9-1-2009

# "Genetics in geographically structured populations: defining, estimating and interpreting FST."

Kent E. Holsinger
*University of Connecticut - Storrs*, kent.holsinger@uconn.edu

Bruce S. Weir
*University of Washington - Seattle Campus*, bsweir@u.washington.edu

## Article Series on *Fundamental Concepts in Genetics*

# Genetics in geographically structured populations: defining, estimating, and interpreting $F_{ST}$


## Kent E. Holsinger

Department of Ecology & Evolutionary Biology, U-3043
University of Connecticut
Storrs, CT 06269-3043
USA

## Bruce S. Weir

Department of Biostatistics
University of Washington
Box 357232
Seattle, WA 98195
USA

e-mail: kent@darwin.eeb.uconn.edu, bsweir@u.washington.edu



## Abstract

**Wright's *F*-statistics, and especially $F_{ST}$, provide important insights into the evolutionary processes that influence the structure of genetic variation within and among populations, and they are among the most widely used descriptive statistics in population and evolutionary genetics. Estimates of $F_{ST}$ can identify regions of the genome that have been the target of selection, and comparisons of $F_{ST}$ from different parts of the genome can provide insight into the demographic history of those populations. For these reasons and others, $F_{ST}$ plays a central role in population and evolutionary genetics, and $F_{ST}$ has wide applications in fields from disease association mapping to forensic science. This article clarifies how $F_{ST}$ is defined, how it should be estimated, how it is related to similar statistics, and how estimates of $F_{ST}$ should be interpreted.**


Nearly every plant or animal species includes many partially isolated populations. Whether as a result of GENETIC DRIFT or divergent natural selection, such populations become genetically differentiated over time. For example, recent analyses based on more than 370 SHORT TANDEM REPEAT LOCI[1] (microsatellites) and 600,000 SNPs[2] suggest that only 5-10% of human genetic diversity is accounted for by genetic differences among populations from major geographical regions. These results indicate that there are far more similarities among geographically distinct human populations than differences. But what does it really mean to say that 5-10% of diversity is accounted for by differences among populations? And how is that figure derived? The short answer is that the estimate of $F_{ST}$ among human populations sampled from these regions is 0.05 for the microsatellite data and 0.10 for the SNP data but that answer

helps only if you understand what $F_{ST}$ is and how it is estimated from data, and what it means to get two different estimates for the same set of populations when we use different genetic markers.

Working independently in the 1940s and 1950s Sewall Wright[3] and Gustave Malécot[4] introduced *F*-statistics as a tool for describing the partitioning of genetic diversity within and among populations. In his remarkable 1931 paper,[5] Wright had already provided a comprehensive account of the processes leading to genetic differentiation among populations. He showed that the amount of genetic differentiation among populations has a predictable relationship to the rates of important evolutionary processes (migration, mutation, and drift). Large populations between which there is much migration, for example, tend to be little differentiated whereas small populations between which there is little migration tend to be greatly differentiated. $F_{ST}$ is a convenient measure of this differentiation, and as a result $F_{ST}$ and related statistics are among the most widely used descriptive statistics in population and evolutionary genetics.

But $F_{ST}$ is more than a descriptive statistic and measure of genetic differentiation. $F_{ST}$ is directly related to the VARIANCE in allele frequency among populations and conversely to the degree of resemblance among individuals within populations. If $F_{ST}$ is small, it means that allele frequencies within each population are very similar; if it is large, it means that allele frequencies are very different. If natural selection favors one allele over others at a particular locus in some populations, $F_{ST}$ at that locus will be larger than at loci where among-population differences are purely a result of genetic drift. Thus, genome scans that compare single-locus estimates of $F_{ST}$ with the genome-wide background may identify regions of the genome that have been subjected to DIVERSIFYING SELECTION.[6-8] Alternatively, if the demographic history of populations affects genetic variation on sex chromosomes differently from that on autosomes, then estimates of $F_{ST}$ derived from sex-chromosome markers may be very different from those derived from autosomal markers.[9]

Estimates of $F_{ST}$ are also important in association mapping of human disease genes and in forensic science. The same evolutionary processes that increase differentiation among populations also increase the similarity among individuals within populations. Thus, $F_{ST}$ must be considered when allele frequencies are compared between "cases" and "controls" to ensure that differences between them are greater than expected by chance. Similarly, the match probability between a suspect and a crime scene sample is specific to the set of people who might reasonably be expected to be sources of the sample. But defining this set is difficult, so a "$\theta$ correction" is applied to population frequencies to accommodate variation among subpopulations. The "$\theta$ correction" depends on the value of $F_{ST}$.

In this review we discuss how $F_{ST}$ is defined, describe approaches for estimating it from data, and illustrate several ways in which analysis of $F_{ST}$ can provide insight into the genetic structure and evolutionary dynamics of populations. In addition, we discuss four statistics that are related to $F_{ST}$ ($G_{ST}$, $R_{ST}$, $\Phi_{ST}$, and $Q_{ST}$), clarify the differences among them, and recommend when each should be used.

These additional statistics partition genetic diversity into within- and among-population components. Of the four, $G_{ST}$ is most closely related to $F_{ST}$, and it has been widely used as a measure of genetic differentiation among populations. As we describe below, however, $G_{ST}$ is an appropriate measure of

genetic differentiation only when the contribution of genetic drift to among population differences is not of interest. As a result, the contexts in which it is useful may be relatively limited. In contrast, $R_{ST}$ (for microsatellite data) and $\Phi_{ST}$ (for molecular sequence data) may be useful in a wide variety of contexts where it is important to account for the mutational "distances" among alleles, and $Q_{ST}$ may be useful in analysis of continuously varying traits.

## Definitions

Wright[3] introduced $F_{ST}$ as one of three interrelated parameters used to describe the genetic structure of diploid populations: $F_{IT}$, the correlation between gametes within an individual relative to the entire population, $F_{IS}$, the correlation between gametes within an individual relative to the subpopulation in which it occurs, and $F_{ST}$, the correlation between gametes chosen randomly from within the same subpopulation relative to the entire population. We describe here how these parameters are defined in terms of the departure of genotype frequencies from Hardy-Weinberg expectations.

### Deriving measures of genetic diversity

It may be easiest to understand *F*-statistics if we first think of statistics that describe departures from Hardy-Weinberg expectation. To make the discussion more concrete, consider two populations segregating for two alleles at a single locus. Label the frequency of allele $A_1$ in population 1 $p_1$, and its frequency in population 2 $p_2$. Also label the frequency of genotype $A_1A_1$ in the first population $x_{11,1}$, of genotype $A_1A_2$ in the first population $x_{12,1}$, and so on. Then the genotype frequencies in the two populations are given by the following set of equations:

$$x_{11,1} = p_1^2 + f_1 p_1(1-p_1)$$
$$x_{12,1} = 2p_1(1-p_1)(1-f_1)$$
$$x_{22,1} = (1-p_1)^2 + f_1 p_1(1-p_1)$$

.

$$x_{11,2} = p_2^2 + f_2 p_2(1-p_2)$$
$$x_{12,2} = 2p_2(1-p_2)(1-f_2)$$
$$x_{22,2} = (1-p_2)^2 + f_2 p_2(1-p_2)$$

Here $f_1$ and $f_2$ are what are often referred to as the within-population inbreeding coefficient, but that term can be misleading. In practice, $f$ is a measure of the frequency of heterozygotes compared to that expected when genotypes are in Hardy-Weinberg proportions. Inbreeding leads to a deficiency of heterozygotes relative to Hardy-Weinberg expectations, so when there is inbreeding in both populations, $f_1$ and $f_2$ will be positive. But if individuals *avoid* inbreeding or if there is HETEROZYGOTE ADVANTAGE, then heterozygotes will be more common than expected under Hardy-Weinberg and $f_1$ and $f_2$ will be negative. In short, $f_1$ and $f_2$ are measures of how different genotype proportions within populations are from Hardy-Weinberg expectations, with positive values indicating a deficiency of heterozygotes and negative values indicating an excess.

Now consider genotype frequencies in a combined sample consisting of a proportion *c* of individuals from the first population and a proportion 1-*c* of individuals from the second population. Just as

genotype frequencies in each population differ from Hardy-Weinberg expectations based on the allele frequency in each population, genotype frequencies in the combined sample will differ from Hardy-Weinberg expectations based on the average allele frequency. Specifically:

$$x_{11} = \pi^2 + F\pi(1-\pi)$$
$$x_{12} = 2\pi(1-\pi)(1-F) \quad ,$$
$$x_{22} = (1-\pi)^2 + F\pi(1-\pi)$$

where $\pi = cp_1 + (1-c)p_2$ is the average allele frequency for $A_1$ in the combined sample and $F$ is the total inbreeding coefficient.[10] A little algebra shows that $F$ can be expressed as

$$(1-F) = (1-f)(1-\theta) \quad , \quad (1)$$

where $f = cf_1 + (1-c)f_2$ is the average within-population departure from Hardy-Weinberg expectations and $\theta$ is a measure of allele frequency differentiation among populations (see Box 1 for a summary of the mathematical notation used in this paper). More generally, we can define $\theta$ as

$$\theta = \frac{\sigma_\pi^2}{\pi(1-\pi)} \quad , \quad (2)$$

where $\sigma^2_\pi$ is the variance in allele frequency among populations. $\pi(1-\pi)$ is the variance in allelic state for an allele chosen randomly the entire population, so it may be regarded as a measure of genetic diversity in the entire population. Thus, $\theta$ can be interpreted as the proportion of genetic diversity that is due to allele frequency differences among populations.

Wright first developed these ideas in the context of a model of discrete populations with each population having the same size and receiving immigrants from all other populations at the same rate,[5] but the statistical argument just developed applies to *any* partitioning in which populations differ in allele frequency, whether or not those populations are discrete.[11] Thus, when we use $\theta$ as a purely descriptive statistic describing the partitioning of genetic diversity among "populations", we need make no assumptions about whether the "populations" we sample are discrete or about the evolutionary processes that may have led to differences among them. Nonetheless, other methods of analysis may be more informative in continuously-distributed populations.[12-14]

## Linking $f$, $\theta$, and $F$ to Wright's $F$-statistics

Using a different approach, Cockerham[10,15] showed that $f$, $\theta$, and $F$ can also be thought of as intraclass correlation coefficients. Using this approach he showed that $f$ is the correlation between alleles within individuals relative to the population to which they belong, $\theta$ is the correlation between alleles within populations relative to the combined population, and $F$ is the correlation between alleles within an individuals relative to the combined population. These are precisely the definitions Wright gave for $F_{IS}$, $F_{ST}$, and $F_{IT}$, respectively. In short, $f$ and $F_{IS}$ can be thought of either as the average within-population departure from Hardy-Weinberg frequencies or as the correlation between alleles within individuals relative to the population to which they belong. $\theta$ and $F_{ST}$ can be thought of either as the proportion of genetic diversity due to allele frequency differences among populations or as the correlations between

alleles within populations relative to the entire population. $F$ and $F_{IT}$ can be thought of either as the departure of genotype frequencies in the combined sample from Hardy-Weinberg expectations or as the correlation between alleles within individuals relative to the combined sample.

In Wright's notation, subscripts refer to a comparison between levels in a hierarchy: *IS* refers to "individuals within subpopulations", *ST* to "subpopulations within total", and *IT* to "individuals within total."[16] The hierarchy in (1) may be extended indefinitely to accommodate such structure. For example, Wright[16] describes variation in the frequency of the Standard chromosome in *Drosophila pseudoobscura* in the western United States at the level of demes (local populations: D), regions (groups of several demes: R), subdivisions (groups of several regions: S), and the total range (T). The corresponding *F*-statistics are related in the same multiplicative way as $f$, $\theta$, and $F$:

$$(1-F_{DT}) = (1-F_{DR})(1-F_{RS})(1-F_{ST}) \quad .$$

In this scheme, $F_{DR}$ measures the amount of differentiation among demes within region, $F_{RS}$ differentiation among regions within subdivisions, and $F_{ST}$ among subdivisions within the total range.

Returning to the examples of genetic differentiation among human populations mentioned in the introduction, we can now see that an estimate for $F_{ST}$ or $\theta$ of 0.05 (from microsatellites) and 0.10 (from SNPs) suggests that only 5-10% of human genetic diversity is a result of genetic differentiation among human populations. What may be surprising is that both estimates are derived from the same set of populations – this indicates that the amount of genetic differentiation among human populations is greater at SNP loci than at microsatellites.

## Estimation

### Statistical sampling

When Wright and Malécot introduced *F*-statistics, they did not distinguish between the parameters defined in the preceding section and the *estimates* of those parameters that we make from data. Not making this distinction is similar to confusing the mean height of the human population with an *estimate* of the mean height calculated from a sample of the population. Estimates of height must account for the variation associated with taking a finite sample from a population. New samples from the same population will have different characteristics. We refer to this variation as *statistical sampling* (Box 2).[17] In the context of *F*-statistics, statistical sampling refers to variation associated with collecting genetic samples from a fixed set of populations that have fixed but unknown genotype frequencies. The magnitude of variation associated with statistical sampling can be reduced by increasing the size of within-population samples.

### Genetic sampling

There is an important difference between estimates made by *F*-statistics and estimates of height. In addition to accounting for statistical sampling, *F*-statistics must also account for differences among the set of populations that might have been sampled. These differences may arise either because the populations from which we sample are only a subset of all populations that could have been sampled (statistical sampling of populations rather than statistical sampling of genotypes within populations) or because the populations from which we sample represent only one possible outcome of an underlying

stochastic evolutionary process. Even if we could take the set of populations we sampled back to a previous point in time and re-run the evolutionary process with all of the same conditions (population sizes, mutation rates, migration rates, and selection coefficients), the genotype frequencies in our new set of populations would differ from those in the populations we actually sampled.[18] This *genetic sampling*[17] is an unavoidable consequence of genetic drift. The magnitude of variation associated with genetic sampling *cannot* be reduced by increasing either the number of individuals sampled within populations or the number of populations sampled. Indeed, it is the characteristics of genetic sampling that estimates of *F*-statistics reveal.

## Approaches to estimating $F_{ST}$

In simple cases, it may make sense to estimate statistical parameters using simple functions of the data, like the sample mean. In more complicated cases, like those presented by *F*-statistics, it is useful to have well-defined approaches to constructing those estimates. Statisticians have developed several different approaches to estimate parameters from data.[19] Three widely used approaches are the method of moments, the method of maximum likelihood, and Bayesian methods.

### *Method of moments estimates*

The method of moments produces an estimate by finding an algebraic expression that makes the expected value of certain sample statistics equal to simple functions of the parameters we are trying to estimate (as explained in more detail below).[19] Method of moments estimates are designed to have low bias in the sense that if samples are taken repeatedly from the same population, the average of the corresponding sample variances will be close to the unknown population variance. They have the additional advantages that they are easy to calculate and that they do not require us to assume anything about the shape of the distribution from which our sample is drawn, other than that it has a mean and variance.

For *F*-statistics, method of moments estimates[17,20,21] are based on an analysis of variance (ANOVA) of allele frequencies. ANOVA is a statistical method that tests whether the means of two or more groups are equal, and can therefore be used to assess the degree of differentiation between populations. Briefly, if the variance among populations is the same as the variance within populations then there is no population substructure. ANOVA calculations are framed in terms of mean squares. Therefore, in practice, one calculates the expected mean square among populations (i.e., the variance of sample allele frequencies about the mean allele frequency over all populations), and the expected mean square within populations (the heterozygosity within populations when genotypes are in Hardy-Weinberg proportions) averaged over all possible samples (statistical sampling) from all possible populations with the same evolutionary history (genetic sampling). These expected values are then equated to observed mean squares calculated from a sample, and the resulting set of equations is solved for the corresponding variance components. Following Cockerham,[10,22] *F*-statistics are defined in terms of these variance components (see Box 3).

### *Maximum likelihood and Bayesian estimates*

In contrast, likelihood and Bayesian estimates are more difficult to calculate and require us to specify the probability distribution from which our sample was drawn. They first require us to specify a

probability distribution from which our sample was drawn. We then calculate a quantity called the LIKELIHOOD that is proportional to the probability of our observed data given those parameters. A maximum-likelihood estimate for the parameters is obtained by finding values of the unknown parameters that maximize that likelihood.[19] In most cases**,** maximum-likelihood estimates are biased. Nonetheless, they typically have a smaller variance and deviate less from the unknown population parameter than the corresponding method of moments estimates.[19] For these and other reasons, the method of maximum-likelihood is the most widely used technique for deriving statistical estimators. [23,24]

Bayesian estimates share many of the advantages associated with maximum-likelihood estimates, because they use the same likelihood to relate the data to unknown parameters. They differ from maximum-likelihood estimates, however, because the likelihood is modified by placing PRIOR DISTRIBUTIONS on unknown parameters, and estimates are based on the POSTERIOR DISTRIBUTION, which is proportional to the product of the likelihood and the prior distributions. Both maximum-likelihood and Bayesian methods suffer the disadvantage that simple algebraic expressions for the estimates are rarely available. Instead, the estimates are obtained through computational methods. Because the MCMC METHODS used for analysis of Bayesian models do not require that a unique point of maximum likelihood be identified, Bayesian estimates can be obtained even in complex models with thousands or tens of thousands of parameters when numerical maximization of the likelihood would be difficult or impossible.[26]

For *F*-statistics, the likelihood approach[27,28] specifies a probability distribution to describe the variation in allele frequencies among populations and a MULTINOMIAL DISTRIBUTION for genotype samples within populations. $\theta$ is related to the variance of the probability distribution describing the among-population distribution of allele frequencies, and genotype frequencies are determined by the allele frequencies in each population and *f*. Estimates are obtained by maximizing the likelihood function with respect to $\theta$, *f*, and the allele frequencies. The Bayesian approach uses the same likelihood function, and after placing appropriate prior distributions on *f*, $\theta$, and allele frequencies, MCMC methods are used to sample from the posterior distributions of *f* and $\theta$.

### *Comparing the methods*

With more than 5000 citations, the moment method described by Weir and Cockerham[20] has been widely used, partly because of its robustness and partly because it is simple to implement. The maximum-likelihood methods also give simple equations when the distribution of allele frequencies among populations is assumed to be normal,[27] but only if sample sizes are equal.[29] Bayesian methods allow probability statements to be made about *F*-statistics and extensions allow the relationship between *F*-statistics and demographic or environmental covariates to be explored in the context of a single model,[30] but implementations may be computationally demanding.

Box 3 uses a simple dataset to illustrate an analysis and the slightly different estimates obtained from each approach. Extensive comparisons of moment and Bayesian estimates of $F_{ST}$ have not been done, but our experience suggests that differences are small when (1) the average number of individuals per population is moderate to large (> 20), (2) the number of populations is moderate to large (> 10-15), and (3) most populations are polymorphic. When differences arise they reflect differences in the treatment

of allele frequency estimates when alleles are rare or sample sizes are small. The Bayesian approach "smooths" population allele frequencies toward the mean,[24] and does so more aggressively when alleles are rare or sample sizes are small. The moment approach treats the sample frequencies as fixed quantities without such smoothing. Simulation results in Levsen et al.[31] are consistent with this interpretation, although they compare Bayesian estimates with estimates of $G_{ST}$,[32] which does not account for genetic sampling.

## Related statistics

Population geneticists have proposed several statistics related to $F_{ST}$. Here we describe four of them: $G_{ST}$, $R_{ST}$, $\Phi_{ST}$, and $Q_{ST}$. Nei[33] introduced $G_{ST}$ as a measure of population differentiation. We discuss its relationship to $F_{ST}$ in Box 2. Haplotype and microsatellite data contain information not only about the frequency with which particular alleles occur but also on the evolutionary distance among them. Statistics like $\Phi_{ST}$ (for haplotype data) and $R_{ST}$ (for microsatellite) data are intended to take advantage of this additional information and to provide greater insight into patterns of relationship among populations. While $F_{ST}$, $\Phi_{ST}$, and $R_{ST}$ all apply to discrete genetic data, $Q_{ST}$ is an analogous statistic for continuously varying traits. Comparing an estimate of $Q_{ST}$ with an estimate of $F_{ST}$ may provide investigators with evidence that natural selection has shaped the pattern of variation in the quantitative trait if the markers used to estimate $F_{ST}$ can be presumed to be selectively neutral.

### $R_{ST}$, $\Phi_{ST}$, and AMOVA

The methods for estimating $f$, $\theta$, and $F$ described above are appropriate for multiallelic data when the alleles are regarded as equivalent to one another. When the data consist of variation at microsatellite loci or of nucleotide sequence (haplotype) information, however, related methods that allow mutation rates to differ between different pairs of alleles may be more appropriate. Excoffier et al.[34] introduced the Analysis of Molecular Variance (AMOVA) for analysis of haplotype variation. AMOVA is based on an analysis of variance framework analogous to the one developed by Weir and Cockerham[20]. The mean squares in an AMOVA analysis are based on a user-specified measure of the evolutionary distance between haplotypes, and AMOVA leads to quantities analogous to classical $F$-statistics (Box 1). Similarly, the mean squares used to calculate $R_{ST}$[35,36] are based on the number of repeat differences between alleles at each microsatellite locus. While the result of both analyses is a partitioning of genetic variance into within- and among-population components analogous to $F_{ST}$, neither has a direct interpretation as a parameter of a statistical distribution. Rather they estimate an index derived from two different statistical distributions: the distribution of allele (haplotype or microsatellite) frequencies among populations and the distribution of evolutionary distances among alleles. Nonetheless, such measures may be thought of as estimating the additional time to common ancestry of randomly chosen alleles that is the result of populations being subdivided,[37,38] provided that the measure of evolutionary distance between any two alleles is proportional to the time since their most recent common ancestor. Extensive simulation studies have shown that estimates of $R_{ST}$ may be unreliable unless a large number of loci are used, [39-41] but unlike $F_{ST}$ the expected value of $R_{ST}$ does not depend on the rate of mutation. Estimates of $\Phi_{ST}$ or $R_{ST}$ may be useful when mutations have contributed substantially to allelic differences among populations, but their usefulness may be limited by the extent to which the mutational model underlying the statistics matches the actual mutational processes in the system.[39]

### $Q_{ST}$ and polygenic variation

Spitze[42] pointed out that another quantity analogous to $\theta$ can be estimated for continuously varying traits. Specifically, we can define

$$Q_{ST} = \frac{\sigma^2_{GP}}{\sigma^2_{GP} + 2\sigma^2_{GI}},$$

where $\sigma^2_{GP}$ is the ADDITIVE GENETIC VARIANCE among populations and $\sigma^2_{GI}$ is the additive genetic variance within populations. $\sigma^2_{GP}$ can be estimated from between-population crosses, and $\sigma^2_{GI}$ can be estimated from within-population crosses. Since the total variance in between-population crosses is $\sigma^2_{GP} + \sigma^2_{GI}$, $Q_{ST}$ is the proportion of additive genetic variance in a trait that is due to among-population differences. If the trait is selectively neutral, if all genetic variation is additive, and if mutation rates at loci contributing to the trait are the same as those at other loci, then we expect $Q_{ST}$ and $F_{ST}$ to be equal.[43,44] Thus, comparing the magnitude of $Q_{ST}$ and $F_{ST}$ may indicate whether a particular trait has been subject to stabilizing ($Q_{ST} < F_{ST}$) or diversifying ($Q_{ST} > F_{ST}$) selection. Because of the large uncertainties associated with estimates of both $Q_{ST}$ and $F_{ST}$, however, such comparisons are likely to be useful only when they are available for a moderately large number of populations (> 20).[45] Furthermore, caution is necessary when suggesting that a $Q_{ST}/F_{ST}$ comparison provides evidence for stabilizing selection, because non-additive genetic variation tends to change $Q_{ST}$, even for a trait that is neutral.[46]

## Applications

*F*-statistics include both $F_{ST}$, which measures the amount of genetic differentiation among populations (and simultaneously the extent to which individuals within populations are similar to one another), and $F_{IS}$, which measures the departure of genotype frequencies within populations from HARDY-WEINBERG PROPORTIONS. Here we focus on applications of $F_{ST}$ for several reasons (see Box 4).

### Estimating migration rates

Wright[5] showed that if all populations in a species are equally likely to exchange migrants and if migration is rare, then

$$F_{ST} \approx \frac{1}{4N_e m + 1},$$

where $m$ is the fraction of each population composed of migrants (the backward migration rate)[47] and $N_e$ is the EFFECTIVE POPULATION SIZE of local populations.[48] Because of this simple relationship, it is tempting to use estimates of $F_{ST}$ from population data to estimate $N_e m$.

Unfortunately, it has been recognized for many years that this simple approach to estimating migration rates may fail.[49] The most obvious reason is that populations are rarely structured so that all populations exchange migrants at the same rate, causing some populations to resemble one another more than others. If differentiation is solely a result of isolation by distance,[50] for example, then the slope of the regression of $F_{ST}/(1-F_{ST})$ on either the logarithm of between-population distance (for populations distributed in 2 dimensions) or the between population distance alone (for populations in a linear

habitat) is proportional to $D_e\delta^2$, where $D_e$ is the effective density of the population ($D_e = N_e$/area) and $\delta^2$ is the mean squared dispersal distance.[51] But if differentiation is the result not only of isolation by distance but also of natural selection, or if the drift-migration process has not reached stationarity, then the slope of this relationship cannot be interpreted as an estimate of migration. Moreover, either a pure migration-drift process or a pure drift-divergence process or a combination of the two could produce the same distribution of allele frequencies. Indeed, either migration-drift or drift-divergence or a combination of the two can account for any pattern of allele frequency differences among populations.[52] Thus, while pairwise estimates of $F_{ST}$ (or $\Phi_{ST}$ or $R_{ST}$) provide some insight into the degree to which populations are historically connected,[37,38] they do not allow us to determine whether that connection is a result of ongoing migration or of recent common ancestry.

And the difficulty goes even deeper than that. Different genetic markers may give different estimates of $F_{ST}$ for a variety of reasons, and to derive an estimate of migration rates from $F_{ST}$ we must assume that the particular set of markers we happen to have chosen bear the expected relationship with $N_em$. This may often be problematic. Differences between $F_{ST}$ estimates from human microsatellites (0.05) and SNPs (0.10), for example, cannot reflect differences in migration rate, because both estimates are derived from the same set of individuals and the same set of populations – the HGDP-CEPH sample.[1,2,53] By incorporating models of the mutational process, COALESCENT-BASED APPROACHES are one way to escape this difficulty.[54-56]

## Inferring demographic history

On the other hand, population-specific or pairwise estimates of $F_{ST}$ may provide insight into the demographic history of populations when estimates are available from many loci. For example, Keinan et al.[9] report pairwise estimates of $F_{ST}$ for 13,600-62,830 autosomal SNP loci and 1100-2700 X-chromosome SNP loci in human population samples from northern Europe, east Asia, and west Africa. Because there are four copies of each autosome in the human population for each three copies of the X chromosome, we expect greater differentiation at X chromosome loci than at autosomal loci. Specifically, for two populations that diverged $t$ generations ago we expect

$$1 - F_{ST} = \left(1 - \frac{1}{2N_e}\right)^t,$$

where $N_e$ is the effective size of the local populations. Thus, if we define $Q$ as $\ln(1 - F_{ST}^{auto})/\ln(1 - F_{ST}^{X})$ we see that $Q$ is approximately $N_e^X / N_e^{auto} = 0.75$.

While $Q$ is approximately 0.75 for comparisons between east Asians and northern Europeans ($Q$ = 0.72±0.05), it substantially smaller for comparisons between west Africans and other populations in the sample ($Q$ = 0.58±0.03 for the comparison with northern Europeans; $Q$ = 0.62±0.03 for the comparison with east Asians). These results suggest either sex-biased dispersal (long-range immigration of males from Africa after non-African populations were initially established) or selection on X-chromosome loci after divergence of African and non-African populations.

## Identifying genomic regions under selection

Similarly, locus-specific estimates of $F_{ST}$ may identify genomic regions that have been subject to selection. The logic is straightforward. The pattern of genetic differentiation at a neutral locus is completely determined by the demographic history of those populations (i.e., the history of population expansions and contractions), the mutation rates at the loci concerned, and the rates and patterns of migration among those populations.[6,57-60] In a typical multilocus sample, it is reasonable to assume that all autosomal loci have experienced the same demographic history and the same rates and patterns of migration. If the loci also have comparable mutation rates and if variation at each locus is selectively neutral, then the allelic variation at each locus represents a separate sample from the same underlying stochastic evolutionary process. Loci showing unusually large amounts of differentiation may mark regions of the genome that have been subject to diversifying selection, while loci showing unusually small amounts of differentiation may mark regions of the genome that have been subject to STABILIZING SELECTION.[58] Several groups have used such genome scans to examine patterns of differentiation in the human genome.

By comparing locus-specific estimates of $F_{ST}$ with the genome-wide distribution, Akey et al. [6] identified 174 regions (out of 26,530 examined) that exhibited what they call "signatures of selection" in the human genome. Of these loci, 156 showed unusually large amounts of differentiation (suggesting diversifying selection) and 18 showed unusually small amounts of differentiation (suggesting stabilizing selection). In contrast, when Weir et al.[7] examined the high resolution Perlegen (ca. 1 million SNPs) and Phase I HapMap (ca. 0.6 million SNPs) data sets in humans to examine locus-specific estimates of $F_{ST}$ they also found large differences in $F_{ST}$ among loci, but their analyses suggested that the very high variance associated with single-locus estimates of $F_{ST}$ precluded using them to detect selection. Both sets of investigators noted a particular problem with single-locus estimates in high-resolution SNP maps: the high correlation between $F_{ST}$ estimates when loci are in strong gametic disequilibrium makes it difficult to determine whether the $F_{ST}$ at any particular SNP is markedly different from expectation.

Even though single-locus estimates of $F_{ST}$ are highly uncertain, simulation studies suggest that when loci are inherited independently, background information at a few hundred loci is sufficient to allow reliable identification of loci subject to selection when a suitable criterion for detecting "outliers" is used. [8,58,61] While few loci are falsely identified as subject to selection when they are neutral, genome scans using $F_{ST}$ may often fail to detect selection when it is present. For example, when a single allele is strongly favored in all populations not only is $F_{ST}$ expected to be nearly zero, but variation is also expected to be nearly non-existent, rendering estimates of $F_{ST}$ either highly unreliable or unobtainable. Similarly, when selection is weak, data from a very large number of loci are needed to recognize $F_{ST}$ at the locus involved as being unusual. More importantly, as mentioned above, high-resolution genome scans must account for the statistical association between closely linked loci. Guo et al.[8] illustrate the use of a CONDITIONAL AUTOREGRESSIVE SCHEME that identified 57 loci showing unusually large amounts of among-population differentiation in a sample of 3000 SNP loci on human chromosome 7 separated by only 860 nucleotides on average. Sixteen of these markers are associated with LEP, a gene encoding a leptin precursor associated with behaviors that influence the balance between food intake and energy expenditure[62]

(Figure 1). Moreover, association studies in one French population had previously suggested a relationship between one of the SNPs identified as an outlier in this study and obesity.[63]

## Forensic science and association mapping

In forensic science, matching genetic profiles from a suspect and a stain left at a crime scene serve as evidence linking the suspect to the crime. To quantify the strength of this evidence it is useful to determine the random match probability, i.e., the probability that the crime scene genetic profile matches the suspect's if the suspect was ***not*** the source of the stain. In some cases two people, the suspect and the person who left the crime sample, may belong to a subpopulation for which there is no specific allele frequency information. In such a case, we can use the "$\theta$-correction''[64] to calculate the match probability based on allele frequency information from a larger population of which this subpopulation is a part. The match probability takes into account allele frequency variation among subpopulations within the wider population for which allele frequencies are available. For example, if the matching profile consisted of a homozygote AA at a single locus, and if $p_A$ is the population frequency of allele A, then the probability that the crime profile is AA given that the suspect is AA and the suspect is not the source of the stain is[65]

$$P(AA \mid AA) = \frac{(3\theta + (1-\theta)p_A)(2\theta + (1-\theta)p_A)}{(1+\theta)(1+2\theta)}$$

There is a similar equation for heterozygotes, and these "$\theta$-correction'' results are multiplied over loci. The 1996 National Research Council report[66] recommended using $\theta = 0.01$ except for very small, isolated subpopulations for which they suggested a value of $\theta = 0.03$ was more appropriate. The practical effect of the "$\theta$-correction'' is that the numerical strength of the evidence against a suspect is reduced. If $p_A$ = 0.01, for example, the uncorrected match probability is 0.0001. With $\theta = 0.01$, on the other hand, the match probability is an order of magnitude larger – 0.0012. With $\theta = 0.03$ it is even larger – 0.0064. Thus, it is much less surprising to see a match when we take account of the population substructure than when we ignore it.

In association mapping, case-control studies compare allele frequencies at genetic markers, generally SNPs, between groups of people with a disease and those who do not have the disease. When frequencies at a marker locus differ between the groups, it is interpreted as evidence for gametic disequilibrium between the marker and a disease-related gene. This, in turn, suggests that the marker and disease-related genes are in close proximity on the same chromosome. As many authors have pointed out, however, population substructure unrelated to disease status could cause exactly the same kind of allele frequency difference.[67-70] The genomic control method is one way to account for population substructure. It uses background estimates of $F_{ST}$ to control for subpopulation differences that are unrelated to disease status.[67,68] If cases and controls have different marker allele frequencies for reasons unconnected with the disease, as would be shown by frequency differences across the whole genome, an uncorrected case-control test would give spurious indications of marker-disease association.

### Relationship to coalescent-based methods

When Kingman introduced the coalescent process to population genetics a little more than 25 years ago,[71,72] it revolutionized the field. Many approaches to analysis of molecular data, particularly molecular sequence and SNP data, now take advantage of the conceptual, computational, and analytical framework that the coalescent provides.[73-79] While *F*-statistics provide only limited insight into rates and patterns of migration, for example, statistics based on the coalescent process can provide insight into rates of mutation, migration, and other evolutionary processes. Coalescent analysis is based on maximizing the likelihood of a given sample configuration or sampling from the corresponding Bayesian posterior distribution. The likelihood is constructed from genealogical histories for the sample that are consistent with the unknown evolutionary parameters of interest, e.g., the size of the population or populations from which the sample was taken, the history of population size changes, mutation rates, recombination rates, or migration rates.[55,80-86] Coalescent analyses are likely to provide relatively precise estimates of effective population size, mutation rates, and migration rates when certain conditions are met: when the model used for analysis is consistent with the actual demographic history of populations from which samples are collected, the actual migration patterns among populations in the sample, and the mutational processes that generated allelic differences in the sample, and when it is reasonable to presume that the drift-mutation-migration process has reached an evolutionary equilibrium.[54,73] But when these assumptions are not reasonable it may not be reasonable to estimate the related evolutionary parameters, and the examples presented above show that analyses based on *F*-statistics may still provide considerable insight.

## Conclusions

Sewall Wright[5] provided a comprehensive account of processes leading to differentiation among populations nearly 80 years ago, but he did not provide the tools empirical population geneticists needed to apply his insights to understanding variation in wild populations. With work on isolation by distance in the plant *Linanthus parryae* in the 1940s,[50,87] the theory of *F*-statistics that he and Gustave Malécot later developed[3,4,16,88] began to emerge. Because of the insight *F*-statistics can provide about processes of differentiation among populations, in the last 50 years they have become the most widely used descriptive statistics in population and evolutionary genetics. From the time population geneticists first began to collect data on allozyme variation[89-94] to recent analyses of SNP variation in the human genome[2,9,95-97] *F*-statistics, and $F_{ST}$ in particular, have been used to investigate processes influencing the distribution of genetic variation within and among populations. Unfortunately, neither Wright nor Malécot distinguished carefully between the *definition* of *F*-statistics and the *estimation* of *F*-statistics. In particular, until Cockerham introduced his indicator formalism[10,22] few, if any, population geneticists understood that estimators of *F*-statistics must take into account both statistical sampling and genetic sampling.

The statistical methodology for estimating *F*-statistics is now well established. With the availability of methods to estimate locus- and population-specific effects on $F_{ST}$,[7,8,27,58,61,98] geneticists now have a set of tools to identify genomic regions or populations with unusual evolutionary histories. Through further extensions of the approach, it is even possible to determine the relationship between the recent evolutionary history of populations and environmental or demographic variables.[99] The fundamentals of

how population size, mutation rates, and migration are related to the genetic structure of populations have been well understood for nearly 80 years. Analyses of *F*-statistics in populations of plants, animals, and microorganisms have broadened and deepened that understanding, but those analyses have mostly been applied to data sets with a relatively small number of loci. The age of population genomics is now upon us.[100,101] The 1000 genomes project (http://www.1000genomes.org) and the HapMap project (http://www.hapmap.org) give a hint of what is to come. In spite of the scale of these projects, much of the data can be understood fundamentally as allelic variation at individual loci. As a result, we expect *F*-statistics to be at least as useful in understanding these massive datasets as they have been in population and evolutionary genetics for most of the last century.

## Box 1

### Mathematical notation

In this box we summarize provide definitions for the mathematical symbols used throughout the paper.

### Among-population allele frequency distribution:

| | |
|---|---|
| $\pi$ | Mean allele frequency |
| $\sigma_\pi^2$ | Variance in allele frequency |

### *F*-statistics

Wright's *F*-statistics and Cockerham's $\theta$ statistics

| Parameter | Definition |
|---|---|
| $F_{IS}$ | Correlation of alleles within an individual relative to the subpopulation in which it occurs; equivalently the average departure of genotype frequencies from Hardy-Weinberg expectations within populations |
| $F_{ST}$ | Correlation of randomly chosen alleles within the same subpopulation relative to the entire population; equivalently the proportion of genetic diversity due to allele frequency differences among populations |
| $F_{IT}$ | Correlation of alleles within an individual relative to the entire population; equivalently the departure of genotype frequencies from Hardy-Weinberg expectations relative to the entire population |
| $f$ | Co-ancestry for alleles within an individual relative to the subpopulation in which it occurs; equivalent to $F_{IS}$ |
| $\theta$ | Co-ancestry for randomly chosen alleles within the same subpopulation relative to the entire population; equivalent to $F_{ST}$ |
| $F$ | Co-ancestry for alleles within an individual relative to the entire population; equivalent to $F_{IT}$ |

### $\Phi$-statistics and $R_{ST}$

$\Phi_{ST}$ from Analysis of Molecular Variance (AMOVA) is used for haplotype data (nucleotide sequence data, mapped restriction site data) and requires a measure of evolutionary distance among all pairs of

haploytpes. $R_{ST}$ for microsatellite data and requires that alleles be labeled according to the number of repeat units they contain.

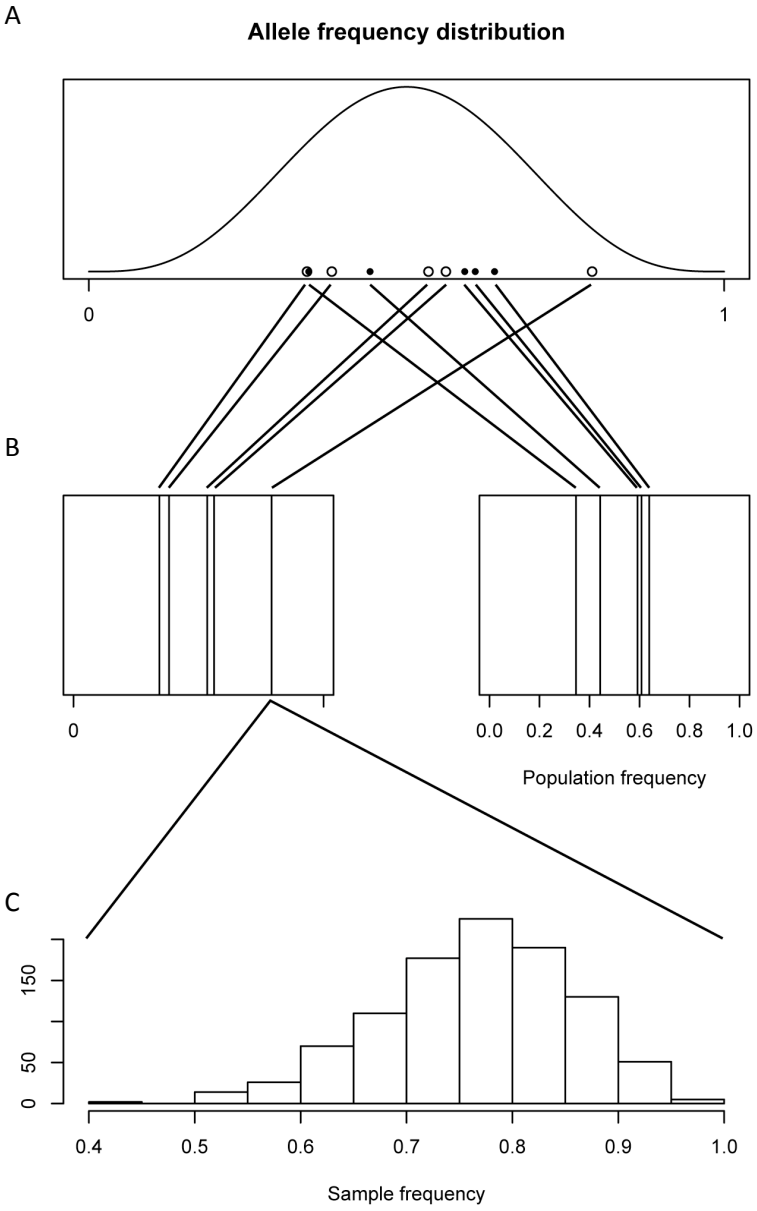
| Parameter | Definitions |
|---|---|
| $\Phi_{IS}$ | Excess similarity of alleles within an individual relative to the subpopulation in which it occurs; analogous to $F_{IS}$ |
| $\Phi_{ST}$ | Excess similarity among randomly chosen alleles within the same subpopulation relative to the entire population, or equivalently the proportion of genetic diversity (measured as the expected squared evolutionary distance between alleles) attributable to differences among populations; analogous to $F_{ST}$ |
| $\Phi_{IT}$ | Excess similarity of alleles within an individual relative to the entire population; analogous to $F_{IT}$ |
| $R_{ST}$ | Excess similarity among randomly chosen alleles within the same subpopulation relative to the entire population, or equivalently the proportion of genetic diversity (measured as the expected squared difference in repeat numbers between alleles) attributable to differences among populations; analogous to $F_{ST}$ |

### Measuring genetic differentiation among populations in quantitative traits:

| Parameter | Definition |
|---|---|
| $\sigma^2_{GI}$ | Additive genetic variance within populations |
| $\sigma^2_{GP}$ | Additive genetic variance among populations |
| $Q_{ST}$ | Proportion of additive genetic variation in entire population due to differences among populations; analogous to $F_{ST}$ |

## Box 2

### Genetic sampling *versus* statistical sampling




Genetic drift leads to differences among populations that are described by the distribution of allele frequencies among populations. The variance of this distribution is directly related to $F_{ST}$ (see equation 2), but in a typical study only a subset of populations are sampled. Thus, in addition to accounting for variation associated with sampling from populations, estimates of *F*-statistics must also account for variation associated with sampling sets of populations from the allele frequency distribution.

### Genetic (or evolutionary) sampling

Panel A shows the distribution of allele frequencies among populations corresponding to a mean allele frequency of $\pi = 0.5$ and $F_{ST} = \theta = 0.1$. If two sets of populations (represented by filled and open circles)

are sampled from this distribution, allele frequencies in the first set of populations (open circles) will differ from those in the second set (closed circles). Panel B shows an example in which two different sets of 5 population frequencies are drawn randomly from the distribution of allele frequencies illustrated in Panel A.

The variation in allele frequencies illustrated in panel A reflects the effect of *genetic* or *evolutionary sampling*. The differences between the sets of samples in panel B reflect the effect of sampling particular populations from the distribution of allele frequencies in panel A and are analogous to those that would be expected in an empirical study if it were repeated on a different set of populations.

### Statistical sampling

Panel C illustrates the more familiar idea of *statistical sampling*. It shows the distribution of *sample* allele frequencies obtained in 1000 samples of size 20 from the population with the largest allele frequency in the population sample on the left in Panel B. *Statistical sampling* refers to the variation in *sample* composition expected under repeated sampling of alleles from a population with a particular allele frequency.

Investigators can control the amount of variation associated with statistical sampling by increasing the number of individuals sampled within populations: the larger the number of individuals sampled, the less sample allele frequencies will differ from the underlying population frequencies. In contrast, investigators cannot control the amount of variation associated with genetic sampling: the variation associated with genetic sampling is an intrinsic property of the underlying stochastic evolutionary process contributing to differentiation among populations.

### The relationship between $F_{ST}$ and $G_{ST}$

Nei[33] introduced the statistic $G_{ST}$ as a measure of genetic differentiation among populations. It is defined in terms of the population frequencies in panel B, not the allele frequency distribution in panel A. In contrast, estimates of $F_{ST}$ account for genetic sampling, and they are intended to reflect properties of the allele frequency distribution in A. As a result, $F_{ST}$ and $G_{ST}$ measure different things. Thus, $G_{ST}$ will be an appropriate measure only when interest focuses on characteristics of particular samples illustrated in panel B. In a typical population study, $\theta$ will be a more appropriate measure of differentiation.

It might seem that similar arguments would apply to exact tests of population differentiation.[102] After all, they use permutations of *sample* configurations to determine whether populations are differentiated from one another. Nonetheless, the permutation test is equivalent to determining whether the allele frequency distribution in A has a variance greater than zero so that exact tests implicitly consider both statistical and genetic sampling effects.

## Box 3

### Comparing methods for estimating $F_{ST}$

To illustrate the differences the calculations involved in method of moment, maximum likelihood, and Bayesian estimates of *F*-statistics, we use data from a classic study on human populations investigating

allele frequency differences at blood group loci. We use a subset of the data originally reported by Workman and Niswander.[103] Their data consists of genotype counts at several loci in native American Papago, and the data were collected from 10 political districts in southwestern Arizona. We report estimates of $F_{IS}$, $F_{ST}$ and $F_{IT}$ derived from the *MN* blood group locus that suggest little departure of genotype frequencies from Hardy-Weinberg expectations within each district and little genetic differentiation among the districts.

## Methods of moment analysis

Analysis of variance on the indicator variable $y_{ij,k}$, where $y_{ij,k}$=1 if allele *i* in individual *j* of population *k* is *M*, produces the table illustrated here gives moment estimates for the variance components of $\sigma_G^2 = 0.160$, $\sigma_I^2 = 0.00511$, and $\sigma_P^2 = 0.0.000667$. Following Cockerham[10]

$$F = \frac{\sigma_P^2 + \sigma_I^2}{\sigma_P^2 + \sigma_I^2 + \sigma_G^2}$$

$$\theta = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_I^2 + \sigma_G^2}$$

$$f = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_G^2}$$

Thus, the moment estimates are $\hat{F} = 0.0348, \hat{\theta} = 0.00402$, and $\hat{f} = 0.0309$. As expected for human populations, there is little evidence that genotype proportions within each political district differ from Hardy-Weinberg expectations ( $\hat{f} \approx 0$). Similarly, there is little evidence of genetic differentiation among political districts ( $\hat{\theta} \approx 0$).

## Bayesian and likelihood analysis

In contrast, current implementations of a Bayesian approach to analyzing these data typically assume independent uniform [0,1] prior distributions for both *f* and $\theta$. The posterior mean of *f* and $\theta$ for these data are 0.050 and 0.019, respectively. The posterior distribution of *f* has a mode near 0, but posterior distribution is relatively broad (95% credible interval 0.0033-0.123), causing the posterior mean of *f* to be larger than the method of moments estimate. Similarly, allele frequency estimates within each population are uncertain, and the estimate of $\theta$ takes this uncertainty into account, suggesting that there is slightly more among-population differentiation than detected with moment estimates. For comparison, the maximum-likelihood estimates are $\hat{F} = 0.0408, \hat{\theta} = 0.00640$, and $\hat{f} = 0.0346$ (obtained by estimating variance components in Gaussian mixed model applied to the indicator variables and using Cockerham's definitions of *F*, *f*, and $\theta$ in terms of the variance components).

| Parameter | Method of moments | Maximum likelihood | Bayesian |
|---|---|---|---|
| $f$ | 0.0309 | 0.0346 | 0.0503 |
| $\theta$ | 0.00402 | 0.00640 | 0.0189 |

| | | | |
|---|---|---|---|
| *F* | 0.0348 | 0.0408 | 0.0683 |

**Table 1. Comparison of point estimates for *F*-statistics derived from the Workman and Niswander data.**

To extend the method of moments approach to multiple alleles and multiple loci, calculations are done separately for every allele at every locus and the sums of squares are combined.[17,27] To extend the likelihood or Bayesian approaches, we make the assumption that $f$ and $\theta$ have the same value at every locus and that genotype counts are sampled independently across loci and populations.[104,105]

## Box 4

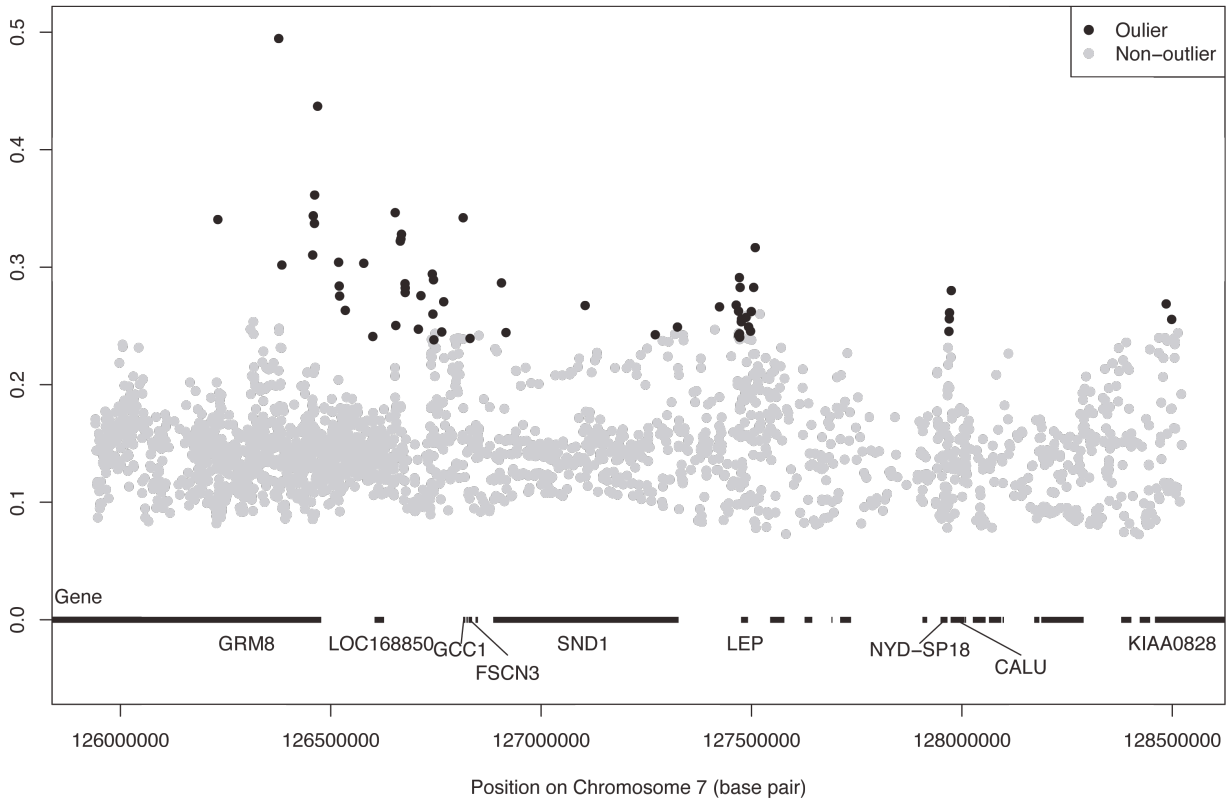
### Why focus on $F_{ST}$?

We focus our discussion on $F_{ST}$ for several reasons. First, $F_{IS}$ is easier to interpret. It is defined with respect to the populations that are included in the sample, either through population-specific estimates or through the average of those estimates. In contrast, $F_{ST}$ is defined and interpreted with respect to the distribution of allele frequencies among *all* populations that could have been sampled, not merely those that happen to have been included in the sample. As a result, estimates of $F_{ST}$ have to account for genetic sampling, introducing a level of complexity and subtlety that requires extra attention.

Second, the application of *F*-statistics to problems in population and evolutionary genetics often centers on estimates of $F_{ST}$. Whether attempting to interpret aspects of demographic history like sex-biased dispersal out of Africa in human populations,[9] to detect regions of the genome that may have been subject to stabilizing or diversifying selection,[8,58,61] or correcting match probabilities in a forensic application for genetic substructure within populations,[106] estimates of $F_{ST}$ often play a crucial role in interpreting genetic data. Estimates of $F_{IS}$ reveal important properties of the mating system within populations, but estimates of $F_{ST}$ reveal properties of the evolutionary process leading to divergence among populations.

Finally, in many populations of animals, and in human populations in particular, within-population departures from Hardy-Weinberg proportions are small. Where present, such departures may reveal more about genetic substructuring within populations than about departures from random mating. Moreover, while estimates of $F_{IS}$ may reveal something about patterns of mating in inbred populations of plants or animals, direct analysis of mother-offspring genotype combinations will usually be more informative and reliable.[107,108]

Figure 1




### Locus-specific estimates of $F_{ST}$ on human chromosome 7

Locus-specific estimates of $F_{ST}$ on human chromosome 7 as inferred from the phase II HapMap data set.[95] Bars indicate the location of known genes. Dark black circles are posterior means for SNPs with estimates detectably different from the genomic background (gray circles). All "outliers" show significantly more differentiation among the four populations in the sample than is consistent with the level of differentiation seen in the genomic background. The excess differentiation suggests that these SNPs are associated with genomic regions in which loci have been subject to diversifying selection among populations. From Guo et al.[8]

## Glossary

ADDITIVE GENETIC VARIANCE: The part of the total genetic variation that is due to the main (or additive) effects of alleles on a phenotype. The additive variance determines the degree of resemblance between relatives and therefore the response to selection.

COALESCENT-BASED APPROACHES: Coalescent-based approaches use statistical properties of the genealogical relationship among alleles under particular demographic and mutational models to make inferences about the effective size of populations and about rates of mutation and migration.

CONDITIONAL AUTOREGRESSIVE SCHEME: A statistical approach developed for analysis of data in which a random effect is associated with the spatial location of each observation and the magnitude of the

random effect is determined by a weighted average of random effects of nearby positions. In most applications, weights are inversely related to the spatial distance between two sample points.

DIVERSIFYING SELECTION: Selection in which different alleles are favored in different populations. It is often a consequence of LOCAL ADAPTATION.

EFFECTIVE POPULATION SIZE ($N_e$): Formulated by Wright in 1931, $N_e$ reflects the size of an idealized population that would experience drift in the same way as the actual (census) population. $N_e$ can be lower than census population size due to various factors, including a history of population bottlenecks and reduced recombination.

GENETIC DRIFT: The random fluctuations in allele frequencies over time that are due to chance alone.

HARDY-WEINBERG PROPORTIONS: A state in which the frequency of each diploid genotype at a locus equals that expected from the random union of alleles. That is: genotypes AA, Aa and aa will be at frequencies $p^2$, $2pq$, and $q^2$.

HETEROZYGOTE ADVANTAGE: A pattern of natural selection in which heterozygotes are more likely to survive than homozygotes.

LIKELIHOOD: A mathematical function that describes the relationship between the unknown parameters of a statistical distribution, e.g., the mean and variance of the allele frequency distribution among populations or the allele frequency in a particular population, and the data. It is directly proportional to the probability of the data given the unknown parameters.

LOCAL ADAPTATION: The situation in which genotypes from different populations have higher fitness in their home environments owing to historical natural selection.

MCMC METHODS: Monte Carlo Markov Chain (MCMC) methods are a computational technique widely used for approximating complex integrals and other functions. In this context MCMC methods are used to approximate the posterior distribution of a Bayesian model.

MULTINOMIAL DISTRIBUTION: A statistical distribution that describes the probability of obtaining a sample with a specified number of objects in each of several categories. The probability is determined by the total sample size and the probability of drawing an object from each category. The binomial distribution is a special case of the multinomial distribution in which there are two categories.

PRIOR DISTRIBUTION: A statistical distribution used in Bayesian analysis to describe the probability that parameters take on a particular value prior to examining any data. It expresses the level of uncertainty about those parameters before the data has been analysed.

POSTERIOR DISTRIBUTION: A statistical distribution used in Bayesian analysis to describe the probability that parameters take on a particular value after the data have been analysed. It reflects both the likelihood of the data given particular parameters and the prior probability that parameters take on particular values.

SHORT TANDEM REPEAT LOCI: Loci consisting of short (2-6 nucleotide) sequences that are repeated multiple times. Alleles at STR loci differ from one another in the number of repeats.

STABILIZING SELECTION: Selection in which either the same allele or the same genotype is favored in different populations.

VARIANCE: A measure of the amount of variation around a mean value.

## Acknowledgment

We are indebted to Rachel Prunier, Kathryn Theiss, and three anonymous reviewers for helpful comments on earlier versions of this paper. The work in the laboratories of the authors was supported in part by grants from the U.S. National Institutes of Health (1 R01 GM 068449-01A1 to K.E.H; 1 R01 GM 075091 to B.S.W).

## References

1. Rosenberg, N.A. et al. Genetic structure of human populations. *Science* **298**, 2381-2385 (2002).
2. Li, J.Z. et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319**, 1100-1104 (2008).
3. Wright, S. The genetical structure of populations. *Annals of Eugenics* **15**, 323-354 (1951).
4. Malécot, G. *Les mathématiques de l'hérédié*, (Masson, Paris, France, 1948).
5. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97-159 (1931).
6. Akey, J.M., Zhang, G., Khang, K., Jin, L. & Shriver, M.D. Interrogating a high-density SNP map for signatures of natural selection *Genome Research* **12**, 1805-1814 (2002).
7. Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M. & Hill, W.G. Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15**, 1468-76 (2005).
8. Guo, F., Dey, D.K. & Holsinger, K.E. A Bayesian hierarchical model for analysis of SNP diversity in multilocus, multipopulation models. *Journal of the American Statistical Association* **164**, 142-154 (2009).
9. Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature Genetics* **41**, 66-70 (2009).
10. Cockerham, C.C. Variance of gene frequencies. *Evolution* **23**, 72-84 (1969).
11. Wahlund, S. Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11**, 65-106 (1928).
12. Sokal, R.R., Oden, N.L. & Thomson, B.A. A simulation study of microevolutionary inferences by spatial autocorrelation analysis. *Biological journal of the linnean society.* **60**, 73 (1997).
13. Sokal, R.R. & Oden, N.L. Spatial autocorrelation analysis as an inferential tool in population genetics. *American Naturalist* **138**, 518-521 (1991).
14. Epperson, B.K. *Geographical Genetics*, (Princeton University Press, Princeton, NJ, 2003).
15. Weir, B.S. & Cockerham, C.C. Mixed self- and random-mating at two loci. *Genetical Research* **21**, 247-262 (1973).
16. Wright, S. *Evolution and the Genetics of Populations. Vol. 4. Variability within and among Natural Populations*, (University of Chicago Press, Chicago, IL, 1978).
17. Weir, B.S. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*, (Sinauer Associates, Sunderland, MA, 1996).
18. Rousset, F. Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**, 371-380 (2002).

19. Casella, G. & Berger, R.L. *Statistical Inference*, (Duxbury, Pacific Grove, CA, 2002).
20. Weir, B.S. & Cockerham, C.C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370 (1984).
21. Excoffier, L. Analysis of population subdivision. in *Handbook of Statistical Genetics* (eds. Balding, D.J., Bishop, M. & Cannings, V.) 271-307 (John Wiley & Sons, Ltd., Chichester, 2001).
22. Cockerham, C.C. Analyses of gene frequencies. *Genetics* **74**, 679-700 (1973).
23. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*, (Springer Verlag, New York, 1985).
24. Robert, C.P. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, (Springer-Verlag, New York, NY, 2001).
25. Lee, P.M. *Bayesian Statistics: An Introduction*, (Edward Arnold, London, 1989).
26. Gelfand, A.E. & Smith, A.F.M. Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409 (1990).
27. Weir, B.S. & Hill, W.G. Estimating F-statistics. *Annual Review of Genetics* **36**, 721-750 (2002).
28. Wehrhahn, C. Proceedings of the ecological genetics workshop. *Genome* **31**, 1098-1099 (1989).
29. Samanta, S., Li, Y.J. & Weir, B.S. Drawing inferences about the coancestry coefficient. *Theoretical Population Biology* **75**, 312-319 (2009).
30. Gaggiotti, O.E. et al. Patterns of colonization in a metapopulation of grey seals. *Nature* **13**, 424-427 (2002).
31. Levsen, N.D., Crawford, D.J., Archibald, J.K., Santos-Geurra, A. & Mort, M.E. Nei's to Bayes': comparing computational methods and genetic markers to estimate patterns of genetic variation in Tolpis (Asteraceae). *Am. J. Bot.* **95**, 1466-1474 (2008).
32. Nei, M. & Chesser, R.K. Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**, 253-259 (1983).
33. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Nat. Acad. Sci. USA* **70**, 3321-3321 (1973).
34. Excoffier, L., Smouse, P.E. & Quattro, J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491 (1992).
35. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-462 (1995).
36. Rousset, F. Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* **142**, 1357-1362 (1996).
37. Slatkin, M. Inbreeding coefficients and coalescence times. *Genetical Research* **58**, 167-175 (1991).
38. Holsinger, K.E. & Mason-Gamer, R.J. Hierarchical analysis of nucleotide diversity in geographically structured populations. *Genetics* **142**, 629-639 (1996).
39. Balloux, F. & Lugon-Molin, N. The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**, 155-165 (2002).
40. Balloux, F., Brunner, F. & Goudet, J. Microsatellites can be misleading: an empirical and simulation study. *Evolution* **54**, 1414-1422 (2000).
41. Gaggiotti, O.E., Lange, O., Rassman, K. & Gliddon, C. A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. *Molecular Ecology* **8**, 1513-1520 (1999).
42. Spitze, K. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. *Genetics* **135**, 467-374 (1993).
43. Lande, R. Neutral theory of quantitative genetic variance in an island model with local extinction and colonization. *Evolution* **46**(1992).

44. McKay, J.K. & Latta, R.G. Adaptive population divergence: markers, QTL and traits. *Trends in Ecology & Evolution* **17**, 285 (2002).
45. O'Hara, R.B. & Merila, J. Bias and Precision in QST Estimates: Problems and Some Solutions. *Genetics* **171**, 1331-1339 (2005).
46. Goudet, J. & Martin, G. Under Neutrality, QST <= FST When There Is Dominance in an Island Model. *Genetics* **176**, 1371-1374 (2007).
47. Notohara, M. The coalescent and the genealogical process in geographically structured population. *J Math Biol* **29**, 59-75 (1990).
48. Charlesworth, B. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195-205 (2009).
49. McCauley, D.E. & Whitlock, M.C. Indirect measures of gene flow and migration Fst (does not equal) 1/(4Nm+1). *Heredity* **82**, 117-125 (1999).
50. Wright, S. Isolation by distance. *Genetics* **28**, 114-138 (1943).
51. Rousset, F. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219-28 (1997).
52. Felsenstein, J. How can we infer geography and history from gene frequencies? *Journal of Theoretical Biology* **96**, 9-20 (1982).
53. Cann, H.M. et al. A Human Genome Diversity Cell Line Panel. *Science* **296**, 261b-262 (2002).
54. Beerli, P. Comparison of Bayesian and maximum-likelihood estimation of population genetic parameters. *Bioinformatics* **22**, 341-345 (2006).
55. Kuhner, M.K. Coalescent genealogy samplers: windows into population history. *Trends Ecol Evol* **24**, 86-93 (2009).
56. Kuhner, M.K. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**, 768-770 (2006).
57. Fu, R., Gelfand, A. & Holsinger, K.E. Exact moment calculations for genetic models with migration, mutation, and drift. *Theoretical Population Biology* **63**, 231-243 (2003).
58. Beaumont, M.A. & Balding, D.J. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980 (2004).
59. Vitalis, R., Dawson, K. & Boursot, P. Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811-1823 (2001).
60. Beaumont, M.A. & Nichols, R.A. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Ser* **263**, 1619 (1996).
61. Foll, M. & Gaggiotti, O. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* **180**, 977-993 (2008).
62. Zhang, Y. et al. Positional cloning of the mouse *obese* gene and its human homologue. *Nature* **372**, 425-432 (1994).
63. Mammès, O. et al. Association of the G-2548A polymorphism in the 5' region of the LEP gene with overweight. *Annals of Human Genetics* **64**, 391-394 (2000).
64. Balding, D.J. & Donnelly, P. How convincing is DNA evidence? *Nature* **368**, 285-6 (1994).
65. Balding, D.J. & Nichols, R.A. DNA match probability calculation: how to allow for population stratification, relatedness, database selection, and single bands. *Forensic Science International* **64**, 125-140 (1994).
66. Council, N.R. *The evaluation of forensic DNA evidence*, (National Academy Press, Washington, DC, 1996).
67. Devlin, B., Roeder, K. & Wasserman, L. Genomic Control, a New Approach to Genetic-Based Association Studies. *Theoretical Population Biology* **60**, 155-166 (2001).
68. Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997-1004 (1999).

69. Pritchard, J.K. & Donnelly, P. Case-control studies of association in structured or admixed populations. *Theor Popul Biol* **60**, 227-37 (2001).
70. Pritchard, J.K. & Rosenberg, N.A. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**, 220-8 (1999).
71. Kingman, J.F.C. On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27-43 (1982).
72. Kingman, J.F.C. The coalescent. *Stoch. Proc. Appl.* **13**, 235-248 (1982).
73. Kuhner, M.K. & Smith, L.P. Comparing Likelihood and Bayesian Coalescent Estimation of Population Parameters. *Genetics* **175**, 155-165 (2007).
74. Wang, J. A Coalescent-Based Estimator of Admixture From DNA Sequences. Vol. 173 1679-1692 (2006).
75. Innan, H., Zhang, K., Marjoram, P., Tavare, S. & Rosenberg, N.A. Statistical Tests of the Coalescent Model Based on the Haplotype Frequency Distribution and the Number of Segregating Sites. *Genetics* **169**, 1763-1777 (2005).
76. Wall, J.D. & Hudson, R.R. Coalescent Simulations and Statistical Tests of Neutrality. *Molecular Biology and Evolution* **18**, 1134-1135 (2001).
77. Nordborg, M. Structured coalescent processes on different time scales. *Genetics* **146**, 1501-14 (1997).
78. Donnelly, P. & Tavaré, S. Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401-421 (1995).
79. Griffiths, R.C. & Tavare, S. Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131-159 (1994).
80. Fearnhead, P. & Donnelly, P. Estimating Recombination Rates From Population Genetic Data. *Genetics* **159**, 1299-1318 (2001).
81. Kuhner, M.K., Beerli, P., Yamato, J. & Felsenstein, J. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**, 439-47 (2000).
82. Kuhner, M.K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393 (2000).
83. Kuhner, M.K. & Felsenstein, J. Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genet Epidemiol* **19 Suppl 1**, S15-21 (2000).
84. Kuhner, M.K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics.* **149**, 429 (1998).
85. Beerli, P. & Felsenstein, J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763-773 (1999).
86. Drummond, A.J., Nicholls, G.K., Rodrigo, A.G. & Solomon, W. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics* **161**, 1307-1320 (2002).
87. Wright, S. An analysis of local variability of flower color in Linanthus parryae. *Genetics* **28**, 139-156 (1943).
88. Malécot, G. *The Mathematics of Heredity*, (W. H. Freeman, San Francisco, 1969).
89. Hamrick, J.L. & Godt, M.J.W. Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London, Series B* **351**, 1291-1298 (1996).
90. Hamrick, J.L. Isosymes and the analysis of genetic structure in plant populations. in *Isozymes in Plant Biology* (eds. Soltis, D.E. & Soltis, P.S.) 87-105 (Dioscorides Press, Portland, OR, 1989).
91. Loveless, M.D. & Hamrick, J.L. Ecological determinants of genetic structure in plant populations. *Annual Review of Ecology & Systematics* **15**, 65-95 (1984).

92. Hamrick, J.L., Linhart, Y.B. & Mitton, J.B. Relationships between life history characteristics and electrophoretically deterctable genetic variation in plants. *Annual Review of Ecology & Systematics* **10**, 173-200 (1979).
93. Gottlieb, L.D. Electrophoretic evidence and plant populations. in *Progress in Phytochemistry*, Vol. 7 (eds. Reinhold, L., Harborne, J.B. & Swain, T.) 1-46 (Pergamon Press, Oxford, 1981).
94. Brown, A.H.D. Enzyme polymorphism in plant populations. *Theor. Popul. Biol.* **15**, 1-42 (1979).
95. Consortium, T.I.H. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
96. Consortium, T.I.H. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).
97. He, M. et al. Geographical Affinities of the HapMap Samples. *PLoS ONE* **4**, e4684 (2009).
98. Balding, D.J. Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology* **63**, 221-230 (2003).
99. Foll, M. & Gaggiotti, O. Identifying the Environmental Factors That Determine the Genetic Structure of Populations. *Genetics* **174**, 875-891 (2006).
100. Begun, D.J. et al. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biology* **5**, e310 (2007).
101. Luikart, G., England, P.R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994 (2003).
102. Goudet, J., Raymond, M., de Meeus, T. & Rousset, F. Testing differentiation in diploid populations. *Genetics* **144**, 1933-1940 (1996).
103. Workman, P.L. & Niswander, J.D. Population studies on southwest indian tribes. II. Local genetic differentiation in the Papago. *American Journal of Human Genetics* **22**, 24-49 (1970).
104. Holsinger, K.E. Bayesian hierarchical models in geographical genetics. in *Hierarchical Modeling for the Environmental Sciences* (eds. Clark, J.S. & Gelfand, A.E.) 25-37 (Oxford University Press, Oxford, 2006).
105. Holsinger, K.E. Analysis of genetic diversity in hierarchically structured populations: a Bayesian perspective. *Hereditas* **130**, 245-255 (1999).
106. Weir, B.S. The rarity of DNA profiles. *Annals of Applied Statistics* **1**, 358-370 (2007).
107. Ritland, K.R. Joint maximum-likelihood estimation of genetic and mating system structure using open-pollinated progenies. *Biometrics*, 33-43 (1988).
108. Thompson, S.L. & Ritland, K. A novel mating system analysis for modes of self-oriented mating applied to diploid and polyploid arctic Easter daisies (*Townsendia hookeri*). *Heredity* **97**, 119-126 (2006).

## Autobiography

Kent Holsinger received his Ph.D. training with Marcus W. Feldman at Stanford University. He was a post-doctoral Fellow in the Miller Institute for Basic Research in Science at the University of California, Berkeley, and he did additional post-doctoral work both with Leslie D. Gottlieb at the University of California, Davis and with Marc Feldman before accepting a faculty position at the University of Connecticut. His research has focused on the evolution of plant mating systems, the conservation biology of rare and endangered species (especially plants), and the development of statistical tools for analysis of genetic diversity in wild populations. More recently, he has become interested in understanding mechanisms underlying evolutionary radiations in the genus *Protea* in southwestern

South Africa (see http://darwin.eeb.uconn.edu/wiki/index.php/Evolutionary_radiations_in_South_African_Proteaceae).

Bruce Weir received his Ph.D. training with Clark Cockerham at North Carolina State University and post-doctoral training with Bob Allard at the University of California, Davis. After a brief time in his native New Zealand he returned to North Carolina State University, where he was a faculty member for 30 years. He is now Professor and Chair of Biostatistics at the University of Washington. His research interests are in statistical genetics with applications to forensic science and, more recently, to association mapping.

## Highlighted references

Wright *Ann. Eugen.* (1951): Develops the explicit framework for analysis and interpretation of *F*-statistics in an evolutionary context.

Malécot (1948): Develops a framework for analysis of genetic diversity in hierarchically structured populations equivalent to Wright's *F*-statistics

Wright *Genetics* (1931): A landmark paper in population genetics in which the impact of population size, mutation, and migration on the abundance and distribution of genetic variation in populations are first quantitatively described.

Cockerham *Evolution* (1969): Develops the first approach for analysis of *F*-statistics recognizing the impact of genetic sampling on estimates of *F*-statistics from population data.

Weir & Cockerham *Evolution* (1984): Develops the ANOVA framework to apply Cockerham's approach to *F*-statistics and provides method of moments estimates for *F*-statistics

Nei *PNAS* (1973): Introduces $G_{ST}$ as a measure of genetic differentiation among populations

Excoffier *Genetics* (1992): Introduces $\Phi_{ST}$ and AMOVA for analysis of haplotype data

Slatkin *Genetics* (1994): Introduces $R_{ST}$ for analysis of microsatellite data

Spitze *Genetics* (1993): Introduces $Q_{ST}$ for analysis of continuously varying trait data

## Online summary

- $F_{IS}$ measures the departure of genotype frequencies within populations from Hardy-Weinberg expectations. Although often referred to as the "within-population inbreeding coefficient", this phrase is misleading. $F_{IS}$ will be negative if there is heterozygote advantage or if individuals avoid inbreeding.

- $F_{ST}$ is a property of the distribution of allele frequencies among populations. It reflects the joint effects of drift, migration, mutation, and selection on the distribution of genetic variation among populations.

- $F_{ST}$ can be used to describe the distribution of genetic variation among any set of samples, but it is most usefully applied when the samples represent relatively discrete units rather than arbitrary divisions along a continuous distribution.

- Statistics related to $F_{ST}$ may be useful for haplotype or microsatellite data if an appropriate measure of evolutionary distance among alleles is available.

- Comparing an estimate of $F_{ST}$ from marker data with an estimate of $Q_{ST}$ from continuously varying trait data might be used to detect selection, but the estimate of $F_{ST}$ may depend on the choice of marker and the estimate of $Q_{ST}$ may differ from neutral expectations if there is a non-additive component of genetic variance.

- Although the simple relationship between $F_{ST}$ and migration rates in Wright's island model makes it tempting to infer migration rates from $F_{ST}$, considerable caution is needed if such an approach is to be used.

- If estimates of $F_{ST}$ from a large number of loci are available, it may be possible to identify certain loci as "outliers" that may have been subject to different patterns of selection or to different demographic processes.

- Case-control studies for association mapping studies must account for the possibility that population substructure accounts for an observed association between a marker and a disease. The genomic control method uses background estimates of $F_{ST}$ to control for such substructure.

- In forensic applications, match probabilities are sometimes calculated for subpopulations lacking specific allele frequency data. A $\theta$-correction, in which $\theta$ is $F_{ST}$, is used to calculate the match probability using allele frequency information from a broader population of which the subpopulation is part.

## Online links

### Software

**ABC4F**: Approximate Bayesian computation for *F*-statistics (http://www-leca.ujf-grenoble.fr/logiciels.htm)
**Arlequin**: Weir & Cockerham *F*-statistics (and many other things; http://cmpg.unibe.ch/software/arlequin3/)
**BayeScan**: Bayesian genome scan for outliers (http://www-leca.ujf-grenoble.fr/logiciels.htm)
**GDA**: Weir & Cockerham *F*-statistics (http://www.eeb.uconn.edu/people/plewis/software.php)
**Genepop**: Weir & Cockerham *F*-statistics (http://kimura.univ-montp2.fr/~rousset/Genepop.htm)
**GESTE**: Bayesian analysis of factors that affect population structure (http://www-leca.ujf-grenoble.fr/logiciels.htm)
**Hickory**: Bayesian *F*-statistics (http://darwin.eeb.uconn.edu/hickory/hickory.html)
**hierfstat**: Weir & Cockerham *F*-statistics for any number of levels in a hierarchy (http://www2.unil.ch/popgen/softwares/hierfstat.htm)

## Course notes

The Wahlund effect and Wright's *F*-statistics (http://darwin.eeb.uconn.edu/eeb348/lecture.php?rl_id=445)

The genetic structure of populations (http://darwin.eeb.uconn.edu/eeb348/lecture.php?rl_id=402)

The genetic structure of populations: a Bayesian approach (http://darwin.eeb.uconn.edu/eeb348/lecture.php?rl_id=403)

Bayesian population genetic data analysis (http://darwin.eeb.uconn.edu/summer-institute/summer-institute.html)