What is genetic differentiation, and how should we measure it – GST, D, neither or both?

ARTICLE *in* MOLECULAR ECOLOGY · JULY 2014

Impact Factor: 6.49 · DOI: 10.1111/mec.12856

Robert Verity
Imperial College London

Richard Alan Nichols
Queen Mary, University of London

Verity R, Nichols R A (2014) What is differentiation, and how should we measure it – G_{ST}, *D*, neither or both?

to be published in the journal *Molecular Ecology.* Please cite this article as doi: 10.1111/mec.12856

What is genetic differentiation, and how should we measure it – G_{ST}, D, neither or both?

Robert Verity[1,2†], Richard A. Nichols[2]

[1] MRC Centre for Outbreak Analysis and Modelling, Imperial College London, London, UK
[2] School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom

[†]Correspondence author. e-mail: r.verity@imperial.ac.uk

Abstract

Estimates of the fixation index, F_{ST}, have been used as measures of population differentiation for many decades. However, there have been persistent voices in the literature suggesting that these statistics do not measure true differentiation. In particular, the statistics Nei's G_{ST} and Wier & Cockerham's θ have been criticised for being 'constrained' to not equal one in some situations that seem to represent maximal differentiation. Here we address the issue of how to evaluate exactly how much information a particular statistic contains about the process of differentiation. This criterion can be used to counter most concerns about the performance of G_{ST} (and related statistics), whilst also being reconciled with the insights of those who have proposed alternative measures of differentiation. In particular, the likelihood-based framework that we put forward can justify the use of G_{ST} as an effective measure of differentiation, but also shows that in some situations G_{ST} is insufficient on its own, and needs supplementing by another measure such as Jost's D or Hedrick's G'_{ST}. This approach will become increasingly important in the future, as greater emphasis is placed on analysing large data sets.

Keywords:

Differentiation, Fixation index, Jost's D, Hedrick's G'_{ST}, Likelihood, Information

Introduction

One of the fundamental observations of genetics is that allele frequencies vary from one location to another in almost all species. The magnitude of this differentiation between demes (partially isolated sub-populations) has traditionally been quantified by the value F_{ST} (Wright, 1943, 1949), which can be estimated from genetic samples using the statistics G_{ST} (Nei, 1973, 1977, Nei and Chesser, 1983), θ (Weir and Cockerham, 1984) and a fleet of alternatives. However, there have been persistent voices suggesting that these methods are fundamentally flawed (e.g. Jost, 2008). One central concern is that these statistics are 'constrained' in the sense that they cannot equal one to indicate maximum differentiation between demes while there is still some polymorphism within demes. This argument, among others, has lead to the rejection of G_{ST} and related statistics by some authors (we refer to G_{ST} for brevity, but our arguments apply equally to related statistics).

This paper asks whether we would actually be justified in rejecting G_{ST} purely because it does not have a fixed range between zero and one, and what might be meant by a 'better' measure? Practical principles for assessing genetic differentiation in real populations are then built on the foundations uncovered by answering these questions. It becomes clear that *differentiation* has acquired different meanings over the years, not all of which are quantities that lie between zero and one. We then use a likelihood-based perspective to ask how informative different measures are about these types of differentiation, and to ask if genetic differentiation is best summarised by the more established measures such as G_{ST}, one of the proposed alternatives, or some combination.

Measurements that make full use of the range 0-1

Nei proposed a coefficient of gene differentiation, G_{ST} (Nei, 1973), which is a function of heterozygosity estimates obtained from a genetic sample drawn from several demes:

$$G_{ST} = \frac{H_{T_e} - H_{S_e}}{H_{T_e}} \quad . \tag{1}$$

(Nei, 1973)

The $H_{.e}$ values are estimates of expected heterozygosity: the probability that two homologous genes sampled from a population would be different alleles. The value H_{Se} is the estimate for two gene copies drawn from different individuals in the same deme, while H_{Te} is for two gene copies drawn at random from the whole population. Nei had broad justifications for specifying this particular ratio, but for our current purposes it is sufficient to note that if the demes have different allele frequencies then the total genetic diversity will exceed the average diversity in a single deme, and hence the value of G_{ST} will be greater than zero.

Hedrick (2005) pointed out that this estimate does not vary between zero and one, but rather between zero and $G_{ST(max)}$, a maximum value that depends on H_{Se} and the number of demes that were sampled (k). It seemed more appropriate to express G_{ST} as a proportion of the maximum possible value:

$$G'_{ST} = \frac{G_{ST}}{G_{ST(\max)}} = G_{ST}\,\frac{(k-1+H_{S_e})}{(k-1)(1-H_{S_e})} \ , \tag{2}$$

(Hedrick, 2005, equation 4b)

(see also Meirmans and Hedrick (2011) for an alternative version corrected for sampling bias). This approach has some apparently desirable consequences: it means that estimates from loci with high mutation rates, such as microsatellites, fall in the same 0-1 range as other loci. Otherwise, their inherently higher H_{Se} values would mean that microsatellite surveys would consistently report lower genetic differentiation than other loci with lower mutations rates, such as SNPs. This is not to say that G'_{ST} is completely insensitive to the value of H_{Se} – for example, when there is no within-deme heterozygosity G'_{ST} will always equal 1, irrespective of the value of H_{Te}.

Jost (2008) suggested a second alternative measure, which, unlike G'_{ST}, can take on values between zero and one irrespective of the H_{Se} value:

$$D = \frac{H_{T_e} - H_{S_e}}{1 - H_{S_e}} \frac{k}{k-1} \ , \qquad (3)$$

(Jost, 2008, equation 12)

which, for direct comparison with G_{ST} and G'_{ST}, can be rearranged as

$$D = \frac{G_{ST} k \, H_{T_e}}{(k-1)(1-H_{T_e})(1-G_{ST})} = \frac{G'_{ST} k \, H_{T_e}}{(k-1+H_{S_e})} \ . \qquad (4)$$

The genetic patterns that constitute 'maximal' differentiation under each of these statistics can be quite different, as can be seen in Figure 1.

The fact that G'_{ST} and D occupy the same 0-1 range, but are different, leads us to a simple conclusion – that having the desired range is an insufficient property, in itself, to guide a choice of one estimate over the other. We might be convinced by different criteria, such as the need to partition diversity in a multiplicative manner (Jost, 2008), yet there are further alternatives such as the statistic I (Dewar et al., 2011) supported by arguments regarding mutual information content. More importantly, we might ask whether *any* single metric can fully capture the evolutionary process of interest, or whether we should abandon the idea of choosing between different statistics completely. In order to adjudicate between these competing arguments it is helpful to ask if there is some underlying property of the populations that could be described as *differentiation*, and whether any of these measures provide useful estimates of it?

What is Differentiation?

Differentiation and inbreeding

One possible description of differentiation is long established, namely Wright's F_{ST}. The value of F_{ST} can be considered to be a parameter: a property of the group of populations being studied. In contrast values such as G_{ST}, G'_{ST} and D are statistics, that is, values calculated from the observed genetic data. The definition of F_{ST} is founded on Wright's earlier work (Wright, 1921b), in which he defines the inbreeding coefficient F as the correlation between the uniting gametes that make up a diploid individual. The term 'correlation' can be understood, in broad terms, by first considering the case where there is no correlation between the gametes that formed a particular individual (F=0). We would then expect the genotypes at neutral loci to occur in Hardy-Weinberg proportions: the proportions that we would expect given independent sampling from the population in question. On the other hand, the parents of a particular individual might be more closely related than average – the offspring of cousins for example – in which case there would be a positive value of F, indicating that the maternal and paternal homologous gene copies are more likely to match, making the individual more likely to be homozygous. In other words, the individual is likely to be genetically differentiated from the deme as a whole.

The value of F is a property of individuals, but a similar logic can be applied to describe the differentiation of demes from each other. In the same way that Wright used a correlation to characterise differentiation of individuals, he specified the inbreeding coefficient F_{ST} to describe this correlation between a pair of gametes drawn at random from a deme. The subscripts S and T indicate that this correlation involves a comparison between subpopulations (demes) and the total population.

A careful, clear, modern definition of what is meant by *correlation* in this context is provided by Balding (2003) in the following form:

$$F^l_{ST} = \frac{E[I^l_{ij} I^l_{i'j'}] - p_l^2}{p_l(1-p_l)} \quad . \qquad (5)$$

(Balding, 2003, equation 9)

This value applies to a particular allele, l, in a single deme, in which the *expected* allele frequency is p_l (more formally, it is the expectation of the allele frequency under the model of interest). In order to understand this notation, consider drawing two homologous gene copies from a deme – if they were both l alleles, then [$I^l_{ij} I^l_{i'j'}$] would take the value one, otherwise zero. To see the link between Wright's F and F_{ST}, imagine we were creating a new individual from these genes; a value of 1 would then indicate an imaginary homozygote, a value of 0 an imaginary heterozygote. Under Hardy-Weinberg assumptions, the expectation, E[$I^l_{ij} I^l_{i'j'}$], would therefore be the expected frequency of homozygotes for this allele in the deme. Notice that the correlation is zero when E[$I^l_{ij} I^l_{i'j'}$] $=p_l^2$. In simple models p_l would be the global allele frequency, in which case the relationship E[$I^l_{ij} I^l_{i'j'}$] $=p_l^2$ indicates that the expected homozygote frequency in the deme matches the global value, which in turn implies the allele frequency in the deme matches the global value – there is no differentiation.

Correlation and the probability of identity by descent: confusion over terms

This precise definition of F_{ST}, as a correlation, can be linked to the alternative perspective of F_{ST} as a description of identity by descent. However, we find that different authors have followed two different paths in reasoning about identity by descent – and that this often-unappreciated divergence leads to two different ways of viewing genetic differentiation.

In order to understand this difference, it is simplest to start with the relationship between genealogy and the individual-centred correlation, F. The basic logic is simple: the correlation between an individual's homologous gene copies is increased by every chain of common ancestry that links the maternal and paternal lineages of that individual; hence the designation of F as the inbreeding coefficient. It follows that if we know the pedigree of an individual then we can compute a value of F for that individual by simply summing over all possible chains of common ancestry. Wright called this the 'method of path coefficients' (Wright, 1921a).

Equivalently, the effect of a pedigree on genetic correlations can be framed in terms of the *probability* of shared ancestry. For example the child of two siblings might have inherited a copy of the same grand-parental gene – with a probability of 0.5 at any one locus. In the absence of any deeper relationships between the grandparents this is the value of *F*. A number of authors arrived at this same line of reasoning via slightly different routes (Haldane and Moshinsky, 1939, Cotterman, 1940, Malécot, 1948), leading to a diverse array of related terms and definitions. These terms were unified to some extent under the common name 'probability of identity by descent', P(IBD), put forward by Crow (1954). Unfortunately this term is not always used in the manner Crow specified, and this has led to ambiguity about the meaning of the parameters *F* and F_{ST}, and hence to confusion as to whether they capture what we would like to estimate as differentiation. This ambiguity can be traced right back to the term's origin. Crow explicitly started with a distinction made by Malécot. He wrote that there are

> two ways in which a pair of alleles may be alike. I shall call them: (a) *alike in state*, i.e., both *A*, both *a*, or both a^x; (b) *identical by descent*, i.e., both derived from a single gene in some common ancestor … it is possible for two genes to be alike in state but not identical [by descent]. Conversely, two alleles which are identical [by descent] may be unlike in state if there has been a mutation since their common origin.
>
> (Crow, 1954, p544)

Hence, under Crow's definition, two gene copies could be identical by descent, but not identical in state, if there has been a mutation since the common ancestor. The concept of P(IBD) is independent of mutation. Malécot also makes this point explicitly (Malécot, 1948, p7, Malécot, 1969, p8). However, in the same text Malécot specifies a coefficient of coancestry for a locus which does not correspond to his earlier definition, or to Crow's P(IBD), since the formula explicitly shows that the coefficient is eroded by mutation:

$$\sum \left(\frac{1-u}{2}\right)^{n_i+p_i} \frac{1+f_A}{2} \ , \tag{6}$$

(Malécot, 1948, p9, Malécot, 1969, p11)

where *u* is the mutation rate. Many modern accounts conform to Crow's definition; they include Crow and Kimura (1970, p65), Hartl and Clark (2007, p259) and Charlesworth and Charlesworth (2010, p36). Others follow Malécot, for example Holsinger (2012, p24) does so explicitly, and Balding's formula (5) does so implicitly. A study could legitimately take either one of these viewpoints, but it is important to be clear which is being used.

F_{ST} with and without mutation

The fixation index F_{ST} is built upon *F*, and so the confusion over whether or not we should condition on a history of no mutations carries over into the study of population differentiation. We can conceive of a mutation-independent version of F_{ST}, and a mutation-dependent version. To clarify the distinction, consider a sub-population that has become isolated from the rest of a species range. Over time, coancestry will build up because breeding is restricted to this subpopulation. Given a sufficient period of isolation, all individuals would trace their ancestry back to their own subpopulation, and so the mutation-independent F_{ST} would tend to one. On the other hand, we would never expect the mutation-dependent F_{ST} to approach one – as there always remains a small chance that one or other lineage has picked up a mutation in the time since the common ancestor. Wright (1931) explored this exact scenario, concluding that the population tends towards an equilibrium between drift and mutation, such that

$$F_{ST} = \frac{1}{1+4N\mu} \ . \qquad (7)$$

Unlike in the mutation-independent view, this value is strictly less than one. (*N.b.* the model can be extended to populations experiencing migration, giving more familiar expressions for F_{ST} e.g. Wright, 1949).

Is G_{ST} constrained?

The statistic G_{ST} is an estimator of the parameter F_{ST}, so a natural question is does it estimate the mutation-independent version or the mutation-dependent version? The answer is apparent in Figure 2, which shows a simulation of the build up of F_{ST} (both the mutation-independent and the mutation-dependent versions) in a model of isolated populations, along with the observed value of G_{ST} calculated on a sample of individuals. It can be seen that G_{ST} is a reliable estimator of the mutation-dependent form of F_{ST}, tending towards Wright's predicted value (7) as time goes on. On the other hand the mutation-independent form of F_{ST} diverges and eventually reaches one when all individuals are descended from the same common ancestor. The correspondence between the mutation-dependent definition of F_{ST} and G_{ST} can be readily understood, since G_{ST} is based on the observed allele frequencies, which are in turn influenced by mutation. Whitlock (2011) gives an excellent account of the properties of F_{ST} and the statistics G_{ST}, G'_{ST} and D, coming to broadly similar conclusions.

There are reasons for wishing to estimate the mutation-independent version of F_{ST}: for example it describes the underlying pattern of ancestry in a sub-population – which can be used to predict patterns at loci with a range of different mutation rates, and which can act as a null distribution in the search for outliers that might be subject to selection (Beaumont & Nichols 1996, Antao *et al* 2008, Foll & Gaggiotti 2008). As well as a description of the probability of identity by descent, it can be viewed as equivalent to a ratio of coalescence times for lineages within and among subpopulations – an approach which makes clear that in non-equilibrium populations F_{ST} is not a simple function of migration rate (Slatkin 1991). In fact, our definition of mutation-independent F_{ST} can be seen as equivalent to 'coalescent F_{ST}' ($F_{ST,coal}$) described by Whitlock (2011).

In the remainder of this paper we ask about the utility of different statistics at estimating the mutation-independent version of F_{ST}. We use the shape of likelihood curves, calculated under a variety of evolutionary models, to evaluate the information provided by Jost's D and Hedrick's $G'_{ST.}$ We find that in some situations they convey complementary information to G_{ST} and are required in addition to it, rather than being effective replacements, while in other situations they contain little or no information.

Methods and results for a simple equilibrium model

In evaluating G_{ST}, D and G'_{ST} we use the criterion of information content: a good measure is one that contains a large amount of information about the underlying quantity of interest. The information contained in a measure can be evaluated using its likelihood function: an expression describing the probability of the observed data as a function of the unknown parameter(s) of the model (Edwards, 1984, provides a clear description of the approach, building on the foundations widely attributed to Fisher, 1921, 1956). When the likelihood curve is sharply peaked around its maximum (the maximum likelihood estimate of the parameters) it can be inferred that the measure used to produce this likelihood function contains a large amount of information about the unknown parameter(s). Conversely, when the likelihood curve falls off in a shallow slope about the maximum it can be inferred that the measure contains relatively little information. Crucially for our approach, a likelihood function can be produced based either on the full dataset, or some compressed version of the data – one or more statistics. It is therefore possible to produce a different likelihood curve based on each of the statistics that we are interested in, and by examining the shape of the resulting curves we can compare the information contained in each statistic directly. In this particular example we focus on estimating the scaled mutation rate $\theta=4N\mu$ (not to be confused with Weir and Cockerham's θ), rather than the level of differentiation, as this allows us to give the most direct demonstration of the method. The problem of estimating the level of differentiation will be addressed in the next section.

As a first step, we constructed a simple island model, in which five isolated subpopulations, each containing N=1000 diploid individuals, were assumed to be at equilibrium between infinite alleles mutation and genetic drift, and with no gene flow between subpopulations. The scaled mutation rate used when generating the data was θ=1 (R script available in online Supplementary Materials). We sampled five diploid individuals from each subpopulation and obtained the raw data given in Table 1. From this data we calculated the following statistical values; G_{ST}=0.520, D=1 and G'_{ST}=1.

Next, we generated likelihood curves based on a number of different levels of compression of the original data. Our primary reason for choosing this particular setup is that it is relatively easy to write down a likelihood function for θ given the complete data set (see **Appendix A** for details). This likelihood function can be written as follows:

$$L(\theta) = \theta^{\sum c_j} \prod_{j=1}^{k} \left[\frac{\Gamma(\theta)}{\Gamma(n_j+\theta)}\right] . \qquad (8)$$

where n_j denotes the number of gene copies sampled from deme j (in this example n_j=10 for all demes), and c_j denotes the total number of different allelic states found in this sample (i.e. the values given in the final column of Table 1).

The first summary statistic that we evaluated was the total number of distinct alleles present in the sample (Σc_j). Returning to the likelihood function (8); notice that the observed data only enter into this function via the value Σc_j (the values n_j depend only on the sample design, and are assumed known). Therefore, as well as being the likelihood function given the complete data set, equation (8) can be considered the likelihood function conditional on knowing just the total number of distinct alleles in the sample.

The second statistic that we evaluated was G_{ST}. Although a likelihood function for θ given G_{ST} cannot be written down in such a straightforward manner, it is relatively easy to reconstruct this likelihood curve via simulation. First, a particular value of θ, denoted θ^*, was chosen from the interval [0, 5]. Second, 50,000 random data sets, of the same size as the observed data, were generated from the equilibrium model under this value of θ^* by simply drawing from (8). Finally, G_{ST} was calculated for each of these random data sets, and compared against the observed value of G_{ST} =0.520. Only those data sets with a value of G_{ST} within a distance ε=0.01 of the true G_{ST} were retained. As long as ε is small enough (strictly in $\lim_{\varepsilon \to \infty}$) the proportion of random data sets retained by this method is a quantity directly proportional to the likelihood, and so by carrying out this method for a range of θ^* values (either drawing uniformly or in our case sampling at fixed intervals) we can reconstruct the likelihood curve to an arbitrary degree of accuracy. This approach has a strong statistical foundation, and is often considered a form of Approximate Bayesian Computation (ABC), although in our case it is the likelihood that we are interested in, rather than the posterior distribution given some informative prior (see Sunnåker et al. (2013) for a modern perspective on this and related approaches).

Finally, we considered the likelihood function for θ given the known values of D or G'_{ST}. Although we could take a simulation-based approach as above, there is in fact no need to do so in this case, as both of these statistics take on a value of one under this model irrespective of the value of θ.

The results of this analysis are shown in Figure 3. Notice that the most sharply peaked likelihood curve is the one produced from the number of unique alleles. Notice also that we obtain the same curve given just this one number as given the entire data set. Surprisingly, this tells us that we can boil down all the data in Table 1 to just a single number, Σc_j=13, without losing any information whatsoever. We can say that the number of unique alleles is a 'sufficient statistic' in this particular scenario – a point made by Ewens (2004, p302). No alternative measure can ever contain more information about θ than this measure, and so the number of unique alleles clearly represents an attractive form of data compression in this context.

The likelihood curve relating G_{ST} to θ has a greater spread than the likelihood function conditional on the full data, and so by choosing G_{ST} as our measure we have lost some information compared with the complete data set. This does not follow directly from the fact that we have boiled the data down from many numbers to just a single number, since we have shown that a single number (Σc_j=13) would suffice.

Finally, we come to the likelihood curves given just the values of Jost's D or Hedrick's G'_{ST}. Both of these likelihood curves are completely flat, as we obtain the same value of D or G'_{ST} for any value of the scaled mutation rate. Thus, we can say that D and G'_{ST} contain exactly zero information in this particular scenario. This does not mean that D and G'_{ST} are worse statistics *per se* – in fact, this analysis could be seen as slightly unfair to D and G'_{ST}, as neither statistic was conceived as an estimator of θ. For this reason we stress that there may be situations in which these statistics are preferable to G_{ST} in terms of information content. Rather, the result in Figure 3 simply demonstrates that D and G_{ST} are not good choices of statistic if our aim is to estimate the scaled mutation rate, and if we are happy to accept this particular model.

Methods and results for non-equilibrium models

Building on the results above, we moved on to consider the problem of estimating mutation-independent F_{ST} in some simple non-equilibrium models. First, we constructed a simple island model of 10 subpopulations, each containing N=100 diploid individuals, with no mutation and no migration between demes. Unlike in the previous example we did not assume that populations were at equilibrium; rather, we were interested in the build up of differentiation over time (see Appendix B for full details of the assumed model, and online Supplementary Materials for the R scripts). We simulated evolution under this model for 100 generations, keeping track of the true value of F_{ST} in the population as a whole (i.e. the true probability of any two randomly chosen gene copies being identical by descent). The final population achieved a level of differentiation of F_{ST}=0.43. We generated our "observed" data by sampling 10 individuals from each subpopulation in the final generation, leading to the following statistical values; G_{ST}=0.356, D=0.150 and G'_{ST}=0.453 (the complete data set can be found in the online Supplementary Materials).

Our objective was to estimate the level of differentiation using only the observed values of the statistics. We did this by simulating 50,000 sets of values of F_{ST}, G_{ST}, D and G'_{ST} from the model above, each time running the model for 1000 generations. By only retaining those values of F_{ST} for which the simulated value of the statistic was within a distance ε=0.01 of the observed value, we were able to reconstruct the likelihood curves for F_{ST} given each of the statistics (this is essentially the same method as that used for the equilibrium case).

The results of this analysis are shown in Figure 4. We can see that G_{ST} does a reasonably good job of characterising the true value of F_{ST} in the demes, while G'_{ST} and D are less informative. If our aim is to critique G_{ST} as a measure of differentiation then this result is perhaps reassuring – as long as the mutation rate of our chosen genetic material is low, we can capture the process of differentiation adequately using the statistic G_{ST} alone.

Moving forward, we considered the problem of estimating F_{ST} in a more challenging model in which mutation was occurring at a high rate. The basic model structure was exactly the same as that above, but incorporating an infinite-alleles model of mutation with scaled mutation rate θ=4. As before, our "observed" data consisted of a sample of 10 individuals per subpopulation taken after 100 generations, at which point the true level of differentiation was F_{ST}=0.433. This data set produced the following statistical values; G_{ST}=0.149, D=0.684 and G'_{ST}=0.731 (again, the complete data set can be found in the online Supplementary Materials).

We then used the simulation-based method to estimate the value of F_{ST} *and* the scaled mutation rate θ given each of the observed statistics, either on their own or in combination. We generated 5000 simulated data sets, each time running the model for 1000 generations. A threshold of ε=0.01 was used when estimating parameters using a single statistic, or ε=0.05 when using two statistics. The results of this analysis can be found in Figure 5. Notice that if we base our inference on the observed value of G_{ST} alone then we cannot distinguish between two possible explanations for the data – equilibrium with a high mutation rate, and non-equilibrium with a low mutation rate. This leads to a ridge in the likelihood distribution, covering a wide range of parameter values. Similarly, using the observed value of D alone we cannot distinguish between populations that are relatively undifferentiated with a high mutation rate, and populations that are completely differentiated with a low mutation rate. However, supplementing G_{ST} with D solves this problem, as it provides us with two separate sources of information. Based on the combined values of G_{ST} and D we can correctly estimate the values of F_{ST} and θ, thereby giving us an accurate picture of the true level of differentiation in the population.

Discussion

By tracing G_{ST} to its origin – the parameter F_{ST}, and further to the inbreeding coefficient F upon which F_{ST} was built – we can identify the root cause of some of the disagreement in the literature around the measurement of population differentiation. We have found that there are two overlapping views regarding the definition of the probability of identity by descent, which in turn have rubbed off on our definitions of F_{ST}, leading to mutation-dependent and mutation-independent versions. The criticism that G_{ST} is constrained is misplaced – at least if the task at hand is to estimate the mutation-dependent version of F_{ST}. If we wish to capture other aspects of the population history, such as the mutation-independent version of F_{ST}, when the mutation rate is relatively high, then we will need to supplement G_{ST} with other measures that capture a different aspect of evolution. No single statistic can be informative about both parameters in this situation, as it is mathematically impossible for a single dimension to fully represent two.

The likelihood framework provides us with an objective and quantitative way of determining the amount of information contained within different measures. Our aim here is one of data compression, in which we seek to retain the important aspects of the data while reducing the number of digits required to describe them. The effectiveness of a particular compression does not have a simple relationship with the number of values that are being thrown away. There may be data compressions that are equivalent in terms of the number of digits retained, but different in terms of information content, as demonstrated in Figure 3.

Finally, and perhaps most importantly, the concept of likelihood is tied to a particular model, and hence it is impossible to quantify the amount of information contained within a measure without referring to a particular model. This sobering fact means that there is no way of evaluating how good a statistic is based solely on the properties of the measure. While we can express certain preferences, in terms of the range of the statistic and how many values we are happy to retain, we cannot say that one statistic is objectively better than another. In order to make value judgements like this we must refer to a particular model or class of models. The fact that the information content of a statistic depends on our chosen model is a consequence of the way that we have defined differentiation – as an unknown parameter describing the past history of a set of populations, rather than as a statistic calculated on the observed data. Thus no statistic *is* differentiation, but some statistics can be used to *infer* differentiation. We find that mutation-independent F_{ST} is a sensible quantity to use as our definition of differentiation, although the general arguments made above are equally valid when applied to alternative definitions.

Under a simple island model with no mutation or migration we found that G_{ST} was a good estimator of F_{ST}, while Jost's D and Hedrick's G'_{ST} were not (Figure 4). This result is perhaps unsurprising, as neither D nor G'_{ST} were designed as estimators of F_{ST}; however, this illustrates the importance of deciding on a definition of differentiation when choosing between statistics. If we accept F_{ST} as our definition then we must conclude that G_{ST} should not be *replaced* by one of these alternative statistics, as we are likely to throw away valuable information. At first sight this might seem to be an argument for using marker loci with low mutation rates, however the different information conveyed by markers with high and different mutation rates can be exploited to obtain information about different periods in a species' evolutionary history (e.g. Nichols & Freeman 2004), although care must be taken to that the mutation model is appropriate to the genetic marker used. Under the same model with a high mutation rate we found that G_{ST} is insufficient on its own to jointly estimate the true level of differentiation and mutation (Figure 5). Supplementing G_{ST} with either G'_{ST} or D solved this problem, providing a distinct source of information that can be used to pull apart the confounded signals. With modern genomic tools providing ready access to a large number of loci with a wide range of mutation rates, even in non-model organisms, it has become practicable to exploit this type of information.

References

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008). LOSITAN: A workbench to detect molecular adaptation based on a F(st)-outlier method. *Bmc Bioinformatics,* **9**.

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology* **63**, 221-230.

Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B-Biological Sciences,* **263**, 1619-1626.

Charlesworth B, Charlesworth D (2010) *Elements of Evolutionary Genetics*. Roberts and Company Publishers, Colorado.

Cotterman CW (1940) *A Calculus for Statistico-genetics*, PhD dissertation, Ohio State University.

Crow JF (1954) Breeding structure of populations. II. Effective population number. *Statistics and mathematics in biology,* **543**, 556.

Crow JF, Kimura M (1970) *An introduction to population genetics theory.* Harper & Row, Publishers, New York, Evanston and London.

Dewar RC, Sherwin WB, Thomas E, Holleley CE, Nichols RA (2011) Predictions of single-nucleotide polymorphism differentiation between two populations in terms of mutual information. *Molecular Ecology,* **20**, 3156-3166.

Edwards AW (1984) *Likelihood*. The Johns Hopkins University Press, Baltimore and London.

Ewens WJ (2004) *Mathematical population genetics: I. Theoretical introduction*. Springer-Verlag, New York.

Fisher RA (1921) On the ``Probable Error'' of a Coefficient of Correlation Deduced from a Small Sample. *Metron,* **1**, 3-32.

Fisher, RA (1956). *Statistical methods and scientific inference.* Oliver and Boyd, Edinburgh.

Foll M, Gaggiotti O (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics,* **180,** 977-993.

Haldane J, Moshinsky P (1939) Inbreeding in Mendelian populations with special reference to human cousin marriage. *Annals of Eugenics,* **9**, 321-340.

Hartl DL, Clark AG (2007) *Principles of Population Genetics*. Sinauer associates, Sunderland.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution,* **59**, 1633-1638.

Holsinger KE (2006) *Lecture Notes in Population Genetics*. Storrs-Mansfield: Dept. Ecology and Evolutionary Biology, University of Connecticut.

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology,* **17**, 4015-4026.

Malécot G (1948) *Les Mathématiques de l'Hérédité*. Masson, Paris.

Malécot G (1969) *The Mathematics of Heredity (Revised, edited and translated by Yermanos, DM)*. W. H. Freeman and Company, San Francisco.

Meirmans, PG, and Hedrick, PW (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources* **11.1**: 5-18.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences,* **70**, 3321-3323.

Nei M (1977) F-statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics,* **41**, 225-233.

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. *Annals of Human Genetics,* **47**, 253-259.

Nichols RA, Freeman KL (2004) Using molecular markers with high mutation rates to obtain estimates of relative population size and to distinguish the effects of gene flow and mutation: a demonstration using data from endemic Mauritian skinks. *Molecular Ecology,* **13**, 775-787.

Pitman J (1996) Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, 245-267.

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research,* **58**, 167-175.

Sunnåker, M, et al. (2013). Approximate bayesian computation. *PLoS computational biology,* **9**, e1002803.

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

Whitlock, MC (2011) G'_{ST} and D do not replace F_{ST}. *Molecular Ecology,* **20.6**, 1083-1091.

Wright S (1921) Correlation and causation. *Journal of agricultural research,* **20**, 557-585.

Wright S (1921b) Systems of mating. I. The biometric relations between parent and offspring. *Genetics,* **6**, 111.

Wright S (1921c) Systems of mating. II. The effects of inbreeding on the genetic composition of a population. *Genetics,* **6**, 124.

Wright S (1931) Evolution in Mendelian populations. *Genetics,* **16**, 97.

Wright S (1949) The genetical structure of populations. *Annals of eugenics,* **15**, 323-354.

Appendices

A) Likelihood at equilibrium under an island model

Let us assume a finite-islands model consisting of *k* perfectly isolated subpopulations, each containing *N* diploid individuals. Further, let us assume that all subpopulations are presently at equilibrium between genetic drift and infinite alleles mutation (with scaled mutation rate $\theta=4N\mu$). Focussing for the moment on the j^{th} subpopulation; the probability of seeing a new (i.e. previously unseen) allele is equal to $\theta/(i+\theta)$, where *i* denotes the number of gene copies already sampled from this subpopulation, while the probability of seeing an old (i.e. previously seen) allele is equal to one minus this quantity, or $i/(i+\theta)$. Thus, given a sample of n_j gene copies drawn from this subpopulation, in which c_j different allelic states are represented, we find that the probability of the sample can be written down as follows:

$$\Pr\,(\text{sample}) \propto \frac{\theta^{c_j}\Gamma(\theta)}{\Gamma(n_j+\theta)} \ . \qquad \text{(A1)}$$

The proportionality symbol has been used here, as we are interested in the likelihood of θ, which is only defined up to a constant of proportionality. Taking the product of (A1) over all *k* subpopulations and collecting terms we obtain

$$L(\theta) = \theta^{\sum c_j} \prod_{j=1}^{k} \left[\frac{\Gamma(\theta)}{\Gamma(n_j+\theta)}\right] \ . \qquad \text{(A2)}$$

The observed data only enter into this equation via the total number of allelic states found in all subpopulations (Σc_j). This quantity can therefore be considered a sufficient statistic in this context – a point made by Ewens (2004, p302).

B) Simulation in non-equilibrium models

When considering the change in genetic composition over time in a simple island model it is necessary that we have a model governing the *initial* makeup of the subpopulations. The exact model that we opt for is unlikely to have a major effect on our conclusions, relative to other driving factors such as the population size or mutation rate, however, we are forced to opt for *some* model otherwise we cannot proceed. In all of the examples above (aside from the equilibrium example, where the initial makeup of the population is irrelevant) a simple model of a large population that became fragmented at time t=0 into *k* subpopulations was used. All 2*Nk* alleles in this large population were drawn from a Chinese restaurant process with concentration parameter α=1 (see Pitman, 1996). Individuals were then randomly assigned into subpopulations, such that there were *N* individuals in each of the *k* subpopulations. This scheme is likely to result in the majority of alleles being shared between subpopulations, but also allows for some alleles to be unique to a particular subpopulation from the outset. All subsequent generations proceeded according to standard Wright-Fisher dynamics, either with or without mutation.

Figures and Tables

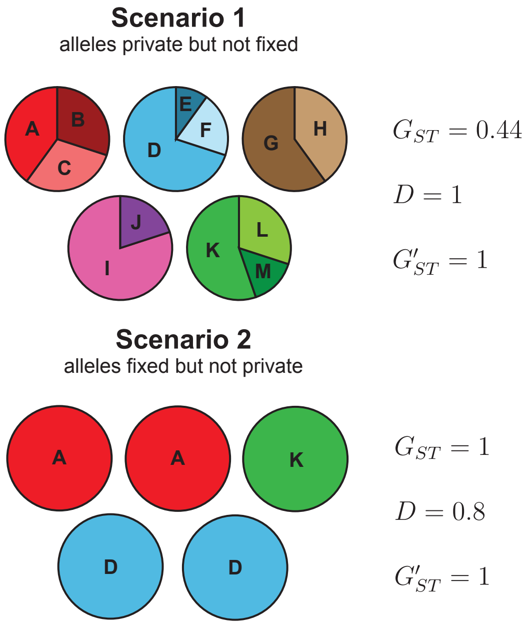

Figure 1: Comparison of different measures of differentiation for two different scenarios. In the first scenario all *k*=5 subpopulations contain only private alleles, and G'_{ST} and *D* both report maximal differentiation. In the second scenario some alleles are shared between subpopulations, and G'_{ST} and *D* disagree about the level of differentiation.

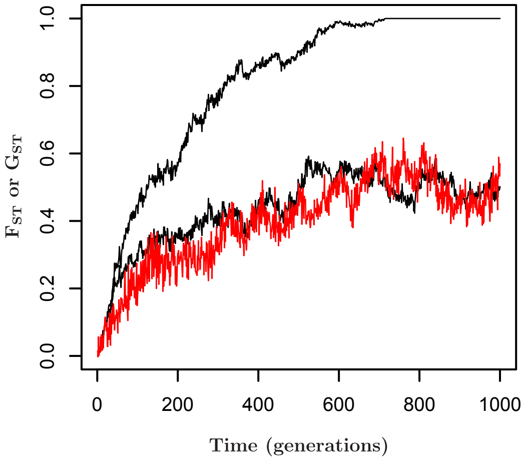

Figure 2: Build up of F_{ST} using the mutation-independent (upper black line) and mutation-dependent (lower black line) definitions, along with G_{ST} (red line) calculated on a random sample of individuals drawn from the population. Simulations were obtained from the finite islands model with 10 subpopulations containing 100 diploid individuals each, a scaled mutation rate of θ=1, and in the case of the G_{ST} calculations a sample size of 10 individuals per subpopulation (drawn independently at each point in time).

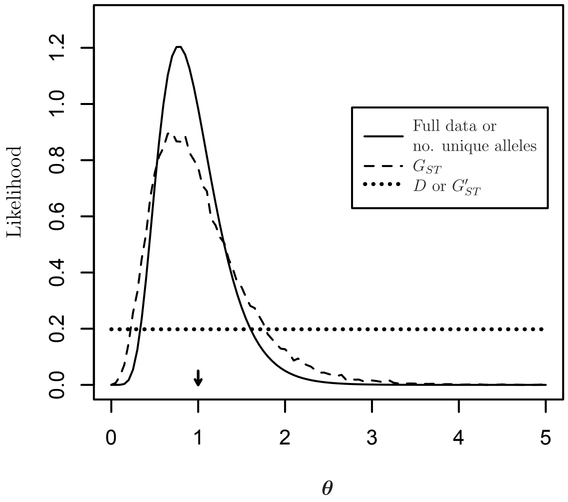

Figure 3: Likelihood curves for the scaled mutation rate (θ=4$N\mu$) under an island model at equilibrium between infinite alleles mutation and drift, and in the absence of migration between demes. Likelihood curves were obtained either analytically or by simulation for various levels of data compression, and have been normalised to have the same area. The true parameter value of θ=1, from which the 'observed' data was generated, is indicated with an arrow.

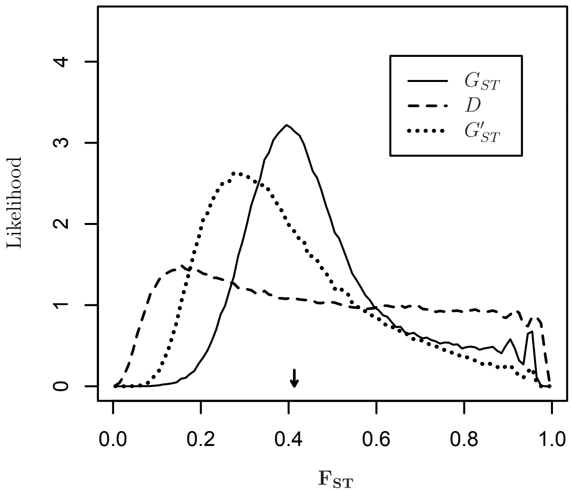

Figure 4: Likelihood curves for the level of differentiation (F_{ST}) reconstructed using simulation based methods with an acceptance distance of ε=0.01. The true value of F_{ST} =0.413 is indicated by an arrow.

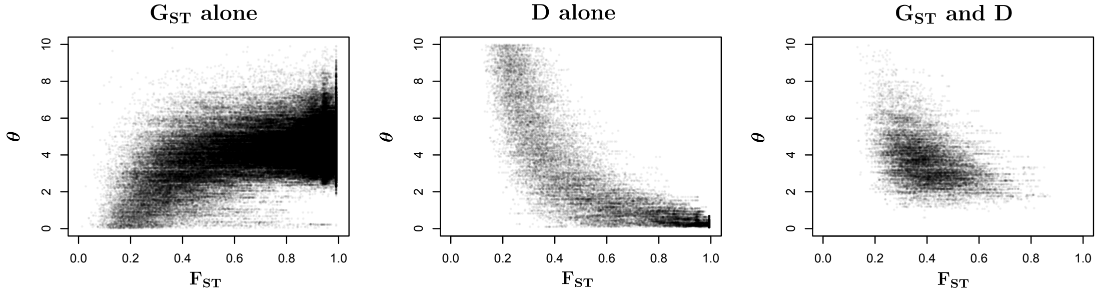

Figure 5: Draws from the bivariate likelihood surface of the level of differentiation (F_{ST}), and the scaled mutation rate (θ=4$N\mu$). In the first panel G_{ST} alone was used to estimate the unknown parameters, in the second panel Jost's D alone was used, and in the third panel both G_{ST} and Jost's D were used. An acceptance distance of ε=0.01 was used when estimating parameters based on G_{ST} or D alone, and a (Euclidian) acceptance distance of ε=0.05 was used for the case of both statistics combined.

Deme	Diploid Genotypes					Allele Counts	Unique Alleles
1	A_1A_1	A_2A_3	A_2A_2	A_2A_1	A_2A_2	$\{3,6,1\}$	$c_1=3$
2	B_1B_2	B_3B_2	B_4B_2	B_2B_2	B_3B_2	$\{1,6,2,1\}$	$c_2=4$
3	C_1C_1	C_1C_1	C_1C_1	C_1C_1	C_1C_1	$\{10\}$	$c_3=1$
4	D_1D_2	D_2D_2	D_2D_2	D_1D_2	D_2D_2	$\{2,8\}$	$c_4=2$
5	E_1E_2	E_3E_3	E_3E_3	E_3E_3	E_1E_3	$\{2,1,7\}$	$c_5=3$

Table 1: Simulated data drawn from island model at mutation-drift equilibrium with θ=1, showing five diploid individuals drawn from each of five subpopulations. All subpopulations contain only private alleles, as indicated by the fact that the alleles of each subpopulation are given a different unique letter. Allele counts and the total number of unique alleles are reported for each subpopulation.