Genetica (2005) 124:71–75
DOI 10.1007/s10709-004-8358-1

Testing population genetic structure using parametric bootstrapping and MIGRATE-N

B. C. Carstens[1,*], A. Bankhead III[2,3], P. Joyce[2,4] & J. Sullivan[1,2]
[1]*Department of Biological Sciences, University of Idaho, USA;* [2]*Initiative in Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, USA;* [3]*Department of Computer Science, University of Idaho, USA;* [4]*Department of Mathematics, University of Idaho, USA; *Present address: Department of Ecology & Evolutionary Biology, 1109 Geddes Ave., Museum of Zoology, Room 1089, University of Michigan, Ann Arbor, MI 48109-1079, USA (Phone: +1-208-885-2550; Fax: +1-208-885-7905; E-mail: bcarsten@umich.edu)*

Received 06 September 2004 Accepted 05 December 2004

Key words: migration, MIGRATE-N, panmixia, parametric bootstrapping, population structure

Abstract

We present a method for investigating genetic population structure using sequence data. Our hypothesis states that the parameters most responsible for the formation of genetic structure among different populations are the relative rates of mutation (μ) and migration (M). The evolution of genetic structure among different populations requires rates of $M \ll \mu$ because this allows population-specific mutation to accumulate. Rates of $\mu \ll M$ will result in populations that are effectively panmictic because genetic differentiation will not develop among demes. Our test is implemented by using a parametric bootstrap to create the null distribution of the likelihood of the data having been produced under an appropriate model of sequence evolution and a migration rate sufficient to approximate panmixia. We describe this test, then apply it to mtDNA data from 243 plethodontid salamanders. We are able to reject the null hypothesis of no population structure on all but smallest geographic scales, a result consistent with the apparent lack of migration in *Plethodon idahoensis*. This approach represents a new method of investigating population structure with haploid DNA, and as such may be particularly useful for preliminary investigation of non-model organisms in which multi-locus nuclear data are not available.

Introduction

The parametric bootstrap has become a useful method for testing hypotheses derived from phylogenetic data because it allows researchers to test a null hypothesis by creating a null distribution for parameters that do not conform to known statistical distributions (Goldman, 1993; Huelsenbeck, Hillis & Jones, 1996; Sullivan, Arellano & Rogers, 2000). In the phylogenetic application, parametric bootstrapping is used to create the null distribution by simulating data on a topology that matches the prediction of the null hypothesis with a model of sequence evolution. The test statistic from the actual data is then compared to this null distribution in order to test its significance. The major limitation of this approach is that it requires a precise topological prediction. This requirement has limited the application of parametric bootstrapping in evolutionary biology to questions that concern relationships among different species, but nearly any question that can be stated in terms of a probabilistic model and a parameter of interest can be tested with a parametric bootstrap. Here we describe an approach for testing for the presence of panmictic population structure using sequence data.

Our basic hypothesis states that the parameters most responsible for the formation of genetic structure among different populations are the relative rates of mutation (μ) and migration (M). The evolution of genetic structure among different populations requires rates of $M \ll \mu$ because

population-specific mutations accumulate under such a relationship. Alternatively, rates of $\mu \ll M$ will not allow genetic differentiation to develop among different demes, resulting in populations that are effectively panmictic (Slatkin, 1980). Our approach to identifying population structure takes advantage of this relationship between the rates of mutation and migration in an indirect way – by simulating data under the assumption of $M \sim \mu$, and using these simulations to construct null distributions that can be used to evaluate the estimated M in samples collected from different geographic locations. By using the parametric bootstrap approach, we have the statistical power to test this hypothesis in spite of the relatively flat likelihood surface around M, which can make the confidence intervals around this parameter unrealistically large (Abdo, Crandall & Joyce, 2004).

Methods

Overview of the migration test of panmixia

The MIGRATE-N test of population structure (MTOPS) was designed to test for the genetic signature of population structure by using coalescent simulations to create a null distribution of the log-likelihood of the data having been produced under a model of panmixia. To implement this procedure we used MIGRATE-N (Beerli, 2002) and TREEVOLVE (v. 1.3; Grassly, Harvey & Holmes, 1999), two programs that are freely distributed via the internet. MIGRATE-N was used to estimate theta ($\theta = 2Ne\mu$) and the scaled migration rate ($M/N_{e)}$, which is equal to the number migrants per generation. We then used a model of sequence evolution and θ to generate 100 simulated datasets under the assumption of panmixia with the program TREEVOLVE.

In order to create the null distribution we used the following process. First, each simulated data set was analyzed with MIGRATE-N under the assumption that all samples were members of a single population. We then removed the resulting θ_i and M_i values from the output files, and reanalyzed each simulated dataset under a model of population structure. This model was enacted by randomizing the simulated samples to match the number of individuals sampled from each population. Markov chains for the second analysis were started with the θ_i value and a migration rate (M) set to a value (5) that was 5× the value expected under panmixia (Wright, 1978). This high migration rate ensures that the Markov chains are started in regions of parameter space consistent with panmixia. To insure adequate mixing of the Markov chains, we used Markov-coupling, where 4 independent chains were allowed to mix according to an adaptive heating scheme (1.0, 1.2, 1.5, 3.0). Ten short chains, of 50,000 generations, and three long chains, of 1.0×10^6 generations were used for all runs of MIGRATE-N. The resulting $-\ln L$ scores were then used to form a null distribution of the log-likelihood of the data having been produced by panmictic populations and the model of sequence evolution. The same two-step process was conducted on the actual data to form the test statistic. A PERL script was written (by A. Bankhead) to automate the steps in this process and is available from the corresponding author.

Empirical study

We applied our MTOPS to data from the plethodontid salamander *Plethodon idahoensis* consisting of 669 bp from the mitochondrial Cytochrome *b* gene (Carstens et al., 2004). For these data, we used DT-MODSEL (Minin et al., 2003) to select the HKY + Γ model of sequence evolution, with equilibrium base frequencies of $\pi_A = 0.306$; $\pi_C = 0.234$; $\pi_G = 0.14$; $\pi_T = 0.32$; transition–transversion ratio = 0.8448); and Γ-distribution shape parameter (α) = 0.285. This model was verified with an absolute goodness-of-fit test ($p = 0.67$; Goldman, 1993), and was used for simulations with TREEVOLVE and analysis with MIGRATE-N.

In order to explore population structure within *P. idahoensis* we divided our samples into 16 data partitions that represented different geographic scales of population structure (Table 1). The most inclusive of these contained samples collected from across the range of *P. idahoensis* ($n = 243$). We analyzed samples from the Clearwater drainage ($n = 89$) and also divided the samples into 8 groups representing the major river drainages inhabited by *P. idahoensis* ($n = 15$–60). On the smallest scale, we analyzed six sets containing samples from restricted portions of certain drainages ($n = 9$–35). This nested approach was designed not because we expected the genetic signature of population

Table 1. Data partitions used in our analysis of 243 Cyt *b* sequences from *P. idahoensis*

	Partition	*n*	#	Structure	θ	N_e	*p*
1	Total	243	8	15 23 51 60 19 38 28 9	0.047304	473040	< **0.01**
2	Clearwater	89	3	15 23 51	0.014548	72740	< **0.01**
3	NFC[1]	51	9	5 5 6 11 3 3 3 7 8	0.007642	38210	= **0.02**
4	Lochsa	23	6	2 2 2 2 5 10	0.004178	20890	= **0.01**
5	Selway	15	2	10 5	0.003106	15530	= **0.02**
6	St. Joe	60	13	2 10 2 2 13 3 4 5 8 2 2 2 5	0.013645	68225	< **0.01**
7	Cd'A[2]	19	5	1 3 7 3 5	0.001835	9175	< **0.01**
8	Kootenai	38	9	4 3 8 2 10 4 2 3 2	0.017954	89770	< **0.01**
9	Columbia	28	4	7 5 9 7	0.04936	24680	= 0.07
10	Freeman	10	2	5 5	0.00402	20100	= **0.03**
11	Upper NFC	35	6	11 3 3 3 7 8	0.00827	41350	= **0.01**
12	HW50-St. Joe	13	2	9 4	0.00217	10850	= **0.01**
13	Upper St. Joe	9	3	2 2 5	0.00366	18300	= **0.04**
14	Kooscanusa	10	4	3 3 3 1	0.01494	74700	= 0.06
15	Revelstoke	16	2	9 7	0.0046	23000	< **0.01**
16	Ione	16	2	7 9	0.004924	24620	= 0.06

Numbers in the left-most column correspond to those in Figure 1. Also shown are the name of the partition, the number of samples in the partition (*n*), the number of populations in the partition (#), the number of samples per each population (structure), θ estimated from each partition (θ), the effective population size if a mutation rate of 1.0×10^{-7} is assumed (N_e), and the *p*-value of the null hypothesis. Significant *p*-values are shown in bold.

[1] North Fork of the Clearwater River.

[2] Coeur d'Alene River.

genetic structure to be found on the largest geographic scales, but because we were interested in identifying an approximate level for which we could reasonably expect populations to be genetically indistinguishable from a randomly mating population. Our *a priori* expectation is that we will be able to reject the null hypothesis of no population genetic structure on all but the smallest geographic scales in *P. idahoensis* because levels of gene flow are low in other plethodontid salamanders (Larson, Wake & Yanev, 1984). We used $p = 0.5$ as a critical value for all tests.

Results and discussion

For most of the data partitions we can reject the model of panmixia using the MTOPS approach (Table 1). The exceptions are the Columbia, Lake Koocanusa, and Ione partitions, where we were not able to reject the null model of no population genetic structure ($p_{COL} = 0.07$; $p_{LK} = 0.06$; $p_{IONE} = 0.06$), indicating that these salamanders are drawn from populations that are effectively panmictic. Our analysis may be complicated by the history of post-Pleistocene expansion in *P. idahoensis*. Expanding populations will produce genealogies that are dominated by the tips of the tree, and MTOPS might interpret a genealogy produced by an expanding population as one with multiple migration events. In species with a demographic history of population-size expansion, MTOPS is likely to be conservative as a test of population structure. We suspect that this has happened in the Columbia partition, which contains the northernmost populations of *P. idahoensis*; populations which live in an area where deglaciation occurred less than 10,000 ybp. Nevertheless, the conservative bias in this approach increases our confidence in the rejection of the model of population panmixia for most of the data partitions, which span wide geographic distances.

The test of panmictic population structure presented here offers new insight into the population structure of *P. idahoensis*. The species is endemic to the northern Rocky Mountains and has a history of post-Pleistocene expansion from

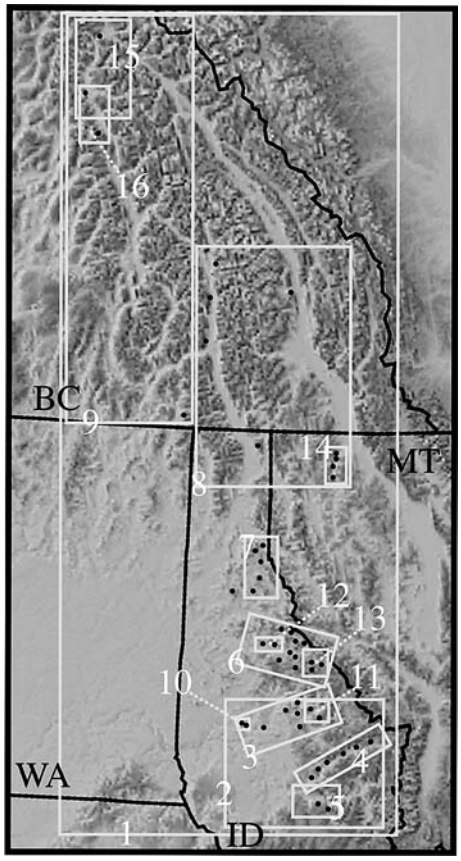

Figure 1. Map of the northern Rocky Mountain region of western North America. States and provinces are identified as follows: ID, Idaho; WA, Washington; MT, Montana; BC, British Columbia. Data partitions 1–16 for the investigation into the population structure of *P. idahoensis* are shown with gray rectangles and identified with white numbers that correspond to those in Table 1. Black dots represent localities from which samples were collected.

the Clearwater drainage, which has complicated previous attempts at inferring population structure (Carstens et al., 2004). Using the MTOPS, we were able to reject a model of random mating among populations on all but the smallest geographic scales, a result consistent with the level of gene flow seen in other plethodontids (Larson, Wake & Yanev, 2004) and a finding with implications for conservation and management of the species. Populations isolated by long geographic distances do not appear to be exchanging migrants, and even in the populations for which we cannot reject panmixia, the actual $-\ln L$ score is close to the critical value. Taken together, these results suggest that *P. idahoensis* should not be expected to migrate at appreciable rates, and that recolonization following local extinction events is unlikely.

The results presented here demonstrate one application of parametric bootstrapping to intraspecific genetics. Parametric bootstraps have not been widely used to test evolutionary models that are not centered on topology in part because computational resources may be prohibitive for these applications. However, as distributed searches become more common parametric bootstrapping is likely to be applied to a variety of biological questions that can be stated in terms of a probabilistic model and a parameter(s) of interest. While MTOPS can be run on a desktop computer, it takes several weeks on a dual-processor Macintosh G4 for any one of the 16 tests presented here. We have conducted the majority of our analyses on a Beowulf cluster with 44 nodes, and distributed portions of the MIGRATE-N searches to different processors. In doing so, computation time is reduced so that any one of the tests presented here can be conducted in a matter of hours. As multi-node clusters become widely available, distributed searches and parametric bootstrapping will be used to implement a wide variety of statistical tests that can increase the inferences possible from genetic data. By enabling researchers to investigate the population structure and test for the genetic signature of population structure without the considerable time and expense involved in creating microsatellite or SNP libraries, the approach outlined above could be a valuable analytical tool for conservation biology because it tests the genetic cohesiveness of populations with organellar DNA data that are easily generated for non-model organisms. It is conceivable that the MTOPS could be used with other molecular data, such as SNPs or microsatellites, but recombination may produce genealogical patterns that are interpreted as migration by MIGRATE-N. Unlike multilocus methods, such as STRUCTURE (Pritchard, Stephans & Donnelly, 2000), this approach can be conducted using data from a single locus. While we expect analyses of data from a single locus to be less statistically powerful than multilocus methods, MTOPS provides a way to conduct a statistical test for population structure using only data from a single locus. As such, MTOPS is particularly useful

for preliminary investigations in a given system; perhaps to justify the development of SNP or microsatellite libraries in non-model organisms.

Acknowledgements

We thank V. Minin, Z. Abdo, and D. Rokyta for discussion pertaining this research. Funding for B. Carstens was provided by the University of Idaho Presidential Fellowship. Funding for computer hardware used for much of the computational aspect of this work was provided with funding from NIH National Center for Research Resources Grants P20 RR16454 for the BRIN Program, and P20 RR16448 for the COBRE program.

References

Abdo, Z., K.A. Crandall & P. Joyce, 2004. Evaluating the performance of likelihood methods for detecting population structure and migration. Mol. Ecol. 13: 837–851.

Beerli, P., 2002. MIGRATE-N: Documentation and Program, Part of Lamarck. Version.1.5. http://evolution.genetics.washington.edu/lamarc.html.

Carstens, B.C., A.L. Stevenson, J.D. Degenhardt & J. Sullivan, 2004. Testing nested phylogenetic and phylogeographic hypotheses in the *Plethodon vandykei* species group. Syst. Biol. 53: 781–792.

Grassly, N.C., P.H. Harvey & E.C. Holmes, 1999. Population dynamics of HIV-1 inferred from gene sequences. Genetics 151: 427–438.

Goldman, N.J., 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36: 182–198.

Huelsenbeck, J.P., D.M. Hillis & R. Jones, 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance, pp. 19–45 in Molecular Zoology: Advances, Strategies, and Protocols, edited by J.D. Ferraris & S.R. Palumbi. Wiley-Liss, New York, NY.

Larson, A., D.B. Wake & K.P. Yanev, 1984. Measuring gene flow among populations having high levels of genetic fragmentation. Genetics 106: 293–308.

Minin, V., Z. Abdo, P. Joyce & J. Sullivan, 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52: 674–683.

Pritchard, J.K., M. Stephans & P. Donnelly, 2000. Inference of population structure using multilocus genotypic data. Genetics 155: 945–959.

Slatkin, M., 1980. Estimating levels of geneflow in natural populations. Genetics 99: 323–335.

Sullivan, J., E. Arellano & D.S. Rogers, 2000. Comparative phylogeography of Mesoamerican highland rodents: Concerted versus independent responses to past climatic fluctuations. Amer. Nat. 155: 755–768.

Wright, S., 1978. Evolution and Genetics of Populations. Volume 4. Variability Within and Among Natural Populations. University of Chicago Press, Chicago IL.