

File S1 - Constraints of Minor Allele Frequency on F_{ST}

Assuming all populations have the same size, we define p_i as the observed frequency of the minor allele in population i , and $p = \frac{\sum_{i=1}^n p_i}{n}$ as the minor allele frequency over all n populations. To simplify calculations, here we analyzed the constraints of minor allele frequency on F_{ST} using the expected value of F_{ST} , instead of the [Weir and Cockerham \(1984\)](#) F_{ST} estimator used in the main text. The expected value of F_{ST} is defined by

$$F_{ST} = \frac{H_T - H_S}{H_T}, \quad (1)$$

where H_T is the heterozygosity in the total population and H_S is the average heterozygosity within subpopulations, defined by

$$H_T = 2p(1 - p) \quad H_S = \frac{\sum_{i=1}^n 2p_i(1 - p_i)}{n}. \quad (2)$$

Maximum F_{ST} for a given p is achieved when H_S is minimal. This happens when all occurrences of the minor allele are concentrated in as few populations as possible and only one population is polymorphic at that site ([Alcala and Rosenberg 2017](#)), i.e. when $\lfloor pn \rfloor$ populations are fixed for the minor allele, and the remainder of minor alleles are all in the same population, with frequency $p^* = pn - \lfloor pn \rfloor$. (The notation $\lfloor \cdot \rfloor$ represents the integer part of the number pn).

All populations that are fixed for either the minor or major allele will not contribute to H_S , since either p_i or $1 - p_i$ will be zero. So the H_S formula in the scenario of maximum F_{ST} can be simplified to

$$H_{S_{maxF_{ST}}} = \frac{2p^*(1 - p^*)}{n}. \quad (3)$$

For example, with $n = 10$ populations and MAF of $p = 0.15$, maximum F_{ST} will be achieved when $\lfloor 1.5 \rfloor = 1$ population is fixed for the minor allele, 1 population has $MAF = p^* = 1.5 - 1 = 0.5$, and the remainder 8 populations are fixed for the major allele. In this case, maximum F_{ST} is 0.8.

To illustrate this constraint imposed by MAF on F_{ST} , we simulated the neutral evolution of SNPs in 10 populations, with virtually no migration among them, allowing SNPs to achieve maximum differentiation among populations. Simulations were performed using the *sim.genot* function of the hierfstat R package ([Goudet 2005](#)). We simulated the neutral evolution of 10,000 bi-allelic loci (SNPs) in 10 populations, each with population size 1000, migration rate of $m = 10^{-5}$ and mutation rate of $\mu = 10^{-8}$, and we used sample sizes of 50, 100 or 1000 individuals. Results were independent of sample size. The low migration rate allowed SNPs to achieve maximum F_{ST} values possible given their MAF. Figure S1 shows the F_{ST} of 10,000 simulated SNPs as a function of their MAFs, as well as the maximum F_{ST} values estimated by replacing the observed values by the expected one in Equations 1-3.

Figure S1 shows that, when all subpopulations are the same size (in the case of Figure S1, $n = 10$), F_{ST} only achieves 1 when MAF is exactly $m \in \{1/n, 2/n, \dots, 1/2\}$. This is because F_{ST} can only achieve 1 when H_S is zero, and H_S can only be exactly zero when all populations are fixed for either the minor or major allele. Maximum values of F_{ST} increase linearly from zero to one as MAF increases from zero to $1/n$. When MAF is between the values of m , maximum F_{ST} is less than 1, which generates the wavy pattern seen in Figure S1.

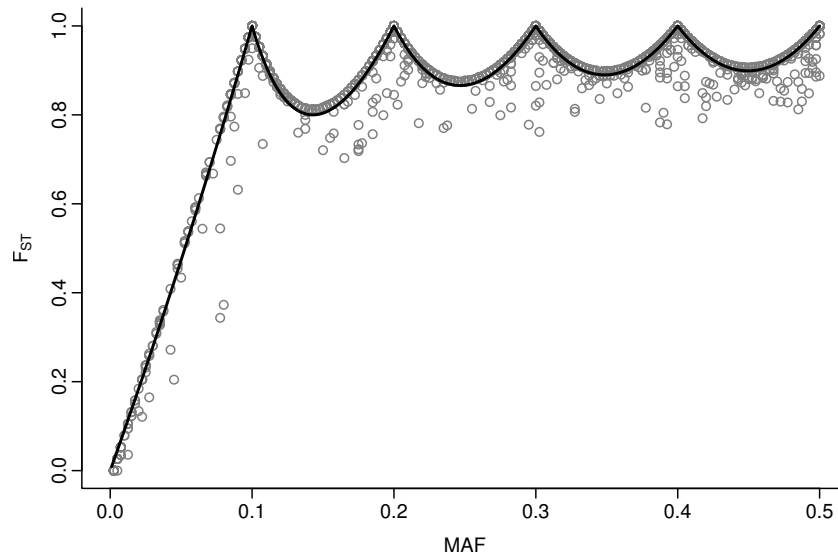


Figure S1 Maximum F_{ST} as a function of MAF for $n = 10$ populations. Black line shows maximum F_{ST} as a function of MAF, calculated using Equations 1-3. Gray points are simulations of biallelic SNPs evolving neutrally in 10 populations of the same size, with low migration among them, which allows them to achieve maximum F_{ST} .

Literature Cited

- Alcala, N. and N. A. Rosenberg, 2017 Mathematical constraints on F_{ST} : biallelic markers in arbitrarily many populations. *Genetics* **206**: 1581–1600.
- Goudet, J., 2005 Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**: 184–186.
- Weir, B. S. and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.