Exegeses on Maximum Genetic Differentiation

François Rousset*,†

*Centre National de la Recherche Scientifique, Institut des Sciences de l'Évolution, Université Montpellier 2, 34095 Montpellier, France, and †Institut de Biologie Computationnelle, 34095 Montpellier, France

ABSTRACT A canon of population genetics concerns the properties of F_{ST}, a descriptor of spatial genetic structure. Interest for F_{ST} arose from Wright's early insights linking F_{ST} to dispersal parameters as well as to his concept of effective population size (*e.g.*, Wright 1938, 1951). Although there is continued interest in this topic, F_{ST} also serves in other applications, such as detecting selected markers in natural populations (Beaumont and Nichols 1996) and more often in routine descriptive works. Remarkably, it is the latter use that seems to attract most discussion. Alternative descriptors have been proposed. Conversely, attempts have been made to draw biological inferences from F_{ST} properties that do not depend on biological processes. A reconsideration of its properties under biological scenarios underlines the weaknesses of such approaches.

In this commentary, François Rousset examines the topic of F_{ST}, a descriptor of spatial genetic structure. The commentary explores the issue of alternative descriptors, which are rejected in lieu of a new interpretation of F_{ST} in Jakobsson *et al.*, "The Relationship Between F_{ST} and the Frequency of the Most Frequent Allele ", published in the February issue of GENETICS.

DISTINCT neutral loci in a genome share a common distribution of genealogical relationships. However, in general these relationships cannot be directly inferred from genetic data, which additionally depend on marker features such as mutation rates and homoplasy (*i.e.*, recurrent mutations to the same allelic type). Different conceptions of this discrepancy have led to sometimes conflicting definitions of F_{ST} and related concepts, such as identity by descent. A definition that is directly applicable to genetic data and that reflects the dependence on marker features can be given in terms of probability of identity in allelic state, Q, or equivalently in term of gene diversity, $H = 1 - Q$. Namely, F_{ST} can be defined as $(Q_w - Q_b)/(1 - Q_b)$, in terms of the probabilities of identity within subpopulations (Q_w) and between subpopulations (Q_b), or equivalently as $(H_b - H_w)/H_b$. According to this definition, F_{ST} values cannot be higher than the within-deme identity, Q_w, which itself depends on mutational features intrinsic to the genetic markers. Interest in F_{ST} stems in part from its relative robustness with respect to these confounding marker-specific factors (*e.g.*, Crow and Aoki 1984), but this robustness is far from perfect.

F_{ST} can be redefined so that there is no such dependence, as $C_{ST} = (T_w - T_b)/T_b$ in terms of the average coalescence times of genes within subpopulations (T_w) and between subpopulations (T_b). This definition is appropriate for generalizing Wright's insights (Slatkin 1991), but the traditional estimators of F_{ST} are not estimators of C_{ST} insofar as they are affected by mutation. This can be addressed in several nonexclusive ways, such as using F_{ST} only in conditions where it is expected to closely approximate C_{ST} or performing likelihood analyses that do not necessarily consider F_{ST} or C_{ST} as model parameters, but can nevertheless provide estimates of them as a byproduct.

However, the fact that F_{ST} cannot always reach 1 independently of Q_w is often perceived as a deficiency, independently of its relationship with C_{ST}. A possible correction is then to divide F_{ST} by its maximum value given Q_w, which is taken to be Q_w itself (Hedrick 2005). However, it is not clear what such a corrected F_{ST} ("G'_{ST}") brings. What one expects from the distribution of a statistic is that it is sensitive to parameters of interest and insensitive to other parameters, but there is no evidence that the corrected F_{ST} has such properties. On the contrary, the correction would bias inferences from F_{ST} in realistic conditions where the uncorrected F_{ST} gives robust answers despite low Q_w (*e.g.*, Leblois *et al.* 2003; Whitlock 2011). Concerns about maximum F_{ST} values, and additional *a priori* arguments about what a differentiation measure should be, have also paved the way for the rejection of F_{ST} as not being a true measure of genetic

doi: 10.1534/genetics.113.152132
Address for correspondence: Institut des Sciences de l'Evolution (UM2-CNRS), Université Montpellier 2, Place Eugène Bataillon, CC 065, 34095 Montpellier cedex 5, France. E-mail: francois.rousset@univ-montp2.fr

differentiation (Jost 2008). The absence of a clear inferential framework raises the same questions for Jost's proposed alternative as for G'_{ST} (Whitlock 2011). Thus, the proposed alternative descriptors deal with maximum values of the descriptors in a cosmetic way, which addresses none of the more substantial issues that make F_{ST} biases a concern.

In a recent paper in *Genetics*, Jakobsson *et al.* (2013) also reject these alternative descriptors and instead consider the relationship between the maximum F_{ST} value given the allele frequency of the most common allele. In essence, a common allele has to be present in different populations, and then F_{ST}, which depends on the variation in allele frequency among populations, cannot take large values. Detailed results demand some thought in the multiallelic case, and Jakobsson *et al.* (2013) further compute the expected maximum F_{ST} under some prior distribution for allele frequencies, not based on a population model. Taking examples from human populations, they emphasize that their calculations also show that highly polymorphic loci should have lower F_{ST} and can explain the relatively low estimates of F_{ST} for microsatellites. Actually, past works have recognized the latter effects at both theoretical and data analytical levels (*e.g.*, Slatkin 1995; Rousset 1996; Estoup *et al.* 1998; Balloux *et al.* 2000). It has also been recognized that these effects contribute to the explanations for differences between single nucleotide polymorphisms (SNPs) and microsatellites in human populations (Payseur and Jing 2009).

Jakobsson *et al.* (2013) emphasize that their results for conditional F_{ST} hold independently of any biological process controlling allele frequency distributions, but a drawback of this model-free approach is that the biological conditions that make F_{ST} biases a concern are not identified. By comparison, earlier studies have addressed this question by taking expectations of gene diversities over distributions of population allele frequencies determined by different biological scenarios. In particular, Jakobsson *et al.* (2013) think that F_{ST}'s among human African populations "underpredict the intuitive level of differentiation" and that this can be explained in terms of the high genetic diversity of these populations. Although one cannot argue about intuitive levels of differentiation, model-based comparisons of F_{ST} to C_{ST} have shown that F_{ST} can approximate C_{ST} well despite high diversity, allowing low-bias inferences of demographic parameters from highly polymorphic markers (*e.g.*, Figure 3 in Leblois *et al.* 2003). Whether F_{ST}'s of highly polymorphic markers will be biased or not can be understood by comparing the distributions of coalescence times of the different pairs of genes from which F_{ST} is defined (Rousset 1996).

Jakobsson *et al.* (2013)'s results concern maximum F_{ST} values, and a similar dependence of conditional average F_{ST} values can be expected. Such values are often considered independent from the allele frequency in the total population, at least when the total population comprises many local populations. However, Jakobsson *et al.* (2013) consider a pair of populations, and expected values of F_{ST} for such a pair, conditional on allele frequency in the same pair, will be lower than average when allele frequency is extreme in a biallelic system. Conversely, they will be higher than average when allele frequency is intermediate, with a variable strength of this dependence according to which frequency (from the pair or the total population) is taken as the conditioning variable (Rousset 2002). If an unbiased F_{ST} estimator is sought, averages over the distribution of the conditioning allele frequency should be computed, and it is not clear why we should look at conditional maximum or expected F_{ST} values. For other problems, such as the detection of selected markers, it may be relevant to consider the conditional distribution of an F_{ST} estimator, for example given the gene diversity (Beaumont and Nichols 1996).

In an inferential perspective, it is well known that the widely used F_{ST} estimators do not use all the information about model parameters in the data. With progress in likelihood methods for inference in population genetics, one could have expected a decreased interest in F_{ST} (*e.g.*, Beerli and Felsenstein 2001). This has not occurred, and may not occur soon for reasons as diverse as authors' fear of rejection for omitting some method, or the persistent difficulties of performing likelihood analyses. Thus, one can expect that statistical properties of F_{ST} and alternative summary statistics will remain a matter for debate, but such debates will be inconclusive as long as they rest only on model-free results.

Acknowledgments

I thank M. Beaumont for inviting this commentary, N. Rosenberg for discussion, and two reviewers for comments.

Literature Cited

Balloux, F., H. Brünner, N. Lugon-Moulin, J. Hausser, and J. Goudet, 2000 Microsatellites can be misleading: an empirical and simulation study. Evolution 54: 1414–1422.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. B 263: 1619–1626.

Beerli, P., and J. Felsenstein, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA 98: 4563–4568.

Crow, J. F., and K. Aoki, 1984 Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA 81: 6073–6077.

Estoup, A., F. Rousset, Y. Michalakis, J.-M. Cornuet, M. Adriamanga *et al.*, 1998 Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). Mol. Ecol. 7: 339–353.

Hedrick, P. W., 2005 A standardized genetic differentiation measure. Evolution 59: 1633–1638.

Jakobsson, M., M. D. Edge, and N. A. Rosenberg, 2013 The relationship between F_{ST} and the frequency of the most frequent allele. Genetics 193: 515–528.

Jost, L., 2008 G_{ST} and its relatives do not measure differentiation. Mol. Ecol. 17: 4015–4026.

Leblois, R., A. Estoup, and F. Rousset, 2003 Influence of mutational and sampling factors on the estimation of demographic

parameters in a “continuous” population under isolation by distance. Mol. Biol. Evol. 20: 491–502.
Payseur, B. A., and P. Jing, 2009 A genomewide comparison of population structure at STRPs and nearby SNPs in humans. Mol. Biol. Evol. 26: 1369–1377.
Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142: 1357–1362.
Rousset, F., 2002 Inbreeding and relatedness coefficients: What do they measure? Heredity 88: 371–380.
Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. 58: 167–175.
Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457–462.
Whitlock, M., 2011 G'_{ST} and D do not replace FST. Mol. Ecol. 20: 1083–1091.
Wright, S., 1938 Size of population and breeding structure in relation to evolution. Science 87: 430–431.
Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354.

Communicating editor: M. Beaumont