Estimating and interpreting F_{ST}: The impact of rare variants

Gaurav Bhatia,[1,2,6,7] Nick Patterson,[2,6,7] Sriram Sankararaman,[2,3] and Alkes L. Price[2,4,5,7]

[1]Harvard–Massachusetts Institute of Technology (MIT), Division of Health, Science, and Technology, Cambridge, Massachusetts 02139, USA; [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA; [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [4]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA; [5]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA

In a pair of seminal papers, Sewall Wright and Gustave Malécot introduced F_{ST} as a measure of structure in natural populations. In the decades that followed, a number of papers provided differing definitions, estimation methods, and interpretations beyond Wright's. While this diversity in methods has enabled many studies in genetics, it has also introduced confusion regarding how to estimate F_{ST} from available data. Considering this confusion, wide variation in published estimates of F_{ST} for pairs of HapMap populations is a cause for concern. These estimates changed—in some cases more than twofold—when comparing estimates from genotyping arrays to those from sequence data. Indeed, changes in F_{ST} from sequencing data might be expected due to population genetic factors affecting rare variants. While rare variants do influence the result, we show that this is largely through differences in estimation methods. Correcting for this yields estimates of F_{ST} that are much more concordant between sequence and genotype data. These differences relate to three specific issues: (1) estimating F_{ST} for a single SNP, (2) combining estimates of F_{ST} across multiple SNPs, and (3) selecting the set of SNPs used in the computation. Changes in each of these aspects of estimation may result in F_{ST} estimates that are highly divergent from one another. Here, we clarify these issues and propose solutions.

[Supplemental material is available for this article.]

Since its introduction by Sewall Wright (1949) and Gustave Malécot (1948), F_{ST} estimation (Weir and Cockerham 1984; Holsinger and Weir 2009) has become a key component of studies of population structure in humans (International HapMap Consortium 2007; Li et al. 2008; The 1000 Genomes Project Consortium 2010; International HapMap 3 Consortium 2010) and other species (Malécot 1948; Wright 1949; Selander and Hudson 1976; Guries and Ledig 1982; Ellstrand and Elam 1993; Palumbi and Baker 1994). Though the utility of F_{ST} and related measures has been subject to recent debate (Jost 2008; Ryman and Leimar 2009), F_{ST} continues to be widely used by population geneticists (Xu et al. 2009; Edelaar et al. 2012; Hangartner et al. 2012).

Despite this widespread use in genetic studies, confusion remains about what F_{ST} is and how to estimate it. Beyond Wright's original description of F_{ST} as a ratio of variances, F_{ST} has been conceptually defined in many ways (Wright 1949; Cockerham 1969; Cavalli-Sforza and Bodmer 1971; Nei 1973; Slatkin 1991; Hudson et al. 1992). Additionally, multiple estimators for F_{ST} have been described in the literature (Nei 1973, 1986; Weir and Cockerham 1984; Hudson et al. 1992; Holsinger 1999; Weir and Hill 2002), often making the correct choice of estimator unclear.

With this diversity of definition and estimation in mind, we consider estimates of F_{ST} published by The 1000 Genomes Project Consortium (2010) of 0.052 for European and East Asian populations and 0.071 for European and West African populations. These are less than half of the published estimates, 0.111 and 0.156, from HapMap3 data (International HapMap 3 Consortium 2010) and may be the result of demography that differentially impacts F_{ST} at rare variants. These estimates have subsequently been used to simulate properties of recent rare variants (Mathieson and McVean 2012), making it imperative to know whether this reduction in F_{ST} is a meaningful result of the inclusion of rare variants or merely an artifact of estimation.

To answer these questions, we examine the issues surrounding F_{ST} estimated on data containing rare variants. We focus our attention on F_{ST} estimation in the context of comparing two populations—potentially with differing amounts of drift since the populations split—using a series of bi-allelic SNPs. We use the definition of Weir and Hill (2002), which allows for population-specific F_{ST}. Using this definition, we divide the issues surrounding estimation into three categories and examine them using both simulated and 1000 Genomes data:

1. Choice of F_{ST} estimator.
2. Combining estimates of F_{ST} across multiple SNPs.
3. Dependence of F_{ST} on the set of SNPs analyzed.

We conclude that the lower F_{ST} estimates reported by The 1000 Genomes Project Consortium (2010) are a consequence of the estimation method that was applied and are not informative for human demographic history. Correcting for differences in estimation method yields F_{ST} estimates of 0.106 for Europeans and East Asians and 0.139 for Europeans and West Africans—much closer to HapMap3 estimates. Overall, our results contradict a recent statement "among human populations, F_{ST} is typically estimated to be <0.1" by Mathieson and McVean (2012), which was based on results from The 1000 Genomes Project Consortium (2010).

Altogether, in the setting of rare variants, a careful protocol for producing F_{ST} estimates is warranted. We provide such a protocol.

[6]These authors contributed equally to this work.
[7]Corresponding authors
E-mail gbhatia@mit.edu
E-mail nickp@broadinstitute.org
E-mail aprice@hsph.harvard.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.154831.113.

23:1514–1521 © 2013, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

Results

Theory

Defining F_{ST}

We use the definition of Weir and Hill (2002) (WH) throughout our manuscript to analyze estimators in the context of comparing two populations at a series of bi-allelic SNPs. In this context, WH define F_{ST} as the correlation between randomly drawn alleles from a single population relative to the most recent common ancestral population:

$$\begin{aligned} E\left[p_i^s \middle| p_{anc}^s\right] &= p_{anc}^s \\ Var\left(p_i^s \middle| p_{anc}^s\right) &= F_{ST}^i p_{anc}^s\left(1-p_{anc}^s\right), \end{aligned} \tag{1}$$

where p_i^s is the allele frequency of the derived allele in population i, at SNP s, p_{anc}^s is the allele frequency of the derived allele in the ancestral population at SNP s, and F_{ST}^i is the population-specific F_{ST} for population i. For a pair of populations, F_{ST} is

$$F_{ST} = \frac{F_{ST}^1 + F_{ST}^2}{2}. \tag{2}$$

Although we use the WH definition of F_{ST} to compare estimation methods, numerous alternate definitions exist in the literature (see Supplemental Material), in part because of confusion regarding Wright's original description of F_{ST}.

Wright (1949) defined F_{ST} as the correlation of randomly drawn gametes from the same population, relative to the total population. However, he did not clearly specify the "total population," leaving subsequent investigators to interpret its meaning. For Nei (1973) the "total population" is the combination of the two population samples. This means that F_{ST} quantifies drift relative to an average of the two population samples. For Cockerham (1969) and WH, the "total" population is the most recent common ancestral population to the two populations being considered. Consistent with those investigators, we view F_{ST} as a parameter of the evolutionary process and not a statistic from observed samples as Nei has described.

To view F_{ST} as a parameter of the evolutionary process, the Cockerham and WH definitions assume that studied SNPs were polymorphic in the ancestral population. This is clear from Equation 1 as $E\left[p_i^s \middle| p_{anc}^s\right] \neq p_{anc}^s$ for SNPs arising from recent mutations. While this assumption does not always hold, we believe that the WH definition provides a valid basis for comparing estimation methods, and also assesses the performance of estimators when this assumption is violated.

By defining only one F_{ST} for both populations in a comparison, Cockerham (1969) and Weir and Cockerham (1984) also assumed that the two populations have experienced identical amounts of drift since splitting. This assumption, which may be unrealistic in many real data sets, was generalized by WH, and motivates our use of the WH definition. In this study, we focus on cases without migration and admixture, though these cases were considered in WH and are the subject of future work (B Weir, pers. comm.).

In addition to the definitions described above, F_{ST} has been related to divergence time, coalescent times, and migration rates. Additionally, likelihood-based definitions view F_{ST} as a parameter of the distribution of allele frequencies in current populations (Balding and Nichols 1995; Nicholson et al. 2002; Balding 2003). Further details are provided in the Supplemental Material.

Choice of F_{ST} estimator

While estimators of F_{ST} handle issues related to finite sample size, we are interested in their behavior in the limit of large sample sizes, or the "quantity being estimated." Most published estimates of F_{ST} are produced using the Weir and Cockerham (WC) (Weir and Cockerham 1984) (>8000 citations) or Nei (Nei 1973) (>5500 citations) estimators. However, we recommend a different estimator motivated by Hudson et al. (1992).

The WC estimator was developed for the case of populations with identical F_{ST}, and if it is used when F_{ST} is not identical for both populations, we demonstrate that the WC quantity being estimated becomes dependent on the ratio of sample sizes M according to (see Methods):

$$\hat{F}_{ST}^{WC} \rightarrow \frac{\left(F_{ST}^1 + F_{ST}^2\right)}{\left(F_{ST}^1 + F_{ST}^2 + 2\frac{1}{(M+1)}\left[M\left(1-F_{ST}^1\right)+\left(1-F_{ST}^2\right)\right]\right)}. \tag{3}$$

We note that this variation with sample size is not due to any flaw in the WC estimator, but rather due to the use of the WC estimator for a purpose different from what was intended. We also note that the WC estimator is often used to produce single SNP estimates of F_{ST} to detect selection. We caution that when sample sizes are very different, the WC estimator can give inflated single SNP estimates of F_{ST}, resulting in false-positive signals of selection (see Supplemental Material).

In the context of the WH definition, the Nei estimator will consistently overestimate F_{ST}, and the degree of overestimation will depend upon the magnitude of F_{ST} values (see Methods):

$$\hat{F}_{ST}^{Nei} \rightarrow \frac{\left(F_{ST}^1 + F_{ST}^2\right)}{2 - \frac{\left(F_{ST}^1 + F_{ST}^2\right)}{2}}. \tag{4}$$

We note that this result, with a maximum value of 2, makes it impossible to view F_{ST} as a correlation.

The Hudson estimator (Hudson et al. 1992; Keinan et al. 2007) produces estimates that are the simple average of F_{ST} according to the WH definition. These estimates are independent of sample sizes even when F_{ST} is not identical across populations. We note that while Hudson did not explicitly provide an estimator of F_{ST}, he did describe a method of estimation that corresponds to the estimator that we explicitly provide here (see Supplemental Material). Thus, we refer to this estimator as the Hudson estimator. Hudson estimates correspond to a simple average of the population specific F_{ST} estimates as given by (see Methods):

$$\hat{F}_{ST}^{Hudson} \rightarrow \frac{\left(F_{ST}^1 + F_{ST}^2\right)}{2}. \tag{5}$$

We note that the Hudson estimator is a simple average of the population-specific estimators proposed by Weir and Hill (2002). We provide comparisons of this estimator to the WC and Nei estimators when applied to simulated data (see Supplemental Material) and empirical data (see below).

Combining estimates of F_{ST} across multiple SNPs

We investigate two approaches for combining estimates of F_{ST} across multiple SNPs. In the first approach, variance components—the numerator and denominator—are averaged separately and the genome-wide estimate of F_{ST} is a "ratio of averages" (Weir and Cockerham 1984; International HapMap 3 Consortium 2010). In the second approach, single SNP estimates of F_{ST} are averaged across SNPs. The resulting "average of ratios" is reported as the genome-wide estimate (The 1000 Genomes Project Consortium 2010) (see Methods).

In the context of the WH definition, the numerator of the Hudson F_{ST} estimator (see Methods) is an unbiased estimator of the variance between populations. The denominator is an unbiased estimator of the total variance in the ancestral population. However, this does not mean that the ratio of the estimators is itself an unbiased estimator of F_{ST}. We are not aware of any unbiased estimator.

While an unbiased estimator is not available, F_{ST} estimates produced using a ratio of these two unbiased estimates will be asymptotically consistent, in the sense that they will converge to the correct underlying value as the number of independent SNPs increases. This is the basis of our recommendation that F_{ST} be estimated as a ratio of averages.

We analyze the effects of choosing an average of ratios in coalescent simulations detailed in the Supplemental Material.

Dependence of F_{ST} on the set of SNPs analyzed

It is well known that population genetic factors can cause variation in F_{ST} estimates, and that ascertainment schemes can alter the properties of studied SNPs (Ramírez-Soriano and Calafell 2008; Albrechtsen et al. 2010). For example, selection can result in differences between F_{ST} estimated on genic and nongenic SNPs (Clark et al. 2005; Barreiro et al. 2008; Hernandez et al. 2011); complex demography can cause F_{ST} to vary with SNP allele frequency (Schaffner et al. 2005) (see below). Indeed, variation in F_{ST} estimates between ascertained classes of SNPs can be used to test a variety of hypotheses about population history (Weir et al. 2005; McVicker et al. 2009). This usage of F_{ST} demonstrates that there is no single correct ascertainment scheme, as F_{ST} is a parameter of both the populations *and* the set of SNPs that are used in the computation.

Though there is no single correct ascertainment scheme, ascertainment in an outgroup may have desirable properties. Outgroup ascertainment guarantees that studied SNPs were polymorphic in the most recent common ancestral population (ignoring recurrent mutation), satisfying an assumption made in the Weir and Hill definition. This leads estimates of F_{ST} to be independent of allele frequency and depend upon time since divergence according to a simple equation (see Supplemental Material, Equation s1).

While we view these as desirable properties, if no reasonable outgroup sample is available, it may become necessary to choose SNPs that are polymorphic in one, both, or either of the populations studied. These choices will affect the estimate of F_{ST} produced and may explain discrepancies in F_{ST} estimates across studies of the same populations.

We explore the effects of various ascertainment schemes on F_{ST} estimates across the allele frequency spectrum in a variety of simulated demographic scenarios (see Supplemental Material).

Other F_{ST} estimators

In addition to the WC, Nei, and Hudson estimators that we analyzed above, we have also analyzed several additional estimators. Our results on each of these estimators are described in detail in the Supplemental Material.

The moment-based estimator of Weir and Hill (2002) (WH) introduced population-specific estimates of F_{ST}. Weir and Hill recommend a sample size weighted average of these estimates, which may result in a wide variation with sample size. However, one could also report these estimates independently or perform a simple average of these estimates.

A separate maximum-likelihood estimator of Weir and Hill (2002) (WH-ML) is based upon a normal approximation to genetic drift. However, the equations provided for the WH-ML estimator are not applicable to the general case of unequal sample size, and the investigators recommend that estimates be "simply averaged across loci," causing WH-ML estimates to vary widely with the inclusion of rare variants.

We evaluated two max-likelihood estimators based on the beta-binomial likelihood using point estimates for the allele frequency in the ancestral population (D Balding, pers. comm.). These estimates perform well for small values of F_{ST}, but do poorly as F_{ST} increases. It may be possible to improve on these methods by integrating over the distribution of ancestral allele frequencies, an interesting direction for future research.

We also considered the beta-binomial MCMC method of Holsinger (1999). However, our simulations suggest that Holsinger estimates increase dramatically if rare SNPs are analyzed. Additionally, the MCMC-based approach imposes a significant computational burden, making the method difficult to apply to modern data sets.

Analysis of 1000 Genomes data

We analyzed data from 1000 Genomes populations (The 1000 Genomes Project Consortium 2010) to illustrate the effects of changes in each of the aspects of estimation described above. We focus largely on the comparison of Utah residents of European ancestry (CEU) and Chinese individuals from Beijing (CHB), as the Yoruba in Ibadan, Nigeria (YRI) sample functions as a natural outgroup for ascertainment of SNPs. This ascertainment has desirable properties (see above).

Choice of F_{ST} estimator

Estimates of F_{ST} for CEU and CHB are 0.106 (s.e. 0.0006), 0.112 (s.e. 0.0006), and 0.107 (s.e. 0.0006) for the WC, Nei, and Hudson estimators, respectively. These estimates were produced over SNPs ascertained as polymorphic in YRI. The higher Nei estimate is expected. In addition, sample sizes for CEU (85 individuals) and CHB (97 individuals) are similar, so we do not expect WC and Hudson estimates to differ.

In order to investigate the effects of sample size variation we selected 14 individuals—the size of the smallest sample (Iberian populations in Spain; IBS) in the 1000 Genomes Consortium data—from both CEU and CHB to produce populations CEU14 and CHB14. Hudson F_{ST} estimates for CEU14 and CHB are similar to those for CHB14 and CEU (see Table 1). However, WC estimates are 0.114 (s.e. 0.0006) and 0.107 (s.e. 0.0006) for CEU14 vs. CHB and CHB14 vs. CEU, respectively. The difference between these estimates is statistically significant (greater than eight standard errors). To verify that this difference is not due to different sets of polymorphic SNPs, we re-estimated F_{ST} restricting to SNPs that were polymorphic in YRI and at least one of CEU14 or CHB14. Re-estimated values of F_{ST} were similar to those above and WC estimates remained discordant (data not shown).

The effect of sample size variation is further exacerbated when ascertainment is performed within the populations studied. For example, in comparing IBS—with a sample size of only 14

Table 1. F_{ST} estimates for pairs of populations in 1000 Genomes

		F_{ST} Estimator					
		WC		Nei		Hudson	
Comparison	Number of SNPs	Est.	Std. error	Est.	Std. error	Est.	Std. error
CEUvCHB	7,799,780	0.107	5.70×10^{-4}	0.112	6.36×10^{-4}	0.106	5.69×10^{-4}
CEUvYRI	17,814,120	0.139	4.97×10^{-4}	0.149	5.79×10^{-4}	0.139	5.00×10^{-4}
CHBvYRI	17,814,120	0.163	5.85×10^{-4}	0.175	6.84×10^{-4}	0.161	5.78×10^{-4}
CEUvCHB14	7,215,431	0.107	6.10×10^{-4}	0.113	7.16×10^{-4}	0.106	6.36×10^{-4}
CHBvCEU14	7,465,953	0.114	6.49×10^{-4}	0.114	7.12×10^{-4}	0.107	6.32×10^{-4}
IBSvYRI	17,814,120	0.121	4.37×10^{-4}	0.145	6.02×10^{-4}	0.131	6.73×10^{-4}
YRIvIBS[a]	7,709,984	0.144	8.06×10^{-4}	0.141	7.77×10^{-4}	0.134	8.43×10^{-4}

Unless otherwise specified, SNPs were ascertained as polymorphic in YRI. These estimates are more concordant with results reported on common SNPs (International HapMap 3 Consortium 2010) than with the results reported by the Genomes Consortium (The 1000 Genomes Project Consortium 2010). Even so, we note that the choice of F_{ST} estimator impacts the resulting estimate. This is evident when comparing CEU14—14 individuals sampled from the CEU population—to CHB and CHB to CEU14. Though these estimates are produced using overlapping sets of SNPs and individuals, the estimates are statistically significantly different when produced using the WC estimator. This difference is underscored when comparing the YRI and IBS populations. The small sample from the IBS population causes WC estimates to change significantly depending on ascertainment in IBS (line 4) or YRI (line 5). The number of SNPs listed indicates the number of SNPs that were polymorphic in the ascertained population (usually YRI) and at least one of the populations studied.
[a] In this case, ascertainment was performed in the IBS sample. In all other cases, ascertainment was performed in YRI.

individuals—to YRI, no reasonable outgroup population exists in the 1000 Genomes data. If we ascertain within one of these populations, WC estimates are 0.121 and 0.144 for ascertainment in YRI and IBS, respectively. These estimates—computed using identical populations and *even identical individuals*—are highly divergent at >25 standard errors apart, whereas Hudson estimates are much more stable (see Table 1). This underscores that F_{ST} estimates can vary substantially based on the choice of estimator.

Regardless of choice of estimator, our estimates of F_{ST} from 1000 Genomes data are relatively close to previously reported values of F_{ST} (see Supplemental Table S1 for all populations). This suggests that while the choice of estimator can impact the resulting value of F_{ST}, it does not explain the disparate results reported by the 1000 Genomes Consortium, and other aspects of estimation may be involved. We consider these in the sections below.

Combining estimates of F_{ST} *across multiple SNPs*

From 1000 Genomes data, we estimated F_{ST} for CEU and CHB as 0.106 (s.e. 0.0006) and 0.072 (s.e. 0.0003) for the ratio of averages and average of ratios, respectively. These estimates were produced over SNPs ascertained as polymorphic in YRI. This suggests that the result reported by the 1000 Genomes Consortium (0.052) may be partially explained by the large reduction in F_{ST} obtained by use of an average of ratios. These results are replicated for several comparisons of populations included in the 1000 Genomes data (see Table 2).

To explore the effect of the rare variants included in sequence data, we compared our results to those obtained using HapMap3 genotypes. We obtain F_{ST} estimates for CEU and CHB of 0.110 (s.e. 0.0010) and 0.089 (s.e. 0.0006) using the ratio of averages and average of ratios, respectively. This suggests that the inclusion of rare variants with low single-SNP F_{ST} estimates in the 1000 Genomes data tends to exacerbate the discrepancy produced by the average of ratios. We expect that this discrepancy will grow with sample sizes and sequencing depth (see Supplemental Fig. S2). Ultimately, using the average of ratios may make estimates incomparable across studies and unrelated to population demographic history.

While the use of the average of ratios clearly results in lower estimates of F_{ST}, these estimates are not as low as those published by the 1000 Genomes Consortium. Below, we explore the possibility that the remaining discrepancy can be accounted for by differences in the set of SNPs analyzed.

Dependence of F_{ST} *on the set of SNPs analyzed*

When estimating F_{ST} for CEU and CHB, we compared the effects of ascertaining in YRI (YRI ascertainment) versus ascertaining SNPs that were polymorphic in CEU, CHB, both populations, or either population (see Table 3). When using an average of ratios, our estimates of F_{ST} were ~0.103 for all of these modified ascertainment schemes. These can be compared to an F_{ST} of 0.106 produced from

Table 2. A comparison of the F_{ST} estimated using 1000 Genomes and HapMap data by either using a ratio of averages or an average of ratios

	Ratio of averages			
	1000 Genomes		HapMap3	
Comparison	Est.	Std. error	Est.	Std. error
CEU-YRI	0.139	5.00×10^{-4}	0.156	9.73×10^{-4}
CEU-CHB	0.106	5.69×10^{-4}	0.110	9.61×10^{-4}
CHB-YRI	0.161	5.78×10^{-4}	0.183	1.13×10^{-4}

	Average of ratios			
	1000 Genomes		HapMap3	
Comparison	Est.	Std. error	Est.	Std. error
CEU-YRI	0.063	1.53×10^{-4}	0.124	6.23×10^{-4}
CEU-CHB	0.072	3.04×10^{-4}	0.089	6.35×10^{-4}
CHB-YRI	0.070	1.70×10^{-4}	0.141	6.93×10^{-4}

It is clear that the average of ratios of F_{ST} results in a significant underestimate of F_{ST}, and use of an average of ratios approach can explain the bulk of the discrepancy between the F_{ST} reported by the 1000 Genomes Consortium and previously reported estimates. The ratio of averages estimates are much more concordant with estimates on HapMap data. We believe that discrepancies between these different data sets are due to the different set of SNPs used in the computation. Finally, use of the average of ratios results in a smaller reduction when applied to HapMap3 data. This is consistent with an average of ratios being sensitive to rare variants that are, in general, excluded from the HapMap set of SNPs.

Table 3. Assessing the effect of ascertainment schemes and combination methods on the resulting F_{ST} estimate for CEU and CHB

Polymorphic in	Ratio of averages		Average of ratios	
CEU	0.104	6.19×10^{-4}	0.056	2.55×10^{-4}
CHB	0.104	6.40×10^{-4}	0.057	2.74×10^{-4}
CEU AND CHB	0.104	7.25×10^{-4}	0.078	4.49×10^{-4}
CEU OR CHB	0.103	5.64×10^{-4}	**0.047**	1.87×10^{-4}

When using a ratio of averages, modified ascertainment results in a small, though statistically significant difference from a value of 0.106 obtained using YRI ascertainment. The effect is much larger when using an average of ratios, and the bolded cell indicates that a permissive ascertainment scheme coupled with an average of ratios can produce a value similar to the estimate of F_{ST} for CEU and CHB published by the 1000 Genomes Consortium.

YRI ascertainment in 1000 Genomes data or 0.110 in HapMap3 data. Though statistically significant, these results suggest that the effects of modified ascertainment are not very large when analyzing human populations using a ratio of averages. This indicates that reasonable estimates of F_{ST} may be produced when comparing populations without access to an outgroup.

However, when using an average of ratios and including all SNPs polymorphic in either CEU or CHB, our estimate changed from 0.072 to 0.047 (s.e. 0.0002), which is similar to the result reported by the 1000 Genomes Consortium. This suggests that much of the discrepancy between previously published estimates of F_{ST} for CEU and CHB and the published 1000 Genomes estimate is explained by using the average of ratios and an ascertainment scheme that includes all SNPs that are polymorphic in either of the two populations. These results are replicated for comparisons of continental populations included in the 1000 Genomes data as we obtained values of 0.056 and 0.063 for comparisons of CEU-YRI and CHB-YRI, respectively.

Separately, we note that when comparing CEU to CHB on the 1000 Genomes data we observed *larger* F_{ST} estimates of 0.108 for the lowest frequency SNPs ($0.0 < \text{MAF} \leq 0.05$) versus estimates of 0.103 for the most common SNPs ($0.45 < \text{MAF} < 0.5$) when ascertaining in CEU. These estimates were 0.131 and 0.097 when ascertaining in CHB (see Fig. 1). Increased F_{ST} for rare variants suggests that bottlenecks are likely to be a stronger influence on F_{ST} estimates for CEU and CHB than recent expansions. Our results also indicate that bottlenecks in the population history of CHB are likely to be stronger than those in the population history of CEU, consistent with the findings of Keinan et al. (2007). This is in contrast to the much lower F_{ST} estimates reported on sequence data by the 1000 Genomes Consortium, which might suggest that expansions are a stronger influence on F_{ST} at rare SNPs.

Under a simple demographic history (i.e., without migration or admixture), this dependence on minor allele frequency is expected to disappear when ascertaining SNPs in an outgroup. When ascertaining in YRI we do not observe any significant dependence on frequency, which suggests that YRI is a reasonable outgroup for the comparison for CEU and CHB.

We note that when ascertaining in YRI, our genome-wide estimate of F_{ST} (0.106) is lower than estimated from HapMap3 (0.110). To investigate whether this difference is due to non-random ascertainment of HapMap3 SNPs, we sampled 10 subsets of SNPs from the 1000 Genomes data that matched the allele frequency spectrum of HapMap3 SNPs (see Supplemental Material). We estimated F_{ST} for CEU and CHB in each of these subsets ranging from 0.106 to 0.107 (s.e. 0.0010). This suggests that HapMap3 SNPs are more highly differentiated than random SNPs, consistent with previous findings on the effects of ascertainment on genotyping arrays (Clark et al. 2005; Albrechtsen et al. 2010).

Recommendations

Choice of F_{ST} *estimator*

Because the Hudson estimator is not sensitive to the ratio of sample sizes and does not systematically overestimate F_{ST}, we recommend that it be used to estimate F_{ST} for pairs of populations. The Hudson estimator for F_{ST} and a corresponding block-jackknife estimator for the standard error of F_{ST} are implemented in the EIGENSOFT software package (EIGENSOFT 4.2 http://www.hsph.harvard.edu/faculty/alkes-price/software/).

Combining estimates of F_{ST} *across multiple SNPs*

Using an average of ratios will result in large reductions in F_{ST} estimates. This effect will be exacerbated when estimating F_{ST} from sequence data. Therefore, we recommend using a ratio of averages.

Dependence of F_{ST} *on the set of SNPs analyzed*

Estimating F_{ST} from SNPs ascertained in an outgroup has the following valuable properties: (1) F_{ST} estimates are expected to be independent of allele frequency in the outgroup, and (2) F_{ST} estimates will relate to divergence time according to Supplemental Equation s1 if there has been no migration or admixture. However, data from a reasonable outgroup is not always available. Additionally, comparison of F_{ST} between ascertained classes of SNPs (e.g.,

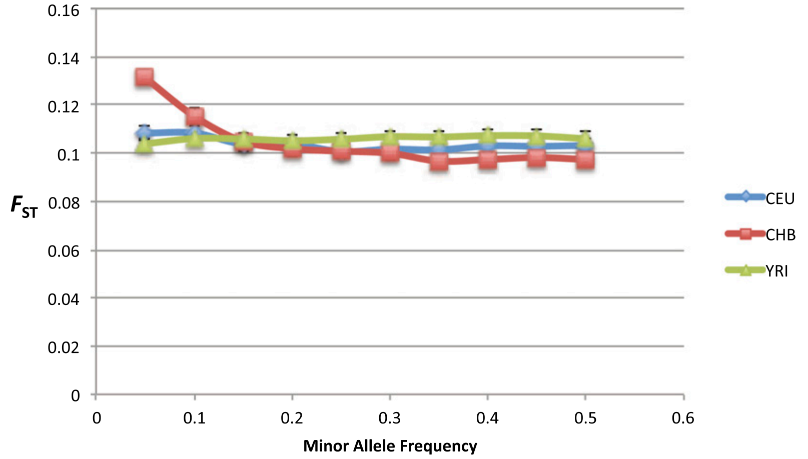

Figure 1. Allele frequency dependence of F_{ST} under different ascertainment schemes. This shows F_{ST} for CEU and CHB as a function of allele frequency when ascertaining in either CEU, CHB, or YRI. The increased F_{ST} for rare variants is consistent with bottlenecks being a stronger force on F_{ST} for CEU and CHB than recent expansion. In fact, this is consistent with a stronger bottleneck in the population history of CHB. We note that this frequency dependence disappears when ascertaining in YRI, suggesting that YRI is a reasonable outgroup for the comparison of CEU and CHB.

genic vs. nongenic) can be used to test a variety of hypotheses regarding population history. Thus, we recommend that future publications of F_{ST} estimates include details of the ascertainment scheme used, including the proportion of SNPs that are polymorphic in each sample.

Discussion

The use of F_{ST} to quantify the genetic distance between populations and to assess differentiation at individual SNPs is widespread. Here, we point out several challenges surrounding F_{ST} and provide a protocol for its robust estimation in the case of two populations and bi-allelic SNPs. We show that the estimator of F_{ST}, the method of combining estimates across SNPs, and the scheme for SNP ascertainment can impact the resulting estimate of F_{ST}. An inappropriate choice for any of these aspects of estimation can lead to widely disparate estimates of F_{ST}, especially in a setting of large numbers of rare variants.

Indeed, the F_{ST} estimate 0.052 for CEU and CHB reported by The 1000 Genomes Project Consortium (2010) underscores the need for a careful analysis. Utilizing the careful protocol set out here, we provide an estimate of 0.106 for CEU and CHB on 1000 Genomes data, which is close to our estimate of 0.110 on HapMap3 (International HapMap 3 Consortium 2010) data. Additionally, we show that when ascertaining for SNPs in one of the two populations studied, rare variants have higher F_{ST} estimates than common variants. This is the exact opposite of the results suggested by the 1000 Genomes data. The difference between these two results changes the conclusions that are drawn about the role of demography in shaping the patterns of differentiation between human populations. In addition to altering genome-wide estimates of F_{ST}, the choice of estimator can introduce inflation at the level of single SNP estimates, potentially making it difficult to interpret high F_{ST} estimates as signals of selection (see Supplemental Material).

Another concern about F_{ST} was considered by Jost (2008), who showed that as heterozygosity becomes large, F_{ST} will naturally approach 0—indicating low differentiation—even if all alleles at a locus are population private. In an effort to avoid this problem, Jost introduced D as an alternate measure of differentiation. However, it has been suggested that Jost's D shares the same problems as F_{ST}, and that these problems are sometimes even more pronounced for Jost's D (Ryman and Leimar 2009). In any case, F_{ST} and related measures "unquestionably provide important insights into population structure" (Jost 2008), particularly for species such as humans, in which heterozygosity is relatively low.

In conclusion, we recommend the use of the Hudson estimator (Hudson et al. 1992; Keinan et al. 2007) of F_{ST} that is independent of sample size. We demonstrate that a ratio of averages is an appropriate method for combining these estimates across multiple SNPs. We also show the value of estimating F_{ST} from SNPs ascertained in an outgroup, though we do not view this as a necessity. We do recommend, however, that future publications of F_{ST} estimates include details of the ascertainment of SNPs.

Methods

Weir and Cockerham's F_{ST} (WC)

Definition

Weir and Cockerham (1984) used the definition provided by Cockerham (1969) of F_{ST} as a ratio of the variance between populations to the total variance in the ancestral population. We analyze this definition in the Supplemental Material.

Estimator

In the setting of population-specific F_{ST}, described by the WH definition, the WC estimator will result in estimates that vary with the ratio of sample sizes (see Supplemental Material for details). In the case of two populations and biallelic SNPs, the WC estimator is

$$\hat{F}_{ST}^{WC} = 1 - \frac{2\frac{n_1 n_2}{n_1+n_2}\frac{1}{n_1+n_2-2}[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]}{\frac{n_1 n_2}{n_1+n_2}(\tilde{p}_1-\tilde{p}_2)^2+\left(2\frac{n_1 n_2}{n_1+n_2}-1\right)\frac{1}{n_1+n_2-2}[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]}, \tag{6}$$

where n_i is the sample size and $\tilde{p}_i$ is the sample allele frequency in population i for $i \in \{1, 2\}$. Then, in the limit of large sample sizes ($n_i - 1 \approx n_i$), we can assume that sample allele frequencies become close to population allele frequencies $(\tilde{p}_i \to p_i)$. We analyze the estimator as the sample sizes increase, but their ratio goes to a constant M (see Supplemental Material for a derivation). In this case, we show (see Supplemental Material) that the estimate tends toward Equation 1 (see Results).

If the sample sizes are equal, $M = 1$, then the estimate becomes

$$\hat{F}_{ST}^{WC} \to \frac{\left(F_{ST}^1 + F_{ST}^2\right)}{2}.$$

Also, when F_{ST} is identical for both populations, i.e., $F_{ST}^1 = F_{ST}^2 = F_{ST}$, it is straightforward to see that $\hat{F}_{ST} \to F_{ST}$, i.e., the estimate will not depend upon the ratio of sample sizes (M). We note that if F_{ST} is identical across populations, weighting by sample sizes will reduce the variance of the estimator. This was the intent of Weir and Cockerham. If the sample sizes are unequal or this assumption does not hold, however, the estimate will depend upon the ratio of sample sizes underlying the limit. Given the complexity of human population history, it is unlikely that this assumption will hold in general. This means that even if large numbers of samples and SNPs are used to estimate F_{ST} for a pair of populations, this estimate may not be comparable across studies with different sample sizes.

We note that when F_{ST} is not identical for both populations, it is possible to estimate F_{ST} separately for each population (i.e., $\hat{F}_{ST}^1, \hat{F}_{ST}^2$) (Weir and Hill 2002). Estimates for those produced according to the method given in Weir and Hill (2002) will not depend on sample size. We focus here on estimating F_{ST} for a pair of populations, as this is a very common use when analyzing human genetic data.

Nei's F_{ST}

Definition

Nei (1986) defined F_{ST} (he used the term G_{ST}) based upon the sample gene diversity between and within populations as

$$F_{ST} = \frac{D_{ST}^{'}}{H_T}, \tag{7}$$

where $D_{ST}^{'}$ is the average gene diversity between populations and H_T is the diversity in the average of the two population samples. We consider this definition in detail in the Supplemental Material.

Estimator

In the case of two populations and bi-allelic SNPs, Nei's estimator is

$$\hat{F}_{ST}^{Nei} = \frac{(\tilde{p}_1 - \tilde{p}_2)^2}{2\tilde{p}_{avg}\left(1 - \tilde{p}_{avg}\right)}, \tag{8}$$

where

$$\tilde{p}_{avg}=\frac{\tilde{p}_1+\tilde{p}_2}{2}$$

and $\tilde{p}_i$ is the sample allele frequency in population i for $i \in \{1, 2\}$. We note that this is Nei's updated estimator and, in the case of two populations, differs from the estimator given in Nei (1973) and Nei and Chesser (1983) by a factor of 2. We use the estimator given in Nei (1986), as it is most closely related to the other estimators considered.

Using the definition of Weir and Hill (2002) we show (see Supplemental Material) that estimates made using Nei's estimator will tend toward Equation 2 (see Results), with a maximum value of 2 as $F_{ST}^1 \rightarrow 1, F_{ST}^1 \rightarrow 1$. This overestimates the average of population-specific F_{ST} values and alters the relation from this average of F_{ST} values to divergence time (see Supplemental Material). Estimates of F_{ST} given for the Nei estimator were generated using the proposed estimator for the numerator (see Supplemental Material) and a simple estimator for the denominator.

Hudson's F_{ST}

Definition

Hudson et al. (1992) defined F_{ST} in terms of heterozygosity. The fundamental difference between these estimators is that for Hudson, the total variance is based upon the ancestral population and not the current sample.

Estimator

Hudson's estimator for F_{ST} is given by

$$\hat{F}_{ST}^{Hudson}=1-\frac{H_w}{H_b}, \tag{9}$$

where H_w is the mean number of differences within populations, and H_b is the mean number of differences between populations. While Hudson did not give explicit equations for H_w and H_b, we cast his description into an explicit estimator (see Supplemental Material for a derivation). The estimator that we analyze is

$$\hat{F}_{ST}^{Hudson}=\frac{(\tilde{p}_1-\tilde{p}_2)^2-\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1-1}-\frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2-1}}{\tilde{p}_1(1-\tilde{p}_2)+\tilde{p}_2(1-\tilde{p}_1)}, \tag{10}$$

where n_i is the sample size and $\tilde{p}_i$ is the sample allele frequency in population i for $i \in \{1, 2\}$. Analyzing this estimator using the definition of Weir and Hill (2002), we show (see Supplemental Material) that F_{ST} estimated using Hudson's estimator will tend toward Equation 3 (see Results), which is exactly the average of population-specific F_{ST} values that we seek to estimate. This emerges naturally, as the proposed estimator is the simple average of the population-specific estimators given in Weir and Hill (2002). This estimator has the desirable properties that it is (1) independent of sample composition, and (2) does not overestimate F_{ST} (it has a maximum value of 1). We recommend its use to produce estimates of F_{ST} for two populations.

Combining estimates of F_{ST} across multiple SNPs

The Hudson estimator is asymptotically consistent, as the estimators of the variance components involved in the computation of F_{ST} are unbiased in the context of the WH definition. However, as their quotient is not an unbiased estimator of F_{ST}, use of an average of ratios will, in general, result in a biased estimate.

As many rare variants discovered by deep sequencing are population specific, we analyze the effect of this approach in the presence of many such variants. Consider a rare SNP with $p_1 = \varepsilon$, $p_2 = 0$. This yields a single SNP $F_{ST} = \varepsilon$. An estimate produced using an average of ratios will be highly sensitive to rare SNPs of this type and is likely to exhibit dependence on both the sequencing depth and sample size used in the analysis (see Supplemental Fig. S2).

Previous works have examined this choice and advocated for the use of a ratio of averages (Reynolds et al. 1983; Weir and Cockerham 1984). However, in describing the WH-ML method, Weir and Hill recommend that estimates be "simply averaged over loci." We believe that use of an average of ratios can account for the bulk of the discrepancy between the estimates of F_{ST} from The 1000 Genomes Project Consortium (2010) and previously published estimates (International HapMap 3 Consortium 2010) (see Results).

Dependence of F_{ST} on the set of SNPs analyzed

In relating quantities being estimated from current populations to parameters of the evolutionary model, we have calculated expected values given the allele frequency in the ancestral population. This implicitly performs an ascertainment of SNPs that are polymorphic in the ancestral population or, equivalently, in an outgroup population. Provided there is no migration or admixture between populations, the relationship between F_{ST} and divergence time is given in Supplemental Equation s12.

This relationship accounts for changes in effective population size (i.e., bottlenecks or expansions) in the demographic history of the populations being compared. Additionally, ascertainment in an outgroup renders the estimate independent of the allele frequency spectrum in the outgroup. Therefore, with this type of ascertainment scheme, estimates should be concordant regardless of whether they are produced from rare or common SNPs.

While ascertainment in an outgroup has several helpful properties, in many practical circumstances no data from a reasonable outgroup is available. In these instances, F_{ST} can be estimated using SNPs ascertained in either one of the populations under study. However, in these instances estimates are *not* expected to be independent of allele frequency spectrum or complex demographic scenarios.

Acknowledgments

We are grateful to B. Weir for substantial contributions to this research and D. Reich, S. Sunyaev, S. Myers, N. Zaitlen, J. Wilson, A. Keinan, W. Jin, and D. Balding for helpful discussions. This research was funded by NIH grants T32 HG002295 (G.B.), R01 HG006399 (N.P. and A.L.P.), and R03 HG006170 (G.B. and A.L.P.).

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* **27:** 2534–2547.

Balding DJ. 2003. Likelihood-based inference for genetic correlation coefficients. *Theor Popul Biol* **63:** 221–230.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96:** 3–12.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet* **40:** 340–345.

Cavalli-Sforza LL, Bodmer WF. 1971. *The genetics of human populations*. W.H. Freeman, San Francisco, CA.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15:** 1496–1502.

Cockerham CC. 1969. Variance of gene frequencies. *Evolution* **23:** 72–84.

Edelaar P, Alonso D, Lagerveld S, Senar JC, Björklund M. 2012. Population differentiation and restricted gene flow in Spanish crossbills: Not isolation-by-distance but isolation-by-ecology. *J Evol Biol* **25:** 417–430.

Ellstrand NC, Elam DR. 1993. Population genetic consequences of small population size: Implications for plant conservation. *Ann Rev Ecol Syst* **24:** 217–242.

Guries RP, Ledig FT. 1982. Genetic diversity and population structure in pitch pine (*Pinus rigida* Mill.). *Evolution* **36:** 387–402.

Hangartner S, Laurila A, Räsänen K. 2012. Adaptive divergence in moor frog (*Rana arvalis*) populations along an acidification gradient: Inferences from Q_{ST}–F_{ST} correlations. *Evolution* **66:** 867–881.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, 1000 Genomes Project, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* **331:** 920–924.

Holsinger KE. 1999. Analysis of genetic diversity in geographically structured populations: A Bayesian perspective. *Hereditas* **130:** 245–255.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat Rev Genet* **10:** 639–650.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132:** 583–589.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52–58.

Jost L. 2008. GST and its relatives do not measure differentiation. *Mol Ecol* **17:** 4015–4026.

Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39:** 1251–1255.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.

Malécot G. 1948. *Les mathématiques de l'hérédié*. Masson & Cie, Paris, France.

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44:** 243–246.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5:** e1000471.

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci* **70:** 3321–3323.

Nei M. 1986. Definition and estimation of fixation indices. *Evolution* **40:** 643–645.

Nei M, Chesser RK. 1983. Estimation of fixation indices and gene diversities. *Ann Hum Genet* **47:** 253–259.

Nicholson G, Smith AV, Jonsson F, Gústafsson Ó, Stefansson K, Donnelly P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J R Stat Soc Series B Stat Methodol* **64:** 695–715.

Palumbi SR, Baker CS. 1994. Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Mol Biol Evol* **11:** 426–435.

Ramírez-Soriano A, Calafell F. 2008. FABSIM: A software for generating F_{ST} distributions with various ascertainment biases. *Bioinformatics* **24:** 2790–2791.

Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* **105:** 767–779.

Ryman N, Leimar O. 2009. G_{ST} is still a useful measure of genetic differentiation— a comment on Jost's *D*. *Mol Ecol* **18:** 2084–2087.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* **15:** 1576–1583.

Selander RK, Hudson RO. 1976. Animal population structure under close inbreeding: The land snail *Rumina* in Southern France. *Am Nat* **110:** 695–718.

Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genet Res* **58:** 167.

Weir BS, Cockerham CC. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38:** 1358–1370.

Weir BS, Hill WG. 2002. Estimating *F*-statistics. *Annu Rev Genet* **36:** 721–750.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* **15:** 1468–1476.

Wright S. 1949. The genetical structure of populations. *Ann Hum Genet* **15:** 323–354.

Xu S, Yin X, Li S, Jin W, Lou H, Yang L, Gong X, Wang H, Shen Y, Pan X, et al. 2009. Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet* **85:** 762–774.

Received January 11, 2013; accepted in revised form July 9, 2013.