genes

MDPI

Opinion

Allele Frequency Difference *AFD*–An Intuitive Alternative to F_{ST} for Quantifying Genetic Population Differentiation

Daniel Berner

Department of Environmental Sciences, Zoology, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland; daniel.berner@unibas.ch; Tel.: +41-(0)-61-207-03-28

Received: 21 February 2019; Accepted: 12 April 2019; Published: 18 April 2019

Abstract: Measuring the magnitude of differentiation between populations based on genetic markers is commonplace in ecology, evolution, and conservation biology. The predominant differentiation metric used for this purpose is F_{ST}. Based on a qualitative survey, numerical analyses, simulations, and empirical data, I here argue that F_{ST} does not express the relationship to allele frequency differentiation between populations generally considered interpretable and desirable by researchers. In particular, F_{ST} (1) has low sensitivity when population differentiation is weak, (2) is contingent on the minor allele frequency across the populations, (3) can be strongly affected by asymmetry in sample sizes, and (4) can differ greatly among the available estimators. Together, these features can complicate pattern recognition and interpretation in population genetic and genomic analysis, as illustrated by empirical examples, and overall compromise the comparability of population differentiation among markers and study systems. I argue that a simple differentiation metric displaying intuitive properties, the absolute allele frequency difference *AFD*, provides a valuable alternative to F_{ST}. I provide a general definition of *AFD* applicable to both bi- and multi-allelic markers and conclude by making recommendations on the sample sizes needed to achieve robust differentiation estimates using *AFD*.

Keywords: genetic differentiation; minor allele frequency; population genetics; sample size; single-nucleotide polymorphism

1. Introduction

Biological studies measuring the magnitude of genetic differentiation between populations, for example to explore levels of gene flow between populations, to discover genome regions influenced by natural selection, or to inform decisions in conservation biology, are published on a daily basis. A differentiation metric used frequently in such work is F_{ST}, interpreted broadly as a measure of the proportion of the total genetic variation at a genetic locus attributable to differentiation in allele frequencies between populations [1]. F_{ST} was conceptualized in the middle of the last century as a descriptor of genetic structure among populations [2–4]. Over the subsequent decades, numerous estimators were developed to allow F_{ST} to be calculated with empirical genetic data, based on different assumptions about the sampled study populations and/or the mutation process of the genetic markers [3,5–14]. Aside from some controversy about how to best calculate F_{ST} with multi-allelic genetic markers such as microsatellites [14–21], the fundamental concept shared among the F_{ST} estimators is firmly established in population genetics and genomics; F_{ST} is currently among the most widely used statistics in these fields. In this note, I will argue that despite its popularity, F_{ST} has shortcomings that complicate the analysis of population differentiation, and that a powerful alternative differentiation metric is available.

2. Features of an Appropriate Differentiation Metric

To approach the problems inherent in F_{ST}, we will start from the very beginning and ask what properties a metric of genetic differentiation should exhibit. First, the scale of the metric should range from zero (no genetic differentiation among populations) to one (complete fixation for different alleles). This familiar scale greatly facilitates interpretation and allows for convenient comparisons of differentiation among genetic markers and study systems. F_{ST} estimators satisfy this scale criterion; they are generally designed to range from zero to one.

The second, perhaps even more crucial requirement of an appropriate differentiation metric is that it should show an intuitive and traceable relationship to the magnitude of genetic differentiation between populations, so a researcher can understand and interpret what they are measuring. But what should this relationship look like? The answer to this question cannot be derived from theory but depends on the needs and expectations of the researchers measuring differentiation among their study populations. To develop a sense for these expectations, I performed a qualitative survey involving a total of 15 haphazardly chosen colleague researchers (advanced postdocs and faculties) having years of experience in population genetics and/or evolutionary genomics, including both empiricists and theoreticians. I confronted these researchers with a graphic displaying a continuum of symmetrically increasing genetic differentiation between samples of nucleotides (n = 40) drawn from two hypothetical populations at a single-nucleotide polymorphism (SNP) (X-axis, ranging from no to complete differentiation). I then asked them to specify the corresponding magnitude of population differentiation (*Y*-axis) an ideal metric of differentiation should exhibit if such a metric was to be invented from scratch. Specifically, the respondents were presented in Figure 1a, with the upper panel of the figure left blank.

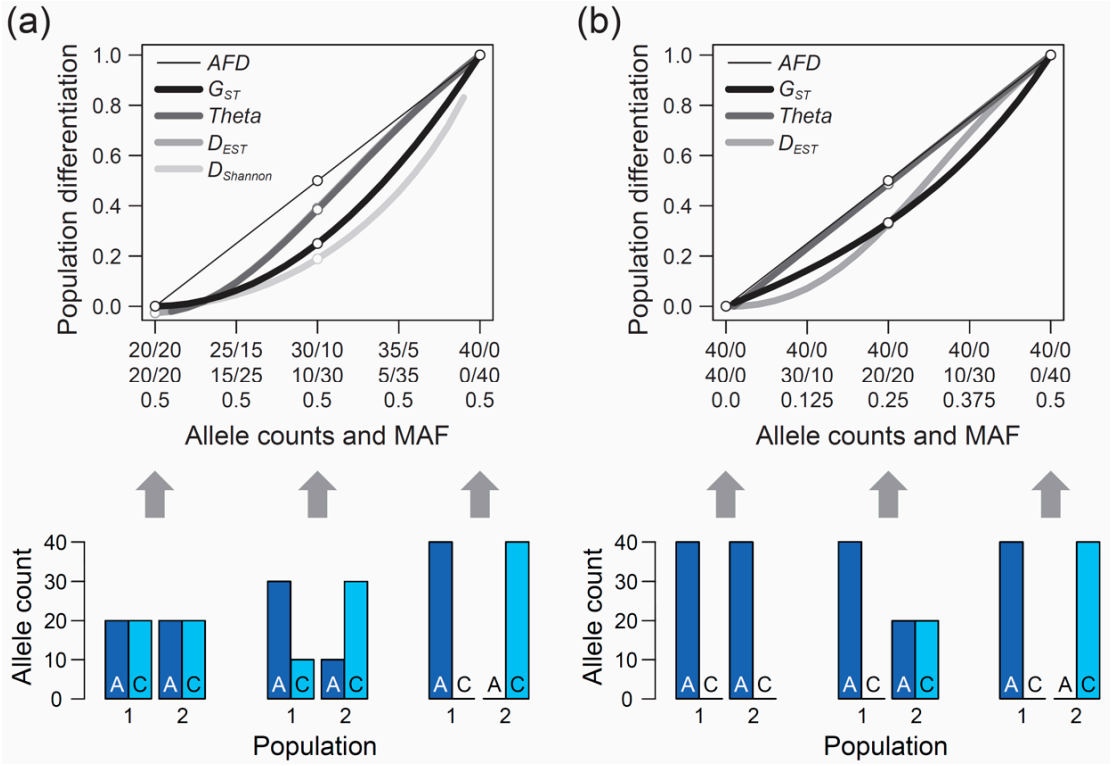

Figure 1. Population differentiation expressed by different metrics. Magnitude of genetic differentiation at a bi-allelic single-nucleotide polymorphism (SNP) along the continuum of allele frequency differentiation between two populations (top graphs). Differentiation is quantified by the absolute allele

frequency difference (*AFD*), by two popular estimators of F_{ST} (G_{ST} and *Theta*), by D_{EST}, and by Shannon differentiation ($D_{Shannon}$). The X-axis specifies the underlying allele counts in population 1 (first row) and population 2 (second row) for two hypothetical alleles (A, C), assuming a draw of 40 total alleles per population at the exact allele frequencies in each population (no sampling stochasticity). The third row gives the frequency of the less common SNP allele across the pool of the two population samples (i.e., the pooled minor allele frequency, MAF). The SNP is specified to exhibit a maximal MAF in (**a**) and a minimal MAF in (**b**) (for the latter, $D_{Shannon}$ is undefined mathematically). The bar plots on the bottom illustrate the counts of the two alleles for three levels of differentiation (none, intermediate, complete). Note that some metrics are undefined at the endpoints of the differentiation continuum, and that in (**a**), D_{EST} very closely approximates *Theta* and is therefore hidden.

Although considered a qualitative rather than formal investigation, and despite a modest sample size, this survey produced a clear result: among the 15 total researchers, 13 argued that the most intuitive differentiation metric would exhibit a linear relationship from zero to one along this continuum of allele frequency shifts between populations, as shown by the thin black line in Figure 1a. Exactly this relationship is expressed by the absolute allele frequency difference, hereafter *AFD*. For a single bi-allelic marker, *AFD* is easily obtained by arbitrarily defining one of the two alleles as the focal allele and calculating the absolute difference in the frequency of this allele between the populations. ('Allele proportion' would perhaps be a more precise term than 'allele frequency', but I will stick to the latter expression used traditionally in population genetics.) More generally, the calculation of *AFD* between two populations at a genetic polymorphism can be formalized as

$$AFD = \frac{1}{2}\sum_{i=1}^{n}\left|(f_{i1} - f_{i2})\right|$$

where n represents the total number of different alleles observed at the polymorphism, and the f_i-terms specify the frequency of allele i in the two populations (an analogous definition is given verbally in Reference [22]). This formula can also be applied to multi-allelic markers like microsatellites. The focus of this paper, however, lies on standard bi-allelic SNPs, given that this type of polymorphism has become the predominant genetic marker. A worked example of *AFD* calculation for both a bi-allelic SNP and a multi-allelic microsatellite is provided as Analysis S1 in the Supplementary Materials (for applications of *AFD* in recent genomic investigations see References [23–27]).

3. Some Problems with F_{ST}

As suggested above, a substantial proportion of researchers appear to find the linear relationship to continuous genetic differentiation exhibited by *AFD* particularly intuitive and interpretable. Note that throughout this paper, (non-)linearity refers only to the immediate relationship of a given differentiation metric to population allele frequencies, and hence does not imply any specific relationship of the estimator to biological factors influencing allele frequencies, such as gene flow, mutation, selection, population size, or divergence time. Now let us consider how F_{ST} behaves along the continuum of differentiation in allele frequencies. For this, we will initially focus on the two most popular F_{ST} estimators, G_{ST} [6] and *Theta* (θ) [8] (given in more accessible notation by the Formulas (8) and (6) in Reference [28]) and consider other metrics later. I emphasize that the insights emerging from these explorations may not be novel to researchers closely familiar with the theory underlying F_{ST}, but they are clearly under-appreciated by empiricists.

When the populations are undifferentiated genetically, G_{ST} is zero, as one would expect (Figure 1a, left end on the X-axis). Likewise, if the two population samples are monomorphic for alternative alleles, differentiation is at its maximum and G_{ST} exhibits the intuitive value of one (Figure 1a, right end on the X-axis). Between these extremes, however, the relationship between allele frequency change and G_{ST} is non-linear. Specifically, within the domain of low population differentiation, a unit increase in the frequency of the allele A in population 1 and a corresponding increase in the frequency

of C in population 2 causes a negligible increase in G_{ST}. A similar unit allele frequency change, however, drives a disproportionally large increase in G_{ST} when the populations are close to complete differentiation (Figure 1a). *Theta* shows qualitatively similar properties, although the deviation from *AFD* is less pronounced than for G_{ST}, except when differentiation is very weak.

The above numerical investigation assumes that the frequency of the minor (or less common) allele, as determined based on the *pool* of the two populations (hereafter MAF, for minor allele frequency), is maximal (i.e., 0.5 across the entire population differentiation range). It is instructive to also explore the behavior of our focal differentiation metrics when the MAF is minimal (i.e., one population is consistently monomorphic). Under this condition, the relationship between genetic differentiation and *AFD* remains straightforward to interpret. For instance, an allele frequency differentiation exactly intermediate between the absence of differentiation (i.e., both populations are monomorphic for the same allele) and the complete fixation for alternative alleles between the populations still yields an intuitive *AFD* value of 0.5 (Figure 1b). Reducing the MAF, however, has a strong influence on the F_{ST} estimators; their deviation from *AFD* declines. In particular, *Theta* now essentially coincides with *AFD* (the deviation of *Theta* from *AFD* under the full range of allele frequency combinations between the two populations is presented in Figure S1a).

Beside the MAF, the influence of sample size on metrics of differentiation also deserves attention. The formulas underlying the calculation of *AFD* and G_{ST} rely exclusively on allele frequencies and thus ignore the sample sizes used to estimate these frequencies. Therefore, the expected (parametric) value of these differentiation metrics is not dependent on sample size. (Empirical values derived from stochastic real-world samples, however, will be influenced by the precision underlying the estimation of allele frequencies, and hence by sample size, as elaborated in a separate section below). By contrast, the expected value of *Theta* at a given marker does depend on sample size. As long as sample sizes are similar between the two focal populations, the absolute size of these samples has a relatively minor influence on *Theta*, at least for typical (not very small) sample sizes used by empiricists (details not presented). However, imbalance in the size of the samples from the populations can have a dramatic influence on *Theta*. To appreciate this point, we assume that we sample nucleotides at a genetic marker from two populations exhibiting intermediate differentiation in allele frequencies (*AFD* = 0.5). Sample size for the first population is always constant (n = 40 nucleotides, as in Figure 1), whereas sample size for the second population is variable, ranging from 20 to 160 nucleotides. If the MAF is chosen to be minimal, we observe a dramatic decline in *Theta* as sample size for the second population increases (solid line in Figure 2). For example, all else equal, *Theta* declines from 0.59 to 0.42 when increasing sample size for the second population from 40 to 80 (equivalent to 20 and 40 diploid individuals). By contrast, choosing allele frequencies in the two populations such that the MAF is maximal, we find that the influence on *Theta* of sample size imbalance between the populations is reversed in direction, and weaker in magnitude (dotted line in Figure 2). Note that these effects are unrelated to sampling stochasticity, as we assume that our samples always mirror exactly the true population frequency (as in Figure 1).

Collectively, the above explorations allow us to draw a number of important conclusions regarding F_{ST}. First, F_{ST} generally displays a non-linear relationship to continuous population differentiation in allele frequencies (note that G_{ST} is sometimes claimed to display a perfectly linear relationship, and thus to coincide with *AFD*, after square-root transformation. This view is incorrect, as demonstrated in Figure S1b). This non-linearity has a more serious implication than just being unintuitive to many scientists: in several research fields using marker-based inference, small differences in the magnitude of genetic differentiation between population comparisons are highly relevant—and yet, this is exactly the domain in which F_{ST} is least sensitive (Figure 1a,b). For instance, observing average *AFD* of 0.05 versus 0.1 in two different population comparisons may point to an interesting difference between these population pairs in the opportunity for gene flow. However, when expressed as F_{ST}, the corresponding difference in the average magnitude of differentiation between the two population comparisons may appear marginal and not attract a researcher's attention. F_{ST} thus compromises the comparability

of differentiation among markers and among studies in the differentiation range most interesting to many empiricists. This point is reinforced by two investigators involved in the above survey having argued that although a linear increase in a differentiation metric along the range of continuous genetic differentiation appears ideal, one could also imagine a non-linear relationship with *elevated* sensitivity across the lower population differentiation range. F_{ST} behaves exactly opposite to this suggestion.

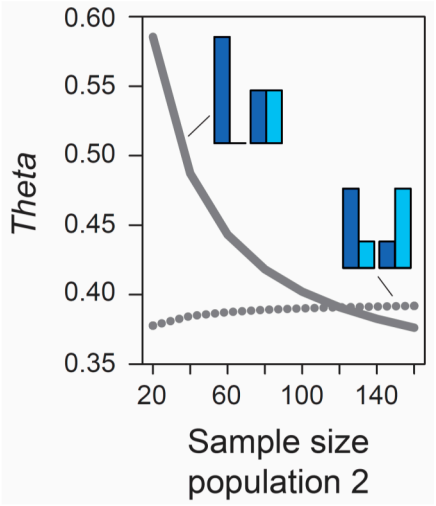

Figure 2. Influence of sample size imbalance between populations on the magnitude of the F_{ST} estimator *Theta* at a single SNP. Sample size for population 1 is always 40 nucleotides, as in Figure 1, but sample size for population 2 varies from 20 to 160 nucleotides (X-axis). Two different MAF levels are considered (minimal, solid line; maximal, dotted line). Sampling occurs deterministically at the exact population allele frequencies illustrated by the bar plots. For both MAF levels and across the full parameter range considered, the magnitude of allele frequency differentiation between the populations is therefore invariably perfectly intermediate (*AFD* is 0.5 throughout).

Our second conclusion is that genetic differentiation values expressed by F_{ST} are contingent on the MAF of the population system (see also References [29–31]). While there may be reasons justifying this behavior in specific analytical contexts, it will appear unintuitive (and often be unknown) to empiricists that a given magnitude of allele frequency differentiation yields different F_{ST} values depending on how balanced the two alleles are in the population pool.

The third conclusion is that at least some F_{ST} estimators, like *Theta*, are sensitive to the balance between populations in the number of nucleotides sampled at a given SNP. This property must be considered a serious nuisance to empiricists, especially those working with high-throughput sequencing data from population pools [32,33]. Even when controlling the number of study individuals in a population genomic experiment tightly, high-throughput sequencers inevitably generate variation in read depth among genomic positions, which will systematically inflate the variance in F_{ST} among markers and among studies when using *Theta* as estimator. Moreover, that the *expected* value of differentiation at a marker should shift when the samples drawn from the focal populations differ in size—even when the allele frequency estimates remain exactly the same—and that this influence of sample size imbalance on F_{ST} is itself contingent on the MAF, can hardly be considered intuitive.

4. Other Differentiation Estimators

So far, our reflections on F_{ST} were based on the commonly used estimators G_{ST} and *Theta*. A number of less widely applied alternative F_{ST} estimators have been introduced, however, hence it is valuable to examine if these estimators share with the former the weaknesses identified above. For this, I repeated the analysis of the magnitude of differentiation along the two continua of differentiation in allele frequencies visualized in Figure 1 with Wright's F_{ST} [3,5] (Formula (1) in Reference [16]) and Phi_{ST}

(Φ) [10]. For the case of a bi-allelic SNP considered here, these metrics yielded values identical to G_{ST}. Similarly, repeating the calculations with Hudson's F_{ST} [11] (Formula (10) in Reference [28]) produced results identical to *Theta*. Clearly, the problems identified for G_{ST} and *Theta* extend to the whole family of F_{ST} estimators. Moreover, D_{EST} (Formula (13) in Reference [14]), proposed as an alternative or complement to F_{ST}, exhibited differentiation values qualitatively similar to the F_{ST} estimators (Figure 1; see also References [34,35]).

Overall, the complexity inherent in F_{ST} (non-linearity, MAF- and sample size-dependence, difference among estimators) makes clear that F_{ST} was not designed as a simple descriptor of allele frequency differentiation. Instead, F_{ST} estimators aim at quantifying progress in population differentiation, or at partitioning genetic variation among hierarchical levels, in the light of specific models of mutation, gene flow, and drift [3,5,11,36–40]. Empiricists using F_{ST}, however, will rarely be aware of the underlying assumptions, and even for those who are, real-world situations will generally not allow evaluating what—if any—evolutionary model is meaningful for any given population pair, genome region, and marker analyzed (for a similar view see Reference [37]). It is therefore not surprising that the question of how F_{ST} should best be interpreted, and what constitutes the optimal F_{ST} estimator in the first place, has been a matter of debate for decades [1,14,16,20,21,29,34,37,40–42]. To conclude, it is not generally clear what quantity F_{ST} measures in empirical contexts—a view in line with the wide-spread sentiment of researchers that an intuitive differentiation metric should behave differently from F_{ST}. Replacing or complementing the complex theory-laden F_{ST} differentiation metrics by a traceable descriptor of differentiation independent from any specific population genetic model thus promises to facilitate the identification and interpretation of patterns in population differentiation, and to increase the comparability among studies and markers. The simple absolute allele frequency difference *AFD* appears adequate for this purpose.

As a potential alternative to *AFD*, I further considered a metric derived from information theory called Shannon differentiation (hereafter $D_{Shannon}$) that has recently been claimed to exhibit a 'straightforward relationship to allele frequency differences' [43] (see also References [44,45]). I explored how this novel metric behaves across the continua of allele frequency differentiation, which revealed two major shortcomings: first, $D_{Shannon}$ exhibits even less sensitivity than F_{ST} in the domain of weak to modest allele frequency differentiation between populations (Figure 1a). Consequently, $D_{Shannon}$ deviates even more strongly from the relationship to allele frequency differentiation considered desirable by many investigators. Second, $D_{Shannon}$ is undefined mathematically as soon as one population is monomorphic, that is, fixed for one allele (hence $D_{Shannon}$ is undefined across the entire differentiation continuum in Figure 1b). Given these problems, it appears doubtful that $D_{Shannon}$ will generally be considered a valuable differentiation metric and adopted widely for empirical analysis.

The latter conclusion also applies to ad hoc differentiation metrics based on *p*-values derived from statistical tests of differentiation between populations at genetic markers (e.g., References [46]). The disadvantage of such metrics is that we generally cannot easily translate a locus-specific *p*-value (i.e., the probability of an observed effect size) quantitatively into progress toward complete genetic differentiation (the effect size itself). In addition, *p*-values are a direct function of sample size, further reducing the comparability among markers and studies.

5. F_{ST} Can Complicate or Mislead the Biological Interpretation of Differentiation Data—Two Examples

In the previous section, the drawbacks of F_{ST} (and related metrics) were exposed based on simple numerical analyses. Given the ubiquity of F_{ST} in empirical research, I next illustrate implications of quantifying population differentiation by F_{ST} in real-world genetic analyses based on two examples from threespine stickleback fish (*Gasterosteus aculeatus* L.).

The first example re-uses SNP data generated through individual-level RAD sequencing (based on *Sbf1* enzyme restriction) in 28 female and 26 male stickleback from a single population inhabiting Misty Lake, Vancouver Island, Canada (the pooled lake and outlet samples from Reference [47];

for background information on this population see References [48–50]). We focus exclusively on SNPs (n = 200) located on chromosome 19, and we ask what distribution (visualized by a simple histogram) the differentiation between the sexes at SNPs along this chromosome will exhibit when quantified by the F_{ST} estimator G_{ST}, and by *AFD*. Our key observation is that both metrics indicate a bi-modal distribution of differentiation values, but that the high-differentiation mode is located in a lower differentiation range for G_{ST} (upper mode at around 0.25–0.3) than for *AFD* (upper mode near 0.5) (Figure 3a). The analytical relevance of this difference in the distribution of differentiation values becomes clear when considering that the threespine stickleback has a chromosomal XY sex determination system, and that the focal chromosome 19 represents the sex chromosome [51]. Crossover between the X and Y gametologs is restricted to a short segment of chromosome 19 [52,53]. Across the rest of the chromosome, the two gametologs represent completely isolated and deeply divergent populations. Consequently, the X and Y have reached (or are close to) fixation for distinct alleles at numerous SNPs. These gametolog-distinctive alleles cause the high-differentiation mode in both histograms, because the females (XX) are homozygous while the males (XY) are heterozygous at these SNPs (confirmed by inspecting allele frequencies in females and males at ten haphazardly chosen SNPs exhibiting *AFD* near 0.5; one example is presented within the box in Figure 3a). In other words, for any SNP allele private to the X, females tend to display a 100% frequency while males display a 50% frequency (note that SNPs with low male read coverage, indicating alignment problems for the Y-derived sequences, were excluded). Obviously, an intuitive differentiation metric—that is, a metric facilitating the understanding of the link between the magnitude of differentiation and the underlying biological cause—should yield a value of 0.5 for such a marker. While *AFD* shows this property, G_{ST} clearly impedes biological interpretation; to understand that the location of the upper differentiation mode in G_{ST} indicates a high abundance of SNPs with (nearly) X- and Y-limited alleles, one needs to be aware of the specific function linking allele frequency differences to G_{ST} (Figure 1a,b). Note that due to the peculiar allele distribution between the sexes, causing the MAF across the pool of the sexes to be minimal, *Theta* fortuitously approximates *AFD* in this specific empirical example (details not presented).

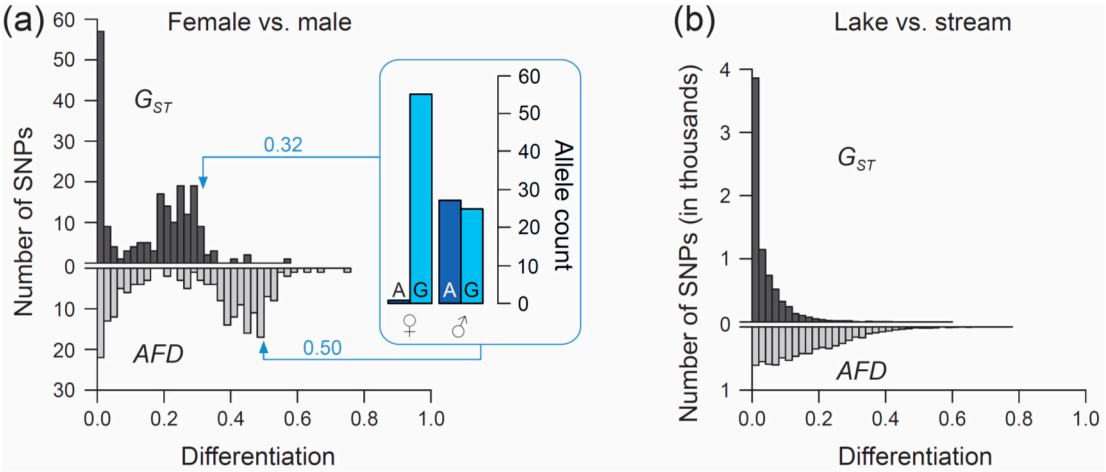

Figure 3. Performance of F_{ST} versus *AFD* in empirical analyses. (**a**) Distribution of the magnitude of differentiation, as measured by the F_{ST} estimator G_{ST} (upper histogram) and by *AFD* (lower histogram), between female and male threespine stickleback across 200 SNPs located on chromosome 19. The box visualizes the sex-specific allele counts at one exemplary SNP representative of the upper mode of each distribution, with G_{ST} and *AFD* for this marker given next to the arrows. (**b**) Distribution of G_{ST} and *AFD* values across 7282 genome-wide SNPs in a lake and stream stickleback population comparison.

For the second example, I re-use SNP data from a young stickleback population pair inhabiting ecologically different but adjacent lake and stream habitats in the Lake Constance basin in Central Europe.

Lake and stream stickleback in the Lake Constance basin occupy different foraging niches [54,55], are under divergent natural selection (as revealed by both marker-based divergence mapping and transplant experiments [56,57]), and show partial sexually based reproductive barriers [58]. We here perform a genetic comparison between the Lake Constance (n = 25 individuals) and the BOH stream (n = 22) population pair from Reference [56] (see also Reference [54]) and examine the distribution of both G_{ST} and *AFD* across genome-wide SNPs. For this, the original raw marker data set (generated by RAD sequencing based on *Nsi1* enzyme restriction; see Reference [56] for details) was subject to the following filters: first, I considered only bi-allelic SNPs represented by at least 32 nucleotides in each population. Second, a SNP was accepted only when located at least 12 nucleotide positions away from its nearest SNP, thus effectively excluding pseudo-SNPs caused by micro-indels within a RAD locus. Finally, only SNPs exhibiting a MAF of at least 0.45 across the population pool were considered. This latter filter dramatically reduced the data set (7282 SNPs remaining), but ensured that only those SNPs having the potential to span almost the full differentiation range entered analysis (i.e., markers with nearly maximal information content *sensu* [59]; markers with a low MAF are constrained to produce low differentiation values and thus bias the distribution of differentiation values, see also Reference [60]).

Comparing the genome scans in the lake-stream stickleback pair performed with G_{ST} and *AFD* as differentiation metrics reveals an important difference: the distribution of SNP-specific G_{ST} values has a striking mode near zero, tapering off steeply into a thin tail of higher differentiation values (strongly 'L-shaped' distribution, Figure 3b; see also Figure 3 in Reference [57]), whereas the *AFD* distribution is more uniform. Likewise, summary point estimates of differentiation differ substantially between the two metrics: genome-wide median G_{ST} is only around 0.02 whereas median *AFD* reaches 0.13, and the highest-differentiation SNP scores only 0.59 with G_{ST} but 0.77 with *AFD* (a scatterplot showing F_{ST} against *AFD* across all SNPs for this population comparisons is presented in Figure S2). These differences in the distribution of differentiation values between the metrics are important because they may stimulate qualitatively different biological interpretations: the F_{ST} distribution would commonly be taken as evidence that most of the genome is homogenized by gene flow between the adjoining populations, with substantial differentiation maintained by strong divergent selection in a few genome regions only [61,62]. But is such a mechanistic interpretation justified? A more cautious view is that for purely mathematical reasons (i.e., the lack of sensitivity in the low-differentiation domain), F_{ST} estimators will return a strongly L-shaped differentiation distribution for any population pair exhibiting weak differentiation—no matter what combination of evolutionary processes this differentiation reflects. Indeed, the *AFD* distribution suggests that appreciable lake-stream differentiation is widespread across the genome, thus questioning simple conclusions about the homogenizing effect of gene flow.

Together, these two empirical examples illustrate that using F_{ST} as a differentiation metric can complicate the recognition and/or interpretation of patterns in population differentiation. The examples further serve as a general warning that in the face of real-world biological complexity, differentiation data alone are unlikely to allow inferring underlying evolutionary processes reliably—no matter what differentiation metric is applied. Combining differentiation data with biogeographic and demographic evidence, and with insights from additional population genetic analyses, will generally be required.

6. AFD—Recommendations for the Application

Given the appeal of *AFD* emerging from both conceptual considerations and empirical analysis, it becomes relevant to explore under what conditions this differentiation metric performs adequately. *AFD* is detached from theoretical assumptions or specific population genetic models, hence the only concerns when estimating population differentiation are that the samples represent the focal populations reliably—an issue of study design, and that sample sizes are large enough to estimate allele frequencies within each population reasonably precisely. To provide a point of reference for the latter criterion, I simulated the consequences of sampling a focal population pair with different

intensities on estimates of *AFD*. Specifically, I modeled two populations with a precisely known allele frequency at a single SNP, choosing these frequencies such that the true parametric *AFD* value was 0, 0.05, 0.1, 0.25, or 0.5. Within each of these five scenarios of increasing population differentiation, I further considered up to five different MAF levels (0.025, 0.05, 0.125, 0.25, and 0.5) across the pool of the two samples, noting that with increasing differentiation, the range of possible MAFs decreases (e.g., with *AFD* = 0.5, the lowest possible parametric MAF is 0.25, see also Figure 1b). For each of the 19 total differentiation-by-MAF combinations considered, I then drew 10,000 replicate samples of equal size from each population, with sample size (i.e., the number of nucleotides) spanning the full range from 1 to 100, and calculated *AFD* between the populations for each replication (an analogous analysis based on G_{ST} is presented as Figure S3).

This analysis revealed that when sample size drops below around 20 nucleotides (corresponding to complete genotype data from 10 diploids) per population, *AFD* tends to become seriously biased upward (Figure 4, 'Simulation'). This bias is most pronounced when both the true magnitude of population differentiation is low and the MAF is high. The reason becomes evident when we assume a SNP (e.g., alleles A and C) completely undifferentiated between two populations (parametric *AFD* = 0) and exhibiting a maximal MAF of 0.5 (i.e., both alleles occur in perfectly balanced proportion in both populations), as in the left bar plot of Figure 1a. If we randomly draw just two nucleotides from each population at this SNP, it is not unlikely ($p = 0.125$) to draw two identical alleles from one population and two opposite alleles from the other, and hence to observe complete differentiation (*AFD* = 1). Such overestimation, however, is not possible when the populations are undifferentiated but the MAF is minimal (i.e., both populations are fixed for the same allele; left bar plot in Figure 1b), or when differentiation is complete (populations fixed for opposite alleles; right bar plots in Figure 1a,b). As a general recommendation, sample sizes of 40–60 nucleotides per population (20–30 diploids) should thus suffice to achieve reasonably accurate estimates of population differentiation, irrespective of the true magnitude and the MAF.

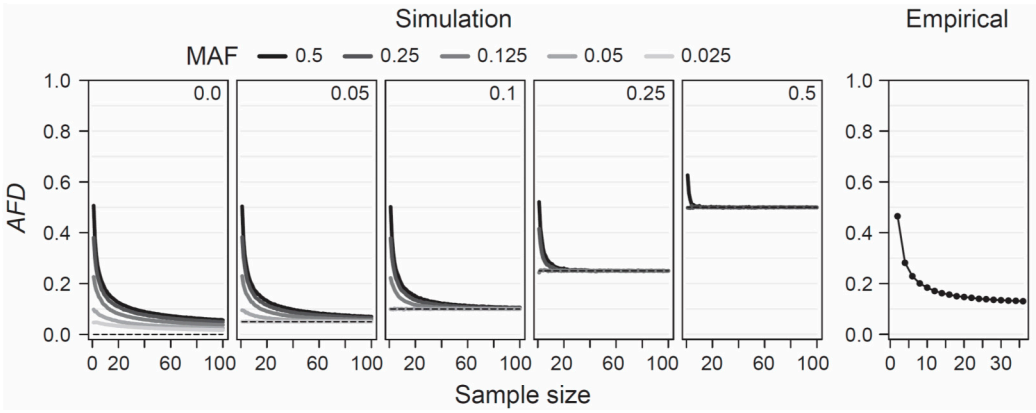

Figure 4. Sample size for *AFD*. Sensitivity of *AFD* to the size of the sample (number of nucleotides) taken from each population, explored by simulation (left) and using empirical population data (right). The simulations consider five different magnitudes of population differentiation (the true parametric differentiation is printed, and plotted as dashed line, inside each box), and up to five different MAF levels for each magnitude of differentiation (indicated by the gray shades of the lines). Note that with increasing differentiation, the possible MAF range becomes increasingly constrained. The lines show mean *AFD* across 10,000 replicate simulations for each sample size level. The empirical analysis shows mean *AFD* across the genome-wide SNPs from the lake-stream stickleback comparison shown in Figure 3b.

To examine this recommendation with empirical data, I again used the SNP data set from the lake-stream stickleback population pair described above. I here assessed how genome-wide mean *AFD* (and G_{ST}; Figure S3) changes when sampling both populations with a sample size ranging from 2 to 36.

Across all sample sizes, I restricted the SNP panel to those represented by a least 36 nucleotides in each population, and I considered only SNPs displaying a MAF of 0.05 or greater across the population pool (a MAF threshold of 0.2 lead to the same conclusions; details not presented). This empirical exploration was in good agreement with the insights from the simulation analysis: the genome-wide mean *AFD* value became relatively stable with sample sizes of around 20–30 nucleotides per population (Figure 4, 'Empirical'). I emphasize that all these conclusions regarding sample size are not specific to *AFD*; F_{ST} shows a very similar sensitivity to sample size and to the associated precision in population allele frequency estimation, as presented in Figure S3a,b.

As a final methodological remark, I highlight that this article has so far considered only the situation in which the number of populations to be compared is exactly two. Although such pairwise population comparisons arguably represent the most common analytical situation, it should be noted that F_{ST} statistics also permit estimating the overall genetic structure across a larger collection of populations. With *AFD* as a differentiation metric, this option is not available. A straightforward ad hoc solution, however, is to simply average multiple *AFD* values for SNPs or genome windows across the multiple population contrasts of interest [27].

7. Conclusions

The purpose of this note was to show that metrics of population differentiation used routinely in the analysis of genetic data—F_{ST} statistics and related metrics—do not necessarily measure the quantities most meaningful in genetic and genomic research. As a point in favor of F_{ST}, one may argue that its long tradition would promote the comparability of differentiation among studies [42]. This view, however, seems overly optimistic; F_{ST} is highly contingent on the specific estimator, is sensitive to the MAF spectrum of the markers, and sometimes to imbalances in sample size. Combined with the general insensitivity across the differentiation range most relevant in many analytical situations—weak population differentiation—F_{ST} falls short of being a reliable standard for measuring genetic differentiation. I argue that in many analytical contexts, the simple absolute allele frequency difference *AFD* will provide a sufficient, meaningful, and robust differentiation metric, thus promoting the discovery of patterns in differentiation, and their interpretation.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/4/308/s1, Figure S1: deviation of *Theta* and of the square-root of G_{ST} from *AFD*, Figure S2: relationship between G_{ST} and *AFD* for an empirical data set, Figure S3: sample size analysis for G_{ST}, Analysis S1: worked example for the calculation of *AFD*.

Funding: Financial support was provided by the Swiss National Science Foundation and by the University of Basel.

Acknowledgments: I am grateful to the many colleagues who responded to my survey and/or shared thoughts about F_{ST}: Simon Aeschbacher, Daniel Bolnick, Josh Van Buskirk, Dieter Ebert, Peter Fields, Simone Fior, Frédéric Guillaume, Christoph Haag, Lukas Keller, Andrew Hendry, Simon Martin, Katie Peichel, Joost Raeymaekers, Marius Roesti, Xavier Thibert-Plante, Claus Wedekind, and Yvonne Willy. William Sherwin offered advice on the calculation of Shannon differentiation. Marius Roesti, Telma G. Laurentino and five anonymous reviewers gave valuable feedback on the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Holsinger, K.E.; Weir, B.S. Genetics in geographically structured populations: Defining, estimating and interpreting FST. *Nat. Rev. Genet.* **2009**, *10*, 639–650. [CrossRef]
2. Wright, S. Isolation by distance. *Genetics* **1943**, *28*, 114–138.
3. Wright, S. The genetical structure of populations. *Ann. Eugen.* **1951**, *15*, 323–354. [CrossRef]
4. Malécot, G. *Les Mathématiques de L'hérédité*; Masson: Paris, France, 1948.
5. Wright, S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution* **1965**, *19*, 395–420. [CrossRef]

6. Nei, M. Analysis of Gene Diversity in Subdivided Populations. *Proc. Natl. Acad. Sci. USA* **1973**, *70*, 3321–3323. [CrossRef] [PubMed]
7. Nei, M.; Chesser, R.K. Estimation of fixation indexes and gene diversities. *Ann. Hum. Genet.* **1983**, *47*, 253–259. [CrossRef]
8. Weir, B.S.; Cockerham, C.C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **1984**, *38*, 1358–1370. [PubMed]
9. Lynch, M.; Crease, T.J. The analysis of population survey data on DNA sequence variation. *Mol. Boil. Evol.* **1990**, *7*, 377–394.
10. Excoffier, L.; Smouse, P.E.; Quattro, J.M. Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics* **1992**, *131*, 479–491.
11. Hudson, R.R.; Slatkin, M.; Maddison, W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **1992**, *132*, 583–589.
12. Slatkin, M. A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. *Genetics* **1995**, *139*, 457–462. [PubMed]
13. Holsinger, K.E. Analysis of genetic diversity in geographically structured populations: A Bayesian perspective. *Hereditas* **1999**, *130*, 245–255. [CrossRef]
14. Jost, L. G(ST) and its relatives do not measure differentiation. *Mol. Ecol.* **2008**, *17*, 4015–4026. [CrossRef] [PubMed]
15. Hedrick, P.W. Perspective: Highly Variable Loci and Their Interpretation in Evolution and Conservation. *Evolution* **1999**, *53*, 313–318. [CrossRef]
16. Hedrick, P.W. A standardized genetic differentiation measure. *Evolution* **2005**, *59*, 1633–1638. [CrossRef] [PubMed]
17. Balloux, F.; Brunner, H.; Lugon-Moulin, N.; Hausser, J.; Goudet, J. Microsatellites can be misleading: An empirical and simulation study. *Evolution* **2000**, *54*, 1414–1422. [CrossRef] [PubMed]
18. Heller, R.; Siegismund, H.R. Relationship between three measures of genetic differentiation GST, DEST and G'ST: How wrong have we been? *Mol. Ecol.* **2009**, *18*, 2080–2083. [CrossRef] [PubMed]
19. Ryman, N.; Leimar, O. GST is still a useful measure of genetic differentiation—A comment on Jost's D. *Mol. Ecol.* **2009**, *18*, 2084–2087. [CrossRef]
20. Meirmans, P.G.; Hedrick, P.W. Assessing population structure: FST and related measures. *Mol. Ecol. Res.* **2011**, *11*, 5–18. [CrossRef]
21. Whitlock, M.C. G'ST and D do not replace FST. *Mol. Ecol.* **2011**, *20*, 1083–1091. [CrossRef]
22. Shriver, M.D.; Smith, M.W.; Jin, L.; Marcini, A.; Akey, J.M.; Deka, R.; E Ferrell, R. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.* **1997**, *60*, 957–964. [PubMed]
23. Turner, T.L.; Bourne, E.C.; Von Wettberg, E.J.; Hu, T.T.; Nuzhdin, S.V. Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nat. Genet.* **2010**, *42*, 260–263. [CrossRef] [PubMed]
24. Stölting, K.N.; Paris, M.; Meier, C.; Heinze, B.; Castiglione, S.; Bartha, D.; Lexer, C. Genome-wide patterns of differentiation and spatially varying selection between postglacial recolonization lineages of Populus alba (Salicaceae), a widespread forest tree. *New Phytol.* **2015**, *207*, 723–734. [CrossRef] [PubMed]
25. Chen, J.; Källman, T.; Ma, X.-F.; Zaina, G.; Morgante, M.; Lascoux, M. Identifying Genetic Signatures of Natural Selection Using Pooled Population Sequencing in Picea abies. *G3 Genes Genomes Genet.* **2016**, *6*, 1979–1989. [CrossRef] [PubMed]
26. Westram, A.M.; Rafajlović, M.; Chaube, P.; Faria, R.; Larsson, T.; Panova, M.; Ravinet, M.; Blomberg, A.; Mehlig, B.; Johannesson, K.; et al. Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evol. Lett.* **2018**, *2*, 297–309. [CrossRef] [PubMed]
27. Haenel, Q.; Roesti, M.; Moser, D.; MacColl, A.D.C.; Berner, D. Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evol. Lett.* **2019**, *3*, 28–42. [CrossRef]
28. Bhatia, G.; Patterson, N.; Sankararaman, S.; Price, A.L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **2013**, *23*, 1514–1521. [CrossRef] [PubMed]
29. Charlesworth, B. Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Boil. Evol.* **1998**, *15*, 538–543. [CrossRef]

30. Noor, M.A.F.; Bennett, S.M. Islands of Speciation or Mirages in the Desert? Examining the Role of Restricted Recombination in Maintaining Species. *Heredity* **2009**, *103*, 439–444. [CrossRef] [PubMed]
31. Cruickshank, T.E.; Hahn, M.W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **2014**, *23*, 3133–3157. [CrossRef]
32. Ferretti, L.; Ramos-Onsins, S.E.; Perez-Enciso, M.; Ramos-Onsins, S.E.; Perez-Enciso, M. Population genomics from pool sequencing. *Mol. Ecol.* **2013**, *22*, 5561–5576. [CrossRef] [PubMed]
33. Gautier, M.; Foucaud, J.; Gharbi, K.; Cezard, T.; Galan, M.; Loiseau, A.; Thomson, M.; Pudlo, P.; Kerdelhue, C.; Estoup, A. Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Mol. Ecol.* **2013**, *22*, 3766–3779. [CrossRef] [PubMed]
34. Jost, L.; Archer, F.; Flanagan, S.; Gaggiotti, O.; Hoban, S.; Latch, E. Differentiation measures for conservation genetics. *Evol. Appl.* **2018**, *11*, 1139–1148. [CrossRef] [PubMed]
35. Alcala, N.; Rosenberg, N.A. G'ST, Jost's D, and FST are similarly constrained by allele frequencies: A mathematical, simulation, and empirical study. *Mol. Ecol.* **2009**, in press.
36. Reynolds, J.; Weir, B.S.; Cockerham, C.C. Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. *Genetics* **1983**, *105*, 767–779. [PubMed]
37. Nei, M. Definition and estimation of fixation indexes. *Evolution* **1986**, *40*, 643–645. [CrossRef]
38. Nagylaki, T. Fixation indices in subdivided populations. *Genetics* **1998**, *148*, 1325–1332.
39. Rousset, F. Genetic differentiation within and between two habitats. *Genetics* **1999**, *151*, 397–407.
40. Alcala, N.; Goudet, J.; Vuilleumier, S. On the transition of genetic differentiation from isolation to panmixia: What we can learn from GST and D. *Theor. Popul. Biol.* **2014**, *93*, 75–84. [CrossRef]
41. Weir, B.S.; Goudet, J. A Unified Characterization of Population Structure and Relatedness. *Genetics* **2017**, *206*, 2085–2103. [CrossRef]
42. Neigel, J.E. Is FST obsolete? *Conserv. Genet.* **2002**, *3*, 167–173. [CrossRef]
43. Sherwin, W.B.; Chao, A.; Jost, L.; Smouse, P.E. Information theory broadens the spectrum of molecular ecology and evolution. *Trends Ecol. Evol.* **2017**, *32*, 948–963. [CrossRef] [PubMed]
44. Dewar, R.C.; Sherwin, W.B.; Thomas, E.; Holleley, C.E.; Nichols, R.A. Predictions of single-nucleotide polymorphism differentiation between two populations in terms of mutual information. *Mol. Ecol.* **2011**, *20*, 3156–3166. [CrossRef]
45. Chao, A.; Jost, L.; Hsieh, T.C.; Ma, K.H.; Sherwin, W.B.; Rollins, L.A. Expected Shannon Entropy and Shannon Differentiation between Subpopulations for Neutral Genes under the Finite Island Model. *PLoS ONE* **2015**, *10*, e0125471. [CrossRef] [PubMed]
46. Turner, T.L.; Hahn, M.W.; Nuzhdin, S.V. Genomic Islands of Speciation in Anopheles gambiae. *PLoS Boil.* **2005**, *3*, e285. [CrossRef]
47. Roesti, M.; Hendry, A.P.; Salzburger, W.; Berner, D. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol. Ecol.* **2012**, *21*, 2852–2862. [CrossRef]
48. Lavin, P.A.; McPhail, J.D. Parapatric lake and stream sticklebacks on northern Vancouver Island: Disjunct distribution or parallel evolution? *Can. J. Zool.* **1993**, *71*, 11–17. [CrossRef]
49. Hendry, A.P.; Taylor, E.B.; McPhail, J.D. Adaptive divergence and the balance between selection and gene flow: Lake and stream stickleback in the misty system. *Evolution* **2002**, *56*, 1199–1216. [CrossRef]
50. Berner, D.; Adams, D.C.; Grandchamp, A.-C.; Hendry, A.P. Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J. Evol. Boil.* **2008**, *21*, 1653–1665. [CrossRef]
51. Peichel, C.L.; Ross, J.A.; Matson, C.K.; Dickson, M.; Grimwood, J.; Schmutz, J.; Myers, R.M.; Mori, S.; Schluter, D.; Kingsley, D.M. The Master Sex-Determination Locus in Threespine Sticklebacks Is on a Nascent Y Chromosome. *Curr. Boil.* **2004**, *14*, 1416–1424. [CrossRef]
52. Ross, J.A.; Peichel, C.L. Molecular Cytogenetic Evidence of Rearrangements on the Y Chromosome of the Threespine Stickleback Fish. *Genetics* **2008**, *179*, 2173–2182. [CrossRef] [PubMed]
53. Roesti, M.; Moser, D.; Berner, D. Recombination in the threespine stickleback genome-patterns and consequences. *Mol. Ecol.* **2013**, *22*, 3014–3027. [CrossRef]
54. Moser, D.; Roesti, M.; Berner, D. Repeated Lake-Stream Divergence in Stickleback Life History within a Central European Lake Basin. *PLoS ONE* **2012**, *7*, e50620. [CrossRef]
55. Lucek, K.; Sivasundar, A.; Seehausen, O. Evidence of Adaptive Evolutionary Divergence during Biological Invasion. *PLoS ONE* **2012**, *7*, e49377. [CrossRef]

56. Roesti, M.; Kueng, B.; Moser, D.; Berner, D. The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun.* **2015**, *6*, 8767. [CrossRef] [PubMed]
57. Moser, D.; Frey, A.; Berner, D. Fitness differences between parapatric lake and stream stickleback revealed by a field transplant. *J. Evol. Boil.* **2016**, *29*, 711–719. [CrossRef] [PubMed]
58. Berner, D.; Ammann, M.; Spencer, E.; Ruegg, A.; Luescher, D.; Moser, D. Sexual isolation promotes divergence between parapatric lake and stream stickleback. *J. Evol. Biol.* **2017**, *30*, 401–411. [CrossRef] [PubMed]
59. Roesti, M.; Salzburger, W.; Berner, D. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Boil.* **2012**, *12*, 94. [CrossRef]
60. Beaumont, M.A.; Nichols, R.A. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. London. Ser. B Boil. Sci.* **1996**, *263*, 1619–1626.
61. Feder, J.L.; Egan, S.P.; Nosil, P. The genomics of speciation-with-gene-flow. *Trends Genet.* **2012**, *28*, 342–350. [CrossRef] [PubMed]
62. Seehausen, O.; Butlin, R.K.; Keller, I.; Wagner, C.E.; Boughman, J.W.; Hohenlohe, P.A.; Peichel, C.L.; Saetre, G.P.; Bank, C.; Brännström, Å.; et al. Genomics and the origin of species. *Nat. Rev. Genet.* **2014**, *15*, 176–192. [CrossRef] [PubMed]