Does G_{ST} underestimate genetic differentiation from marker data?

J. Wang

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Left running head: J. Wang

Right running head: Correlation of G_{ST} and H_S

Key words: Population Structure, Genetic Differentiation, F_{ST}, G_{ST}, Markers, Drift, Mutations

Corresponding author

Jinliang Wang
Institute of Zoology
Regent’s Park
London NW1 4RY
United Kingdom
Tel: 0044 20 74496620
Fax: 0044 20 75862870
Email: jinliang.wang@ioz.ac.uk

Abstract

The widely applied genetic differentiation statistics F_{ST} and G_{ST} have recently been criticised for underestimating differentiation when applied to highly polymorphic markers such as microsatellites. New statistics claimed to be unaffected by marker polymorphisms have been proposed and advocated to replace the traditional F_{ST} and G_{ST}. This study shows that G_{ST} gives accurate estimates and underestimates of differentiation when demographic factors are more and less important than mutations, respectively. In the former case, all markers, regardless of diversity (H_S), have the same G_{ST} value in expectation and thus give replicated estimates of differentiation. In the latter case, markers of higher H_S have lower G_{ST} values, resulting in a negative, roughly linear correlation between G_{ST} and H_S across loci. I propose that the correlation coefficient between G_{ST} and H_S across loci, r_{GH}, can be used to distinguish the two cases and to detect mutational effects on G_{ST}. A highly negative and significant r_{GH}, when coupled with highly variable G_{ST} values among loci, would reveal that marker G_{ST} values are affected substantially by mutations and marker diversity, underestimate population differentiation, and are not comparable among studies, species and markers. Simulated and empirical datasets are used to check the power and statistical behaviour, and to demonstrate the usefulness of the correlation analysis.

Introduction

A species rarely breeds at random throughout its whole range to form a homogenous unit. Frequently a species is genetically structured in space, subdivided into subunits called demes, races, subpopulations, … Delineating the spatial genetic structure by dividing a species into subunits and quantifying the genetic differentiation among the subunits is important in many biological fields such as evolution, conservation, human medicine and forensics. The subdivision can be made based on natural (e.g. rivers) or artificial (e.g. dams or highways) boundaries, on geographical locations, or on genetic data (e.g. Pritchard *et al*. 2000), and the differentiation can be measured from marker data using Wright’s (1943) F_{ST}, Nei’s (1973) G_{ST} and related statistics such as Weir & Cockerham’s (1984) θ and Slatkin’s (1995) R_{ST}. The development and wide application of highly polymorphic markers such as microsatellites made these statistics ever more popular, but also caused some confusion and concern. The most popular differentiation statistics, F_{ST} and G_{ST}, are believed to underestimate population differentiation when calculated from markers of high diversity (e.g. Nagylaki 1998; Hedrick

2005; Jost 2008), and for this reason alternative statistics were proposed and advocated to replace them (Hedrick 2005; Jost 2008; Meirmans & Hedrick 2011). The new differentiation statistics, however, are criticized for their lack of biological meaning and applications, their marker dependency but drift independency, and so on (see Ryman & Leimar 2009; Whitlock 2011; Wang 2012).

The claim that F_{ST} and G_{ST} underestimate population differentiation is made from both theoretical and empirical grounds. The mathematical definition of $G_{ST} = (H_T - H_S) / H_T$ suggests that it cannot take values larger than the average within subpopulation homozygosity, $1 - H_S$ (Jin & Chakraborty 1995; Nagylaki 1998; Hedrick 1999, 2005). This constraint is true both mathematically and biologically. Both F_{ST} and G_{ST} are inherently constrained by H_S, as they signify the amount of genetic variation between populations (V_B) as a proportion of the total variation V_T, which is composed of within (V_W) and between (V_B) population variation. A high H_S means a high V_W, and necessarily a low V_B as a proportion of V_T (i.e. low F_{ST} and G_{ST}). However, the constraint imposed on F_{ST} and G_{ST} by H_S does not necessarily mean they are always marker H_S dependent and underestimate differentiation from markers of high H_S, as claimed by some authors (e.g. Nagylaki 1998; Hedrick 1999, 2005; Jost 2008). On the empirical grounds, some studies showed that G_{ST} based on highly polymorphic microsatellites is usually lower than G_{ST} based on weakly polymorphic allozyme loci (e.g. Sanetra & Crozier 2003), and is obviously too low for highly differentiated subspecies (e.g. Balloux *et al*. 2000; Carreras-Carbonell *et al*. 2006). These empirical evidences are true for these particular systems, but do not suggest that F_{ST} and G_{ST} calculated from highly polymorphic markers must always underestimate population differentiation in all circumstances.

Are F_{ST} and G_{ST} dependent on marker diversity? Do they always underestimate population differentiation from markers of high diversity (e.g. microsatellites)? Under which set of conditions do they provide marker dependent (and thus biased) and marker independent (and thus accurate) estimates of population differentiation? Is it possible to detect whether F_{ST} and G_{ST} values calculated from a set of markers underestimate differentiation or not? In this paper, I will use a combination of analytical modelling, simulated data and empirical data to answer these questions. I show G_{ST} is independent of H_S when mutation rate (u) is small relative to migration rate (m) or drift ($1/2N$). Otherwise, G_{ST} decreases nearly linearly with an increase in H_S. The results suggest a test for the presence or absence of mutational effects on G_{ST}. If single-locus G_{ST} values are highly variable and the correlation between single-locus

G_{ST} and H_S values is significantly negative, then the observed G_{ST} values are substantially affected by mutations, are locus specific, and seriously underestimate the differentiation due to population demography. If the correlation is insignificant, then the observed single-locus G_{ST} values are unaffected by mutations and are marker independent. They can then be averaged to give an overall estimate of the genetic differentiation caused by demography only. Simulations and empirical data are analysed to check the power and statistical properties of the correlation and regression analyses.

Method

The relationship between G_{ST} and H_S is investigated by analyses of standard population genetics models of migration, drift and mutation. The results are then verified by analyses of simulated and empirical datasets.

Theory

Following most previous studies of F_{ST}, I assume a population under the finite island model of migration (Wright 1931) and the infinite allele model of mutation (Kimura & Crow 1964) for mathematical tractability. The results and conclusions are, however, applicable qualitatively to populations under other migration models, such as Wright's (1943) isolation by distance or neighbourhood model and Kimura & Weiss's (1964) stepping stone model, and under other mutation models, such as stepwise mutation model for microsatellites or allozymes (Ohta & Kimura 1973).

Under the finite island model with migration rate m among s subpopulations of effective size N, and under the infinite allele model for a neutral locus with mutation rate u, the recurrence equations for the expected homozygosity within a subpopulation, J_0, and between two subpopulations, J_1, is (Nei 1975; Li 1976)

$$J_{0(t+1)} = d\big(a\big(c + (1-c)J_{0(t)}\big) + (1-a)J_{1(t)}\big), \tag{1}$$

$$J_{1(t+1)} = d\big(b\big(c + (1-c)J_{0(t)}\big) + (1-b)J_{1(t)}\big), \tag{2}$$

where $b = m(2-m)/s$, $a = (1-m)^2 + b$, $c = 1/(2N)$ and $d = (1-u)^2$. Equivalently, J_0 and J_1 are the probabilities that two genes taken at random from within a subpopulation and from different subpopulations, respectively, are identical in state. The complements, H_S =1- J_0 and H_1 =1- J_1, give the expected (i.e. assuming random union of gametes) heterozygosity or gene diversity (Nei 1973) within and between subpopulations. The total expected heterozygosity or gene diversity in the entire population is $H_T = (H_S + (s - 1)H_1)/s = 1 - J_1 - (J_0 - J_1)/s$ (Nei 1975). Given H_T and H_S, G_{ST} is calculated by $G_{ST} = 1 - H_S/H_T$ (Nei 1973). Using (1) and (2), we can calculate recurrently the values of H_S, H_T and G_{ST} at each generation, given parameters *m, N, u, s* and initial gene identities $J_{0(0)}$ and $J_{1(0)}$.

Under the joint action of mutation, migration and drift at rates *u, m,* and 1/(2*N*) respectively, the gene diversity (H_S, H_T) and its distribution (G_{ST}) will reach equilibrium values. G_{ST} attains its equilibrium value much faster than H_S and H_T, because it is determined by the strongest (in terms of rate) while H_S and H_T are determined by the weakest among the forces of mutation, migration and drift. The equilibrium gene identity values are (Nei 1975; Li 1976)

$$J_{0(\propto)} = cd(a - (a - b)d)/G, \tag{3}$$

$$J_{1(\propto)} = cdb/G, \tag{4}$$

where $G = 1 - d(a(1 - c) + 1 - b) + d^2(a - b)(1 - c)$. The equilibrium gene diversity and differentiation values, $H_{S(\propto)}$, $H_{T(\propto)}$ and $G_{ST(\propto)}$, can be calculated using (3) and (4). The expression for $G_{ST(\propto)}$ is complicated, but can be simplified approximately to (Takahata & Nei 1984)

$$G_{ST(\propto)} \approx 1/\left[1 + 2N\left(\frac{s}{s-1}\right)\left(\frac{1}{(1-m)^2(1-u)^2} - 1\right)\right]. \tag{5}$$

When *m, u* « 1, (5) is further simplified to (Takahata & Nei 1984)

$$G_{ST(\propto)} \approx 1/\left[1 + 4N\left(\frac{s}{s-1}\right)(m + u)\right]. \tag{6}$$

When *s*→∝, (6) again reduces to the equilibrium F_{ST} of the infinite island model of Wright (1969, page 291), indicating that F_{ST} and G_{ST} are equivalent (Nei 1977; Takahata & Nei 1984).

Although several studies have used similar models to investigate the impact of mutations on F_{ST} and G_{ST} (e.g. Ryman & Leimar 2008; Whitlock 2011), none has examined the direct relationship between G_{ST} and H_S. Herein I will use equations (1-6) to explore this relationship in populations in both equilibrium and non-equilibrium conditions under different parameter (m, u, N, s) combinations. This is important as both G_{ST} and H_S are estimable from marker data, and examining the observed patterns of G_{ST} and H_S at a set of marker loci sheds light on the possible impact of mutations on G_{ST}.

Simulations

Simulated data typical of those encountered in practice were generated to test whether the correlation analysis of single locus estimates of G_{ST} and H_S could be used to detect the effect of mutations on G_{ST} when it is present, and whether the analysis does not falsely detect the effect of mutations when it is absent. The behaviour and power of the correlation analysis were investigated by analysing simulated data with varying sampling intensities (of individuals from a subpopulation, of subpopulations, and of markers), different population properties (N, s, m, u) and different mutation and migration models.

The simulations considered the finite island model as described above, and a one-dimensional circular stepping stone model (Kimura & Weiss 1964). In the latter model, a number of s subpopulations are arranged in a circle and each subpopulation receives a proportion $m/2$ of its individuals from each of its two neighbouring subpopulations. In both models, each subpopulation is composed of N diploid monoecious individuals. At each discrete generation, the events are mutations, migrations and reproductions occurring in that order. Mutations are assumed to follow either the infinite allele model or the stepwise mutation model. For the former, a mutation always generates a novel allele the population has never seen before. For the latter, the mutated allele increases or decreases in size by 1 repeat with an equal probability of 0.5. For both models, the number of new mutations at a locus in each subpopulation at each generation was sampled from a Poisson distribution with parameter value $2Nu$. For each new mutation, a gene was drawn at random from the $2N$ genes and was changed according to the mutation model. Reproduction is assumed to be random union of gametes, such that selfing and outbreeding occur at rates $1/N$ and $1-1/N$ respectively, and the effective size is equal to the census size for each subpopulation.

An ancestral population was assumed to be the same as the subdivided population described above except for population size and structure. It was unsubdivided and had a size $N_A = rsN$, where r=0.5, 1 and 2 such that it had equilibrium genetic diversity smaller than, close to, and larger than the subdivided population respectively. The ancestral population was maintained for a large number of generations for it to reach mutation-drift equilibrium at a neutral locus with mutation rate u (which was variable among a number of L loci). It was then subdivided into *s* subpopulations of size *N*, which were maintained as described above for *g* (=100, 200, 400) generations or for a sufficiently large number of generations, in the order of Max(1/*u*, 1/*m*, 2*N*), to reach mutation-drift-migration equilibrium. A sample of *M* individuals was then taken at random from each of *R* (≤*s*) randomly selected subpopulations, and each sampled individual was genotyped at a number of *L* loci.

The genotype data were then used to calculate Nei & Chesser's (1983) nearly unbiased estimators of H_S, H_T, and thus G_{ST},

$$\widehat{H}_S = \frac{2\widetilde{M}}{(2\widetilde{M}-1)R}\sum_{j=1}^{R}\left(1-\sum_{i=1}^{k}x_{ij}^2\right),$$

$$\widehat{H}_T = 1-\sum_{i=1}^{k}\left(\frac{1}{R}\sum_{j=1}^{R}x_{ij}\right)^2+\frac{\widehat{H}_S}{2\widetilde{M}R},$$

$$\widehat{G}_{ST} = 1-\widehat{H}_S/\widehat{H}_T,$$

where x_{ij} is the frequency of allele *i* in the sample from subpopulation *j*, *k* is the number of alleles observed in the set of samples from the *R* subpopulations, and $\widetilde{M}$ is the harmonic mean sample sizes (≡*M* in the simulations).

The estimates $\widehat{H}_S$ and $\widehat{G}_{ST}$ were then used to calculate their correlation coefficient r_{GH} across loci. The significance of r_{GH} was tested by a permutation analysis in which $\widehat{H}_S$ and $\widehat{G}_{ST}$ were both randomized across loci before calculating r_{GH} in 10^6 replicates. The proportion of replicates in which r_{GH} was smaller than the r_{GH} value calculated from the original data was taken as the *p* value. The correlation coefficient was taken as statistically significant when $p<0.001$. A significant negative correlation r_{GH} indicates that $\widehat{G}_{ST}$ has been affected by mutations and thus underestimates the differentiation caused purely by demography (drift and migration). Otherwise, markers with different levels of diversity $\widehat{H}_S$ are equally differentiated, they all give the same G_{ST} expected from the impact of drift and migration only, and the single locus G_{ST} estimates can be averaged to give a better (in precision) overall estimate of differentiation.

Too many parameter combinations, due to the numerous parameters and the numerous plausible values of each parameter, are involved in determining $\widehat{H}_S$ and $\widehat{G}_{ST}$ that a realistic simulation study can only consider a small fraction of them. I studied the effect of each parameter in isolation of others each time by varying the values of the focal parameter only (see Table 1). For each parameter combination, a number of 100 replicate datasets were generated and analysed. The analysis results were reported as the mean correlation coefficient between $\hat{G}_{ST}$ and $\hat{H}_S$, $\bar{r}_{GH}$, and the proportion of replicates with a statistical significant (at p<0.001) r_{GH} among the 100 replicates.

The simulation program was checked by comparing the simulated against the predicted values of several quantities to make sure it worked properly. First, the effective size of the entire population in the finite island model is $N_e = sN/(1\text{-}F_{ST})$ (Wright 1943; Wang & Caballero 1999), where F_{ST} can be replaced by G_{ST}. This theoretical prediction was compared with that estimated from the simulated pedigrees, using the formula $\frac{1}{2N_e} = \frac{\theta_{t+1}-\theta_t}{1-\theta_t}$ where t the generation is large and θ_t is the average coancestry at generation t for all individuals in the entire population. Second, the predicted values of H_S, H_T and G_{ST} by (3-4) were compared with the corresponding observed values for an equilibrium population under infinite allele and finite island models. In all situations investigated, the predicted and estimated (observed) values fitted very well.

Empirical data

The simulation model may be too simple to reflect the reality. In a real population, both m and N may vary over space and time, and migrations and mutations may not follow the ideal models assumed in the simulations. Supplementing simulations, therefore, I also analysed several recently published empirical datasets to demonstrate the use of the proposed correlation analysis.

Atlantic Salmon: To investigate the genetic structure of Atlantic salmon populations in the entire North American range of the species, Moore *et al*. (2014) sampled 9142 individuals from 153 populations and genotyped each individual at 15 microsatellite loci. They also sampled 1080 individuals from 50 populations and genotyped each individual at 3192 SNP loci. The two datasets were analysed separately in the present study of the relationship between G_{ST} and H_S.

Blacknose sharks: Using 23 microsatellites and mtDNA sequences, Portnoy *et al*. (2014) investigated the genetic structure and barriers to gene flow of 10 blacknose shark populations sampled (651 individuals in total) from the western North Atlantic Ocean. It was found that the F_{ST} values at the 23 microsatellite loci between the Bahamas and any of the other populations were more than an order of magnitude greater than the values between any two of the other populations. Therefore, G_{ST} and H_S values were calculated for each locus in the 2 alternative population structures, the 10- and 2-population (Bahamas and the rest) models in the present study.

Mediterranean shore crab: Schiavina *et al*. (2014) investigated the genetic structure of the Mediterranean shore crab (*Carcinus aestuarii*) in the Adriatic Sea (central Mediterranean), using 11 polymorphic microsatellites in 431 individuals collected from eight sites. One locus, Cae30, has only 5 alleles and a gene diversity of H_S =0.1, much lower than the locus with the 2nd lowest diversity, which has 13 alleles and a H_S=0.77. So Cae30 was excluded as an obvious outlier from the G_{ST} and H_S correlation analysis.

Blacktip reef sharks: To understand the genetic structure of blacktip reef sharks (*Carcharhinus melanopterus*), Vignaud *et al*. (2014) sampled 758 individuals from 15 sites (4 widely separated locations in the Indo-Pacific and 11 islands in French Polynesia) widely distributed in the Indian and Pacific Oceans. Each sampled individual was genotyped at 17 microsatellite loci. Three loci (cil169, cli107 and cli12) were found to deviate significantly from Hardy-Weinberg equilibrium and were suspected to contain null alleles (Vignaud *et al*. 2014). The three loci were excluded from their original genetic analysis. Herein I investigated the impact of mutations on the estimated differentiation among these shark populations by analysing the relationship between G_{ST} and H_S, using both the entire set of 17 loci and the selected subset of 14 loci.

Copper rockfish: Using 17 microsatellite DNA loci, Dick *et al*. (2014) assessed the genetic diversity of and the differentiation among ten populations of copper rockfish (*Sebastes caurinus*) representing paired samples of outer coast and the heads of inlets in five replicate sounds on the west coast of Vancouver Island, British Columbia. The sample size per population varies between 30 and 105. I calculated the G_{ST} and H_S values at each of the 17 loci among the 10 populations, and tested whether the marker differentiation is affected by mutations or not.

Results

Analytical results

Equation (6) suggests that G_{ST} at neutral loci is determined by the joint action of migration, mutation and drift occurring at rates *m*, *u*, and 1/(2*N*) respectively. The relative impact of each evolutionary force on G_{ST} is determined by its rate as a proportion of the total rate, $m+u+1/(2N)$. When subpopulations are small such that drift is the dominating force (i.e.

1/(2*N*) » *u*+*m*), then $H_S \rightarrow 0$ (i.e. fixation) and $G_{ST(\propto)} \rightarrow 1$ in equilibrium conditions. When mutation is weak relative to drift and migration (i.e. *u*«1/(2*N*) + *m*), then $G_{ST(\propto)} \approx 1/\left[1 + 4N\left(\frac{s}{s-1}\right)m\right]$, which suggests that $G_{ST(\propto)}$ reflects demography only and all loci with varying but small *u* have the same expected G_{ST}. In contrast, for loci with a high *u* in a population with a large *N* and a small *m* (i.e. *u* » 1/(2*N*) + *m*), $G_{ST(\propto)}$ becomes locus (or mutation) dependent and covaries with locus specific H_S (below). In such a case, marker based $G_{ST(\propto)}$ has little bearing on population demography, the $G_{ST(\propto)}$ value calculated from one set of loci can hardly be congruent with that from another set of loci, and it is incomparable among studies, species and loci.

Figure 1 plots the equilibrium G_{ST} as a function of H_S, calculated by (5) and (3) respectively, for different parameter combinations of *u*, *m* and *N*, assuming *s*=10. When differentiation is expected to be small due to either strong migration (*m*≥0.01) or weak drift (*N*≥2500), G_{ST} keeps constant and does not vary with H_S in its entire range of [0, 1] caused by widely varying *u* values in range of [10^{-6}, 10^{-2}]. The observation disproves the belief that G_{ST} underestimates differentiation and becomes H_S dependent when H_S is high (e.g. Nagylaki 1998; Hedrick 1999, 2005; Jost 2008). High H_S values (say 0.95) do constrain G_{ST} to small values with a maximum of 1- H_S, but do not necessarily lead to underestimated and locus-varying G_{ST}. What is relevant is the main mechanism (determined by the relative strengths of mutation, drift and migration) leading to the observed high H_S, not the observed high H_S *per se*. A high H_S is usually due to a high *u* or/and a high *N*. However, as long as *m* is much higher than *u*, G_{ST} is virtually independent of H_S.

When drift is strong (i.e. *N* small) and migration is weak relative to mutations, G_{ST} decreases almost linearly with an increasing H_S due to an increasing *u* (Figure 1). Only in this situation is the belief that G_{ST} covaries with H_S (e.g. Nagylaki 1998; Hedrick 1999, 2005; Jost 2008) certified. For the parameter combination *N*=250, *m*=0.001, and *s*=10 in Figure 1, for example, G_{ST} keeps almost a constant value of 0.45 when *u* varies between 10^{-6} and 3×10^{-6} that leads to a H_S varying between 0 and 0.5. With $u>3\times10^{-6}$ and thus H_S >0.5, G_{ST} begins to decrease linearly with an increasing H_S (or *u*). Similar results are obtained with other values of the number of subpopulations (*s*).

Many generations, in the order of 1/*m*, 1/*u* or 2*N* whichever is the smallest, are required for a subdivided population to reach the equilibrium differentiation. Natural populations may never reach such equilibrium as *m* and *N* are constantly changing. It is thus important to check whether the above observations (Figure 1) also apply to non-equilibrium populations. Figure 2 plots G_{ST} as a function of H_S at generations 50, 200 and 1000 since the subdivision. Mutation rate (*u*) is assumed to vary from 10^{-6} to 10^{-2}, and the initial gene diversity is assumed to be $J_{0(0)} = J_{1(0)}$ and to take values $rJ_{0(\infty)}$, where r=1, 0.5 and 0.25. The relationship between G_{ST} and H_S in a non-equilibrium population is similar to that in an equilibrium population (Figure 1). Whenever $u \ll 1/(2N) + m$, G_{ST} does not vary with H_S (or *u*). Depending on *u* as well as *N* and *m*, H_S can freely vary in almost the entire range of [0,1] without affecting the value of G_{ST}. Otherwise, G_{ST} decreases nearly linearly with an increasing H_S (or *u*). The further away a population departs from the equilibrium, the less affected it is by mutations because the latter require time to accumulate. When N=250, for example, mutations start to have a substantial impact on G_{ST} at generations 50, 200 and 1000 when $H_S \geq 0.8$ ($u \geq 0.0015$), $H_S \geq 0.5$ ($u \geq 0.00001$) and $H_S \geq 0.3$ ($u \geq 0.00003$) respectively. Initial gene identities (or diversities) seem to have little effect on the relationship between G_{ST} and H_S at any generation.

Simulation results

Confirming the analytical results presented above, simulations show that, when mutations are strong relative to migrations (m=0.001), G_{ST} estimates vary among loci that have different *u* and thus different H_S, and are negatively correlated with H_S (Figure 3). This is true for the finite island and stepping stone migration models, and for the infinite allele, finite allele and stepwise mutation models. This is also true no matter the population is at mutation-drift-migration equilibrium (Figure 3) or not (data not shown). The negative correlation in stepping stone migration model and infinite allele mutation model is stronger than that in other migration and mutation models. In contrast, when mutations are weak relative to migrations (m=0.01), G_{ST} estimates are small and are almost constant among loci with different *u* and thus different H_S. This is shown for an equilibrium population under different migration and mutation models (Figure 3), but is also true for non-equilibrium populations (data not shown).

When migrations are weak relative to mutations such that G_{ST} is substantially affected by *u* and becomes negatively correlated with H_S, a modest sampling effort is needed to detect the correlation for different migration and mutation models (Figure 4). This is also true for populations that have not reached mutation-drift-migration equilibrium (data not shown). Setting the statistical significance at a conservative level of p<0.001, the false detection rate of mutational effects is low (generally below 7%), while the power is generally above 60% except when less than 10 loci and less than 4 subpopulations are used in the analysis. In agreement with the results in Figure 3, the correlation analysis is less powerful for the island migration model coupled with the stepwise or 2-allele mutation model than other models. While the power increases with the numbers of sampled loci and sampled subpopulations (Figure 4), it is little affected by the number of sampled individuals per subpopulation, M, as long as $M > 10$. This is not surprising because the population is highly differentiated for the parameter combinations and just a few individuals per subpopulation would allow for a good estimate of G_{ST}.

Empirical analysis

The Atlantic salmon data clearly show an extremely strong negative correlation (r =-0.953) between G_{ST} and H_S estimates among the 15 microsatellites (Figure 5A), with a p value of 0.0×10^{-6}. These markers are highly polymorphic, with H_S varying between 0.66 and 0.94 and with the number of observed alleles varying between 15 and 91. Compatible with a substantial impact of mutations, these markers have low but highly variable G_{ST} values, varying between 0.02 and 0.09 with a mean of 0.045 and a coefficient of variation of 0.629. These single locus G_{ST} values are all highly significant, as determined by permutation (permuting individuals among subpopulations) tests.

In contrast, the correlation between G_{ST} and H_S estimates of the 3192 SNPs (Figure 5B) is positive and small (r=0.044), with a p value of 0.993 which is insignificant. H_S values distribute nearly uniformly in the range [0, 0.5]. While most SNPs have G_{ST} values of about 0.1, quite a few outliers show G_{ST} values well above 0.4. The mean G_{ST} is 0.099 for the 3192 SNPs and is 0.091 when the outlier SNPs with $G_{ST} > 0.3$ are removed. Both values are much larger than the mean G_{ST} across the 15 microsatellites which is 0.045. The comparison between SNPs and microsatellites further verifies that the differentiation at microsatellites is greatly impacted by mutations and thus underestimates the underlying population differentiation due to demography.

The blacknose sharks have highly variable single-locus G_{ST} values, with the highest being 0.35 and 0.18 and the lowest being 0 and 0 for the 2- and 10-population models respectively (Figure 5C). Among the 23 microsatellites, G_{ST} and H_S estimates are moderately negatively correlated, with a correlation coefficient of -0.41 (p=0.017) and -0.43 (p=0.007) for the 2- and 10-population models respectively. None of the correlations are significant at *p*=0.001, but there is a clear trend of less differentiation at more polymorphic marker loci which indicates that mutations might have reduced the G_{ST} values at these loci.

The differentiation calculated from each of the 10 microsatellites is low (G_{ST} <0.04) among the 8 Mediterranean shore crab populations (Figure 5D). Nevertheless, G_{ST} and H_S estimates are highly negatively correlated, with a correlation coefficient of -0.80 and a small *p* value (0.010). It is likely that mutations have substantially impacted on the G_{ST} estimates from these microsatellites, and thus the underlying population differentiation due to demography may well be underestimated by these microsatellites.

The 17 microsatellites in blacktip reef sharks are highly variable in diversity, with the number of observed alleles varying from 4 to 48 and the H_S varying from 0.15 to 0.89. The G_{ST} values among the 15 populations estimated from the 17 loci are also highly variable, from 0.04 to 0.41 (Figure 5E). The 3 loci showing deviation from Hardy-Weinberg equilibrium are apparently not outliers in terms of both diversity and differentiation. The single locus G_{ST} and H_S estimates are highly negatively correlated, with a correlation coefficient of -0.890 (p=0.000×10^{-6}) and -0.913 (p=0.000×10^{-6}) for the entire set of 17 loci and the subset of 14 loci respectively. In this system, mutations are highly likely to have reduced the differentiation of the microsatellites; the underlying population differentiation due to drift and migration should be higher than the average G_{ST} value calculated from these microsatellites.

The differentiation measured by G_{ST} at each of the 17 microsatellites is low among the 10 copper rockfish populations (Figure 5F). Except for locus Sra11-103 which has a G_{ST} = 0.09, single locus G_{ST} values are below 0.05. The overall mean G_{ST} across loci is 0.027, very close to the F_{ST} value 0.031 obtained by Dick *et al*. (2014). Single locus G_{ST} and H_S estimates are not correlated, with a correlation coefficient of 0.011 and a *p* value of 0.649. It can be concluded confidently that mutations have no impact on these G_{ST} estimates, and all markers, regardless of polymorphisms, should have the same expected differentiation which is equivalent to the population differentiation. The average G_{ST} across loci, 0.027, should be an unbiased estimate of the population differentiation due to demography.

Discussion

The claim that F_{ST} and G_{ST} are dependent on marker H_S and underestimate population differentiation when calculated from highly polymorphic (i.e. high H_S) markers (e.g.

Nagylaki 1998; Hedrick 2005; Jost 2008) can be misleading. It has led to the conclusion that these traditional statistics should be either "corrected" for H_S (e.g. Hedrick 2005) or replaced by new statistics such as D (Jost 2008). The claim creates lots of confusions, as if F_{ST} and G_{ST} should be independent of H_S to be correct measures of differentiation. As Wright (1978, p.82) explicitly stated, however, F_{ST} (the same for G_{ST}) measures "the amount of differentiation among subpopulations, relative to the limiting amount under complete fixation". Complete fixation means each subpopulation is fixed with an allele (i.e. all individuals in a subpopulation have the same homozygous genotype), which is not necessary to be unique among subpopulations. Fixation results in H_S=0, and the maximal differentiation of F_{ST} =1 is achieved only at H_S =0. For this reason, Wright (1951) also called his F_{ST} a fixation index, among other fixation indices of F_{IS} and F_{IT}. The quantity H_S measures the *absolute* distance from complete fixation (i.e. H_S =0), and naturally constrains F_{ST}, which measures the *relative* (to total diversity H_T) or standardized distance from complete fixation. The definition of $G_{ST} = 1 - H_S/H_T$ (Nei 1973) makes the functional relationship between absolute (i.e. H_S) and relative (i.e. G_{ST}) differentiations explicit. Therefore, both F_{ST} and G_{ST} legitimately depend on, or more precisely, are constrained by H_S. This relationship is true both mathematically and biologically, and does not inherently cause F_{ST} and G_{ST} to underestimate differentiation for markers with high H_S.

More precisely, F_{ST} and G_{ST} become marker dependent and underestimate population differentiation only when migration rate is lower than mutation rate. Otherwise, they provide accurate estimates of population differentiation regardless of marker H_S. In a population with low migration rates (i.e. $m < u$), a marker with a higher u is expected to have a higher H_S (or absolute differentiation) and a correspondingly lower G_{ST} (or relative differentiation) in both equilibrium and many non-equilibrium conditions (Whitlock 2011; this study). It should be emphasized that a high u does not necessarily lead to a high H_S, and *vice versa*. This is because it is the quantity uN rather than u that determines H_S. A marker with a small u in a population with a large N can still harbour a high H_S, and a marker with a large u in a population with a small N can still have a low H_S. The statement that microsatellites, because of their high allelic polymorphisms and high H_S, must always underestimate differentiation is imprecise. Such markers show less differentiation than less polymorphic markers (e.g. SNPs) only when migration is weak ($m<u$), as illustrated by Figures 1 and 2.

This study reveals that whenever $m<u$ and thus mutations have a substantial impact, single locus G_{ST} values decrease almost linearly with single locus H_S. This is true in both

equilibrium (Figure 1) and non-equilibrium populations, as verified by simulations under different migration and mutation models (Figure 3). It is not surprising that the pattern observed under the ideal island migration model and the infinite allele mutation model applies to other migration and mutation models, because G_{ST} and F_{ST} are defined as descriptive statistics without any predefined demographic and mutation models. Mutations act to increase genetic diversity (H_S and H_T) and thus to decrease differentiation among subpopulations, no matter they occur in the finite or infinite allele model or in the stepwise mutation model (Wright 1943). Migrations, in contrast, tend to redistribute genetic diversity evenly among subpopulations. Thereby they tend to reduce the difference between H_S and H_T and thus to reduce G_{ST}, no matter they occur in the island model, stepping stone model or the isolation-by-distance model.

The simulations confirm that a correlation analysis between single locus G_{ST} and H_S estimates can be used to detect the mutational effects on differentiation. Under typical sampling intensities, the analysis has sufficient power to identify the mutational effect when it is present, and it does not falsely detect the mutational effect when it is absent (Figure 4), when the significance level is chosen as p=0.001. A higher significance p value (say, 0.05 or 0.01) leads to higher powers, but also higher false detect rates. Under the finite island and infinite allele models (first row in Figure 4), for example, the power (when m=0.001) and false detection rate (when m=0.01) increase to 86.7% and 11.8% respectively when p=0.01, and to 90.7% and 30.0% respectively when p=0.05. A good balance between type I and II errors is achieved at p=0.001, which leads to a false detection rate being always below 7% irrespective of the widely varying sampling intensities of the number of subpopulations, the number of individuals per subpopulation, and the number of loci and polymorphisms (Figure 4).

Two out of the five empirical microsatellite datasets (Figures 5A, 5E) show strong evidence (a high negative r_{GH} value and a small p value) that mutations have reduced G_{ST} estimated from microsatellites, two datasets (Figures 5C, 5D) indicate a similar trend with higher uncertainties, and the remaining dataset (Figure 5F) shows no detectable effect of mutations on G_{ST}. It is noticeable that the copper rockfish populations (Figure 5F) have high and widely variable H_S values across the 17 microsatellites, the highest H_S being 0.936. These H_S values are similar to those of the microsatellites in Atlantic salmon populations (Figure 5A) and the blacktip reef shark populations (Figure 5E). Yet, contrasting patterns of G_{ST} and H_S were observed among the three species. This again verifies the theory and simulation based conclusion that a high H_S does not necessarily lead to marker dependent G_{ST}, and does not necessarily result in underestimation of population differentiation. In situations where the correlation between G_{ST} and H_S has a high uncertainty (e.g. Figure 5C), collection of more data (by sampling more subpopulations, loci, and individuals) may confirm or reject the hypothesis that G_{ST} in a study system is affected by H_S or mutations. In contrast, the analysis of a big SNP dataset (Figure 5B) does not detect any mutational effect. The correlation between single locus G_{ST} and H_S values, 0.044, is small and positive, and clearly indicates no mutational effects on G_{ST}. The results are understandable because the u for SNPs can be several orders smaller than that for microsatellites, and as a result is more likely to be smaller than migration rate m.

The five empirical microsatellite datasets were taken from the most recent literature at random with regard to the relationship between G_{ST} and H_S, which was revealed only after the correlation analyses. If this small sample of datasets represents the reality, then we may conclude that underestimation of differentiation by microsatellites could be a common problem (Hedrick 1999, 2005; Jost 2008). A meta-analysis of many more microsatellite datasets as exemplified in this study is required for a solid conclusion. However, while microsatellites do underestimate differentiation in some (or many) situations, they can also yield unbiased estimates in situations where migration is high as shown for the copper rockfish populations (Figure 5F). The assertion that all microsatellites of high H_S underestimate differentiation and therefore all G_{ST} estimates should be standardized (Hedrick 2005) or abandoned and replaced by new differentiation statistics (Jost 2008) is unjustified. In addition to the problems shown before (Ryman & Leimar 2009; Whitlock 2011; Wang 2012), these new statistics are also marker diversity dependent as shown below.

It is notable that several authors have conducted a meta-analysis of the relationship between G_{ST} and H_S across species/populations (Heller & Siegismund 2009; Meirmans & Hedrick 2011). They found that the estimated G_{ST} is always smaller than the maximum value of 1- H_S, as expected, and shows a weak negative correlation with H_S. It should be pointed out that the correlation analysis proposed in my study is fundamentally different from that in these meta-analyses. In the latter, the correlation is at the species level, where G_{ST} and H_S are average values across loci for each species. Because different species may have experienced different evolutionary forces and demography such that their G_{ST} values differ, it is unclear what the hypothesis these meta-analyses are trying to prove or disapprove, except for the functional relationship G_{ST} <1- H_S which should however always be true from the definition of G_{ST}. The presence of a negative correlation between G_{ST} and H_S does not prove that G_{ST} is underestimated and is marker dependent because of mutational effects. The absence of the correlation does not prove that mutations have negligible effects and G_{ST} is unbiased and marker independent. In my study, the correlation is between single locus values of G_{ST} and H_S within a species (population). The hypothesis, clearly defined and supported by theory and simulations, is that G_{ST} values should be similar across markers of different H_S if mutations are unimportant (when $u<m$), resulting in an r_{GH} not different from 0. Otherwise (i.e. $u>m$), G_{ST} values should decrease with markers showing an increasing H_S, resulting in a highly negative correlation between G_{ST} and H_S.

G_{ST} calculated from a locus measures the genetic differentiation among subpopulations at the locus due to the combined effect of all evolutionary forces (Nei 1973). Selection directly influences F_{ST} and G_{ST}, as Wright (1943) illustrated with several different types of selection. In principle, a negative correlation between H_S and G_{ST} can also be generated for markers closely linked with a locus under strong selection for spatially different alleles (which causes a decrease in H_S and an increase in G_{ST}) or/and for spatially different allele combinations (which causes an increase in H_S and a decrease in G_{ST}). Although my correlation analysis assumes the absence of selection, it should be robust in most applications. First, frequently only a few microsatellites (<30) are used in calculating F_{ST} or G_{ST}, and the chance of any of them being under selection or being linked to loci under selection strong enough (compared with other evolutionary forces) for detection is slim. Second, with genomic dense markers such as SNPs, it is highly likely that a small fraction of the loci are under strong selection. The correlation analysis should however still be robust because the vast majority of loci are neutral and a few selected loci should not affect the overall relationship between H_S and G_{ST}.

This study focusses on the widely applied differentiation statistic G_{ST} (Nei 1973). Other differentiation statistics or estimators such as θ (Weir & Cockerham 1984), D (Jost 2008) and G'_{ST} (Hedrick 2005) could also be affected by mutations and yield marker (H_S) dependent estimates. All these statistics measure differentiation at marker loci due to the collective actions of all evolutionary forces, including mutations. When mutations are important (i.e. $u>m$), therefore, differentiation estimates are expected to be different among loci. Some statistics, like D which is claimed to outperform G_{ST} for highly polymorphic markers (Jost 2008), are even more problematic and produce marker dependent differentiation estimates even when mutation rate is small relative to migration rate. For the data simulated in finite island and infinite allele models, finite island and stepwise mutation models, and stepping stone and stepwise mutation models shown in Figure 3, for example, the correlation coefficient between *D* and H_S, r_{DH}, is 0.43, 0.30, and 0.22 respectively when *m*=0.001, and is 0.71, 0.24 and 0.26 respectively when *m*=0.01. The correlation is always positive and substantially high, even in the situation where mutation is very weak relative to migration and G_{ST} is uncorrelated with H_S. Similarly highly positive r_{DH} values are obtained for all of the empirical datasets. For the Atlantic salmon SNP dataset, r_{DH} is 0.73 while r_{GH} is only 0.04. This means *D* always increases with H_S, even for markers with low mutation rate (e.g. SNPs) and low diversity, and for a population with a high migration rate.

Slatkin's (1995) R_{ST} provides unbiased estimates of population differentiation regardless of the mutation rates or diversity of markers. A mutation does not erase the evolutionary history of a gene when it occurs in some models such as the stepwise model. Mutations occurring in these models are accommodated by R_{ST}, which therefore measures differentiation purely due to population demography (*m* and *N*). Unfortunately, however, R_{ST} is sensitive to violations of the assumed mutation models and have a higher sampling variance than G_{ST} (Balloux & Lugon-Moulin 2002). Unless many (say in the hundreds) markers are used, R_{ST} may have a lower accuracy than G_{ST}.

What are the uses of a correlation analysis on G_{ST} and H_S? What we are usually interested are population level forces such as migration (or isolation) and drift, which have roughly the same effect on all loci in the genome, and population differentiation, which depends on population level forces and is estimated by all loci mainly controlled by population level forces. G_{ST} always faithfully reflects the differentiation at the marker loci, no matter the loci are governed primarily by population demography (*m* and *N*) or locus specific forces such as selection and mutation. Marker G_{ST} provides an unbiased and good estimate of population differentiation only when these markers are not significantly affected by locus specific forces. The correlation analysis essentially tests whether different markers give replicated or different estimates of G_{ST}, or whether or not population level forces are much more important than locus specific forces in shaping the marker diversity and distribution. A highly negative correlation between G_{ST} and H_S values indicates that 1) the migration rate must be low, lower than the mutation rate; 2) the marker G_{ST} may well underestimate population differentiation; 3) another set of markers with lower (higher) polymorphisms may well yield a higher (lower) estimate of G_{ST}; 4) the marker G_{ST} should be used cautiously in comparisons across species, studies and sets of loci. If the correlation between G_{ST} and H_S

values among loci is small and non-significant, then these single locus G_{ST} estimates should be marker (diversity) independent and can be averaged to provide a good estimate of population differentiation.

A computer program, **CoDiDi** (**Co**rrelation between **Di**versity and **Di**ferentiation), is written to calculate single locus G_{ST} and H_S values, to test whether a single locus G_{ST} value is significantly different from 0 or not by permutations, and to calculate and test the significance of the correlation between G_{ST} and H_S. The correlation analyses of all of the simulated and empirical data presented in this study were conducted by this program, freely available from the website: http://www.zsl.org/science/software/CoDiDi.

Acknowledgments

I thank PW Hedrick, PG Meirmans, N Ryman, Robin Waples, MC Whitlock and an anonymous referee for valuable comments on earlier versions of this manuscript.

References

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, **11**, 155-165.

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

Carreras-Carbonell J, Macpherson E, Pascual M (2006) Population structure within and between subspecies of the Mediterranean triplefin fish *Tripterygion delaisi* revealed by highly polymorphic microsatellite loci. *Molecular Ecology*, **15**, 3527–3539.

Dick S, Shurin JB, Taylor EB (2014) Replicate divergence between and within sounds in a marine fish: the copper rockfish (*Sebastes caurinus*). *Molecular Ecology,* **23**, 575-590.

Hedrick PW (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution*, **53**, 313–318.

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

Heller R, Siegismund H (2009) Relationship between three measures of genetic differentiation G_{ST}, D_{EST} and G'_{ST} : how wrong have we been? *Molecular Ecology*, **18**, 2080–2083.

Jin L, Chakraborty R (1995) Population structure, stepwise mutation, heterozygote deficiency and their implications in DNA forensics. *Heredity*, **74**, 274-285.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology,* **17**, 4015–4026.

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725-738.

Kimura M, Weiss G (1964) The stepping-stone model of population structure and the decrease of genetic correlation with distance. *Genetics,* **49**, 561-576.

Li W-H (1976) Effect of migration on genetic distance. *American Naturalist*, **110**, 841–847.

Meirmans PG, Hedrick PW (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources,* **11**, 5–18.

Moore JS, Bourret V, Dionne M *et al*. (2014) Conservation genomics of anadromous Atlantic salmon across its North American range: outlier loci identify the same patterns of population structure as neutral loci. *Molecular Ecology*, **23**, 5680-5697.

Nagylaki T (1998) Fixation indices in subdivided populations. *Genetics*, **148**, 1325–1332.

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

Nei M (1975) *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam, Netherlands.

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

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

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, **22**, 201-204.

Portnoy DS, Hollenbeck CM, Belcher CN *et al*. (2014) Contemporary population structure and post-glacial genetic demography in a migratory marine species, the blacknose shark, *Carcharhinus acronotus*. *Molecular Ecology,* **23**, 5480-5495.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics,* **155**, 945-959.

Ryman N, Leimar O (2008) Effect of mutation in genetic differentiation among nonequilibrium populations. *Evolution*, **62**, 2250–2259.

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

Sanetra M, Crozier R (2003) Patterns of population subdivision and gene flow in the ant *Nothomyrmecia macrops* reflected in microsatellite and mitochondrial DNA markers. *Molecular Ecology*, **12**, 2281–2295.

Schiavina M, Marino IAM, Zane L, Melià P (2014) Matching oceanography and genetics at the basin scale. Seascape connectivity of the Mediterranean shore crab in the Adriatic Sea. *Molecular Ecology*, **23**, 5496-5507.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–462.

Takahata N, Nei M (1984) F_{ST} and G_{ST} statistics in the finite island model. *Genetics*, **107**, 501-504.

Vignaud TM, Mourier J, Maynard JA *et al*. (2014). Blacktip reef sharks, *Carcharhinus melanopterus*, have high genetic structure and varying demographic histories in their Indo-Pacific range. *Molecular Ecology*, **23**, 5193-5207.

Wang J (2012) On the measurements of genetic differentiation among populations. *Genetics Research*, **94**, 275-289.

Wang J, Caballero A (1999) Developments in predicting the effective size of subdivided populations. *Heredity*, **82**, 212-226.

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

Whitlock MC (2011) G_{ST} and *D* do not replace F_{ST}. *Molecular Ecology*, **20**, 1083–1091.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97-159.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114-138.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Wright S (1969) *Evolution and the Genetics of Populations, Vol. II. The Theory of Gene Frequencies.* University of Chicago Press, Chicago.

Wright S (1978) *Evolution and the Genetics of Populations, Vol. IV. Variability Within and Among Natural Populations*. University of Chicago Press, Chicago.

J. Wang is interested in developing population genetics models and methods of analysis of empirical data to address issues in evolutionary and conservation biology.

Data accessibility

The computer program for simulating genotype data under different migration and mutation models, and for estimating H_S, G_{ST} and their correlation: Dryad DOI: 10.5061/dryad.733s9.

The 6 empirical datasets were retrieved from Dryad with DOIs:
http://dx.doi.org/10.5061/dryad.sb601; http://dx.doi.org/10.5061/dryad.vv277;
http://dx.doi.org/10.5061/dryad.r0d1q; http://dx.doi.org/10.5061/dryad.th4h5;
http://dx.doi.org/10.5061/dryad.s489b

The input files of the 6 empirical datasets for **CoDiDi** analysis: Dryad DOI: 10.5061/dryad.733s9.

Table 1 Parameter ranges in simulations

Migration model	Mutation Model	*t*	*m*	*u*	*M*	*R*	*L*
FIM, SSM	IAM, SWM, FAM	200, ∝	0.01,0.001	10^{-5}~10^{-3}	10, 20, 40, 80, 160	2, 3, 4, 6, 8, 10, 12	5, 10, 15, 20, 30

The size (*N*) and number (*s*) of subpopulations are fixed at 250 (or 1000) and 20, respectively. The finite island model (FIM) and circular stepping stone model (SSM) for migrations are considered for neutral loci under infinite allele model (IAM), stepwise model (SWM) or finite allele model (FAM) for mutations. For FAM, 2 alleles are considered to mimic SNPs. Symbols *t*, *m*, *u*, *M*, *R*, *L* represent number of generations when sampling occurs, migration rate, mutation rate, number of individuals sampled from a subpopulation, number of sampled subpopulations, and number of sampled loci, where *t*=∝ indicates a population at mutation-drift-migration equilibrium.

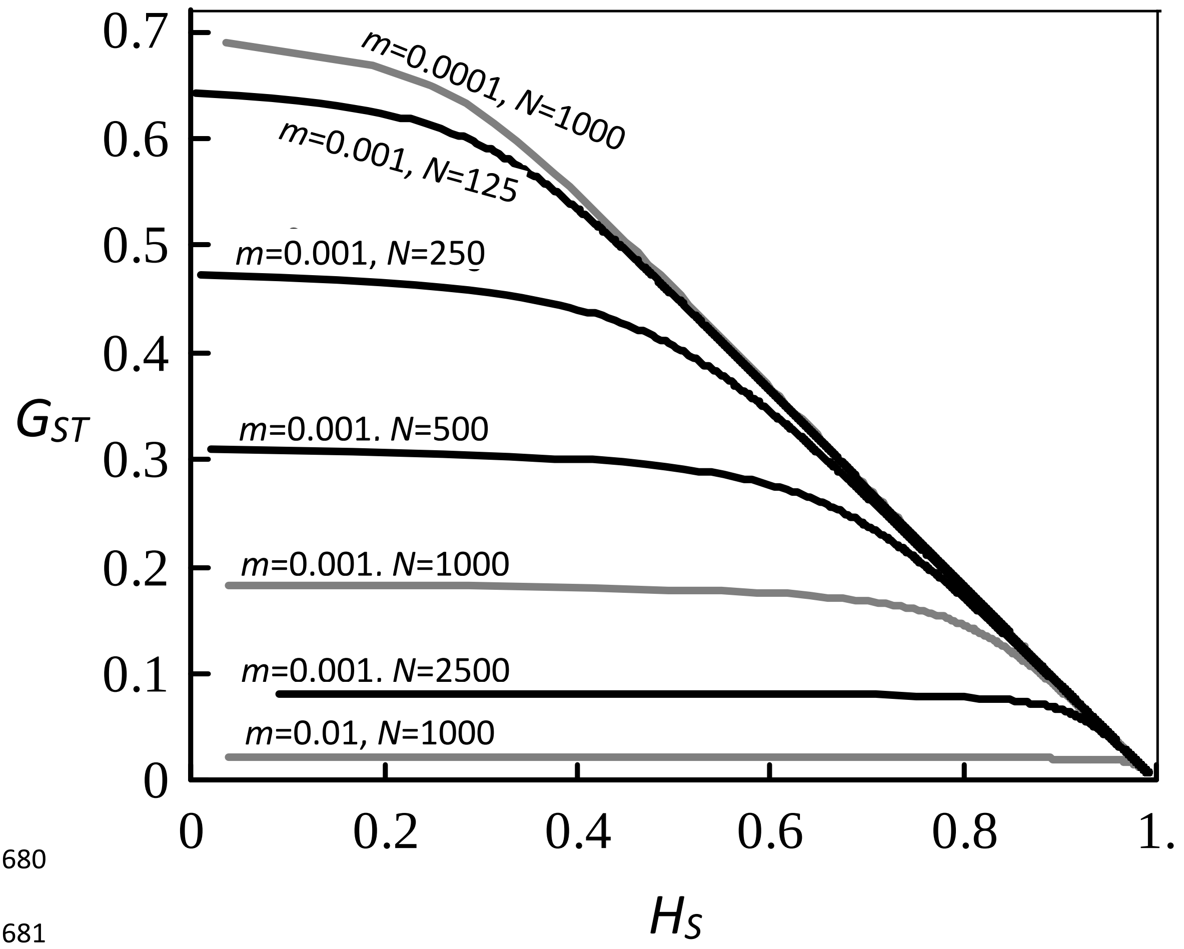

Fig. 1 G_{ST} as a function of H_S in equilibrium populations. The G_{ST} (y axis) and H_S (x axis) values for a population in a finite island model with s=10 subpopulations at mutation-drift-migration equilibrium were calculated for various parameter values of subpopulation size (N), migration rate (m), and mutation rate (u), where u ranges from 10^{-6} (left side of x axis) to 10^{-2} (right side of x axis).

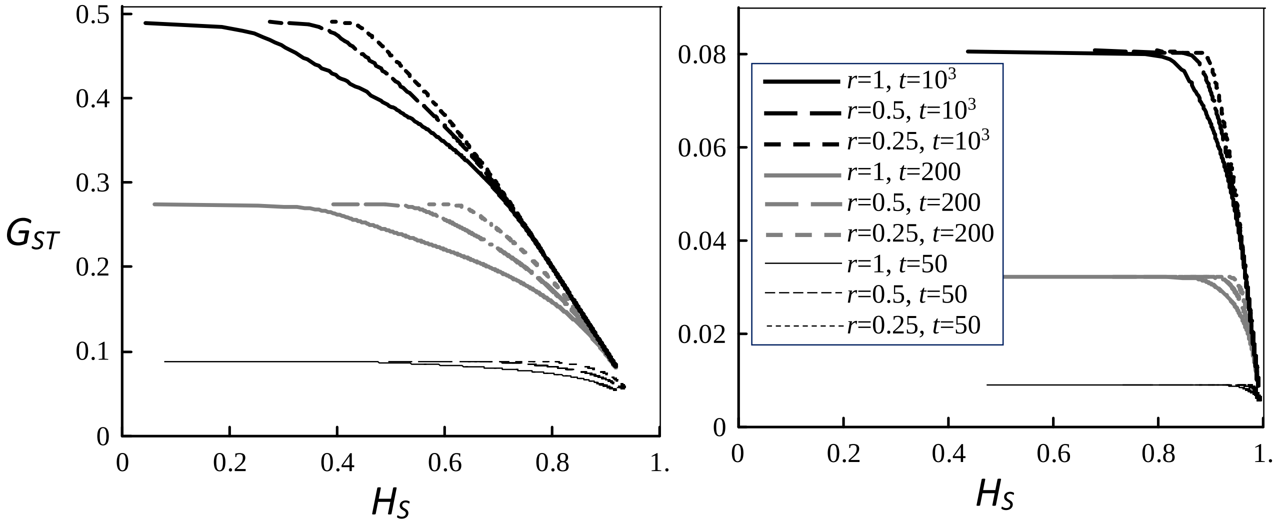

Fig. 2 G_{ST} as a function of H_S in non-equilibrium populations. The G_{ST} (y axis) values are plotted against H_S (x axis) values at different generations (t=50, 200, 1000) for a population in a finite island model with s=10 subpopulations, assuming parameter values of N=250 (left panel) or 1000 (right panel), m=0.001, and a variable u ranging from 10^{-6} (left side of x axis) to 10^{-2} (right side of x axis). The initial probability of gene identity is assumed to be $rJ_{0(\propto)}$, where r=1, 0.5 and 0.25 and $J_{0(\propto)}$ is the equilibrium value of J_0 given parameters N, m, u, s.

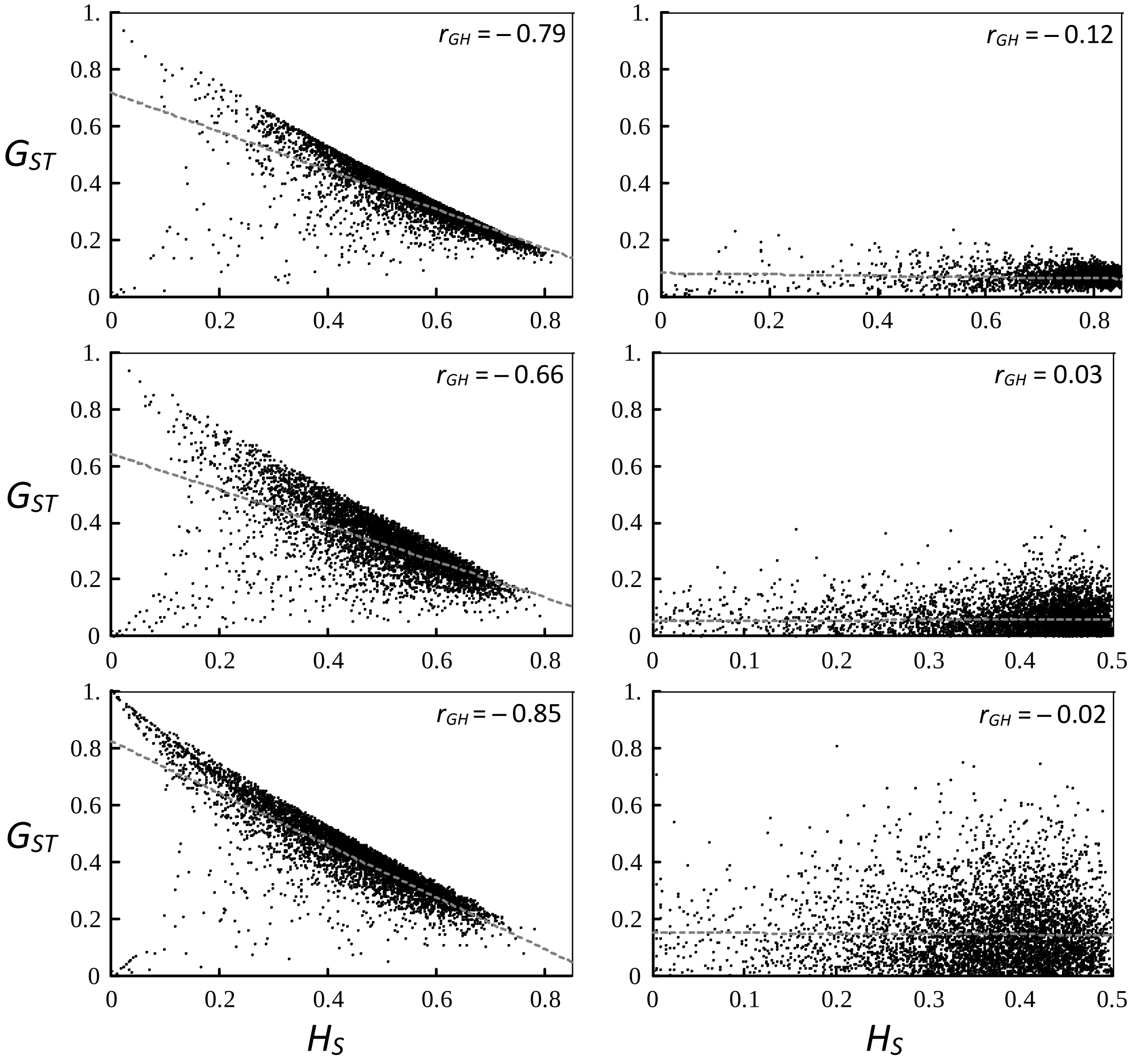

Fig. 3 Scatter graphs of G_{ST} (y axis) and H_S (x axis) estimates in mutation-drift-migration equilibrium populations. The population parameters are N=250, s=20, u is taken at random from a uniform distribution in the range [10^{-5}, 10^{-3}], and migration rate is either m=0.001 (left column) or m=0.01 (right column). The population is assumed to follow the finite island and infinite allele models (first row), finite island and stepwise mutation models (second row), or stepping stone and stepwise mutation models (third row). For each graph, 5000 replicate simulated datasets (loci) were generated to estimate G_{ST} and H_S, using R=4 (out of s=20) randomly sampled subpopulations and M=50 (out of N=250 or 1000) randomly sampled individuals per subpopulation. The correlation between the G_{ST} and H_S estimates for each graph is shown at the right corner of the graph.

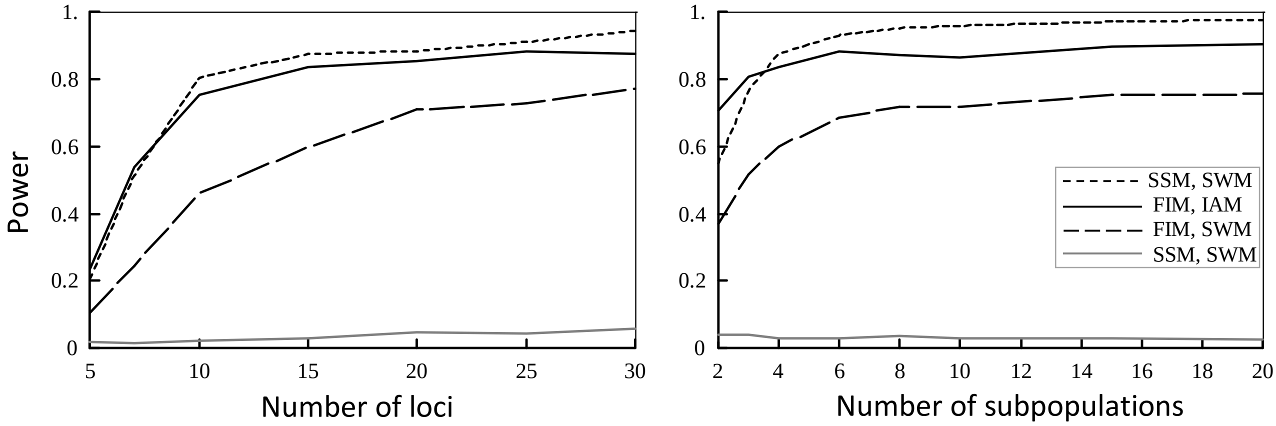

Fig. 4 Power of correlation analysis of G_{ST} and H_S estimates in mutation-drift-migration equilibrium populations. The population parameters are N=250, s=20, and u is taken at random from a uniform distribution in the range [10^{-5}, 10^{-3}]. The numbers of sampled subpopulations and loci are 4 and variable for the left panel, or variable and 15 for the right panel. Migration rate is either m=0.001 (black continuous, black broken and black dotted lines) or m=0.01 (grey continuous lines). The population is assumed to follow the finite island model (FIM) and infinite allele model (IAM), finite island model and stepwise mutation model (SWM), or stepping stone model (SSM) and stepwise mutation model. For each parameter combination, the proportion of 1000 replicate datasets in which the correlation coefficient between G_{ST} and H_S, estimated using 40 individuals per sampled subpopulation, is statistically significant at p<0.001 is plotted (on y axis) as a function of the number of sampled loci (left panel) or the number of sampled subpopulations (right panel) (on x axis). The black lines show the power in detecting mutational effects on G_{ST} when such effects exist (i.e. when migrations are weak relative to mutations, m=0.001), and the grey lines show the false detection rates when mutational effects are absent (i.e. when migrations are strong relative to mutations, m=0.01).

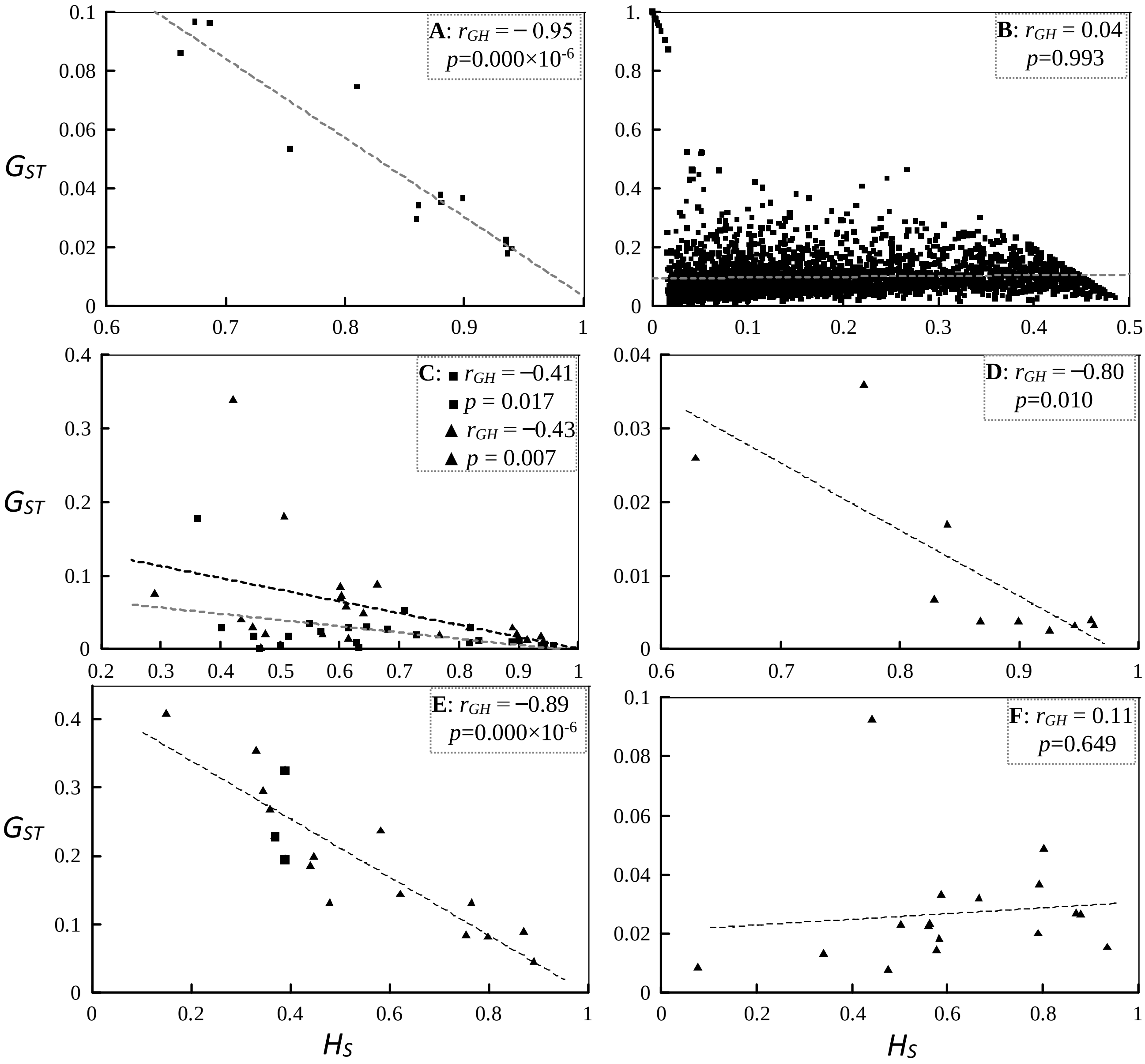

Fig. 5 The relationship between single locus G_{ST} and H_S estimates in empirical datasets. The correlation coefficient between G_{ST} and H_S and the p value for each dataset are shown at the top right corner of each graph, and the grey dotted lines show the fitted regression of G_{ST} on

H_S. Graphs A and B show the results for the 15 microsatellites and 3129 SNPs respectively in North American Atlantic salmon populations. Graph C shows the results for the 23 microsatellites in the blacknose shark populations, where each triangle and each square shows the pair of G_{ST} and H_S values estimated from a single marker in the 2- and 10-population models, respectively. Graph D shows the results for the 10 microsatellites in eight Mediterranean shore crab populations. Graph E shows the results for the 17 microsatellites in 15 blacktip reef shark populations, where each triangle and each square represents a single marker without and with deviation from Hardy-Weinberg equilibrium. Graph F shows the results for the 17 microsatellites in 10 copper rockfish populations.