1 Appendix

We begin by giving new formulae for estimation of F_{st}. Suppose we have a biallelic marker in two populations in Hardy-Weinberg equilibrium. Choose the variant allele, and suppose that the allele has population frequency p_1, p_2 in populations 1 and 2 respectively Set $q_i = 1 - p_i$. Then we can define Wright's F_{st} as

$$F_{st} = N/D \tag{1}$$

where

$$\begin{aligned} N &= p_1(q_2 - q_1) + p_2(q_1 - q_2) && (2)\\ D &= p_1q_2 + q_1p_2 = N + p_1q_1 + p_2q_2 && (3) \end{aligned}$$

This is a definition of F_{st}, a parameter measuring divergence at a given locus, *not* a sample statistic. In this paper we are only interested in divergence measures of biallelic markers and the theory will always assume the populations are homogeneous.

Suppose we have a set S of markers $A_k(k = 1, \ldots M)$. For marker k we define now $N^{[k]}$ and $D^{[k]}$ in the obvious way. We now *define* $F(S) = F_{st}$ for the marker set S by

$$F(S) = \frac{N(S)}{D(S)} \tag{4}$$

where

$$\begin{aligned} N(S) &= \frac{\sum_{k=1}^{M} N^{[k]}}{M} && (5)\\ D(S) &= \frac{\sum_{k=1}^{M} D^{[k]}}{M} && (6) \end{aligned}$$

Given the form of equation (4) it is highly desirable to find unbiased estimators of $N^{[k]}, D^{[k]}$ else the bias will eventually dominate the estimate. Fix for now, marker k, and suppose the population frequencies are p_1, p_2 for the variant allele, and we observe allele counts a_1, a_2 for the variant allele, b_1, b_2 for the reference allele. Take $n_i = a_i + b_i,\ i = 1, 2$. $N = N^{[k]}$ is defined as $(p_1 - p_2)^2$. A naive estimator for N is

$$X = (a_1/n_1 - a_2/n_2)^2$$

We calculate the bias of X. Writing

$$X = ((a_1/n_1 - p_1) - (a_2/n_2 - p_2) + (p_1 - p_2))^2$$

Then

$$\begin{aligned} E(X) &= (p_1 - p_2)^2 + Var(a_1/n_1|p_1) + Var(a_2/n_2|p_2) && (7)\\ &= (p_1 - p_2)^2 + p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2 && (8) \end{aligned}$$

Define $h_1 = p_1(1-p_1)$ ($2h_1$ is the heterozygosity at the marker for population 1). Then a natural estimator for h_1 is

$$\hat{h}_1 = \frac{a_1(n_1 - a_1)}{n_1(n_1 - 1)} \tag{9}$$

It is easy to check that $\hat{h_1}$ is unbiased. Similarly define h_2 for population 2, with a corresponding estimator $\hat{h}_2$ This is enough to show that:

$$\hat{N} = (a_1/n_1 - a_2/n_2)^2 - \hat{h_1}/n_1 - \hat{h_2}/n_2 \tag{10}$$

is an unbiased estimator for N. Now

$$D = N + h_1 + h_2$$

which shows

$$\hat{D} = \hat{N} + \hat{h}_1 + \hat{h}_2 \tag{11}$$

is an unbiased estimator for D.
By the Lehmann-Scheffé theorem [1, Theorem 4.2.2] $\hat{N}$ and $\hat{D}$ are uniformly minimum variance unbiased estimators. No longer fixing a marker and writing $\hat{N}^{[k]}$ for our estimator of $N^{[k]}$, and so on, we see that a natural estimator for $F(S)$ is

$$\hat{F} = \frac{\sum_k \hat{N}^{[k]}}{\sum_k \hat{D}^{[k]}} \tag{12}$$

Note that (12) does *not* give an unbiased estimator. However the law of large numbers does imply that as sample size or the number of unlinked markers become large we get an estimator that is asymptotically consistent.

Given our assumptions, our estimates of $N^{[k]}$, $D^{[k]}$ are exactly unbiased both here and in the section below. Our formulae are different from those of Weir and Cockerham [6], at least when population sample sizes differ.

1.1 Estimators in the presence of inbreeding

The estimators above are not correct if there is inbreeding. We continue to assume that within a population there is no structure, but no longer assume that the pair of chromosomes of each sample are unrelated. Thus we may have excess homozygosity compared with Hardy-Weinberg equilibrium.

We extend our theory to this case. We give estimators of N, D that are unbiased, without explicitly estimating the inbreeding coefficients. Let x_0, x_1, x_2 be the number of samples of population 1 with 0, 1, 2 copies of the variant allele. Let y_0, y_1, y_2 be the corresponding numbers for population 2. Let

$$\begin{aligned} s &= x_0 + x_1 + x_2 \\ t &= y_0 + y_1 + y_2 \end{aligned}$$

We will require that $s, t > 1$. In the notation of the previous section:

$$\begin{aligned} a_1 &= x_1 + 2x_2 \\ a_2 &= y_1 + 2y_2 \\ n_1 &= 2s \\ n_2 &= 2t \end{aligned}$$

which will lead to estimators for N, D. In the presence of inbreeding, these estimators are incorrect. Note however that if we pick alleles randomly from each diplotype, then we will obtain valid unbiased estimators. We can of course then obtain more efficient estimators by averaging over our choice of alleles.

Select an allele at random from each diploid genotype. Let u be the allele count for population 1, and v be the count for population 2. From equation (10) we want to compute expected values of:

$$\begin{aligned} X &= (u/s - v/t)^2 \\ \hat{h}_1 &= \frac{u(s-u)}{s(s-1)} \\ \hat{h}_2 &= \frac{v(t-v)}{t(t-1)} \end{aligned}$$

when our estimator for N is

$$\hat{N} = E(X) - E(\hat{h}_1)/s - E(\hat{h}_2)/t \tag{13}$$

For X, we see that u has mean $x_1/2 + x_2$ and variance $x_1/4$. Similarly v has mean $y_1/2 + y_2$ and variance $y_1/4$. It follows that

$$E(X) = \left(\frac{x_1 + 2x_2}{2s} - \frac{y_1 + 2y_2}{2t} \right)^2 + \frac{x_1}{4s^2} + \frac{y_1}{4t^2}$$

For $E(\hat{h})$ we need the expected value of $u(s-u)$. Standard binomial coefficient identities show that

$$E(u(s-u)) = x_0 x_2 + (x_0 + x_2)x_1/2 + x_1(x_1 - 1)/4$$

Now it follows that:

$$E(\hat{h}_1) = \frac{x_0 x_2 + (x_0 + x_2)x_1/2 + x_1(x_1 - 1)/4}{s(s-1)} \tag{14}$$

$$E(\hat{h}_2) = \frac{y_0 y_2 + (y_0 + y_2)y_1/2 + y_1(y_1 - 1)/4}{t(t-1)} \tag{15}$$

We now can apply equation (13) to obtain $\hat{N}$. For $\hat{D}$ we have, using $D = N + h_1 + h_2$ the equation

$$\hat{D} = \hat{N} + E(\hat{h_1}) + E(\hat{h_2}) \tag{16}$$

These formulae are slightly different from those of [5] who correct for inbreeding by directly estimating an inbreeding 'fixation index' (see below) and state that their estimates of the numerator N and denominator D are only 'approximately unbiased'. (see their equation (8)). We now obtain, using estimates over many markers

$$\hat{F} = \frac{\sum_k \hat{N}^{[k]}}{\sum_k \hat{D}^{[k]}} \tag{17}$$

where $\hat{N}^{[k]}, \hat{D}^{[k]}$ are the estimators above, robust to inbreeding, for marker k. Just as before, the estimator of (17) is not unbiased but asymptotically consistent as the number of unlinked markers becomes large.

The same ideas lead to a simple estimator of the inbreeding coefficient, p_I. the probability, in a sample from a population, that the two alleles at a locus are identical by descent (IBD). For our case, with an assumed homogeneous population, this is the same as Wright's fixation index F. (See [4, page 154]). Consider population 1, with the same notation as above. Let H be the probability that two alleles from an individual are heterozygous. Then

$$H = (1 - p_I)h$$

so that $p_I = (h - H)/h$. An unbiased estimator of H is

$$\hat{H} = \frac{x_1}{s}$$

Thus we obtain a natural estimate of p_I:

$$\hat{p}_I = \frac{\sum(\hat{h} - \hat{H})}{\sum \hat{h}} \tag{18}$$

where we sum over all SNPs in our data.

We have not yet worked out the theory, but it would appear that these estimators of F_{st} have, in the absence of inbreeding, standard errors that are only a little increased from the 'optimal' estimators using equations (10, 11).

2 f-statistics

We now discuss our f-statistics. f_4 is the simplest. We have 4 distinct populations W, X, Y, Z. An allele has population frequencies w, x, y, z respectively We observe counts w_0, w_1 of the allele and the complementary allele in a sample from population W. Similarly we observe counts $x_0, x_1; y_0, y_1; z_0, z_1$. We will assume that the total count for each population is at least 2. Thus the natural (naive) estimator of w is

$$w' = \frac{w_0}{(w_0 + w_1)}$$

with similar definitions of x', y', z'. We wish to form unbiased estimates of quantities such as $(w-x)(y-z)$ which we term an *f_4-statistic*. It is easy to see that the naive estimate

$$f_4(W,X,Y,Z) = (w'-x')(y'-z')$$

indeed is an unbiased estimator. Next suppose we want an estimator *(f_3-statistic)* for $(w-x)(w-y)$ where w appears twice. Consider the naive estimator: $q = (w'-x')(w'-y')$ Then we can write q as

$$q \quad = \quad ((w'-w)-(x'-x)+(w-x))((w'-w)-(y'-y)+(w-y))$$

This shows that the bias of q is $E(w'-w)^2$. Let $n_W = w_0 + w_1$ be the total allele count for W. Then

$$E(w'-w)^2 = \frac{w(1-w)}{n_W}$$

Define $h_W = w(1-w)$
($2\ h_W$ is the heterozygosity at the marker for population W). Then a natural estimator for h_W is $\hat{h}_W =$ defined analogously to h_1.

$$f_3(W,X,Y) = (w'-x')(w'-y') - \hat{h}_W/n_W$$

and f_3 is an unbiased estimator of $(w-x)(w-y)$. Similarly we can define

$$f_2(W,X) = (w'-x')(w'-x') - \hat{h}_W/n_W - \hat{h}_X/n_X$$

and show that $f_2(W,X)$ is an unbiased estimator of $(w-x)^2$.

In applications we always wish to compute weighted sums of the f-statistics across many markers. Unbiasedness is critical here ensuring convergence of our average f-statistic to the average we would obtain by using the true allele frequencies.

2.1 The Denominator

For $f = F_{st}$ we have shown how to compute estimators for marker k $\hat{N}^{[k]}, \hat{D}^{[k]}$. Our estimate $\hat{F}$ for F is now simply:

$$\hat{F} = \frac{\sum_k \hat{N}^{[k]}}{\sum_k \hat{D}^{[k]}}$$

For our f-statistics we have some choices. Our key idea is that the denominator should not be population dependent. All our statistics are valid under any reasonable choice, and what we did was the following.

We picked an outgroup (Hapmap Yoruba (YRI)), chosen as a 'neutral' population relative to the non-African populations studied here.

1. For our graph calculations in Figure 4 we wanted to mimic our F_{st} estimates closely, and the f-statistics are

$$\frac{s \sum_k \hat{N}^{[k]}}{\sum_k \hat{D}^{[k]}}$$

where $D^{[k]}$ is $p_k(1-p_k)$, and p_k is the empirical frequency of the variant allele at marker i in YRI. We require $0 < p_k < 1$. Here, s is an arbitrary scalar, unimportant for the analysis. Our f-statistics have no denominator and so are in some sense 'dimensionless'. (Of course when we apply a statistical test, such as a Z-score, the statistic is invariant to scaling). The raw f-statistics are dependent on irrelevant quantities, such as the allelic spectrum of ascertained markers, and thus are not comparable across different data sets. We chose to scale f_2 to minimize the deviations from F_{st} by least squares, considering all pairs of populations in the analysis being carried out. We then rescale f_3, f_4 using the same scale factor. This makes all our quantities have units on the same scale as F_{st} and make our inferences interpretable as genetic drift. The only effect of the outgroup here is to force our markers to be polymorphic in our YRI samples — this is appropriate for our purposes, as 'private' alleles are not of interest here.

2. For our 4-population test we use the formula:

$$\hat{f} = \sum_k \hat{N}^{[k]} / \hat{D}^{[k]}$$

where $\hat{D}^{[k]}$ is defined as above. This seemed to give more sensitivity. Note that *any* weighting of the $\hat{N}^{[k]}$ is statistically valid (here we use weights $1/\hat{D}^{[k]}$), at least if the weights are chosen only using outgroup data. We do not yet understand what 'optimal' weights would be, in terms of statistical power.

3. For our 3-population test where we are estimating $(p_X - p_Y)(p_X - p_W)$ the population X plays a distinguished role in this expression (and indeed we are testing here the genetic history of X). We therefore set $\hat{D}^{[k]}$ to be an unbiased estimate of the heterozygosity at marker k for population X (using (9). We use

$$\hat{f}_3 = \frac{\sum_k \hat{N}^{[k]}}{\sum_k \hat{D}^{[k]}}$$

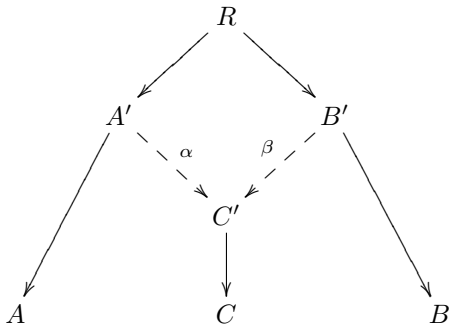
In all cases standard errors (and statistical significance) are estimated through a weighted block jackknife [3, 2]. We use a block size of $5cM$.

2.2 Expected values of our f-statistics

.

We can calculate expected values, at least for simple demographies, involving

populations splits and admixture events (but not yet migrations occurring continuously in time). We give an illustration for our f_3-statistics. Consider a demography:



Here, populations A', B' split from a root population R. C' then was formed by admixture in proportions $\alpha : \beta$ ($\beta = 1-\alpha$). Modern populations A, B, C are then formed by drift from A', B', C'. We want to calculate the expected value of $f_3(C; A, B)$ That is we want

$$F_3 = E(f_C - f_A, f_C - f_B)$$

where f_A, f_B, f_C are allele frequencies in A, B, C respectively. We see by orthogonality of drifts that

$$F_3(C; A, B) = E(f_{C'} - f_A, f_{C'} - f_B) + E((f_{C'} - f_C)^2).$$

($f_{C'}$ is the allele frequency in C') which we will write as

$$F_3(C; A, B) = F_3(C'; A, B) + F_2(C, C') \tag{19}$$

Now, label alleles at a marker 0, 1. Then picking chromosomes from our populations independently we can write

$$F_3(C'; A, B) = E(c' - a)(c'' - b))$$

where a, b, c', c'' are alleles in populations A, B, C'. However c' originated from A' with probability α and so on. Thus:

$$\begin{aligned} F_3(C'; A, B) &= E(c' - a)(c'' - b) \\ &= \alpha^2 E(a' - a)(a'' - b) + \\ &\quad \beta^2 E(b' - a)(b'' - b) + \\ &\quad \alpha\beta E(a' - a)(b' - b) + \\ &\quad \alpha\beta E(b' - a)(a' - b) \end{aligned}$$

where a', a'' are independently picked from A' and b', b'' from B'. The first 3 terms vanish Further

$$E(b' - a)(a' - b) = -E((a' - b')^2)$$

and we obtain, using (19):

$$F_3(C;A,B) = F_2(C,C') - \alpha\beta F_2(A',B') \tag{20}$$

This last equation will have a negative right-hand side if there is little drift between C, C', α is not close to 0 or 1 and A', B' have substantially drifted. Note that drift between A' and A and also between B' and B is immaterial here.

It is worth commenting that this is very specifically a test for admixture of population C and complex demography in the history of A and B does not effect the validity of the test. For example suppose we have two modern populations A, B formed by recent admixture of populations A_0, B_0. In an obvious notation $A = w_1A_0 + w_2B_0$, $B = v_1A_0 + v_2B_0$ where

$$w_1 + w_2 = v_1 + v_2 = 1$$

Then by a similar argument to that above we find that

$$\begin{aligned} f_3(C;A,B) &= f_3(C; w_1A_0 + w_2B_0, v_1A_0 + v_2B_0) \\ &= w_1v_1f_2(C,A_0) + \\ &\quad w_2v_2f_2(C,B_0) + \\ &\quad (w_1v_2 + w_2v_1)f_3(C;A_0,B_0) \end{aligned}$$

and so $f_3(C;A,B) < 0$ implies $f_3(C;A_0,B_0) < 0$. The complex recent admixture has weakened the test, but not removed the validity.

2.3 f_3 and f_4 statistics can be formed from f_2.

From the identity

$$(a-b)^2 = ((c-a)-(c-b))^2 = (c-a)^2 + (c-b)^2 - 2(c-a)(c-b)$$

It follows that

$$2f_3(C;A,B) = f_2(C,A) + f_2(C,B) - f_2(A,B) \tag{21}$$

Next, writing

$$d - b = c - b - (c - d)$$

It follows that

$$f_4(C,A;D,B) = f_3(C;A,B) - f_3(C;A,D) \tag{22}$$

Also $f_4(E,A;D,B) = f_4(C,A;D,B) - f_4(C,E;D,B)$. This shows that given knowledge of all the f_2 statistics, then all f_3, f_4 statistics can be computed. Conversely, fix a population C and suppose we know $f_3(C;A,B)$ for all populations A, B. and also $f_2(C,A)$ for every A. Then equation (21) shows that $f_2(A,B)$ is determined for all A, B and therefore all f_3, f_4 statistics are determined.

In calculations it is convenient to pick a basis for the f-statistics. Two natural bases are:

1. $f_2(A, B)$ for all A, B.

2. For a fixed C (usually an outgroup) $f_3(C; A, B)$ and $f_2(C, A)$ for all A, B.

References

[1] P. J. Bickel and K.A. Doksum. *Mathematical statistics: Basic Ideas and selected topics*. Holden-Day, 1977.

[2] F.M.T.A. Busing, E. Meijer, and R. van der Leeden. Delete-m jackknife for unequal m. *Statistics and Computing*, 9:3–8, 1999.

[3] H R Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17:1217–1241, 1989.

[4] M. Nei. *Molecular evolutionary genetics*. Columbia University Press, 1987.

[5] Nei, M. and Chesser, B.K. Estimation of fixation indices and gene diversities. *Ann Hum Genet*, 47:253–259, 1983.

[6] B.S. Weir and C. C. Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.