RESEARCH ARTICLE

An Improved F_{st} Estimator

Guanjie Chen[1]*, Ao Yuan[2], Daniel Shriner[1], Fasil Tekola-Ayele[1], Jie Zhou[1], Amy R. Bentley[1], Yanxun Zhou[3], Chuntao Wang[3], Melanie J. Newport[4], Adebowale Adeyemo[1], Charles N. Rotimi[1]

1 Center for Research on Genomics and Global Health, NHGRI, NIH, Bethesda, Maryland, United States of America, **2** Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC, United States of America, **3** Suizhou Central Hospital, Suizhou, Hubei, China, **4** Brighton and Sussex Medical School, Falmer, Brighton, United Kingdom

* chengu@mail.nih.gov

OPEN ACCESS

Citation: Chen G, Yuan A, Shriner D, Tekola-Ayele F, Zhou J, Bentley AR, et al. (2015) An Improved F_{st} Estimator. PLoS ONE 10(8): e0135368. doi:10.1371/journal.pone.0135368

Editor: Francesc Calafell, Universitat Pompeu Fabra, SPAIN

Received: November 12, 2014

Accepted: July 21, 2015

Published: August 28, 2015

Data Availability Statement: The real dataset from the Wolaita were obtained in 2007 with written informed consent that limits data access only to the investigators. Therefore, public deposition of data would breach ethical compliance because participants did not give consent for deposition of their genotypes. Please contact Dr. Fasil Tekola Ayele (fasil.ayele2@nih.gov) for real data detail informaiton.

Funding: This research was supported by the Intramural Research Program of the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute, the National Institute of

Abstract

The fixation index F_{st} plays a central role in ecological and evolutionary genetic studies. The estimators of Wright ($\hat{F}_{st_1}$), Weir and Cockerham ($\hat{F}_{st_2}$), and Hudson *et al*. ($\hat{F}_{st_3}$) are widely used to measure genetic differences among different populations, but all have limitations. We propose a minimum variance estimator $\hat{F}_{st_m}$ using $\hat{F}_{st_1}$ and $\hat{F}_{st_2}$. We tested $\hat{F}_{st_m}$ in simulations and applied it to 120 unrelated East African individuals from Ethiopia and 11 subpopulations in HapMap 3 with 464,642 SNPs. Our simulation study showed that $\hat{F}_{st_m}$ has smaller bias than $\hat{F}_{st_2}$ for small sample sizes and smaller bias than $\hat{F}_{st_1}$ for large sample sizes. Also, $\hat{F}_{st_m}$ has smaller variance than $\hat{F}_{st_2}$ for small F_{st} values and smaller variance than $\hat{F}_{st_1}$ for large F_{st} values. We demonstrated that approximately 30 subpopulations and 30 individuals per subpopulation are required in order to accurately estimate F_{st}.

Introduction

The fixation index F_{st} is widely used as a measure of population differentiation due to genetic structure. Wright [1, 2] defined F_{st} as the ratio of the observed variance of allele frequencies between subpopulations to the expected variance of allele frequencies assuming panmixis. Wright's estimator of F_{st} is biased, because *a priori* expected allele frequencies are unknown and the numerator and denominator terms in the equation are not independent. In practice, various frameworks have been proposed to improve estimation of F_{st}. Weir and Cockerham used an analysis of variance (ANOVA) approach to estimate within- and between-population variance components [3, 4]. Weir and Cockerham's estimator is widely used because their estimator can describe the genetic population structure in a single summary statistic, is asymptotically unbiased with respect to sample size, and can compensate for overestimates particularly at low levels of genetic differentiation unlike Wright's estimator [5]. However, it can be upwardly biased unless adjustment is done for intralocus sampling error, the number of subpopulations sampled, time of divergence, *etc.* [6]. In the present study, we propose a method that improves F_{st} estimation by combining Wright's and Weir and Cockerham's estimators to

Diabetes and Digestive and Kidney Diseases, the Center for Information Technology, and the Office of the Director at the National Institutes of Health (Z01HG200362). The data from the Wolaita came from a study supported by the Wellcome Trust (Grant #079791). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

achieve a minimum variance estimate. For comparison, we also include Hudson et al.'s estimator [7], which recently has been recommended by Bhatia et al. [8]. We demonstrate application of our modified estimator in analysis of real data.

Methods

For a diallelic marker, let p be the true minor allele frequency in the total population. Let the true subpopulation allele frequencies be $p_1, \ldots, p_r$ in $r \geq 2$ subpopulations. Let σ^2 be the true population variance in allele frequencies across subpopulations. Suppose the observed sample frequencies are $\hat{p}_1, \ldots, \hat{p}_r$ and the sample sizes are $n_1, \ldots, n_r$. Let $n = \sum_{j=1}^{r} n_j$ and $\bar{n} = \sum_{j=1}^{r} n_j / r$. Let ϑ be the difference in allele frequencies, such that for two subpopulations, $\hat{\vartheta} = \hat{p}_1 - \hat{p}_2$.

Wright's F_{st} [2] is defined as

$$F_{st} = \frac{\sigma^2}{p(1-p)}$$

and is estimated as

$$\hat{F}_{st_1} = \frac{\sum_{j=1}^{r} (\hat{p}_j - \bar{p})^2}{r\bar{p}(1-\bar{p})},$$

with

$$\bar{p} = \sum_{j=1}^{r} \hat{p}_j / r.$$

For the special case of two subpopulations, Rosenberg et al. [9] showed that by algebraic rearrangement

$$F_{st} = \frac{(p_1 - p_2)^2}{(p_1 + p_2)(2 - (p_1 + p_2))}.$$

Thus, F_{st} is a function of the difference in allele frequencies and is proportional to ϑ^2.

Weir and Cockerham's estimator [4], assuming a random union of gametes or equivalently no individual-level inbreeding, is based on

$$\bar{p} = \sum_{j=1}^{r} n_j \hat{p}_j / n,$$

$$S^2 = \frac{1}{(r-1)\bar{n}} \sum_{j=1}^{r} n_j (\hat{p}_j - \bar{p})^2,$$

$$T_1 = S^2 - \frac{1}{2\bar{n} - 1} \left[\bar{p}(1-\bar{p}) - \frac{r-1}{r} S^2 \right],$$

$$n_c = \frac{1}{r-1} \left(\sum_{j=1}^{r} n_j - \frac{\sum_{j=1}^{r} n_j^2}{\sum_{j=1}^{r} n_j} \right),$$

and

$$T_2 = \frac{2n_c - 1}{2\bar{n} - 1}\bar{p}(1-\bar{p}) + \left[1 + \frac{2(r-1)(\bar{n} - n_c)}{2\bar{n} - 1}\right]\frac{S^2}{r},$$

yielding

$$\hat{F}_{st_2} = \frac{T_1}{T_2}.$$

The definition of F_{st} of Hudson et al. [7] is

$$F_{st_3} = 1 - \frac{H_W}{H_B}, \qquad H_W = \frac{1}{r}\sum_{j=1}^{r} 2p_j(1-p_j), \qquad H_B = \frac{1}{r(r-1)}\sum_{i\neq j}^{r} 2p_i(1-p_j).$$

Given observed sample estimates $\hat{p}_1, \ldots, \hat{p}_r$, $\hat{H}_W = r^{-1}\sum_{j=1}^{r} 2\hat{p}_j(1-\hat{p}_j)$ is a biased estimate of H_W, because

$$E(\hat{p}_j(1-\hat{p}_j)) = E(\hat{p}_j - \hat{p}_j^2) = E(\hat{p}_j) - E(\hat{p}_j^2)$$
$$= p_j - \left(p_j^2 + \frac{p_j(1-p_j)}{2n_j}\right) = p_j(1-p_j)\left(1 - \frac{1}{2n_j}\right).$$

An unbiased estimate of $p_j(1 - p_j)$ is thus given by $[2n_j/(2n_j - 1)]\hat{p}_j(1-\hat{p}_j)$. However, $\hat{H}_B = [r(r-1)]^{-1}\sum_{i\neq j}^{r} 2\hat{p}_i(1-\hat{p}_j)$ is an unbiased estimate of H_B if $Cov(\hat{p}_i, \hat{p}_j) = 0$, i.e., under the null hypothesis. Therefore, we estimate F_{st} by

$$\hat{F}_{st_3} = 1 - \frac{(r-1)\sum_{j=1}^{r}\frac{2n_j}{2n_j-1}\hat{p}_j(1-\hat{p}_j)}{\sum_{i\neq j}^{r}\hat{p}_i(1-\hat{p}_j)},$$

which is a ratio of unbiased estimates. This estimator generalizes Bhatia et al.'s [8] version of Hudson et al.'s [7] estimator for $r > 2$.

Note that under the null hypothesis of $p_1 = \cdots = p_r$, both $\bar{n}\hat{F}_{st_1}$ and $\bar{n}\hat{F}_{st_2}$ are asymptotically zero. Our goal is to construct an estimator based on a linear combination of $\bar{n}\hat{F}_{st_1}$ and $\bar{n}\hat{F}_{st_2}$ such that the new estimator has the smallest variance among all such linear combinations. Let σ_1^2 and σ_2^2 be the asymptotic variances of $\bar{n}\hat{F}_{st_1}$ and $\bar{n}\hat{F}_{st_2}$, and σ_{12} be the asymptotic covariance. We propose the following weighted version of $\hat{F}_{st}$:

$$\hat{F}_{st_m} = \hat{F}_{st}(a) = a\hat{F}_{st_1} + (b-a)\hat{F}_{st_2}, a > 0,$$

where $b > 0$ is a fixed number to be chosen later. We choose $a = a_0$ such that $Var(\hat{F}_{st}(a))$ is minimized:

$$a_0 = \arg\min_{a>0}\{a^2\sigma_1^2 + (b-a)^2\sigma_2^2 + 2a(b-a)\sigma_{12}\}. \tag{1}$$

It is seen that $Var(\hat{F}_{st}(a_0)) \leq \min\{Var(\hat{F}_{st_1}), Var(\hat{F}_{st_2})\}$ and hence is more precise in estimation. From the proof of the Proposition we see that Eq (1) is equivalent to,

$$a_0 = \arg\min_{a>0}\{a^2 + (b-a)^2\delta^2 + 2a(b-a)\delta\}, \qquad \delta = \lim_{n\to\infty}\bar{n}/n_c. \tag{2}$$

which gives, with $b = (\delta - 1)/(\delta + 1)$,

$$a_0 = \arg \min_{a>0} \{a^2(\delta - 1)^2 - 2ab\delta(\delta - 1) + \delta^2\} = \frac{\delta}{\delta + 1}.$$

At the end of the proof of the following Proposition, we show that $\delta \geq 1$ with equality if and only if $n_1 = \cdots = n_r$. When $n_1 = \cdots = n_r$, we have $\bar{n} = n_c$ and $\hat{F}_{st}(a_0) = \frac{1}{2}\hat{F}_{st_1} - \frac{1}{2}\hat{F}_{st_2}$. Let $\xrightarrow{D}$ denote convergence in distribution.

Proposition. *Assume that* $0 < p_0 < 1$ *and that the* n_j*'s are not all equal (so that* $\delta > 1$*). If* $p_1 = \cdots = p_r$, *with*

$$\delta = \lim_{n\to\infty} \bar{n}/n_c$$

and

$$\gamma_j = \lim_{n\to\infty} n_j/n,$$

we have

$$\bar{n}\hat{F}_{st}(a) + \delta(1 - a) \xrightarrow{D} \frac{a + \delta(1 - a)}{(r - 1)p_0(1 - p_0)} \sum_{j=1}^{r} \lambda_j \chi_j^2,$$

where $\lambda_1, \ldots, \lambda_r$ *are the eigenvalues of* $\Omega'^{1/2}\,\mathbf{B}\Omega^{1/2}$, $\Omega = (\omega_{ij})_{r\times r}$ *with* $\omega_{ij} = p_0(1 - p_0)$ *if* $i = j$ *and* $\omega_{ij} = 0$ *if* $i \neq j$, $\Omega^{-1/2}$ *is the square root of* Ω: $\Omega = \Omega'^{1/2}\Omega^{1/2}$, *and* $\mathbf{B} = \sum_{j=1}^{r} \gamma_j \mathbf{b}_j \mathbf{b}_j'$, $\mathbf{b}_j = (-\gamma_1, \ldots, -\gamma_{j-1}, (1 - \gamma_j), -\gamma_{j+1}, \ldots, -\gamma_r)'$.

In the above Proposition, take $a = a_0$, then $a_0+\delta(1 - a_0) = 0$ and $\delta\,(1 - a_0) = -\delta/(\delta - 1)$, and we get

Corollary 1. *Under conditions of the Proposition,*

i.

$$\bar{n}\hat{F}_{st}(a_0) + \frac{\delta}{\delta + 1} \xrightarrow{D} \frac{2\delta}{(r - 1)(\delta + 1)p_0(1 - p_0)} \sum_{j=1}^{r} \lambda_j \chi_j^2.$$

ii. *If* a = *1, then*

$$\bar{n}\hat{F}_{st_1} = \bar{n}\hat{F}_{st}(1) \xrightarrow{D} \frac{1}{(r - 1)p_0(1 - p_0)} \sum_{j=1}^{r} \lambda_j \chi_j^2.$$

iii. *If* a = *0, then*

$$\bar{n}\hat{F}_{st_2} = \bar{n}\hat{F}_{st}(0) + \delta \xrightarrow{D} \frac{\delta}{(r - 1)p_0(1 - p_0)} \sum_{j=1}^{r} \lambda_j \chi_j^2.$$

Simulations Under the Balding-Nichols model [10], the allele frequency in each of r subpopulations conditional on p and F_{st} is a random deviate from the beta distribution $\beta\left(\frac{1-F_{st}}{F_{st}}p, \frac{1-F_{st}}{F_{st}}(1 - p)\right)$, which has mean p and variance $p(1 - p)F_{st} = \sigma^2$.

Simulation 1. This simulation was designed to estimate bias in the worst case scenario of two subpopulations. We evaluated the relationships between $\hat{F}_{st}$ and F_{st} and between $\hat{F}_{st}$ and $\hat{\vartheta}$. First, given the true average allele frequency p for $r = 2$, F_{st} reaches its maximum value for p_j values of 0 and $2p$. The estimator $\hat{F}_{st} = \frac{S^2}{\bar{p}(1-\bar{p})+\frac{S^2}{r}}$ [4] yields a constrained range for $\hat{F}_{st}$ from 0 to $2p$. Therefore, we first assigned the true value for p by drawing a random uniform deviate from the interval (0, 0.5) and the true value for F_{st} by independently drawing a random uniform deviate from the interval (0, $2p$). Conditional on the true values of p and F_{st}, we randomly generated p_j from the beta distribution. We next assigned the number of individuals per subpopulation $n_j = [5, 10, 20, 50, 100, 110]$. We then randomly drew alleles from the binomial distribution $Bin(2n_j, p_j)$. We generated 10,000 independent replicate data sets. Based on the above formulae, the four estimators $\hat{F}_{st_1}$, $\hat{F}_{st_2}$, $\hat{F}_{st_3}$, and $\hat{F}_{st_m}$ were calculated. Linear regression models were used to evaluate the relationship between F_{st} and $\hat{F}_{st}$ and between $\hat{F}_{st}$ and $\hat{\vartheta}^2$. We assessed the fit in a linear regression model with the F-test, r^2, and the root mean squared error (RMSE), which is the square root of the sum of the variance and the squared bias.

Simulation 2. This simulation was designed to evaluate variance under sampling conditions approaching unbiasedness, i.e., large numbers of subpopulations and individuals per subpopulation. We evaluated the relationships between $\hat{F}_{st}$ and the number of subpopulations (r) and between $\hat{F}_{st}$ and the number of individuals per subpopulations (n_j). Conditional on the average allele frequency p, F_{st}, the number of subpopulations $r = [5, 10, 20, 50, 100, 250]$, and the number of individuals per subpopulation $n_j = [5, 10, 20, 50, 100, 250, 1000]$, we randomly generated r allele frequencies as in Simulation 1 and calculated $\hat{F}_{st_1}$, $\hat{F}_{st_2}$, $\hat{F}_{st_m}$, and $\hat{F}_{st_3}$.

Application to data

We included genotype data from a total of 120 unrelated individuals from the Wolaita (WETH) ethnic group from southern Ethiopia who served as controls in a genome-wide association study of podoconiosis [11]. The Wolaita ethnic group speaks an Omotic language, and comparison with HapMap African populations has shown that it has the closest genetic similarity with the Maasai from Kenya and the lowest genetic similarity with the Yoruba in Nigeria [12]. Genotyping was performed by deCODE Genetics using the Illumina HumanHap 610 Bead Chip, which assays > 620,000 single-nucleotide polymorphisms (SNPs). Of the 551,840 autosomal SNPs in the raw genotype data, we excluded 39,249 SNPs that had a minor allele frequency of < 0.05, 378 that were missing in > 0.05 of individuals, and 321 that had a Hardy-Weinberg p-value < 0.001. The remaining 511,892 SNPs were merged with genotype data for ASW ($n = 49$), CEU ($n = 112$), CHB ($n = 84$), CHD ($n = 85$), GIH ($n = 88$), JPT ($n = 86$), LWK ($n = 90$), MKK ($n = 143$), MXL ($n = 50$), TSI ($n = 88$), and YRI ($n = 113$) in HapMap phase 3, release 2, which contained 1,440,616 SNPs. A total of 464,642 SNPs were common to both of WETH and HapMap data sets. $\hat{F}_{st_1}$ $\hat{F}_{st_2}$, $\hat{F}_{st_m}$, and $\hat{F}_{st_3}$ were calculated per marker.

Results

Simulation 1: We first compared $\hat{F}_{st}$ with the true F_{st} for the worst-case scenario of $r = 2$. For small sample sizes, $\hat{F}_{st_1}$ was the least biased estimator, followed by $\hat{F}_{st_m}$, $\hat{F}_{st_2}$, and $\hat{F}_{st_3}$ (Table 1). For large sample sizes, $\hat{F}_{st_m}$ and $\hat{F}_{st_1}$ were comparably good, and $\hat{F}_{st_2}$ and $\hat{F}_{st_3}$ were identically worse (Table 1). None of the four estimators was strongly sensitive to equal vs. unequal sample sizes (Fig 1). When $\hat{\vartheta}$ was close to 0, $\hat{F}_{st_3}$ yielded the most negative estimates, followed by $\hat{F}_{st_2}$ and $\hat{F}_{st_m}$. As expected, all four estimators showed a quadratic relationship with $\hat{\vartheta}$ (Fig 1). With

Table 1. $\hat{F}_{st}$ **vs.** F_{st} **and** $\hat{\vartheta}^2$ **for two subpopulations.**

	$\hat{F}_{st_1}$	$\hat{F}_{st_2}$	$\hat{F}_{st_m}$	$\hat{F}_{st_3}$
$n = 5$				
$\hat{F}_{st}$ vs. F_{st}				
Root MSE	0.2213	0.2241	0.2228	0.2241
Squared bias	0.0465	0.0481	0.0473	0.0484
r^2	0.1096	0.0869	0.0979	0.0869
F-test	12310	9522	10848	9522
$\hat{F}_{st}$ vs. $\hat{\vartheta}^2$				
Root MSE	0.0340	0.1002	0.0657	0.1127
Squared bias	0.0008	0.0089	0.0036	0.0114
r^2	0.9855	0.9129	0.9549	0.9129
F-test	6.808×10^6	1.048×10^6	2.116×10^6	1.048×10^6
$n = 1000$				
$\hat{F}_{st}$ vs. F_{st}				
Root MSE	0.2102	0.2105	0.2101	0.2105
Squared bias	0.0415	0.0419	0.0416	0.0419
r^2	0.1981	0.1958	0.1986	0.1958
F-test	24701	24344	24778	24344
$\hat{F}_{st}$ vs. $\hat{\vartheta}^2$				
Root MSE	0.0233	0.0706	0.0455	0.0706
Squared bias	0.0002	0.0041	0.0015	0.0041
r^2	0.9913	0.9355	0.9700	0.9355
F-test	1.140×10^7	1.450×10^6	3.236×10^6	1.450×10^6

doi:10.1371/journal.pone.0135368.t001

respect to $\hat{\vartheta}^2$, by all four measures $\hat{F}_{st_1}$ was the best estimator whereas $\hat{F}_{st_2}$ was the worst estimator (Table 1).

An assessment of bias by the total sample size (n_1 and n_2) for $r = 2$ is presented in Fig 2. $\hat{F}_{st_1}$ was biased and this bias was constant across total sample size, as expected given that this estimator does not account for n_j. In contrast, $\hat{F}_{st_2}$, $\hat{F}_{st_m}$, and $\hat{F}_{st_3}$ were less biased as the total sample size increased. When the total sample size exceeded 30, $\hat{F}_{st_3}$ was the least biased estimator; otherwise, $\hat{F}_{st_1}$ was the least biased estimator. For $r = 2$, the magnitude of bias for all four estimators was constant when the total sample size was at least 60.

Simulation 2: Given $p = 0.2$, $F_{st} = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$, $n = 1000$ individuals, and $r = 200$ subpopulations, mean $\hat{F}_{st}$ values are presented in Table 2. The means for $\hat{F}_{st_1}$, $\hat{F}_{st_2}$, $\hat{F}_{st_3}$, and $\hat{F}_{st_m}$ equaled the expected values, consistent with all four estimators being asymptotically unbiased. First, we investigated the relationship between F_{st} and the variance of the four estimators. Given $p = 0.2$ and $F_{st} < 0.5$, $\hat{F}_{st_1}$ had the smallest variance, followed by $\hat{F}_{st_m}$ and $\hat{F}_{st_2}$ (Fig 3). Given $p = 0.2$ and $F_{st} > 0.5$, $\hat{F}_{st_2}$ had the smallest variance, followed by $\hat{F}_{st_m}$ and $\hat{F}_{st_1}$. Similar results were obtained for $p = 0.1$, 0.3, 0.4, and 0.5 (S1 Table).

Second, we investigated how the number of subpopulations and the number of individuals per subpopulation affected bias. When the number of subpopulations was approximately 30, no matter the number of individuals per subpopulation, bias was stable (Fig 3). For $r > 30$ and

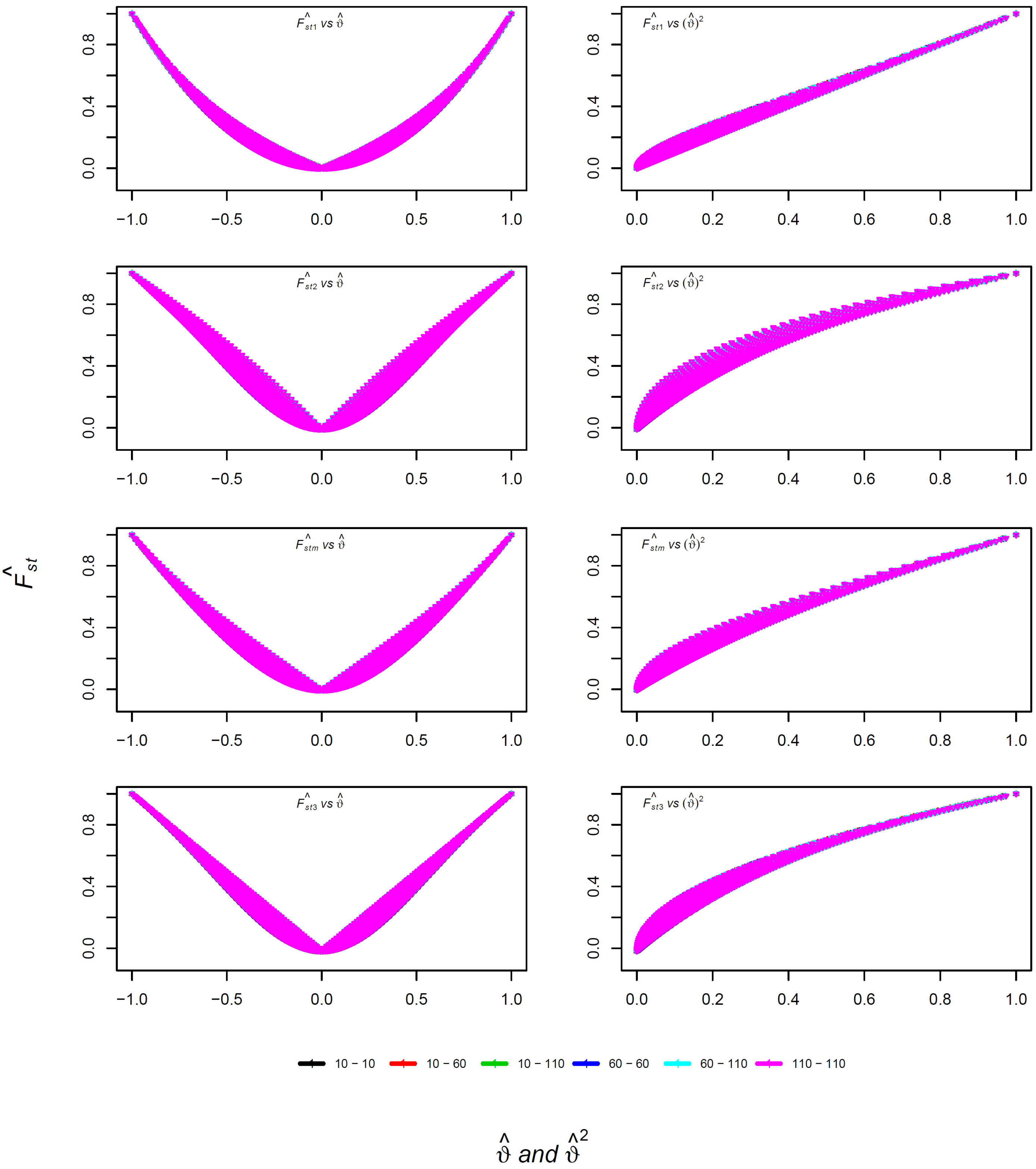

Fig 1. The relationship between $\hat{F}_{st}$ and $\hat{\vartheta}$ for simulated data. The x-axis shows the difference of allele frequencies between two subpopulations $\hat{\vartheta}$ (left plots) and $\hat{\vartheta}^2$ (right plots); the y-axis shows $\hat{F}_{st}$ values for Wright's (top row), Weir and Cockerham's (second row), the modified (third row), and Hudson et al.'s estimators (bottom row), and the legend indicates the sample sizes n_1 (before hyphen) and n_2 (after hyphen).

doi:10.1371/journal.pone.0135368.g001

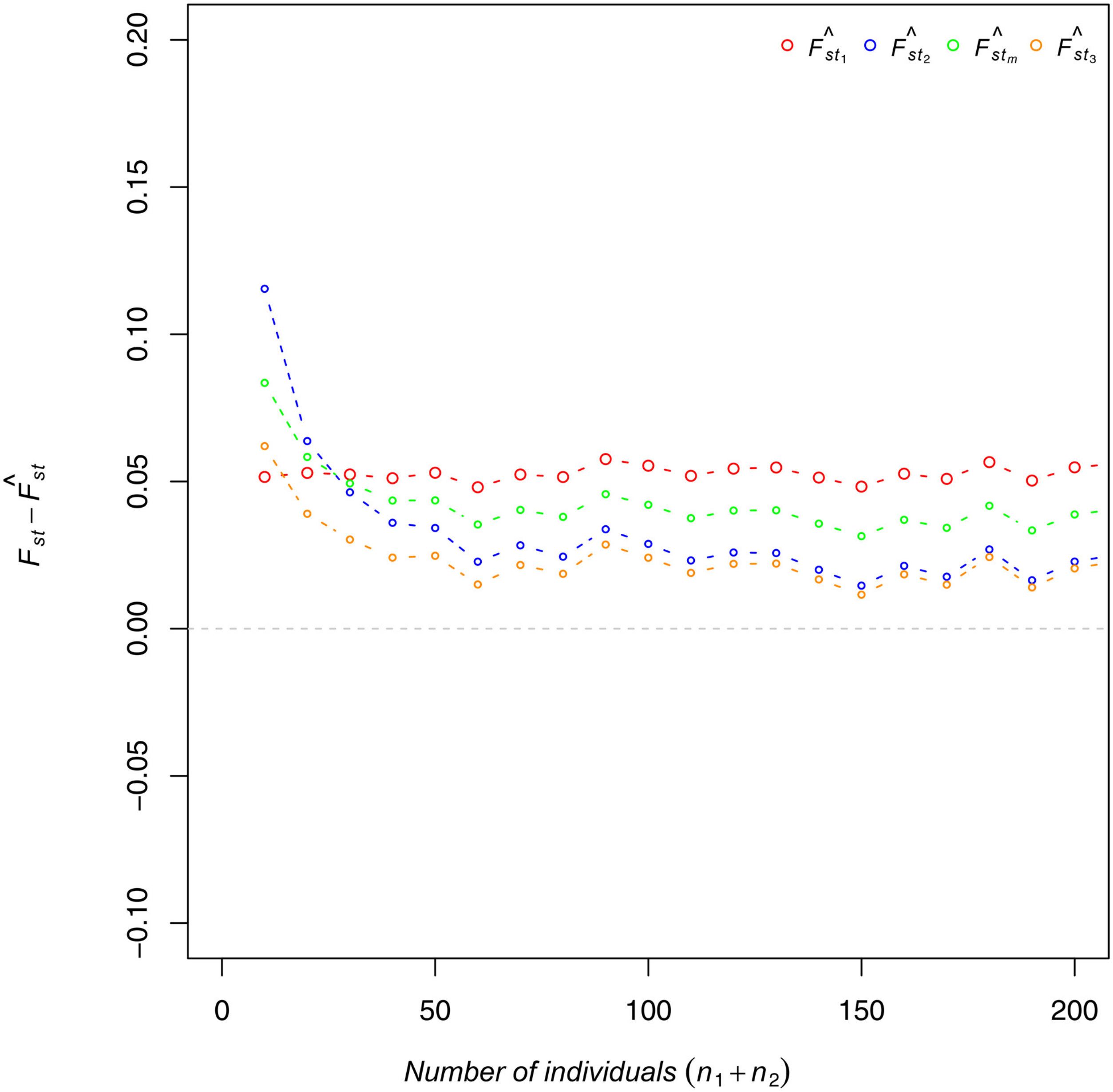

Fig 2. Bias as a function of total sample size. The x-axis shows the total sample size ($n_1 + n_2$). The y-axis shows $F_{st} - \hat{F}_{st_1}$ (red), $F_{st} - \hat{F}_{st_2}$ (blue), $F_{st} - \hat{F}_{st_m}$ (green), and $F_{st} - \hat{F}_{st_3}$ (orange) for $r = 2$.

doi:10.1371/journal.pone.0135368.g002

small n_j, all four estimators were biased, with the order of $\hat{F}_{st_1} < \hat{F}_{st_3} < \hat{F}_{st_m} < \hat{F}_{st_2}$. For $r > 30$ and large n_j, all four estimators were unbiased. For $n_j > 30$, all four estimators were stable and bias decreased as r increased, with $\hat{F}_{st_3}$ the best estimator and $\hat{F}_{st_1}$ the worst estimator (Fig 4).

Application to Data: The means and variances of $\hat{F}_{st}$ values between the WETH and 11 samples in HapMap 3 are presented in Table 3. The WETH sample was closest to the MKK sample, consistent with shared Cushitic and Nilo-Saharan ancestry [13]. $\hat{F}_{st_2}$ and $\hat{F}_{st_m}$ yielded the same order for all pairs of relationships and all four estimators yielded the same order of relationships

Table 2. Means, Variances, and MSEs of $\hat{F}_{st}$ in simulation 2.

True F_{st}		$\hat{F}^{*}_{st_1}$	$\hat{F}_{st_2}$	$\hat{F}_{st_m}$	$\hat{F}_{st_3}$
0.1	Means	9.95E-02	9.95E-02	9.95E-02	9.99E-02
	Variances	9.40E-05	9.49E-05	9.44E-05	9.48E-05
	MSE	9.43E-05	9.51E-05	9.47E-05	9.48E-05
0.2	Means	1.99E-01	1.99E-01	1.99E-01	2.00E-01
	Variances	3.36E-04	3.38E-04	3.37E-04	3.38E-04
	MSE	3.37E-04	3.39E-04	3.38E-04	3.38E-04
0.3	Means	2.99E-01	2.99E-01	2.99E-01	3.00E-01
	Variances	6.27E-04	6.30E-04	6.28E-04	6.29E-04
	MSE	6.29E-04	6.30E-04	6.29E-04	6.29E-04
0.4	Means	3.98E-01	3.99E-01	3.99E-01	3.99E-01
	Variances	9.12E-04	9.15E-04	9.14E-04	9.14E-04
	MSE	9.15E-04	9.16E-04	9.15E-04	9.14E-04
0.5	Means	4.98E-01	4.99E-01	4.98E-01	4.99E-01
	Variances	1.11E-03	1.11E-03	1.11E-03	1.11E-03
	MSE	1.11E-03	1.11E-03	1.11E-03	1.11E-03
0.6	Means	5.98E-01	5.99E-01	5.99E-01	6.00E-01
	Variances	1.18E-03	1.18E-03	1.18E-03	1.18E-03
	MSE	1.18E-03	1.18E-03	1.18E-03	1.18E-03
0.7	Means	6.98E-01	6.99E-01	6.99E-01	6.99E-01
	Variances	1.12E-03	1.12E-03	1.12E-03	1.11E-03
	MSE	1.12E-03	1.12E-03	1.12E-03	1.11E-03
0.8	Means	7.98E-01	7.99E-01	7.99E-01	7.99E-01
	Variances	8.67E-04	8.63E-04	8.65E-04	8.62E-04
	MSE	8.69E-04	8.64E-04	8.66E-04	8.63E-04
0.9	Means	8.99E-01	8.99E-01	8.99E-01	8.99E-01
	Variances	4.81E-04	4.77E-04	4.79E-04	4.77E-04
	MSE	4.81E-04	4.78E-04	4.79E-04	4.77E-04

* Means, variances, and Mean Squared Errors (MSEs) from 200 subpopulations with 1000 individuals per subpopulation, given $p = 0.2$.

doi:10.1371/journal.pone.0135368.t002

for the five HapMap samples closest to the WETH sample. The order of the means was $\hat{F}_{st_1} < \hat{F}_{st_m} < \hat{F}_{st_2} < \hat{F}_{st_3}$. $\hat{F}_{st_m}$ was approximately 30% larger than $\hat{F}_{st_1}$ and approximately 20% smaller than $\hat{F}_{st_2}$, which has corresponding effects on divergence time estimates. Given that $\hat{F}_{st_2}$ and $\hat{F}_{st_m}$ are less downward biased than $\hat{F}_{st_1}$ for these sample sizes (Fig 2), the larger values are more likely to be correct.

Discussion

F_{st} is directly related to the variance in allele frequencies among subpopulations. The dependence of F_{st} on allele frequencies and genetic diversity has been observed [14]. In our study, an approximately linear relationship between $\hat{F}_{st_1}$, $\hat{F}_{st_2}$, $\hat{F}_{st_m}$, and $\hat{F}_{st_3}$ with the squared difference of allele frequencies ($\hat{\vartheta}^2$) was observed, as expected. By simulation, we found that all four estimators were unbiased for large numbers of subpopulations and individuals per subpopulation but that no one estimator was uniformly better than the others. For $F_{st} < 0.5$, $\hat{F}_{st_1}$ had smaller

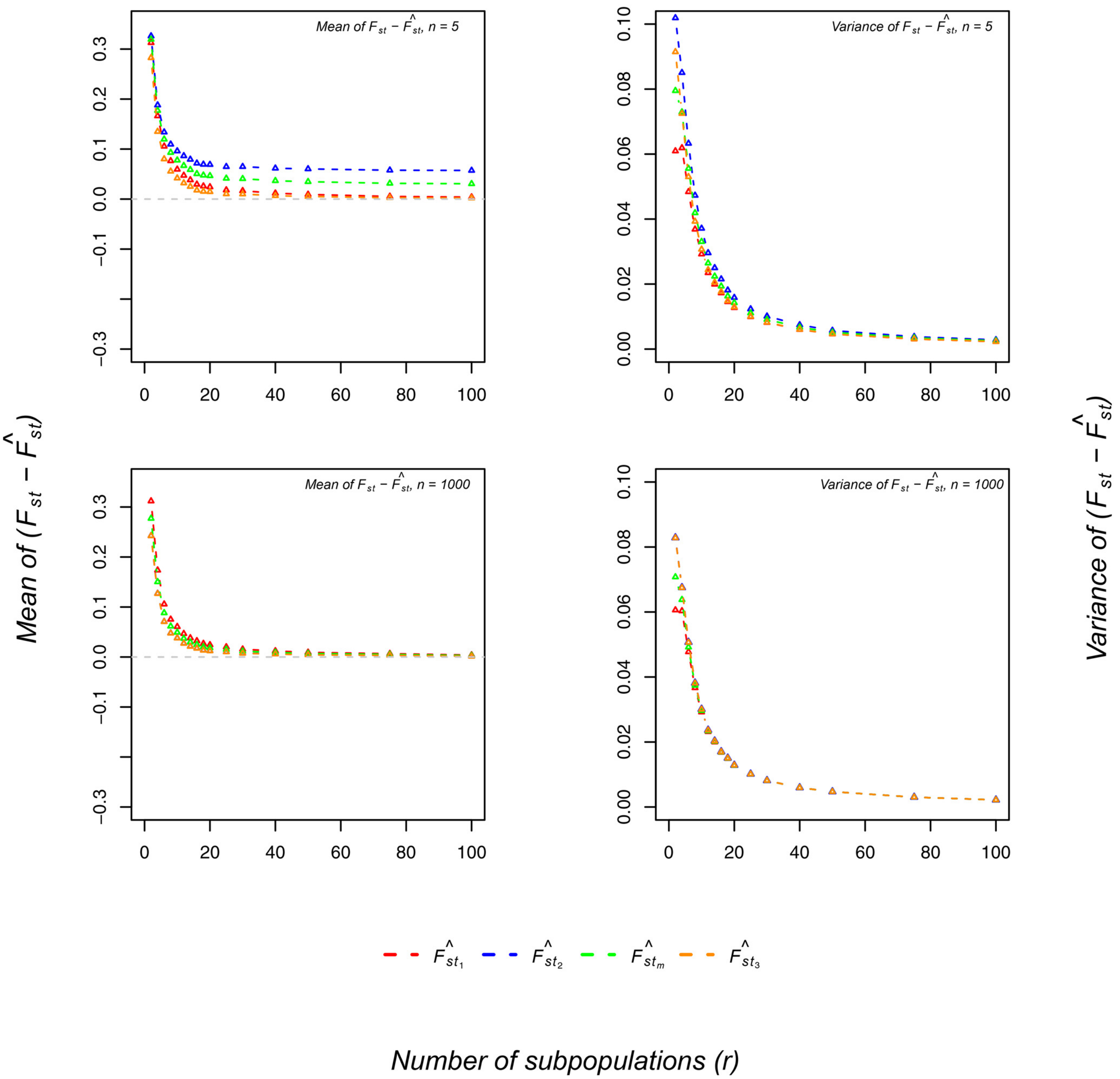

Fig 3. Effect of the number of subpopulations on bias. The x-axis shows the number of subpopulations. The y-axis shows the mean (left) and variance (right) of $F_{st} - \hat{F}_{st_1}$ (red), $F_{st} - \hat{F}_{st_2}$ (blue), $F_{st} - \hat{F}_{st_m}$ (green), and $F_{st} - \hat{F}_{st_3}$ (orange) values, given $F_{st} = 0.5$ and average allele frequency $p = 0.2$. The top plot represents 5 individuals per subpopulation and the bottom plot represents 1000 individuals per subpopulation.

doi:10.1371/journal.pone.0135368.g003

variances and MSE values. For $F_{st} > 0.5$, $\hat{F}_{st_2}$ had smaller variances and MSE values. For $F_{st} \approx 0.5$, $\hat{F}_{st_1}$, $\hat{F}_{st_2}$, and $\hat{F}_{st_m}$ had similar variance and MSE values.

The numbers of individuals and markers have been reported to affect F_{st} estimation [5]. We found that the number of subpopulations was more important than the number of individuals per subpopulation. Estimation of F_{st}, both in terms of means and variances, stabilized with approximately 30 subpopulations, regardless of the number of individuals per subpopulation.

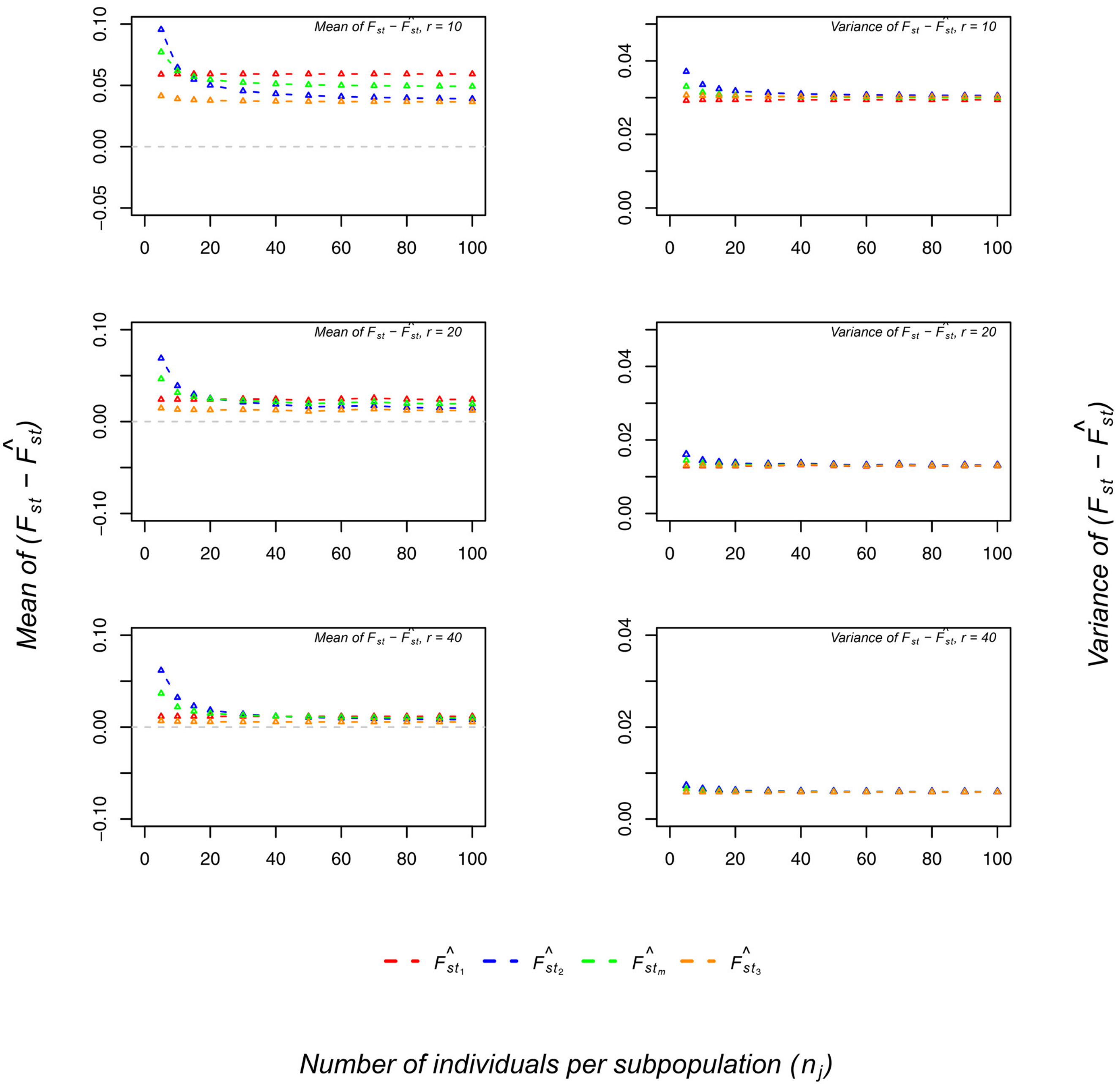

Fig 4. Effect of the number of individuals per subpopulation on bias. The x-axis shows the number of individuals per subpopulation. The y-axis shows the mean (left) and variance (right) of $F_{st} - \hat{F}_{st_1}$ (red), $F_{st} - \hat{F}_{st_2}$ (blue), $F_{st} - \hat{F}_{st_m}$ (green), and $F_{st} - \hat{F}_{st_3}$ (orange) values, given $F_{st} = 0.5$ and an average allele frequency $p = 0.2$. From top to bottom, the plots represent the number of subpopulations $r = 10$, 20, and 40, respectively.

doi:10.1371/journal.pone.0135368.g004

This behavior occurs because there are r estimates of $\hat{p}_j$ with which to estimate p and σ^2. Estimation was biased for $r = 2$ and improved as r increased, according to the Central Limit Theorem. Estimation was biased for $n_j < 30$ and improved as n_j increased (except for Wright's estimator), also according to the Central Limit Theorem. Our proposed estimator is a minimum variance combination of Wright's and Weir and Cockerham's estimators and is less

Table 3. $\hat{F}_{st}$ between WETH and HapMap 3 samples.

	$\hat{F}^*_{st_1}$	$\hat{F}^*_{st_2}$	$\hat{F}^*_{st_m}$	$\hat{F}^*_{st_3}$
ASW	0.0155 (0.0005)	0.0222 (0.0016)	0.0185 (0.0008)	0.0226 (0.0016)
CEU	0.0368 (0.0023)	0.0630 (0.0067)	0.0499 (0.0042)	0.0632 (0.0067)
CHB	0.0624 (0.0059)	0.1012 (0.0139)	0.0815 (0.0094)	0.1044 (0.0146)
CHD	0.0629 (0.0059)	0.1021 (0.0141)	0.0822 (0.0095)	0.1052 (0.0147)
GIH	0.0359 (0.0022)	0.0603 (0.0063)	0.0479 (0.0039)	0.0611 (0.0064)
JPT	0.0634 (0.0060)	0.1029 (0.0142)	0.0828 (0.0096)	0.1060 (0.0149)
LWK	0.0210 (0.0008)	0.0343 (0.0023)	0.0276 (0.0014)	0.0351 (0.0024)
MKK	0.0081 (0.0002)	0.0121 (0.0005)	0.0101 (0.0003)	0.0122 (0.0005)
MXL	0.0371 (0.0023)	0.0601 (0.0067)	0.0474 (0.0039)	0.0612 (0.0067)
TSI	0.0344 (0.0020)	0.0578 (0.0058)	0.0459 (0.0036)	0.0586 (0.0060)
YRI	0.0264 (0.0011)	0.0451 (0.0035)	0.0358 (0.0021)	0.0454 (0.0036)

* Shown are means (variances) of $\hat{F}_{st}$.

doi:10.1371/journal.pone.0135368.t003

biased than Weir and Cockerham's estimator for small samples sizes and less biased than Wright's estimator for large sample sizes.

Conclusion

A modified F_{st} estimator is proposed, which combines Wright's and Weir and Cockerham's estimators. It splits the difference in biases present in Wright's and Weir and Cockerham's estimators. We propose the routine use of this new and improved estimator of F_{st} as a way to reduce the biases and limitations of the classical estimators. We demonstrated that, in order to estimate F_{st} accurately, at least 30 subpopulations and 30 individuals per subpopulations are required.

Appendix

Proof of the Proposition

As $\bar{n} \to \infty$, $\bar{n}\hat{F}_{st_1}$ is asymptotically a chi-squared random variable, $S^2 \xrightarrow{P} 0$,
$\bar{n}T_1 \sim \bar{n}S^2 - \bar{p}(1-\bar{p}) = \bar{p}(1-\bar{p})\left(\bar{n}\hat{F}_{st_1} - 1\right)$, and

$$1 + \frac{(r-1)(\bar{n}-n_c)}{\bar{n}-1} = \frac{\sum_{j=1}^{r} n_j^2/(r\bar{n}) - 1}{\bar{n}-1} = O(1).$$

Thus

$$T_2 = [(n_c - 1)/(\bar{n} - 1)]\bar{p}(1-\bar{p}) + O_p(1)$$

and

$$\bar{n}\hat{F}_{st_2} = \frac{\bar{n}-1}{n_c-1}(\bar{n}\hat{F}_{st_1} - 1) + O_p(1).$$

Let

$$\delta = \lim_{n\to\infty} \bar{n}/n_c$$

and σ_1^2 be the asymptotic variance of $\bar{n}\hat{F}_{st_1}$, then the asymptotic variance of $\bar{n}\hat{F}_{st_2}$ is $\delta^2\sigma_1^2$, and the asymptotic covariance of $(\bar{n}\hat{F}_{st_1}, \bar{n}\hat{F}_{st_2})$ is $\delta\sigma_1^2$. Now we have

$$\bar{n}\hat{F}_{st}(a) = [a + \delta(1-a)]\bar{n}\hat{F}_{st_1} - \delta(1-a) + O_p(1).$$

If $p_1 = \cdots = p_2$, $\bar{p}(1-\bar{p}) \xrightarrow{P} p_0(1-p_0)$. Note the $\hat{p}_j$'s are independent, and

$$\bar{n}\hat{F}_{st_1} = \frac{1}{(r-1)p_0(1-p_0)}\sum_{j=1}^{r} n_j(\hat{p}_j - \bar{p})^2 + O_p(1).$$

Let $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_r)'$, $\mathbf{p}_0 = (p_0, \ldots, p_0)'$, $\gamma_j = n_j/n$, and $\mathbf{b}_j = (-\gamma_1, \ldots, -\gamma_{j-1}, (1-\gamma_j), -\gamma_{j+1}, \ldots, -\gamma_r)'$, then $\mathbf{b}'_j\mathbf{p}_0 = [(1-\gamma_j) + \sum_{i\neq j}^{r}\gamma_i]p_0 = 0$ for $j = 1, \ldots, r$, and so by the Central Limit Theorem,

$$\sqrt{n}(\hat{p}_j - \bar{p}) = \sqrt{n}\mathbf{b}'_j\hat{\mathbf{p}} = \sqrt{n}\mathbf{b}'_j(\hat{\mathbf{p}} - \mathbf{p}_0) \xrightarrow{D} N(0, \tau_j^2),$$

where $\tau_j^2 = \mathbf{b}'_j\Omega\mathbf{b}_j$, $\Omega = (\omega_{ij})_{r\times r}$ with $\omega_{ij} = p_0(1-p_0)$ if $i = j$ and $\omega_{ij} = 0$ if $i \neq j$.

Now we have, with $\mathbf{B} = \sum_{j=1}^{r}\gamma_j\mathbf{b}_j\mathbf{b}'_j$,

$$\begin{aligned}\bar{n}\hat{F}_{st}(a) + \delta(1-a) &= \frac{a + \alpha(1-a)}{(r-1)p_0(1-p_0)}\sum_{j=1}^{r} n_j(\hat{p}_j - \bar{p})^2 + O_p(1)\\ &= \frac{a + \alpha(1-a)}{(r-1)p_0(1-p_0)}\sum_{j=1}^{r}\gamma_j n\mathbf{b}'_j(\hat{\mathbf{p}} - \mathbf{p}_0)(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{b}_j + O_p(1)\\ &= \frac{a + \alpha(1-a)}{(r-1)p_0(1-p_0)}n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{B}(\hat{\mathbf{p}} - \mathbf{p}_0) + O_p(1).\end{aligned}$$

Let $\Omega^{-1/2}$ be the square root of Ω: $\Omega = \Omega'^{1/2}\Omega^{1/2}$, $\lambda_1, \ldots, \lambda_r$ be all the eigenvalues of $\Omega'^{1/2}\mathbf{B}\Omega^{1/2}$, and $\Lambda = diag(\lambda_1, \ldots, \lambda_r)$, then there is an orthogonal normal matrix $\mathbf{Q}$ such that $\Omega'^{1/2}\mathbf{B}\Omega^{1/2} = \mathbf{Q}'\Lambda\mathbf{Q}$, and so

$$n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{B}(\hat{\mathbf{p}} - \mathbf{p}_0) \xrightarrow{D} \sum_{j=1}^{r}\lambda_j\chi_j^2,$$

where the χ_j^2's are independent chi-squared random variables with one degree of freedom. This gives the desired result.

Lastly, we prove

$$\delta = \lim_{n\to\infty}\frac{\bar{n}}{n_c} \geq 1, \quad \text{with" = "if and only if } n_1 = \cdots = n_r.$$

In fact,

$$(r-1)(\bar{n} - n_c) = \frac{\sum_{j=1}^{r} n_j^2}{\sum_{j=1}^{r} n_j} - \frac{1}{r}\sum_{j=1}^{r} n_j = \frac{r\sum_{j=1}^{r} n_j^2 - (\sum_{j=1}^{r} n_j)^2}{r\sum_{j=1}^{r} n_j}.$$

It is known that for $r = 1$ or 2, $r\sum_{j=1}^{r} n_j^2 - (\sum_{j=1}^{r} n_j)^2 \geq 0$ with "=" if and only if $n_1 = \cdots = n_r$. Now we use induction to prove this is true for all integer r. In fact, suppose the above conclusion is true for some integer $r > 2$, then for integer $r + 1$,

$$A_n := (r+1)\sum_{j=1}^{r+1} n_j^2 = r\sum_{j=1}^{r} n_j^2 + \sum_{j=1}^{r} n_j^2 + (r+1)n_{r+1}^2,$$

and

$$B_n := \left(\sum_{j=1}^{r+1} n_j\right)^2 = \Big(\sum_{j=1}^{r} n_j\Big)^2 + 2n_{r+1}\sum_{j=1}^{r} n_j + n_{r+1}^2.$$

Since by assumption $r\sum_{j=1}^{r} n_j^2 \geq \big(\sum_{j=1}^{r} n_j\big)^2$,

$$A_n - B_n \geq \sum_{j=1}^{r} n_j^2 + (r+1)n_{r+1}^2 - 2n_{r+1}\sum_{j=1}^{r} n_j - n_{r+1}^2$$

$$= \sum_{j=1}^{r} n_j^2 + rn_{r+1}^2 - 2n_{r+1}\sum_{j=1}^{r} n_j = \sum_{j=1}^{r}(n_j^2 + n_{r+1}^2 - 2n_j n_{r+1}) \geq 0$$

with "=" if and only if $n_1 = \cdots n_{r+1}$, since $n_j^2 + n_{r+1}^2 - 2n_j n_{r+1} = (n_j - n_{r+1})^2 \geq 0$, with "=" if and only if $n_j = n_{r+1}$.

This gives $\delta \geq 1$ with "=" if and only if $n_1 = \cdots = n_r$.

Supporting Information

S1 Table. Means, Variances, and MSEs of $\hat{F}_{st}$ in simulation 2.
(PDF)

Acknowledgments

We thank Professor Gail Davey for allowing us to use the Wolaita GWAS dataset in this study.

Author Contributions

Conceived and designed the experiments: GC AY DS CNR. Analyzed the data: GC AY DS FT JZ AB YZ CW AA. Contributed reagents/materials/analysis tools: MJN. Wrote the paper: GC AY DS FT AB AA CNR.

References

1. Wright S. Genetical structure of populations. Nature 1950; 66(4215): 247–249. doi: 10.1038/166247a0
2. Wright S. The genetical structure of populations. Ann Eugen 1951; 15(4): 323–354. PMID: 24540312
3. Cockerham CC. Variance of Gene Frequencies. Evolution 1969; 23(1): 72–84. doi: 10.2307/2406485
4. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution 1984; 38(6): 1358–1370. doi: 10.2307/2408641
5. Willing E-M, Dreyer C, van Oosterhout C. Estimates of Genetic Differentiation Measured by F_{st} Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. PLOS ONE 2012; 7(8): e42649. doi: 10.1371/journal.pone.0042649 PMID: 22905157
6. Waples RS. Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. J Hered 1998; 89(5): 438–450. doi: 10.1093/jhered/89.5.438
7. Hudson RR, Slatkin M, Maddison WP. Estimation of Levels of Gene Flow from DNA Sequence Data. Genetics 1992; 132(2): 583–589. PMID: 1427045
8. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting F_{ST}: The impact of rare variants. Genome Res 2013; 23: 1514–1521. doi: 10.1101/gr.154831.113 PMID: 23861382

9. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of Genetic Markers for Inference of Ancestry. Am J Hum Genet 2003; 73(6): 1402–1422. doi: 10.1086/380416 PMID: 14631557

10. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for identity and paternity. Genetics 1995; 96: 3–11.

11. Tekola Ayele F, Adeyemo A, Finan C, Hailu E, Sinnott P, Burlinson ND, et al. HLA class II locus and susceptibility to podoconiosis. N Engl J Med 2012; 366(13): 1200–1208. doi: 10.1056/NEJMoa1108448 PMID: 22455414

12. Tekola-Ayele F, Adeyemo A, Aseffa A, Hailu E, Finan C, Davey G, et al. Clinical and pharmacogenomic implications of genetic variation in a Southern Ethiopian population. Pharmacogenomics J 2014; 15(1): 101–108. doi: 10.1038/tpj.2014.39 PMID: 25069476

13. Shriner D, Tekola-Ayele F, Adeyemo A, Rotimi CN. Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. Sci Rep 2014; 4: 6055 doi: 10.1038/srep06055 PMID: 25116736

14. Jakobsson M, Edge MD, Rosenberg NA. The Relationship between F_{st} and the Frequency of the Most Frequent Allele. Genetics 2013; 193(2): 513–528. doi: 10.1534/genetics.112.144758