# Fst
# The algorithm used in `snpStats`

David Clayton

April 17, 2024

## $F$ statistics for diversity of groups

There is a very large literature on this topic and the author does not claim any great expertise. The purpose of this vignette is simply to document the method of calculation implemented in `snpStats`.

We shall start by introducing some notation. Let:

| | |
|---|---|
| $N_g$ | Number of chromosomes in group $g$ |
| $N_{sg}$ | Number of chromsomes in group $g$ observed for SNP $s$ |
| $N_s$ | Number of chromosomes observed for SNP $s$, all groups |
| $p_{sg}$ | Allele (relative) frequency for SNP $s$ in group $g$ |
| $p_s$ | Overall allele frequency for SNP $s$ |

and let:

$$\begin{aligned} Y_s &= \frac{N_s}{N_s - 1} p_s(1 - p_s) \\ X_{sg} &= \frac{N_{sg}}{N_{sg} - 1} p_{sg}(1 - p_{sg}) \\ X_s &= \sum_g W_g X_{sg} \end{aligned}$$

where $W_g$ are group-specific weights (see below).

The value returned for the F statistic for SNP $s$ is

$$F_s = 1 - \frac{X_s}{Y_s} = \frac{Y_s - X_s}{Y_s}.$$

A natural combined value over all SNPs is obtained by summing numerators and denominators of the SNP-specific values Denoting summation by a + subsscript,

$$F = \frac{Y_+ - X_+}{Y_+},$$

which can also be written as a weighted mean of the SNP-specific values, with $Y_s$ as weights:

$$F = \frac{1}{\sum_s Y_s} \sum_s Y_s F_s.$$

There appear to be two ways of looking at this index, leading to different weights when group sizes are unequal.

### Pairwise differences

One rationale suggests that the index can be written

$$\frac{D_T - D_W}{D_T}$$

where $D_T$ is the probability that two chromosomes, sampled at random from the total population, differ and $D_W$ is the probability that two chromosomes, sampled at random from the same subpopulation, differ. For a single SNP, $s$, $Y_s$ is an unbiased estimator for $D_T$ and $X_{sg}$ is an unbiased estimator for $D_W$ *within subgroup* $g$. With this rationale, the weights, $\{W_g\}$ should reflect the numbers of distinct pairwise comparisons within each group:

$$W_g = \frac{N_g(N_g - 1)}{\sum_g N_g(N_g - 1)}$$

### The analysis of variance

Another rationale would seem to be in terms of partition of the total variance, specifically the ratio of the between-group variance to the total variance (this seems to be the thrust of a series of papers by Cockerham). $Y_s$ is then an unbiased estimate of the total variance of SNP $s$ and $X_{sg}$ is an unbiased estimator of its variance in group $g$. But the natural weights are then

$$W_g = \frac{N_g}{\sum_g N_g}.$$

(Cockerham also seems to have considered an unweighted analysis but that could be inefficient if group sizes differ substantially).

## An example

Here we show the results of these calculations using the HapMap data discussed in other vignettes. These data were constructed by re-sampling individuals from two groups of HapMap subjects, the CEU sample (of European origin) and the JPT+CHB sample (of Asian origin), these groups being identified by the variable `stratum` in the subject support data frame.

We first use the pair-wise difference weights, first calculating the SNP-specific values, followed by the weighted average across all SNPs:

```
> data(for.exercise)
> f1 <- Fst(snps.10, subject.support$stratum, pairwise=TRUE)
> weighted.mean(f1$Fst, f1$weight)

[1] 0.06795229
```

We now compare this result with that obtained with the alternative (AOV) weights:

```
> f2 <- Fst(snps.10, subject.support$stratum, pairwise=FALSE)
> weighted.mean(f2$Fst, f2$weight)

[1] 0.06767651
```

Here there is little difference between the two values, since the group sizes are nearly the same:

```
> table(subject.support$stratum)

    CEU JPT+CHB
    494     506
```

In other cases the two weighting schemes could lead to different answers. In such situations, the preference of this author is for the analysis of variance weights and, accordingly, this has been set as the default action.