QUANTITATIVE GEOGRAPHY AND GENOMICS

Spatial analysis to detect signatures of selection along a gradient of altitude in the common frog (*Rana temporaria*)

Stéphane JOOST[1] and Aurélie BONIN[2]

[1] Istituto di Zootecnica, Università Cattolica del S.Cuore, Piacenza, Italy
[2] Laboratoire d'Ecologie Alpine, CNRS-UMR 5553, Université Joseph Fourier, Grenoble, France

ABSTRACT

We applied a spatial approach to detect regions of the genome of the common frog (*Rana temporaria*) which are possibly selected along an altitude gradient. The identification of selected regions in the genome is important as it gives the possibility to understand which genes are driven by natural selection. In the field of population genetics, several statistical methods using molecular data were developed to reveal genomic regions under selection. Here, we tackled the issue from an environmental point of view by using a spatial analysis method (SAM) recently developed. The method is based on the spatial coincidence concept and has recourse to GIS, environmental data and molecular data to simultaneously process many univariate logistic regression models. This research showed that there is a strong correspondence between spatial analysis results and those obtained with a standard population genomics approach.

KEYWORDS

Spatial analysis, logistic regression, genetics, natural selection, landscape genomics

INTRODUCTION

Detecting signatures of natural selection within the genome of organisms is key since it may allow a greater understanding which genes are being shaped by natural selection. In general, regions of the genome that are under selection are likely to be of functional importance, and inferences regarding selection may provide important information [1]. The discovery of such genomic regions is the keystone of promising applications in medical research, conservation biology and selective breeding [2]. Here, we implemented a spatial analysis method (SAM) recently developed to look for possible signatures of natural selection in a wide-spread European amphibian, the common frog (*Rana temporaria*), with an altitude gradient as a framework. The approach resorts to the geographical coordinates of sampling places and to multiple logistic regressions to measure statistical association between an environmental parameter (altitude) and specific molecular markers. In common frog, several traits are under natural selection along an altitude gradient [3]. We also analysed the same data set with a standard population genomics approach in order to compare the results. Other applications of the SAM implied up to more than one hundred environmental parameters [4].

BIOLOGICAL MODEL

The common frog (*Rana temporaria*) is the most wide-spread amphibian in Europe, occurring from Northern Spain to subartic Scandinavia [5]. This large latitude range is coupled with a wide altitudinal range as the common frog can be found from the sea level to an altitude of 2600m in the Alps [5]. Across this large distribution area, many traits were shown to be subject to natural selection [5], in particular along an altitude gradient in the Alps [3][5].

DATA

The method requires a georeferenced data set made up of one or more environmental variables describing the sampling locations, and a data set containing the genetic characteristics of the investigated animals.

Environmental data

The amphibians were sampled in the North French Alps. Samples (i.e., adult frog fingers and tadpoles) were collected in six breeding sites scattered throughout a 7'000 km2 area: two low altitude localities (i.e., about 400 m above sea level), two intermediate altitude localities (i.e., about 1'100 m) and two high altitude localities (i.e., about 2'100 m) to cover the whole altitude gradient. Altitude was recorded using an altimeter.

Molecular data

Biotechnologies allow identifying specific sequences of DNA in the genome called *markers*, which act as genetic milestones. These molecular markers are not necessarily genes, but they can be statistically associated with tracked genes. Each individual has a particular genetic profile at these markers, which is shaped at the population level by evolutionary forces like natural selection. Several methods exist to find out markers within the genome [6]. For the biological model mentioned in this paper, the genome was scanned using AFLP (Amplified Fragment Length Polymorphism) markers, which can be obtained easily and are randomly distributed in the genome [7]. The precise AFLP procedure is described in [5]. AFLP profiles were recorded in a matrix as presence (1) or absence (0) of AFLP bands for each individual.

METHODS

The common frog was studied from both the SAM and the population genomics point of view in order to compare the results provided by the two approaches.

Spatial Analysis Method (SAM)

Logistic regression is used to provide a measure of the association level between the frequency of molecular markers at a sampling place and the environmental parameters favouring or not their increase in frequency. The principle is to assess the significance of the models constituted by all possible [marker ⇆ environmental variable] pairs, and to highlight the markers involved in the most significant models as possibly under selection. To verify if a model with the examined environmental variable explain the observed distribution of molecular markers better than a model with a constant only, the significance of the coefficients is evaluated by statistical tests. According to Hosmer and Lemeshow [8], we resorted to the Likelihood ratio (G) and Wald statistical tests to achieve the comparison of observed to predicted values and determine the significance of the models. A model is considered significant when both tests presented above reject the corresponding null hypothesis. In that case, the marker under consideration is possibly under natural selection. Molecular data sets may contain a lot of markers and many different environmental parameters are likely to describe the sampling locality. Thus many simultaneous univariate models have to be run in order to detect markers submitted to natural selection. When one wishes to simultaneously test several hypotheses at a common significance level α, the probability of rejecting at least one of the hypotheses being tested that is in fact true (type I error) is typically much in excess of α [9]. Several methods allow correcting this multiple hypotheses testing effect and we applied the conservative Bonferroni correction [10] which implies to divide the wanted significance threshold α by the number of processed models. SAM method is described in details in [4].

Population genomics approach

To assess the results provided by the SAM, we resorted to a theoretical approach developed by Beaumont and Nichols [11] which uses computer simulations (Dfdist software, modified from [11]) to model the behaviour of neutral markers under a specific model of population structure. It is based on the principle that

genetic differentiation between populations is expected to be higher for markers under selection than for the rest of the genome, considered as neutral.

RESULTS

Overall, 138 individuals were analysed at AFLP 392 markers. However, markers absent or present at almost all sampling places were ignored as they were deemed not to be informative.

SAM

Among the 364 markers finally analysed, the SAM highlighted 46 different markers significant for at least one single test, with a significance threshold set to 2.74E05, i.e., 99% confidence level (CL) including Bonferroni correction. With this CL, 20 markers were significant with both Wald and G tests. Marker 301 was standing out very strongly, all tests being very significant (CL=2.74E-11). Other detected markers were 320 (CL=2.74E-10), 214 (CL=2.74E-9), 337 and 357 (CL=2.74E-8), 180, 84, 328 and 179 (CL=2.74E-7), 385, 233 and 354 (CL=2.74E-6), and 62, 390, 58, 248, 3, 271, 228 and 265 (CL=2.74E-5).

Population genomics

Analyses with Dfdist were conducted with localities grouped according to their altitude category (i.e., low, intermediate or high). Only two markers (301 and 84) were detected as outliers with Dfdist at the 99% CL. They thus appeared as the best candidates for selection along an altitude gradient, as also revealed by the SAM. At the 95% CL, 4 other markers (97, 228, 250 and 388) stood outside the confidence interval and were also detected by the SAM, by the G test only for markers 97 and 388.

DISCUSSION

There is a good correspondence between SAM results and those based on the population genomics approach. Statistical signals exist, which associate markers and altitude, and the same markers do behave in a particular way in comparison with the theoretical distribution of neutral markers. The atypical behaviour of these markers is thus highlighted by two independent methods based on completely different hypotheses. Apparently, the SAM detects more markers than Dfdist, but we noticed that when gradually lowering the confidence level in Dfdist, this method was also able to identify those “missing” outliers. We do not master the SAM’s sensitivity yet, and further studies will be necessary to establish precise relationships between population genomics approaches and these statistical measures of association.

The many markers detected can be explained because the study is based on altitude only as environmental parameter, and given that many traits may respond to this many-sided selection pressure, it is reasonable to expect that several genomic regions are the targets of selection in relation with altitude [5].

In opposition with classical theoretical methods developed in population genetics, the SAM is free of any constraint about the genetic structure of the population because it proceeds at the individual and not at the population level. The SAM presents thus two remarkable advantages which are i) its easy use with organisms for which it is not obvious to define the geographical limits of the populations, and ii) its absence of preliminary hypotheses on the genetic structure of the populations.

The best candidates detected in this study (301, 84) are only regions of the genome *possibly* under natural selection. The ultimate proof for a marker and an associated gene as being under natural selection requires functional evidence. To show that the SAM is actually able to detect markers under selection, one possibility is to start from using a marker associated with a gene known to be involved in a functional process and to try to detect it. A preliminary study (C. Parisod and S. Joost, unpublished results) on a plant (*Biscutella laevigata*) in the Swiss Prealps highlighted an AFLP marker known to be linked to a gene (AGAMOUS) involved in flowering time, and associated with a coherent environmental parameter which is

the number of frozen days during flowering time. This illustrates the role this method can play in hunting for functional genes in the future.

Although the method is based on a basic spatial coincidence concept, the SAM perfectly illustrates the role quantitative geography may play in a life sciences challenging issue, likely to improve our understanding of the genetic mechanisms of evolution.

ACKNOWLEDGEMENTS

We would like to thank Claude Miaud, Olivier Marquis, Christian Miquel, Ludovic Gielly, Delphine Rioux, François Pompanon and Pierre Taberlet help with the sampling, the experiments or the analyses.

REFERENCES

1. **Nielsen, R.**, Molecular Signatures of Natural Selection. Annual Review of Genetics, Nr. 39, 2007, pp. 197-218.

2. **Luikart, G. et al**, The power and promise of population genomics: From genotyping to genome typing. Nature Reviews Genetics, Nr. 4, 2003, pp. 981-994.

3. **Miaud, C.**, **Merilä, J.**, Local adaptation or environmental induction? Causes of population differentiation in alpine amphibians. Biota, Nr. 2, 2001, pp. 31-50.

4. **Joost, S. et al**, A Spatial Analysis Method (SAM) to detect candidate loci for selection: towards a landscape genomics approach of adaptation, submitted to Molecular Ecology.

5. **Bonin, A. et al**, Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). Molecular Biology and Evolution. Nr. 23, 2006, pp. 773-783.

6. **Avise, J.C.,** Molecular Markers, Natural History, and Evolution, Sinauer, Sunderland, 2004.

7. **Vos, P. et al**, AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research, Nr. 23, 1995, pp. 4407-4414.

8. **Hosmer, D.W., Lemeshow, S.**, Applied logistic regression, John Wiley & Sons, New York, 2000.

9. **Shaffer, J.P.**, Multiple Hypothesis-Testing. Annual Review of Psychology, Nr. 46, 1995, pp. 561-584.

10. **Narum, S.R.**, Beyond Bonferroni: less conservative analyses for conservation genetics. Conservation Genetics, Nr. 7, 2006, pp. 783-787.

11. **Beaumont, M.A, Nichols, R.A.**, Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London Series B-Biological Sciences, Nr. 263, 1996, pp. 1619-1626.

AUTHORS INFORMATION

Stéphane JOOST, PhD
Istituto di Zootecnica, Università Cattolica del S.Cuore, via E. Parmense 84, 29100 Piacenza, Italy
stephane.joost@econogene.eu

Aurélie BONIN, PhD
Diversity Arrays Technology
PO Box 7141, Yarralumla, ACT 2600, Australia
a.bonin@DiversityArrays.com