^2 = 2(0.5 \cdot 0.5) - 2(0.25) = 0.5 - 0.5 = 0$.

$F_{ST} = 0.25 / 0.25 = 1$. Wright's heterozygosity would be

$2\overline{pq}(1 - F) = 2(0.5)(0.5)(1 - 1) = 0$

A less extreme case

	Genotypic frequencies			Allelic frequencies	
	AA	*Aa*	*aa*	*pA*	*qa*
Population 1	0.25	0.50	0.25	0.50	0.50
Population 2	0.81	0.18	0.01	0.90	0.10
Pooled HW	0.49	0.42	0.09	0.70	0.30
Wahlund	0.53	0.34	0.13		
Wright	0.53	0.34	0.13		

$s_p^2 = \overline{p_i^2} - \overline{p}^2 = \dfrac{0.25 + 0.81}{2} - 0.7^2 = 0.53 - 0.49 = 0.04.$

$H_T = 2\overline{p}\overline{q} - 2s_p^2 = 2(0.7 \cdot 0.3) - 2(0.04) = 0.42 - 0.08 = 0.34.$

$F_{ST} = 0.04 / 0.21 = 0.19$. Wright's heterozygosity would be

$2\overline{pq}(1 - F) = 2(0.7)(0.3)(1 - 0.19) = 0.42(0.81) = 0.34$

The Wahlund Effect with more than two alleles

The situation becomes somewhat more complicated when there are more than two alleles, because in addition to the variance in allele frequencies across demes, some pairs of alleles might covary in frequency. Although this situation still leads to an overall deficiency of heterozygotes, a slight excess of heterozygotes may exist for pairs of alleles that positively covary. For a theoretical analysis of this problem see Nei (1965)

pop	*AA*	*BB*	*CC*	*AB*	*AC*	*BC*
1	p_1^2	q_1^2	r_1^2	$2p_1q_1$	$2p_1r_1$	$2q_1r_1$
2	p_2^2	q_2^2	r_2^2	$2p_2q_2$	$2p_2r_2$	$2q_2r_2$
Wahlund	$\bar{p}^2+\sigma_p^2$	$\bar{q}^2+\sigma_q^2$	$\bar{r}^2+\sigma_r^2$	$2\overline{pq}+2COV_{pq}$	$2\overline{pr}+2COV_{pr}$	$2\overline{qr}+2COV_{qr}$

For Example:

pop	*AA*	*BB*	*CC*	*AB*	*AC*	*BC*	sum
1	0.010	0.490	0.040	0.140	0.040	0.280	1.000
2	0.640	0.010	0.010	0.160	0.160	0.020	1.000
mean	0.325	0.250	0.025	0.150	0.100	0.150	1.000
pooled HW	0.203	0.160	0.023	0.360	0.135	0.120	1.000
difference	0.123	0.090	0.003	-0.210	-0.035	0.030	0.000
adjusted	0.325	0.250	0.025	0.150	0.100	0.150	1.000

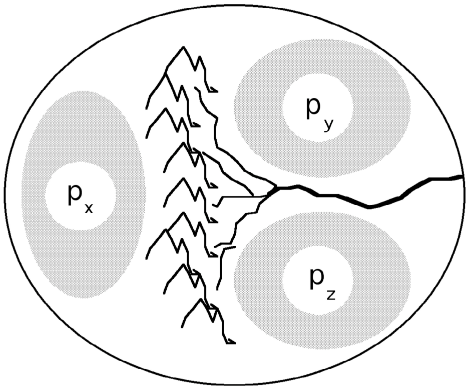
pop	*p(A)*	*p(B)*	*r(C)*	sum
1	0.100	0.700	0.200	1.000
2	0.800	0.100	0.100	1.000
mean	0.450	0.400	0.150	1.000
var(*p*)	0.123	0.090	0.003	
covariances		*B*	*C*	
	A	-0.105	-0.018	
	B		0.015	

With *n* multiple alleles it is best to estimate heterozygosity as $H = 1 - \sum_{i=1}^{n} p_i^2$.

Convince your self that this equals 2*pq* in the 2-allele case.

WRIGHT'S F-STATISTICS

Envision a subdivided population (right).



Genetic variation in subdivided population must be considered at three levels: (*I*) individuals within subpopulations; (*S*) subpopulations; (*T*) the total population as if there were no subdivision.

Wright showed that panmictic indices in a subdivided population can be related in the following way:

$$P_{IT} = P_{IS}P_{ST}$$

Since $P = (1 - F)$, $$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$$

Solving for F_{IT}, $$F_{IT} = F_{IS} + (1 - F_{IS})F_{ST}$$

F_{IT} is the deviation from Hardy-Weinberg proportions in the total population.

F_{IS} is the average deviation from Hardy-Weinberg proportions in subpopulations. It is most often due to inbreeding.

F_{ST} is Wright's standardized variance. It is due to the variance among demes.

Lets see what these terms are about by looking at the following examples:

Population	*AA*	*Aa*	*aa*	*p*	*q*	*F*
Population 1	0.25	0.50	0.25	0.50	0.50	0.00
Population 2	0.35	0.30	0.35	0.50	0.50	0.40
Subdivided	0.30	0.40	0.30	0.50	0.50	0.20
Pooled	0.25	0.50	0.25	0.50	0.50	
	$F_{ST} = 0.0$	$F_{IS} = 0.2$	$F_{IT} = 0.2$			
Population 1	0.25	0.50	0.25	0.50	0.50	0.00
Population 2	0.49	0.42	0.09	0.70	0.30	0.00
Subdivided	0.37	0.46	0.17	0.60	0.40	0.04
Pooled	0.36	0.48	0.16	0.60	0.40	
	$F_{ST} = 0.04$	$F_{IS} = 0.0$	$F_{IT} = 0.04$			
Population 1	0.25	0.50	0.25	0.50	0.50	0.00
Population 2	0.53	0.34	0.13	0.70	0.30	0.20
Subdivided	0.39	0.42	0.19	0.60	0.40	0.13
Pooled	0.36	0.48	0.16	0.60	0.40	
	$F_{ST} = 0.04$	$F_{IS} = 0.10$	$F_{IT} = 0.14$			

Real examples (SMOUSE and LONG 1988):

Locus	F_{IS}	F_{IT}	F_{ST}
The Yanamama			
Rh	-0.0465	-0.0138	0.0312**
Duffy	-0.0034	0.0329	0.0363**
MN	0.0242	0.0841	0.0614**
Pgm-1	-0.0157	0.0271	0.0416**
Mean (7 loci)	-0.0157	0.0271	0.0416**
The Gainj and Kalam			
Mean (5 loci)	0.0392	0.0628	0.0225**

** Significant at 0.01

THE WORKMAN-NISWANDER (1970) TEST FOR HETEROGENEITY

The statistical significance of the standardized variance F_{ST} can be determined from the following relationship:

$$\chi^2 = \frac{2N\sigma_p^2}{\overline{p}\overline{q}} = 2NF_{ST}$$

where $\overline{p}$ and $\overline{q}$ are weighted means, σ_p^2 is the weighted variance of p, and N is the total sample size. *Degrees of freedom* are (k - 1)(n - 1), where k is the number of populations and n the number of alleles.

With weighted means and variances, this procedure is equivalent to doing a χ^2 contingency test.

Genotype counts	*AA*	*Aa*	*aa*	N	*p*(*A*)	*q*(*a*)
population 1	475	89	5	569	0.9130	0.0870
population 2	233	385	129	747	0.5696	0.4304
total	708	474	134	1316	0.7181	0.2819
Allele counts		*N(A)obs*	*N(a)obs*	row total	*N(A)exp*	*N(a)exp*
	pop1	1039	99	1138	817.2	320.8
	pop2	851	643	1494	1072.8	421.2
	col. total	1890	742	*G* = 2632	1890.0	742.0

For a contingency test the expected values in each cell are determined as $E_{ij} = (R_i \cdot C_j)/G$, where R_i is the total for row i, and C_j is the total for column j, and G is the grand total.

$\sum \chi^2 = 375.8$, and $df = (R - 1)(C - 1) = 1$; highly significant

Note:

$$\sigma_p^2 = \sum w_i p_1^2 - \bar{p}^2 = \frac{569(0.9130)^2 + 747(0.5696)^2}{1316} - 0.7181^2 = 0.0289$$

$$\bar{p}(1 - \bar{p}) = (0.7181)(0.2819) = 0.2024$$

$$F_{ST} = \frac{0.0289}{0.2024} = 0.1428\text{; and } \chi^2 = 2NF_{ST} = 2(1316)(0.1428) = 375.8$$

PARTITIONING OF GENETIC DIVERSITY

Nei (1973) also showed how you could use these statistics to partition the total genetic diversity to within and between subpopulation components.

$$H_T = \bar{H}_S + V_{ST}$$

where V_{ST} is the variance among subpopulations. Dividing both sides by H_T, we get

$$1 = \frac{\bar{H}_S}{H_T} + \frac{V_{ST}}{H_T} = \frac{\bar{H}_S}{H_T} + G_{ST}$$

where G_{ST} is roughly equivalent to F_{ST}, and $G_{ST} = \frac{V_{ST}}{H_T}$

The two terms can be interpreted as the proportion of the total diversity due to heterozygosity within subpopulations and due to variance among subpopulations.

Examples:

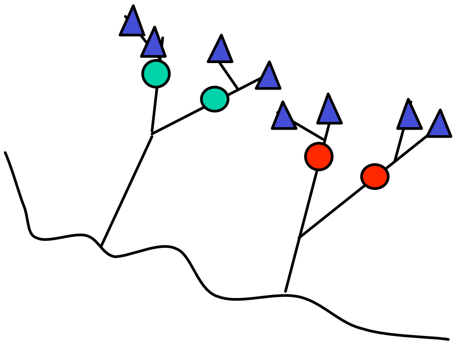
Species	(within) $\bar{H}_S / H_T$	(among) G_{ST}
Homo sapiens	0.93	0.07
Dipodomys ordii	0.30	0.70

The diversity due to differences among major racial groups of humans is relatively small when compared with the diversity due to differences between local populations of a desert rat.

HIERARCHICAL ANALYSIS

Using the relationship $P_{IT} = P_{IS}P_{ST}$, you can see how this type of analysis can be extended to higher levels in hierarchical population structures. Imagine fish demes distributed in local tributaries (L) that lead to different river systems (R). Now,

$$P_{IT} = P_{IL}P_{LR}P_{RT}$$



You can solve this just like before and partition diversity within and between each level.

$$H_T = \overline{H}_L + V_{LR} + V_{RT}$$

Species	within local tributaries $\overline{H}_L / H_T$	between tributaries within rivers V_{LR} / H_T	between rivers V_{RT} / H_T
Poeciliopsis occidentalis	0.21	0.26	0.53
Oncorhynchus clarki lewisi	0.42	0.25	0.33
Oncorhynchus mykiss	0.85	0.08	0.07

ESTIMATING F_{ST}

$$F_{ST} = \frac{\overline{p^2} - \overline{p}^2}{\overline{p}(1-\overline{p})} \quad \text{or} \quad F_{ST} = \frac{s_p^2}{\overline{p}(1-\overline{p})}$$

(Wright 1951)

$$G_{ST} = \frac{\overline{H}_T - \overline{H}_S}{\overline{H}_T}$$

(Nei 1977)

$$\hat{\theta} = \frac{s_A^2 - \frac{1}{\overline{n}-1}\left[\tilde{p}_{A\cdot}(1-\tilde{p}_{A\cdot}) - \frac{r-1}{r}s_A^2\right]}{\frac{n_c-1}{\overline{n}-1}\tilde{p}_{A\cdot}(1-\tilde{p}_{A\cdot}) + \left[1 + \frac{(r+1)(\overline{n}-n_c)}{\overline{n}-1}\right]\frac{s_A^2}{r}}$$

(Weir and Cockerham 1984)

All can be estimated using *Arlequin* (Schneider *et al.* 2000).

For microsatellite data use R_{ST}. a good program called RST-Calc is available (Goodman 1997).

REFERENCES

GOODMAN, S., 1997 Rst Calc: a collection of computer programs for calculating estimates of genetic differentition from microsatellite data and a determining their significance. Molecular Ecology **6:** 881-885.

NEI, M., 1965 Variation and covariation of gene frequencies in subdivided populations. Evolution **19:** 256-258.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences, USA **70:** 3321-3323.

NEI, M., 1977 F-statistics and analysis of gene diversity in subdivided populations. Annals of Human Genetics **41:** 225-233.

SCHNEIDER, S., D. ROESSLI and L. EXCOFFIER, 2000 Arlequin, a software package for population genetics data analysis, pp. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva.

SMOUSE, P. E., and J. C. LONG, 1988 A comparative F-statistics analysis of the Yanamama of lowland South America and the Gainj and Kalam of highland New Guinea, pp. 32-46 in *Proceeding II International Conference on Quantitative Genetics*, edited by B. S. WEIR, G. EISEN, M. M. GOODMAN and G. NAMKOONG. Sinauer Associates, Sunderland, Massachusetts.

WAHLUND, S., 1928 Zusammensetzung von Populationen und Korrelationerscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas **11:** 65-106.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358-1370.

WORKMAN, P. L., and J. D. NISWANDER, 1970 Population studies on Southwestern Indian tribes. II. Local genetic differentiation in the Papago. American Journal of Human Genetics **22:** 24-49.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97-159.

WRIGHT, S., 1951 The genetical structure of populations. Annals of Eugenics **15:** 323-354.