COMPUTER PROGRAM NOTE

A program to compare genetic differentiation statistics across loci using resampling of individuals and loci

BRYAN D. NEFF and BONNIE A. FRASER

Department of Biology, University of Western Ontario, 1151 Richmond Street, London, ON, Canada N6A 5B7

Abstract

Comparisons of genetic differentiation across populations based on different loci can provide insight into the evolutionary patterns acting on various regions of genomes. Here, we develop a program to statistically compare population genetic differentiation statistics (F_{ST} or G'_{ST}) calculated from different loci. The program employs a routine that resamples either or both of individuals and loci and calculates a bootstrap confidence interval in the statistics. Resampling individuals is important when fewer than 25 individuals are sampled per population and when confidence intervals are required for individual loci. Resampling loci provides confidence intervals for sets of loci, such as a set presumed to be neutral, but can be anticonservative if fewer than 20 loci are analysed. We demonstrate the program using previously published data on the genetic differentiation at a major histocompatibility complex locus and at microsatellite loci across 10 populations of the guppy (*Poecilia reticulata*).

Keywords: bootstrapping, F_{ST}, G'_{ST} microsatellite, G_{ST}, major histocompatibility complex, population structure

Wright's F_{ST} and its variants are used extensively in the field of molecular ecology to assess population genetic differentiation and structure. These statistics partition genetic diversity to within- and between-populations with the intent of calculating the proportion of the total diversity that can be found among populations (Wright 1951). In theory, F_{ST} varies between 0 and 1, where a value of 0 implies that there is no diversity among populations and hence no population genetic structure, whereas a value of 1 implies that all diversity lies among populations and there is maximal population genetic structure (i.e. all populations are genetically unique). F_{ST} indices have been widely used in studies of population genetics, conservation genetics and phylogeography.

Comparisons of F_{ST}-estimates calculated from different loci across a set of subpopulations can provide insight into the evolutionary patterns acting on various regions of genomes. For example, F_{ST}-estimates based on the loci of the major histocompatibility complex (MHC), which code for peptides involved in the vertebrate immune response, have been compared to estimates

based on microsatellite loci, which are presumed to be effectively neutral (reviewed by Bernatchez & Landry 2003). In one such study, Sommer (2003) found that population differentiation at the MHC was lower than that observed at microsatellite loci in two populations of Malagasy giant jumping rats (*Hypogeomys antimena*) and inferred that balancing selection was occurring at the MHC. Balancing selection is thought to make populations more similar by selecting for rare migrants and therefore increasing the effective migration rate (Schierup *et al.* 2000; Muirhead 2001). Conversely, Landry & Bernatchez (2001) found populations of Atlantic salmon (*Salmo salar*) occurring in different habitat types had higher than expected population divergence at the MHC when compared to microsatellite loci and concluded that varying selection pressure from the different habitat types was occurring at the MHC.

Comparisons of F_{ST}-estimates across loci generally fall into two categories comprising model- and empirical-based approaches (Luikart *et al.* 2003). Model-based approaches rely on a likelihood function or a simulated null (neutral) distribution of F_{ST}-values (e.g. Beaumont & Nichols 1996; Beaumont & Balding 2004), whereas empirical-based approaches rely on an observed null distribution of F_{ST}-values or generate the distribution

Correspondence: Bryan D. Neff, Fax: +1 519 850 2532; E-mail: bneff@uwo.ca

using a bootstrap approach (e.g. Goudet 2001). With both approaches, a locus that may be under selection is identified when a corresponding F_{ST}-estimate falls outside of the null distribution; e.g. outside the 95% confidence interval of F_{ST}-values associated with neutral loci. The disadvantage of model-based approaches is that they are dependent on assumptions of population parameters and mutational dynamics that can bias the likelihood function or simulated distribution of neutral loci (Storz 2005). Empirical-based approaches, on the other hand, typically rely on resampling loci presumed to be neutral to generate confidence intervals around the corresponding F_{ST}-values. Those confidence intervals can be anticonservative or have discrete or otherwise peculiar properties when 20 or fewer loci are used in the analysis (Lewontin & Krakauer 1973; van Dongen 1995; also see Whitlock 2008). Conversely, by not resampling across individuals, a second source of error on the F_{ST}-estimate is ignored. Such sampling error can be particularly important when fewer than 25 individuals are sampled from within populations (Beaumont & Nichols 1996).

There has been considerable debate over the accuracy of Wright's original F_{ST}-formulation and several derivatives of the index have been proposed (e.g. Hedrick 2005; Jost 2008; Heller & Siegismund 2009). The debate has centred on the accuracy of the index and its derivatives at calculating the proportion of total genetic diversity that is found among populations. For example, Hedrick (2005) concisely shows that Nei's (1973) commonly used formulation of Wright's F_{ST} (referred to as G_{ST}) is always $\leq 1 - H_S$, where H_S is the average Hardy–Weinberg heterozygosity across subpopulations. Thus, although F_{ST} should range from 0 to 1, it can be constrained to values much lower than 1 when loci that are highly polymorphic within populations are used. Hedrick (2005) proposed a variant of G_{ST} that he called G'_{ST}. Hedrick's index simply divides G_{ST} by a correction factor, $G'_{ST(max)}$, which represents that theoretical maximum G_{ST}-value based on a given locus. The correction factor thus standardizes the G_{ST}-estimate for within-population locus variability and ensures that the values of G'_{ST} range from 0 to 1.

In this study, we develop a program that uses the empirical-based approach to compare an F_{ST}- or G'_{ST}-estimate from a locus presumed to be under selection to a null distribution based on neutral loci. The null distribution is generated by resampling either or both of individuals and loci. We use the F_{ST}-formulation developed by Ronfort *et al.* (1998) for autotetraploid species (their equation 26). The formulation is based on the diploid analysis of Cockerham (1969, 1973) and uses an analysis of variance approach. It can be used to calculate F_{ST} for either diploid or tetraploid species by defining the ploidy level. Our program can also be used to conduct a comparison using values corrected for locus variability. This

correction is accomplished by dividing the F_{ST}-values by $G'_{ST(max)} = (1 - H_S)$ (see equation 3 in Hedrick 2005).

The program begins by calculating F_{ST} or G'_{ST} for each of two locus-types and for each possible population pairwise comparison. The program then resamples individuals or loci or both as defined by the user. Individuals are resampled from within populations with replacement until the original sample sizes for each population are produced. Similarly, loci are resampled with replacement until the original number of loci is matched (the same set of resampled loci are used for all populations). Using the resampled data set, the two statistics are again calculated for each locus-type. For each population pair, a comparison is then made between the F_{ST}- or G'_{ST}-estimates for the two locus-types, and the locus-type associated with the higher or lower value is recorded. The routine is repeated for a total of 1000 replicates, from which the mean and median are calculated and the 95% confidence interval is determined from the 25th and 975th value in a ranked list. The proportion of comparisons in which one locus-type was either higher or lower than the other type is also reported. This proportion can serve as a one-tailed *P*-value for the null hypothesis that one locus-type is either higher or lower than the other type. The program was written in the C++ programming language and is available as a downloadable executable file from http://publish.uwo.ca/~bneff/software.htm#Fst. There is also a help file and sample input files available on the website.

To demonstrate the program, we use data from the MHC class IIB and six microsatellite loci from 10 populations of the guppy (*Poecilia reticulata*). Details of the genotyping and sequencing methods can be found in Fraser *et al.* (2009) and Suk & Neff (2009). The guppy is a tropical freshwater fish native to Trinidad and neighbouring portions of northeastern South America (Magurran 2006). The 10 populations were collected from five rivers and three drainages in the northern range of Trinidad. Populations within rivers encompass upper and lower regions, which are separated by physical barriers, such as waterfalls, that substantially reduce gene flow and create distinct genetic clusters (Crispo *et al.* 2006). Furthermore, two of the drainages that we collected from, the Caroni and Oropouche, diverged an estimated 2.5 Ma (Magurran 2006). Thus, the 10 populations span a wide range of divergence times as well as degrees of connectivity, which can affect gene-flow patterns.

We use our program to calculate and statistically compare F_{ST}- and G'_{ST}-estimates from the two locus-types for all 45 pairwise combinations of the 10 populations. We have previously reported on comparisons using the F_{ST}-estimator and resampling of individuals only (Fraser *et al.* 2009). We now also examine resampling by loci only and by both individuals and loci, and we calculate both F_{ST} and G'_{ST}. The data are used to illustrate the effect of

each sampling approach and estimator on the variance in the estimates as well as the proportion of the population pairwise comparisons in which the MHC estimate is identified as significantly different from that of the microsatellite loci. A discussion of the evolutionary significance of the differences in the population genetic structure of the MHC and microsatellite loci is presented in Fraser *et al.* (2009). Here, we instead focus on the effect of the resampling level (individuals and/or loci) and the estimator (F_{ST} or G'_{ST}) on the variance and statistical comparisons across locus-types.

We found 43 different MHC class IIB alleles in 142 individuals across the 10 populations (see Figure 2 in Fraser *et al.* 2009). Within populations, the mean number of MHC class IIB alleles was 9.5 (range: 4–15), mean expected heterozygosity was 0.62 (range: 0.23–0.85), and mean observed heterozygosity was 0.42 (range: 0.14–0.64) (Table 1). For the six microsatellites, the mean number of alleles within populations was 7.3 (range: 2–11), mean expected heterozygosity was 0.63 (range: 0.15–0.85), and mean observed heterozygosity was 0.55 (range: 0.14–0.75) (Table 1). Across populations, the observed heterozygosity at the MHC was highly correlated with the mean heterozygosity at the microsatellites ($r = 0.71$, $n = 10$, $P = 0.022$), albeit it was marginally higher at the microsatellites than at the MHC (paired t-test: $t_9 = 2.73$, $P = 0.023$).

The distribution of the sizes of the 95% confidence intervals for the microsatellite F_{ST}- and G'_{ST}-estimates from the 45 population pairwise comparisons is presented in Fig. 1. For both estimators, the size of the confidence intervals differed among the resampling approaches with the smallest intervals occurring with resampling individuals, intermediate intervals with resampling of loci, and
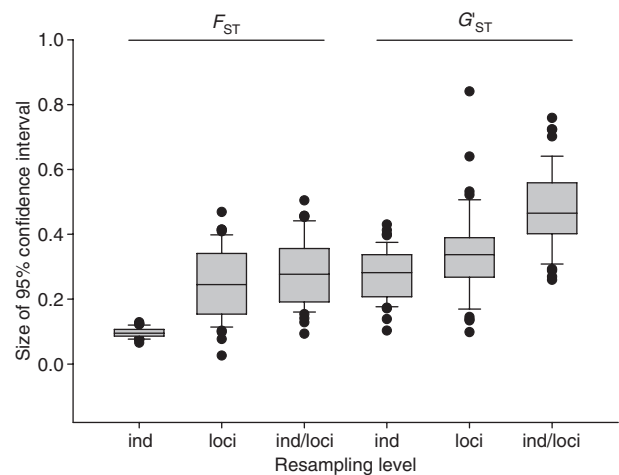

Fig. 1 Bootstrap resampling level and sizes of the 95% confidence intervals for F_{ST}- or G'_{ST}-estimates from population pairwise comparisons in the guppy (*Poecilia reticulata*). The three sampling levels comprise individuals, loci and both individuals and loci. The box plots denote 25, 50 and 75 percentiles, whiskers denote the 10 and 90 percentiles, and filled circles denote data outside the 10–90 percentile range. The plots are based on the 45 pairwise comparisons of 10 populations from which the size of the confidence interval was calculated as the difference between the 97.5 and 2.5 percentiles.

the largest intervals with resampling of both individuals and loci (ANOVA: F_{ST}, $F_{2,132} = 54.7$, $P < 0.001$; G'_{ST}, $F_{2,132} = 39.1$, $P < 0.001$). Indeed, for the F_{ST}-estimator, the confidence intervals associated with resampling loci were more than double than those associated with resampling individuals (Fig. 1).

Focusing on the data from resampling individuals (MHC) and both individuals and loci (microsatellites), we found that the F_{ST} of the MHC class IIB and F_{ST} of the

Table 1 Summary of the MHC class IIB locus and six microsatellite loci sampled in 10 populations of the guppy (*Poecilia reticulata*)

	LA	UA	LG	UG	LQ	UQ	LT	UT	LY	UY
Location	PS 938	PS 931	PS 909	PS 893	PS 971	PS 970	QS 002	QS 003	PS 807	PS 833
	786	817	744	848	805	806	784	809	948	876
MHC										
N	16	15	11	14	15	13	13	17	14	14
K	14	12	13	4	6	8	9	15	9	5
H_E	0.85	0.59	0.78	0.29	0.66	0.64	0.74	0.80	0.63	0.23
H_O	0.56	0.53	0.64	0.14	0.33	0.54	0.38	0.47	0.43	0.21
Microsatellite loci										
N	21	19	19	19	38	39	21	19	20	19
K	9.5	5.0	9.6	2.0	11.3	7.8	9.8	7.5	7.5	2.8
H_E	0.78	0.46	0.85	0.15	0.79	0.70	0.78	0.78	0.69	0.35
H_O	0.63	0.46	0.69	0.14	0.72	0.60	0.63	0.75	0.55	0.28

For each population, data include map coordinates, number of individuals (N), number of MHC class IIB or average number of microsatellite alleles (K), expected heterozygosity (H_E) and observed heterozygosity (H_O). The populations are lower Aripo (LA), upper Aripo (UA), lower Guanapo (LG), upper Guanapo (UG), lower Quare (LQ), upper Quare (UQ), lower Turure (LT), upper Turure (UT), lower Yarra (LY) and upper Yarra (UY).

microsatellites were positively correlated across the 45 pairwise comparisons ($r = 0.37$, $n = 45$, $P = 0.011$). Similarly, G'_{ST}-estimates of the MHC were positively correlated with the G'_{ST}-estimates of the microsatellites, albeit the correlation was not significant ($r = 0.19$, $n = 45$, $P = 0.20$). As expected, the mean F_{ST} from the 45 pairwise comparisons was significantly lower than the mean G'_{ST}-value for both the MHC (mean $F_{ST} = 0.092 \pm 0.058$ SD; mean $G'_{ST} = 0.16 \pm 0.09$; paired t-test: $t_{44} = 12.1$, $P < 0.001$) and the microsatellites (mean $F_{ST} = 0.31 \pm 0.17$; mean $G'_{ST} = 0.66 \pm 0.20$; paired t-test: $t_{44} = 23.4$, $P < 0.001$). Furthermore, the majority of the F_{ST}- and G'_{ST}-values calculated from the MHC were less than those from the microsatellites (Fig. 2). Twenty-nine of the 45 (64%) population pairwise MHC F_{ST}-estimates were significantly lower ($P < 0.05$ based on the randomization routine) than the F_{ST}-estimates for the microsatellite loci. Similarly, 38 of the 45 (84%) MHC G'_{ST}-estimates were significantly lower than the G'_{ST}-estimates for the microsatellite loci. A greater number of the pairwise comparisons were significant in the latter analysis because the locus variability was marginally higher at the microsatellites than the MHC. Thus, the correction factor employed in the G'_{ST}-estimates increased the original F_{ST}-estimates more so for the microsatellites than the MHC. By comparison, when only individuals were resampled, the MHC F_{ST} and G'_{ST} were significantly lower than the microsatellite estimates in, respectively, 33 (76%) and 40 (89%) of the 45 pairwise comparisons, and when only loci were resampled, 44 (98%) and 45 (100%) of the comparisons were significant for the F_{ST} and G'_{ST} respectively (data not shown).

Our analysis highlights several properties of the resampling level on the variance in F_{ST}-estimates. First, as expected, the variance in the estimates associated with resampling loci was large because we utilized only six microsatellite loci. Lewontin & Krakauer (1973) predicted that the variance among neutral loci should be $2 \times \overline{F_{ST}^2}$ for pairwise comparisons, which across our 45 comparisons gives a mean expected variance of 0.23. Examining the variance associated with the resampling of loci as an estimate of the among locus variance, we found it to be consistent with the prediction at 0.24. Second, variance introduced by resampling individuals was relatively small compared to the variance from resampling loci. Beaumont & Nichols (1996) used coalescent simulations of the Island Model to examine the effect of sample size on the variance in F_{ST}-estimates and concluded that there is little value in sampling more than 25 individuals per population. Conversely, Petit & Pons (1998) used variance estimators and numerical data to show that resampling individuals can provide inflated estimates of the true variance and thus can be anticonservative (the authors did not determine the expected bias associated
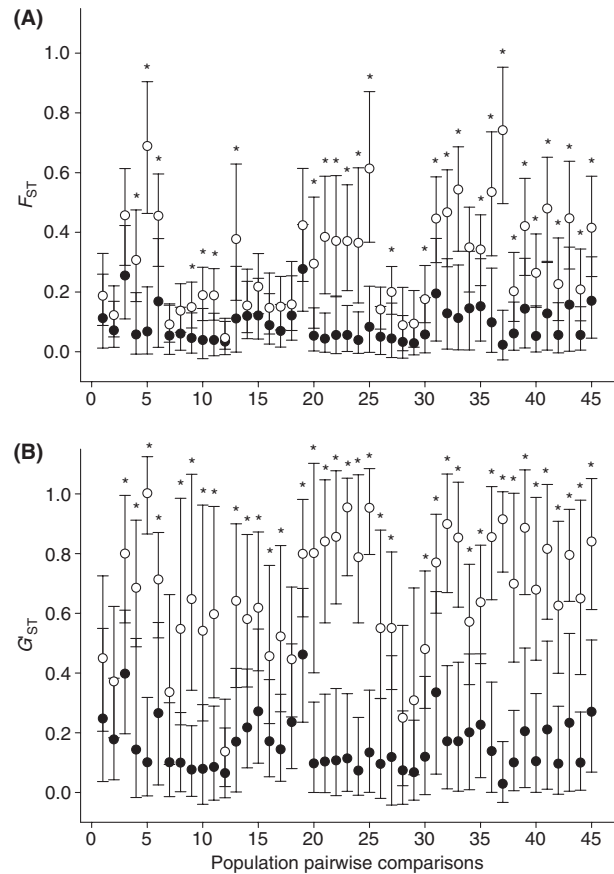

Fig. 2 Comparison of population genetic differentiation based on neutral loci (six microsatellite loci) and MHC class IIB for each pairwise comparison of 10 populations of the guppy (*Poecilia reticulata*). Mean F_{ST} (A) or G'_{ST} (B) values are presented with the microsatellite loci denoted by open circles and the MHC class IIB denoted by filled circles. Error bars indicate 95% confidence intervals as estimated by resampling of individuals (MHC) or both individuals and loci (microsatellites) with replacement 1000 times (*$P < 0.05$). The 45 pairwise population comparisons are (see Table 1, for population names): LA–UA, LA–LG, LA–UG, UA–LG, UA–UG, LG–UG, LQ–UQ, LQ–LT, LQ–UT, UQ–LT, UQ–UT, LT–UT, LY–UY, LA–LQ, LA–UQ, LA–LT, LA–UT, LA–LY, LA–UY, UA–LQ, UA–UQ, UA–LT, UA–UT, UA–LY, UA–UY, LG–LQ, LG–UQ, LG–LT, LG–UT, LG–LY, LG–UY, UG–LQ, UG–UQ, UG–LT, UG–UT, UG–LY, UG–UY, LQ–LY, LQ–UY, UQ–LY, UQ–UY, LT–LY, LT–UY, UT–LY and UT–UY.

with resampling loci). Regardless, our samples sizes within populations averaged 23 (range: 19–39) for the microsatellite loci and thus sampling variance from individuals was expected to be low. Indeed, the variance averaged only 0.0006, which was <0.5% of the expected microsatellite F_{ST}-values.

Our analysis also suggests that caution is warranted when comparing uncorrected F_{ST}-values across locus-types, especially when variance is estimated from resampling only loci. Hedrick (2005) has previously shown that F_{ST}-estimates are restricted to a maximum value of $1 - H_S$

instead of the expected maximum value of 1. In our study, although H_S was similar at the MHC and microsatellite loci, it was significantly higher at the latter loci. Consequently, the correction increased the F_{ST}-estimates associated with the microsatellites more so than the MHC, and population pairwise comparisons revealed an additional nine significant differences (38 vs. 29 of 45 comparisons). Furthermore, if we had resampled only loci, all 45 comparisons would have been identified as being significantly different based on the G'_{ST}-estimator. This common approach of comparing MHC and microsatellite loci can be anticonservative because it assumes no error in the MHC estimate (i.e. a single value for differentiation at the MHC is used for each population pairwise comparison). Although our results are probably particularly sensitive to sampling variance at the MHC because of the relatively small sample sizes for this locus—we had a mean of 14 individuals per population—they serve to highlight the importance of resampling at both the individual and locus levels to avoid potentially false identification of loci under selection.

In conclusion, we have developed a computer program that allows the statistical comparison of genetic differentiation at different locus-types using population pairwise comparisons. The program employs a routine that resamples individuals and/or loci. Although such bootstrap approaches can be anticonservative when it comes to estimating parameter variance, we recommend resampling at both levels to incorporate both sources of variance in the confidence estimates. We have highlighted the effect of resampling at the various levels by examining differences in genetic differentiation at the MHC and microsatellite loci in 10 populations of the guppy.

Acknowledgements

References

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

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

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Cockerham CC (1969) Variance of gene frequencies. *Evolution*, **23**, 72–84.

Cockerham CC (1973) Analysis of gene frequencies. *Genetics*, **74**, 679–700.

Crispo E, Bentzen P, Reznick DN, Kinnison MT, Hendry AP (2006) The relative influence of natural selection and geography on gene flow in the guppies. *Molecular Ecology*, **15**, 49–62.

van Dongen S (1995) How should we bootstrap allozyme data? *Heredity*, **74**, 445–447.

Fraser BA, Ramnarine IW, Neff BD (2009) Selection at the MHC class IIB locus across guppy (*Poecilia reticulata*) populations. *Heredity*, doi: 10.1038/hdy.2009.99.

Goudet J (2001) *FSTAT: A program to estimate and test gene diversities and fixation indices, version 2.9.3*. Institut d'Ecologie, Universite de Lausanne, Laussane, Switzerland.

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

Heller R, Siegismund HR (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.

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

Landry C, Bernatchez L (2001) Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **10**, 2525–2539.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. *Genetics*, **154**, 175–195.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Magurran AE (2006) *Evolutionary Ecology: The Trinidadian Guppy*. Oxford University Press, Oxford.

Muirhead CA (2001) Consequences of population structure on genes under balancing selection. *Evolution*, **55**, 1532–1541.

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

Petit RJ, Pons O (1998) Bootstrap variance of diversity and differentiation estimators in a subdivided population. *Heredity*, **80**, 56–61.

Ronfort J, Janczewski E, Bataillon T, Rousset F (1998) Analysis of population structure in autotetraploid species. *Genetics*, **150**, 921–930.

Schierup M, Vekemans X, Charlesworth D (2000) The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetic Research*, **76**, 51–62.

Sommer S (2003) Effects of habitat fragmentation and changes of dispersal behaviour after a recent population decline on the genetic variability of noncoding and coding DNA of a monogamous Malagasy rodent. *Molecular Ecology*, **12**, 2845–2851.

Storz JF (2005) Using genome scans of NDA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Suk HY, Neff BD (2009) Microsatellite genetic differentiation among populations of the Trinidadian guppy. *Heredity*, **102**, 425–434.

Whitlock MC (2008) Evolutionary inference from Q_{ST}. *Molecular Ecology*, **17**, 1885–1896.

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