BMC Evolutionary Biology

METHODOLOGY ARTICLE Open Access

Uninformative polymorphisms bias genome scans for signatures of selection

Marius Roesti, Walter Salzburger and Daniel Berner*

Abstract

Background: With the establishment of high-throughput sequencing technologies and new methods for rapid and extensive single nucleotide (SNP) discovery, marker-based genome scans in search of signatures of divergent selection between populations occupying ecologically distinct environments are becoming increasingly popular.

Methods and Results: On the basis of genome-wide SNP marker data generated by RAD sequencing of lake and stream stickleback populations, we show that the outcome of such studies can be systematically biased if markers with a low minor allele frequency are included in the analysis. The reason is that these 'uninformative' polymorphisms lack the adequate potential to capture signatures of drift and hitchhiking, the focal processes in ecological genome scans. Bias associated with uninformative polymorphisms is not eliminated by just avoiding technical artifacts in the data (PCR and sequencing errors), as a high proportion of SNPs with a low minor allele frequency is a general biological feature of natural populations.

Conclusions: We suggest that uninformative markers should be excluded from genome scans based on empirical criteria derived from careful inspection of the data, and that these criteria should be reported explicitly. Together, this should increase the quality and comparability of genome scans, and hence promote our understanding of the processes driving genomic differentiation.

Keywords: Allele frequency distribution, F_{ST}, Gasterosteus aculeatus, Genetic marker, Hitchhiking, Population differentiation, Singleton

Background

A major challenge in evolutionary biology is to understand how natural selection acts on molecular genetic variation [1-4]. One approach to studying the consequences of selection at the genomic level is the application of genome scans that screen a collection of polymorphic genetic marker loci for their extent of differentiation between multiple (typically two) populations occupying ecologically distinct environments. Loci or genomic regions displaying particularly high population differentiation (usually quantified by an F_{ST} estimator [5]) relative to some differentiation baseline (reflecting primarily neutral drift) are interpreted as either being directly under divergent selection, or exhibiting genetic hitchhiking along with a quantitative trait locus (QTL) under divergent selection [6-9]. Genome scans therefore have the potential to illuminate the link between ecological selection and molecular variation, and hence to contribute to our understanding of adaptive diversification. This is particularly true if information from genome scans is integrated with complementary lines of evidence such as QTL mapping [10].

Genome scans can be performed in different ways, depending on the genomic resources available for a focal research system. On the one hand, reference-free (anonymous) scans are carried out without information on the physical genomic position of a marker locus. Here the F_{ST} value for each locus is treated as an independent data point and is evaluated against a baseline distribution derived from the entire data set *e.g.*, [11-14]. Loci exhibiting extreme F_{ST} values relative to the baseline ('outlier loci') are then interpreted as being directly or indirectly influenced by divergent selection. (Note that we here use divergent selection in a broad sense, including situations where an allele is selected in one environment but neutral in the other.) On the other hand, reference-based genome scans map loci physically to an

* Correspondence: daniel.berner@unibas.ch
Zoological Institute, University of Basel, Vesalgasse 1, Basel CH-4051, Switzerland

available genome *e.g.*, [15-18]. This offers a great advantage: loci occurring in the same genomic neighbourhood, and consequently exhibiting some physical linkage, will tend to display correlated F_{ST} values that can be integrated by taking a sliding window approach. This allows not only the identification of genomic regions displaying exceptionally high population differentiation, but also exploring the number and physical extent of such regions [3]. Moreover, depending on the marker resolution, outlier regions may be screened for candidate genes potentially targeted by divergent selection.

Inferences drawn from both reference-free and reference-based genome scans obviously depend on the availability of reliable polymorphism data. The objective of our study is to highlight a potential problem with polymorphism data sets that can introduce bias to genome scans and lead to incorrect interpretations of genomic differentiation, or the lack thereof. The problem lies in F_{ST} being sensitive not only to the extent of genetic differentiation among populations, but also to the allele frequency distribution. Specifically, very low F_{ST} values (*i.e.*, near zero, or even negative values, depending on the formula used for calculation) at a polymorphic marker locus can arise for two different reasons: first, when the locus' polymorphism involves alleles segregating at relatively even frequencies in both populations, but the frequency distribution of the alleles does not differ between the populations (upper example in Table 1). For such a locus, inferring the absence of population differentiation would generally be reasonable.

Second, a very low (or negative) F_{ST} value will also arise if the alleles at a marker locus exhibit an extremely skewed frequency distribution. That is, if a locus is nearly monomorphic in both populations but contains an alternative allele segregating at very low frequency such that this allele occurs only once or a few times in the entire data set (lower example in Table 1). Such a locus is *constrained* to display a very low F_{ST} value between the populations [11]. However, inferring the absence of population differentiation from this F_{ST} value is problematic. The reason is that such rare alleles primarily represent relatively recent mutations, most of which will experience rapid stochastic loss [20]. Markers with a very low minor allele frequency therefore lack the adequate sensitivity to capture the historical signatures of drift and hitchhiking, the key processes in genome scans.

Table 1 Differentiation between two populations, as quantified by Weir and Cockerham's F_{ST} estimator theta [19]

	Genotypes population A			Genotypes population B			F_{ST}
Informative polymorphism	TT	TC	CC	TT	TC	CC	
	5	10	5	5	10	5	-0.026
Uninformative polymorphism	TT	TC	CC	TT	TC	CC	
	20	0	0	19	1	0	0.000

Other F_{ST} estimators produce qualitatively similar results), given informative and uninformative single nucleotide polymorphism at a marker locus (two alleles are present, T and C).

To illustrate this point, imagine that a novel QTL allele arises in the neighborhood of a nearly monomorphic marker. This QTL allele is unlikely to be linked to the rare allele at the marker. If the QTL allele is favored by selection and increases in frequency within the population where it arose, hitchhiking along with the QTL will produce only a very minor (if any) allele frequency shift at the marker locus (Figure 1A). Population differentiation at the QTL will therefore not be visible at the linked marker. A clear signature of hitchhiking, however, will be seen if the marker displays a more balanced allele frequency distribution (Figure 1C; or if the QTL allele happens to be linked to the rare marker allele, Figure 1B). A similar inconsistency in differentiation between selected QTL and associated markers with highly skewed allele frequency distribution also occurs in the situation where selection acts on standing variation (soft sweep; [21]).

Of course, in addition to the situation where a *natural* allele segregates at very low frequency within populations, a highly skewed allele frequency distribution at a locus can also arise artificially during marker data acquisition. For instance due to PCR replication or sequencing error. The locus then produces a minimal F_{ST} value although correctly no F_{ST} value would be calculated because the locus is not polymorphic. However, many strategies exist to avoid such technical errors (including achieving high sequencing coverage, or re-sequencing; see also [23] and references therein). Our paper is therefore primarily concerned with biological polymorphisms.

To summarize, there are two fundamentally different causes for minimal F_{ST} values in genome scan data sets: polymorphisms with relatively even allele frequency distribution, but without population differentiation, *versus* polymorphisms with extremely skewed allele frequency distribution unable to pick up population differentiation. Hereafter, we refer to these forms of polymorphisms as 'informative' *versus* 'uninformative'. We emphasize, however, that we restrict this crude classification to genome scans searching for signatures of selection in the form of elevated differentiation. Markers with highly skewed allele frequency distributions might well be informative in other analytical contexts, such as the estimation of mutational or demographic parameters based on allele frequency spectra [24,25].

If uninformative polymorphisms are abundant in a marker data set used for a genome scan (and they

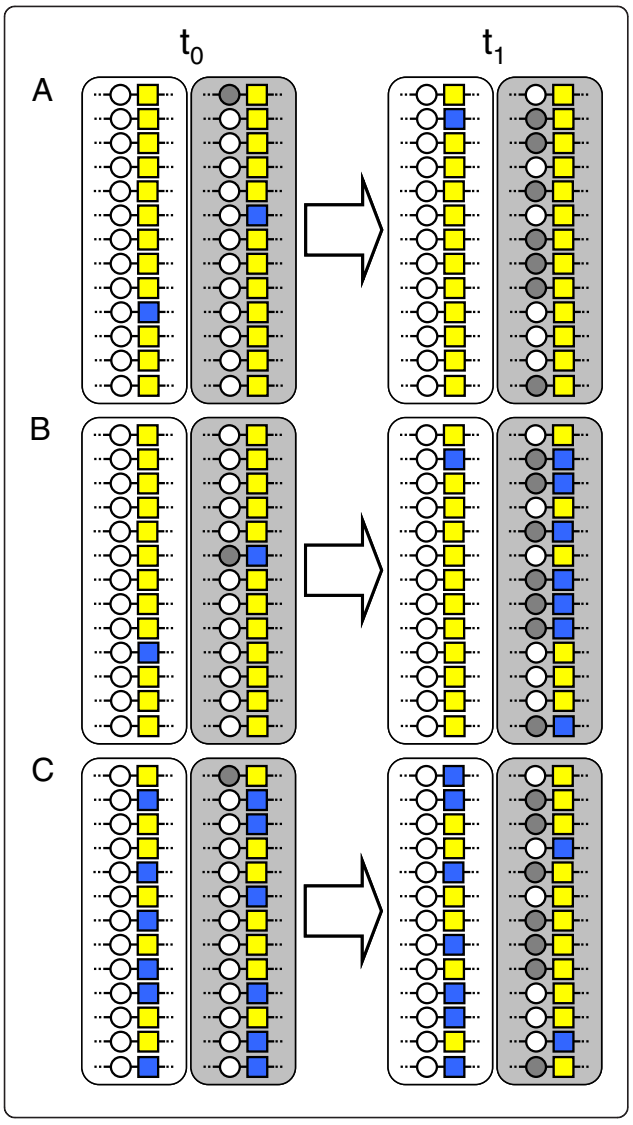

Figure 1 Informative and uninformative markers in genome scans. Two populations derived from the same ancestral population occupy ecologically distinct environments (white and gray boxes) at t_0,. Circles represent an ecologically important QTL with two alleles under divergent selection; white and gray alleles are favored in the white and gray environment. Squares represent a neutral marker with two alleles (yellow and blue). The marker is tightly physically linked to the QTL. In **A**), both initial (t_0) populations display a very low frequency for the blue marker allele. A novel adaptive QTL allele arising in the gray habitat will therefore likely be associated with the frequent yellow marker allele. When sampling the populations at t_1, after a period of selection that has increased the frequency of the gray QTL allele in the gray environment, no signature of selection is visible at the marker locus because hitchhiking along with the favored QTL allele has not materially changed the allele frequency distribution at the marker (F_{ST} [22] approximates zero at both t_0 and t_1). In **B**), the initial conditions (t_0) are as in A), except that the novel adaptive QTL allele happens to be linked to the rare blue marker allele. At t_1, selection at the QTL will be visible at the marker ($F_{ST} = 0.22$) because the blue allele has hitchhiked to high frequency. In **C**), the initial (t_0) allele frequency distribution at the marker is relatively even in both populations ($F_{ST} = 0$). At t_1, the marker exhibits a clear signature of selection ($F_{ST} = 0.13$) because the yellow allele has increased in frequency by hitchhiking. In both B) and C) but not in A), we would consider the marker locus informative at t_1 based on its minor allele frequency across both samples, and consider the marker for a genome scan for the signature of selection (see text).

generally will, see below), we can predict a number of undesirable consequences: in both reference-free and reference-based approaches, the estimated overall baseline differentiation, which is considered to reflect the effect of drift, will be biased downward. As a consequence of this bias in the baseline, the number of loci considered outliers driven by divergent selection in reference-free genome scans may be inflated. By contrast, in sliding window type of scans, the magnitude of among-population differentiation in genomic regions influenced by selection will be weakened, or in the worst case erased. Both effects can lead to incorrect conclusions about the genomic consequences of divergent selection. We emphasize that these problems will arise irrespective of the specific estimator used to quantify population differentiation, or the method chosen for outlier detection. That is, uninformative marker loci will also influence sophisticated methods that estimate F_{ST} for a locus by taking into account genome-wide differentiation and locus-specific sample size [14], or approaches based on P-values from locus-specific significance tests (*e.g.*, [16]).

It would thus seem straightforward to eliminate uninformative marker loci from polymorphism data sets prior to performing a genome scan, as reflected in Beaumont and Nichols' [11] recommendation to preferably use loci with high heterozygosity for such analyses. However, a screen of 24 recent genome scan papers based on single nucleotide polymorphisms (SNPs), including most such studies currently available, suggests that the above issue is not generally recognized. (Note that our paper focuses on SNPs because this marker type is becoming standard in population genomics; but the conclusions hold for any type of marker.) Only three studies report marker filtering according to some minor allele frequency threshold ([18,26,27]; the latter study excluded singleton loci only, *i.e.*, markers with the minor allele occurring only a single time). It is therefore possible that patterns reported and conclusions drawn in many genome scan studies are unreliable to some extent. Given that genome scans are becoming increasingly easy to perform owing to the advent of high-throughput sequencing technology [28], new techniques for extensive SNP discovery (in particular restriction site associated DNA (RAD) sequencing [29]), and automated data analysis pipelines, the problem of bias arising from

uninformative marker loci deserves wide recognition. A first goal of our study is therefore to use extensive SNP data from lake and stream population of stickleback fish to demonstrate that uninformative marker loci indeed have the potential to bias results from genome scans. Our second goal is to show that such bias can be avoided through careful inspection of the data set and subsequent exclusion of uninformative marker loci based on empirical criteria.

Methods

Our study uses SNP data from threespine stickleback (*Gasterosteus aculeatus*) populations occurring in lake and stream habitats within two independently colonized drainages. The first is the Lake Constance drainage in Switzerland (the 'COW' lake-stream population pair from [30]), hereafter called the 'Constance system'. The divergence between the lake and stream population in this system appears to be recent (a few hundred years). The second is the Boot Lake drainage on Vancouver Island, Canada (the Boot sites 'L' and 'S2' in [31]), hereafter called the 'Boot system'. Lake-stream divergence in this system is more ancient (thousands of years). Lake and stream stickleback are known to experience divergent selection [31,32], and the specific population pairs were chosen because they differ in the magnitude of habitat-related phenotypic and neutral genetic (microsatellite) divergence (stronger divergence in the older Boot system than in the younger Constance system). For further details on the locations and populations see [30,31]. All samples were taken with permission from the British Columbia Ministry of Environment (permit number NA06-20791), and the fisheries authority of the canton Thurgau.

For SNP detection, we Illumina-sequenced RAD [29] derived from 27 stickleback specimens from each of the four sites (*i.e.*, one lake and one stream site in two drainages; total N = 108). Library preparation essentially followed the method described in [17]. In short, DNA was digested by using the *Sbf1* restriction enzyme and barcode-ligated for each individual separately. Amplified barcoded DNA was then single-end sequenced on an Illumina genome analyzer IIx with 76 cycles in libraries of 18 pooled individuals each. The Illumina short reads (sequenced RAD sites; deposited at the NCBI Short Read Archive, accession number SRP007695) were parsed by individual barcode, and for each individual separately aligned to the stickleback genome (Ensembl database version 63.1, assembly Broad S1) using Novoalign v2.07.06 (http://novocraft.com). Alignment to a unique genome position was enforced, effectively eliminating sequences derived from repeated elements. The average sequence coverage per individual and RAD site was 27 and 31 for the lake and stream sample in the Constance system, and 30 and 11 for the Boot system. Alignments were converted to BAM format using Samtools v0.1.11 [33]. For each individual and RAD site, we then determined the consensus diploid genotype if ten or more replicate reads were available, or a haploid consensus genotype if replication was below ten. This threshold was chosen because for polymorphic nucleotide positions, we identified heterozygote diploids based on a binomial test with insufficient power at low replication. This test involved calculating the binomial likelihood of the observed frequency distribution of the SNP alleles under the null hypothesis of heterozygosity (*i.e.*, assuming a probability of 0.5 for both alleles). Positions were considered heterozygous if the likelihood was greater than 0.01. Consensus genotyping was quality-aware in that bases with a greater than 0.01 calling error probability were excluded from the binomial test.

To find SNP markers and calculate genome-wide lake-stream population differentiation within each of the two systems, we pooled individual consensus genotypes from the lake and stream sample for each RAD site. If at least 27 genotypes were available from *each* of the two habitats, we proceeded with F_{ST} calculation. In other words, a RAD site was considered only if each individual contributed at least one haploid consensus genotype on average to the site's genotype pool. For F_{ST} calculation, the genotype pool for each RAD site was screened base by base for polymorphisms. If a variable position occurred, we calculated F_{ST} based on haplotype diversity (equation 7 in [22]). For RAD sites exhibiting multiple SNPs, we retained only the highest F_{ST} value observed across all variable base positions. (Using the average F_{ST} value across all positions, or selecting a single SNP at random, produced very similar results supporting identical conclusions.) Negative F_{ST} values were rounded to zero, as commonly done.

The above F_{ST} calculation considered *any* type of SNPs. To explore the effect of informative *versus* uninformative markers, we repeated the above F_{ST} calculation protocol by imposing the restriction that the minor (less frequent) allele had to occur at least n times in the lake-stream genotype pool, where n spanned the range from two to ten in increments of one. (The above default F_{ST} calculation represents the case with $n = 1$.) For each calculation series, we then computed the number of resulting SNPs, and the mean F_{ST} value across all SNPs. We also visualized genomic differentiation by a sliding window approach using local polynomial fitting (LOESS) implemented in R (R Development Core Team [34]; 2nd order polynomial with band width of 0.4; using simpler polynomials and different band widths did not alter our conclusions). All post-sequencing analysis except for alignment and file conversion was coded in the R

language, making use of the Bioconductor packages ShortRead [35], Biostrings, and Rsamtools.

Results

In both the Constance and Boot stickleback population pair, raising the threshold for the minimal required count of the minor SNP allele (n) had a dramatic influence on the number of polymorphic marker loci available for F_{ST} calculation. Most strikingly, the number of SNPs dropped by 46.5% (from 19,729 to 10,546) and 34% (from 16,729 to 7,546) in the Constance and Boot system when singleton loci were excluded by setting n to two (Figure 2A). Increasing n from two to ten, however, had a relatively minor effect on the number of polymorphic loci. Our stickleback data sets thus exhibit a very high proportion of singleton loci, as generally found in empirical studies (*e.g.*, [36-39]). The genomic location of these singleton loci did not show any systematic association with chromosome position (details not).

Including these uninformative marker loci in the genome scan led to the consequences predicted above. First, baseline differentiation was substantially lower than the differentiation obtained when setting n to two or greater (Figure 2B). For instance, genome-wide F_{ST} increased by 17% and 20% in the Constance and Boot system when raising n from one to two. In absolute terms, this shift was more dramatic in the Boot system displaying the higher overall differentiation between the populations. Second, F_{ST} profiles obtained from sliding window analyses including all markers ($n = 1$) were strikingly flatter than those from analyses excluding uninformative polymorphisms. These two consequences are visualized for a segment of chromosome seven (Figure 3), which is representative of what we found throughout the genome. For that specific genomic region, analyses with and without uninformative marker loci might lead to qualitatively different conclusions about the magnitude and physical extent of population differentiation. For example, in the Constance system, a large segment ranging approximately from 12–14 mb displays elevated differentiation, as revealed when using informative markers only. This differentiation is certainly substantial, given the low baseline differentiation in that young system (Figure 2B), and might indicate ongoing divergent selection in that genomic region. Nevertheless, elevated differentiation within that region would probably not be recognized when tolerating uninformative markers in the sliding window analysis.

Note that in Figure 3, we define informative marker loci as those with the minor allele occurring at least four times ($n = 4$), resulting in an average inter-locus distance of 53 kb and 63 kb for the Constance and Boot system. This minor allele threshold eliminated bias associated with uninformative marker loci relatively effectively; choosing higher thresholds had a relatively minor influence on the sliding window profiles.

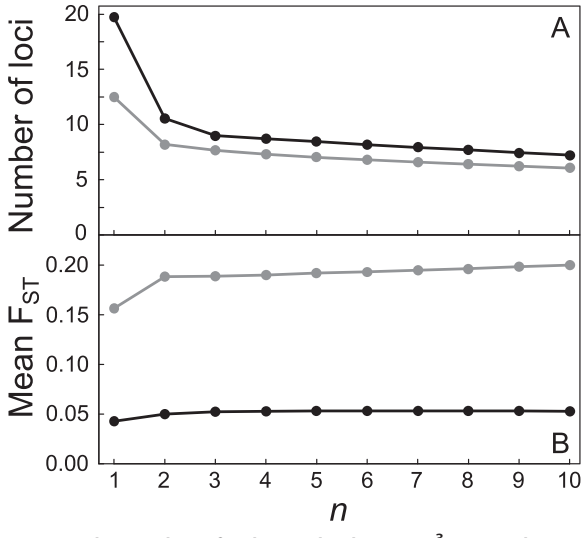

Figure 2 The number of polymorphic loci (x 10^3) (A), and mean F_{ST} across all loci (B), for different minor allele count thresholds (n) in the Constance (black) and Boot (gray) lake-stream stickleback system. This threshold specifies the minimum number of times the minor SNP allele at a locus had to occur in the pooled lake and stream sample for a polymorphic locus to remain in the data set.

Discussion

Our empirical analysis demonstrates that abundant uninformative polymorphisms in a genome scan data set can bias the estimated baseline differentiation, and hence affect conclusions about the genomic signatures of selection.

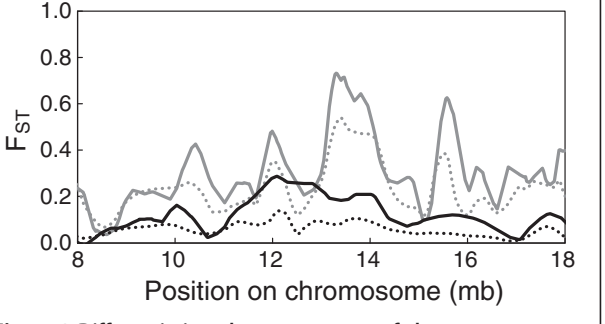

Figure 3 Differentiation along a segment of chromosome seven between the lake and stream stickleback population from the Constance (black) and Boot (gray) system. Sliding window analysis was performed by local polynomial fitting of F_{ST} values for data sets with the allele frequency threshold n set to one (*all* SNPs in the data sets considered; dotted lines), and n set to four (at least four copies of the minor allele required across the pooled lake and stream sample in each system; solid lines). Note the relatively flat differentiation profiles with $n = 1$.

In our stickleback data set, uninformative polymorphisms (essentially in the form of singleton loci) were very abundant. Illumina sequencing type one errors (*i.e.*, wrong base calls despite high indicated base call quality) in RAD sequences poorly replicated at the individual level might contribute to this pattern [23,39]. To examine this possibility, we inspected 50 randomly chosen SNPs exhibiting zero F_{ST} from the full data set accepting any type of polymorphisms (*i.e.*, minor allele count threshold $n = 1$) for each lake-stream system. As expected, a high proportion of these markers were singleton loci (Constance: 28 [56%]; Boot: 35 [70%]). For the Boot system with lower replication per locus in the stream sample (see above), 15 of the 35 singleton loci represented unreplicated RAD sequences. For these loci, the minor allele is likely a sequencing error.

However, all but one of the singleton alleles in the Constance system represented consensus genotypes integrating multiple (2–26, mean: 9.1) replicate RAD sequences. Hence, the bulk of the uninformative marker loci in our data clearly *cannot* be attributed to sequencing error, because the probability of multiple identical errors at a specific nucleotide position at a given RAD site is practically zero. The abundance of rare SNP alleles therefore represents a real biological feature of the studied stickleback populations (acknowledging a small potential contribution from PCR artefacts). This is not unexpected: theory consistently predicts a skew toward polymorphisms with low minor allele frequency, and hence a high proportion of singleton polymorphisms, under a broad range of demographic and selective conditions [24,36,40-44]. Bias associated with uninformative polymorphisms is therefore of general importance to genome scan studies, and not specific to our empirical system. Our analysis also raises a caveat regarding marker densities; the effective number of markers providing relevant information in genome scans might often be dramatically lower than the number reported.

In the present study, excluding singleton polymorphisms had the greatest influence on the results. Reliable quantification of differentiation patterns, however, might require substantially more stringent minor allele frequency thresholds. (Note that such marker filtering also effectively eliminates any sequencing and PCR error from the data.) Bradbury *et al.* [27], for instance, excluded SNPs exhibiting an overall minor allele frequency of 0.25 or less, and a similar threshold was adopted in a recent lake-stream stickleback study carried out in our lab [45]. To obtain a guideline for marker filtering, the latter RAD-based study evaluated the strength of the correlation in F_{ST} values between 'sister' RAD sites (*i.e.*, DNA sequences flanking the *same* restriction site in the genome) in relation to increasingly stringent minor allele frequency thresholds (see Appendix S2 in the Supporting information to [45]). The rationale was that if an F_{ST} value provided by a given marker reliably quantifies the consequences of drift and selection in a genomic region, then another extremely tightly linked marker should yield a similar F_{ST} value. This approach, however, requires tightly physically linked markers data and substantial population differentiation (otherwise the correlation in F_{ST} between linked will remain poor even with stringent marker filtering).

Conclusions

Given the rapidly increasing feasibility and popularity of genome scans for signatures of selection, researchers should be aware that uninformative polymorphisms need to be excluded from data sets. This is not achieved by just avoiding technical errors, as a high prevalence of nearly monomorphic loci is a general biological feature of samples from natural populations. We suggest that a reasonable strategy to define and eliminate uninformative polymorphisms should be chosen by inspecting the allele frequency distribution of the polymorphisms, and by assessing the influence of different marker filtering thresholds on the genomic patterns of interest, or appropriate statistics (such as the correlation of F_{ST} between sister RAD sites). Also, the approach taken to eliminate uninformative polymorphisms should be reported explicitly. Together, this should increase the quality and comparability of genome scans, and hence promote our understanding of the processes shaping genomic differentiation.

Competing interest

The authors declare no competing interest.

Acknowledgment

Field work was aided by Anne-Catherine Grandchamp and, for the Boot system, supported financially by Andrew Hendry. Roman Kistler (fisheries authority of the canton Thurgau) permitted sampling of the Constance specimens. Paul Etter and Bill Cresko kindly shared their experience and protocol for RAD library preparation. Brigitte Aeschbach and Nicolas Boileau facilitated wet lab work, Ina Nissen performed Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich, and Lukas Zimmermann provided IT support. The bioinformatics pipeline benefited from modifications to Novoalign by Colin Hercus, and from coding suggestions by Martin Morgan and Hervé Pagès. Matthieu Foll, Markus Pfenninger, and three anonymous reviewers provided valuable suggestions that improved the paper. WS was supported financially by the European Research Council (Starting Grant 'INTERGENADAPT'), the Swiss National Science Foundation, and the University of Basel. DB was supported by the Swiss National Science Foundation (Ambizione grant PZ00P3_126391 / 1) and the Research Found of the University of Basel. We kindly thank all these people and institutions.

Authors' contributions

DB and MR conceived the study; WS provided materials and infrastructure; MR and DB generated the sequence data; DB and MR analyzed the data; DB wrote the paper, with input from MR and WS. All authors read and approved the final manuscript.

Received: 24 January 2012 Accepted: 22 June 2012
Published: 22 June 2012

References

1. Wu CI: **The genic view of the process of speciation.** *J Evol Biol* 2001, **14**:851–865.
2. Mitchell-Olds T, Willis JH, Goldstein DB: **Which evolutionary processes influence natural genetic variation for phenotypic traits?** *Nat Rev Genet* 2007, **8**:845–856.
3. Nosil P, Funk DJ, Ortiz-Barrientos D: **Divergent selection and heterogeneous genomic divergence.** *Mol Ecol* 2009, **18**:375–402.
4. Schluter D: **Evidence for ecological speciation and its alternative.** *Science* 2009, **323**:737–741.
5. Holsinger KE, Weir BS: **Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}.** *Nat Rev Genet* 2009, **10**:639–650.
6. Lewontin RC, Krakauer J: **Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms.** *Genetics* 1973, **74**:175–195.
7. Maynard Smith J, Haigh J: **Hitch-hiking effect of a favorable gene.** *Genet Res* 1974, **23**:23–35.
8. Beaumont MA: **Adaptation and speciation: what can Fst tell us?** *Trends Ecol Evol* 2005, **20**:435–440.
9. Storz JF: **Using genome scans of DNA polymorphism to infer adaptive population divergence.** *Mol Ecol* 2005, **14**:671–688.
10. Stinchcombe JR, Hoekstra HE: **Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits.** *Heredity* 2008, **100**:158–170.
11. Beaumont MA, Nichols RA: **Evaluating loci for use in the genetic analysis of population structure.** *Proc R Soc Lond B* 1996, **263**:1619–1626.
12. Beaumont MA, Balding DJ: **Identifying adaptive genetic divergence among populations from genome scans.** *Mol Ecol* 2004, **13**:969–980.
13. Foll M, Gaggiotti O: **A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective.** *Genetics* 2008, **180**:977–993.
14. Excoffier L, Hofer T, Foll M: **Detecting loci under selection in a hierarchically structured population.** *Heredity* 2009, **103**:285–298.
15. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a high-density SNP map for signatures of natural selection.** *Genome Res* 2002, **12**:1805–1814.
16. Turner TL, Hahn MW, Nuzhdin SV: **Genomic islands of speciation in *Anopheles gambiae*.** *PLoS Biol* 2005, **3**:e285.
17. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.** *PLoS Genet* 2010, **6**:e1000862.
18. Lawniczak MKN, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, *et al*: **Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences.** *Science* 2010, **330**:512–514.
19. Weir BS, Cockerham CC: **Estimating F-statistics for the analysis of population-structure.** *Evolution* 1984, **38**:1358–1370.
20. Ewens W: *Mathematical population genetics*. New York: Springer; 1979.
21. Hermisson J, Pennings PS: **Soft sweeps: molecular population genetics of adaptation from standing genetic variation.** *Genetics* 2005, **169**:2335–2352.
22. Nei M, Tajima F: **DNA polymorphism detectable by restriction endonucleases.** *Genetics* 1981, **97**:145–163.
23. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet* 2011, **12**:443–451.
24. Marth GT, Czabarka E, Murvai J, Sherry ST: **The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations.** *Genetics* 2004, **166**:351–372.
25. Keightley PD, Eyre-Walker A: **Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies.** *Genetics* 2007, **177**:2251–2261.
26. Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O'Brien SJ, Smith MW: **Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations.** *PLoS One* 2008, **3**:e1712.
27. Bradbury IR, Hubert S, Higgins B, Borza T, Bowman S, Paterson IG, Snelgrove PVR, Morris CJ, Gregory RS, Hardie DC, *et al*: **Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature.** *Proc R Soc B* 2010, **277**:3725–3734.
28. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, **24**:133–141.
29. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA: **Rapid SNP discovery and genetic mapping using sequenced RAD markers.** *PLoS One* 2008, **3**:e3376.
30. Berner D, Roesti M, Hendry AP, Salzburger W: **Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents.** *Mol Ecol* 2010, **19**:4963–4978.
31. Berner D, Grandchamp A-C, Hendry AP: **Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions.** *Evolution* 2009, **63**:1740–1753.
32. Berner D, Adams DC, Grandchamp A-C, Hendry AP: **Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology.** *J Evol Biol* 2008, **21**:1653–1665.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **Genome Project Data P: The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.
34. R Development Team: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2010.
35. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R: **ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data.** *Bioinformatics* 2009, **25**:2607–2608.
36. Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST: **Genetic traces of ancient demography.** *Proc Nat Acad Sci USA* 1998, **95**:1961–1967.
37. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: **Natural selection has driven population differentiation in modern humans.** *Nat Genet* 2008, **40**:340–345.
38. Geiler KA, Harrison RG: **A Δ11 desaturase gene genealogy reveals two divergent allelic classes within the European corn borer (*Ostrinia nubilalis*).** *BMC Evol Biol* 2010, **10**:112.
39. Keightley PD, Halligan DL: **Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans.** *Genetics* 2011, **188**:931–940.
40. Nei M, Li W-H: **The transient distribution of allele frequencies under mutation pressure.** *Genet Res* 1976, **28**:205–214.
41. Li WH: **Maintenance of genetic variability under joint effect of mutation, selection and random drift.** *Genetics* 1978, **90**:349–382.
42. Fu YX: **Statistical properties of segregating sites.** *Theor Pop Biol* 1995, **48**:172–197.
43. Eberle MA, Kruglyak L: **An analysis of strategies for discovery of single-nucleotide polymorphisms.** *Genet Epidemiol* 2000, **19**:29–35.
44. Evans SN, Shvets Y, Slatkin M: **Non-equilibrium theory of the allele frequency spectrum.** *Theor Pop Biol* 2007, **71**:109–119.
45. Roesti M, Hendry AP, Salzburger W, Berner D: **Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs.** *Mol Ecol* 2012, **21**:2852–2862.

doi:10.1186/1471-2148-12-94
Cite this article as: Roesti *et al.*: Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology* 2012 **12**:94.