# Documentation for *qpfstats*

Nick Patterson

June 18, 2020

## 1 Introduction

We document a new program *qpfstats* in the ADMIXTOOLS suite. The output program is now used in some circumstances, by *qpGraph*, *qpWave* and *qpAdm* . Another new program, using *qpfstats* is *qpfv* briegly documented in *README.qpfstats*.

## 2 The central idea

Theoretical $F$-statistics and empirical $f$-statistics both form a linear vector space of dimension $n(n-1)/2$. See [1] for a detailed explanation. One basis, and what we use in ADMIXTOOLS is all statistics of form

$$f(O; A, B)$$

where $O$ is an outgroup and $A, B \neq O$. If we have $n$ populations, there are $n-1$ $f_2$ statistics and $(n-1)(n-2)/2$ $f_3$ statistics in the basis. Of course, any $f$-statistic $f_i$ can be written as

$$\hat{f}_i = \sum_j c_{ij} \hat{b}_j \tag{1}$$

where $\{b_j\}$ are $f$-statistics in the basis, and $c_{ij}$ are fixed coefficients, independent of data. But this is only true if *all* $f$-statistics are evaluated on the same SNP set, as will be the case if the data is complete, with no missing genotypes. In practice, requiring this may make the resulting set of SNPs too small, especially with low coverage ancient DNA, and analysts using ADMIXTOOLS often use the option

```
allsnps: YES
```

in which case equation (1) is not true. The old software just ignores this which is not at all satisfactory.

We can write

$$\hat{f}_i = \sum_j c_{ij} b_j + \sigma_i n_i$$

where $n_i$ is a noise variable, mean 0, variance 1. We can estimate $\sigma_i$ by the block jackknife. [The noise terms $n_i$ are not independent, but we ignore this in the current implementation.] Set $c'_{ij} = c_{ij}/\sigma_i$. Then we obtain an estimate of $b_j$ by solving the system of equations

$$\sum_j c'_{ij} b_j \approx \hat{f}_i/\sigma_i$$

by least squares, to obtain estimates $\hat{b}_j$ We can then compute an error covariance on $\hat{b}_j$, again by using the block jackknife.

Thus we make 2 passes over the data, the first being to estimate $\sigma_i$ and the second to make the final estimates of basis $f$-statistics and covariance. Standard theory then yields an estimate (and covariance) for any set of $f$-statistics.

As an example we were interested in testing if there were genetic differences between a very low coverage Harappa sample for damaged and undamaged reads. Let $H_D$ be calls from damaged reads, $H_U$ from undamaged. We want to compute

$$f_4(Out, X; H_U, H_D)$$

for a variety of $X$. but a direct computation fails as there were almost no SNPs with data in common. but if we pick another population $C$ with nearly complete data, then we can write

$$f_4(Out, X; H_U, H_D) = f_4(Out, X; C, H_D) - f_4(Out, X; C, H_U)$$

and obtain an estimate. We used this trick in [2]. The work described in this note can be thought of as a generalization.

# References


[1] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, Nov 2012.

[2] Vasant Shinde, Vagheesh M Narasimhan, Nadin Rohland, Swapan Mallick, Matthew Mah, Mark Lipson, Nathan Nakatsuka, Nicole Adamski, Nasreen Broomandkhoshbacht, Matthew Ferry, et al. An ancient Harappan genome lacks ancestry from steppe pastoralists or iranian farmers. *Cell*, 179(3):729–735, 2019.