CHAPTER 9

Hitchhiking Mapping:
Limitations and Potential for the Identification of Ecologically Important Genes

Christian Schlötterer*

Abstract

A recent series of publications demonstrated that identification of genomic regions subjected to positive selection (hitchhiking mapping) is possible and could be applied in an ecological context. This review focuses on the use of microsatellite markers in genome scans for the identification of beneficial mutations. The pitfalls and potential of the lnRθ test statistic are discussed as well as different approaches for the identification of the molecular change(s) underlying an observed selective sweep.

Introduction

Ecological genomics encapsulates a recent trend to apply high-throughput genomic tools to questions in ecology and evolution.[1] Progress in genomics technology has shifted the focus from the analysis of a small number of candidate genes to multiple genomic regions in several populations. While this research area is still in its infancy, already a considerable number of studies have demonstrated the enormous potential of multilocus approaches for the identification of genomic regions bearing ecologically important loci/alleles.[2,3]

This approach, which has been termed hitchhiking mapping[3,4] or selection mapping,[5] relies on a very simple population genetic principle. Theory predicts that a beneficial mutation is either lost quickly due to genetic drift or becomes fixed in the population. Importantly, not only the beneficial mutation increases in frequency, but also other, neutral variants linked to the target of selection (hitchhiking[6]). Thus, as a consequence of the spread of a beneficial mutation in a population, the allele frequency spectrum is significantly distorted from neutral expectations in a genomic region around the target of selection. Population genetics has devised a range of different approaches to use this change in allele frequency spectrum for the identification of past episodes of nonneutral evolution.[7,8]

Many of these classic neutrality tests, such as Tajima's D,[9] are affected by demographic effects, such as bottlenecks and admixture, preventing the use of a nominal P-value for the identification of selected loci.[10] When a large number of loci are surveyed, however, it is possible to build an empirical distribution of any test statistic used to quantify the distortion in allele frequency.[11] Rather than relying on a nominal P-value, genomic regions possibly subjected to selective forces are then identified as loci in the tails of the empirical distribution. While this approach eliminates the effects of demographic events, its disadvantage is that even in the absence of selection, genomic regions will be (falsely) identified as targets of selection, due to their location in the tail of the distribution.

*Christian Schlötterer—Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, A-1210 Wien, Austria, Europe. Email: Christian.schloetterer@vu-wien.ac.at

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber.

Here, statistical approaches of hitchhiking mapping using microsatellite variation will be introduced. The general limitations of the hitchhiking mapping approach will be demonstrated using microsatellite data and possible approaches to overcome them will be discussed.

Microsatellites—A Widely Used Genetic Marker

Since their introduction in 1989[12-14] microsatellites have developed into one of the most commonly used genetic markers.[15] A microsatellite consists of a tandem repetition of one repeat motif, such as $(GT)_n$ or $(GAC)_n$. Due to DNA replication slippage, a mutation process specific to tandemly repeated DNA, the copy number of the repeat units changes at a high rate (up to 10^{-2}).[16,17] Using locus-specific PCR primers flanking a microsatellite, the variation in repeat number can be easily detected.

Most microsatellite mutations encompass either the gain or loss of a single repeat unit, but larger changes in repeat number have also been described. Nevertheless, the mutation process of microsatellites can be well-approximated by the stepwise mutation model, which was originally introduced to describe the evolution of proteins.[18] The interpretation of microsatellite variability data, in particular the comparison across loci, is significantly complicated by pronounced locus specific mutation rates.[19,20]

The lnRθ Statistic

The Concept

One of the possible consequences of a selective sweep is a reduction in variability at a genomic region subjected to a selective sweep. Thus, genome scans for targets of selection can aim for the identification of genomic regions bearing microsatellites that have less variability than expected under neutrality. The large variation in microsatellite mutation rates, however, significantly complicates the interpretation of variability patterns, as it is not possible to distinguish whether a locus has low levels of variability due to a selective sweep or a low mutation rate. In an attempt to overcome the problem of locus specific mutation rates, lnRθ has been proposed as a means of identifying microsatellite loci, which show a more pronounced reduction in variability.[21,22] Rather than analyzing microsatellite variability in one population only, the lnRθ statistic requires polymorphism data from two populations. Assuming that the microsatellite mutation rate does not differ among populations, the expectation for lnRθ is the same for all microsatellite loci, independent of the mutation rate.

$$\ln\left[\mathrm{E}(R\theta)\right] = \ln\left[\mathrm{E}\left(\frac{\frac{1}{2}\theta_{Pop1}}{\frac{1}{2}\theta_{Pop2}}\right)\right] = \ln\left[\mathrm{E}\left(\frac{\left(2N_{e_{Pop1}}\mu\right)}{\left(2N_{e_{Pop2}}\mu\right)}\right)\right] \cong \ln\left[\frac{\mathrm{E}\left(2N_{e_{Pop1}}\mu\right)}{\mathrm{E}\left(2N_{e_{Pop2}}\mu\right)}\right] \qquad \text{(ref. 23)} \quad (1)$$

Two different estimators for θ could be used: variance in repeat number (V) and gene diversity (H, expected heterozygosity):

$$\theta = 2V = 4N_e\mu = \frac{2}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad \text{(ref. 23)} \quad (2)$$

x_i is the repeat number of allele i.

$$\theta = \left(\left(\frac{1}{1-H_{Pop1}}\right)^2 - 1\right)\frac{1}{8\mu} \qquad \text{(ref. 18)} \quad (3)$$

θ estimates based on the variance in repeat number have a larger variance and are thus less sensitive to identify loci subjected to a selective sweep than θ estimates based on gene diversity.