MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) doi: 10.1111/1755-0998.12209

HZAR: hybrid zone analysis using an R software package

ELIZABETH P. DERRYBERRY,*†[1] GRAHAM E. DERRYBERRY,‡[1] JAMES M. MALEY†§ and ROBB T. BRUMFIELD†

*Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, LA 70118, USA, †Museum of Natural Science and Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA, ‡Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA, §Moore Laboratory of Zoology, Occidental College, 1600 Campus Road, Los Angeles, CA 90041, USA

Abstract

We present a new software package (HZAR) that provides functions for fitting molecular genetic and morphological data from hybrid zones to classic equilibrium cline models using the Metropolis–Hastings Markov chain Monte Carlo (MCMC) algorithm. The software applies likelihood functions appropriate for different types of data, including diploid and haploid genetic markers and quantitative morphological traits. The modular design allows flexibility in fitting cline models of varying complexity. To facilitate hypothesis testing, an autofit function is included that allows automated model selection from a set of nested cline models. Cline parameter values, such as cline centre and cline width, are estimated and may be compared statistically across clines. The package is written in the R language and is available through the Comprehensive R Archive Network (CRAN; http://cran.r-project.org/). Here, we describe HZAR and demonstrate its use with a sample data set from a well-studied hybrid zone in western Panama between white-collared (*Manacus candei*) and golden-collared manakins (*M. vitellinus*). Comparisons of our results with previously published results for this hybrid zone validate the HZAR software. We extend analysis of this hybrid zone by fitting additional models to molecular data where appropriate.

Keywords: clines, hybrid zones, *Manacus*, manakin, R package

Received 20 August 2012; revision received 27 November 2013; accepted 29 November 2013

Introduction

Speciation geneticists have long made use of interspecific crossing experiments in the laboratory to identify phenotypic traits and genetic loci (so-called speciation genes) that contribute to reproductive isolation (Coyne & Orr 2004). For organisms that are not especially fecund nor can be bred easily in captivity, hybrid zones provide a natural speciation genetics laboratory. Hybrid zones are geographic regions where genetically divergent taxa meet and hybridize (Barton & Hewitt 1985, 1989; Harrison 1990). These regions are thought to be stable over evolutionary time and can be maintained by environmental gradients (Ender 1977) or by a balance between selection against hybrids and the dispersal of parentals into the zone (Barton 1979, 2001; Barton & Gale 1993). In hybrid zones, genomic regions are exchanged differentially through recombination such that some regions could be impermeable to introgression (i.e. the movement of alleles from one taxon to another) while other regions introgress freely (Barton 1979; Harrison 1990; Rieseberg *et al.* 1999). Patterns of introgression can be measured using cline analyses where the transition, or cline, of genetic or morphological traits is estimated across a hybrid zone.

Cline theory provides a conceptual framework to understand the forces maintaining hybrid zones and to help infer the relevant evolutionary parameters describing the introgression of traits across hybrid zones. Genes or phenotypic traits characterized by transition clines that are narrow relative to the dispersal capabilities of the organism are thought to contribute more to reproductive isolation than genes and traits characterized by wider clines. Hybrid zones thus provide a powerful venue in which to examine the forces contributing to reproductive isolation between taxa and to explore the evolutionary potential of introgression (Barton 1979; Rieseberg *et al.* 1999).

The geographic structure of hybrid zones can be analysed by modelling the shape of clines for specific traits. These clines estimate changes in the population frequency of characters (alleles or phenotypes) along

Correspondence: Robb T. Brumfield, Fax: 225-578-3075; E-mail: Robb@lsu.edu

[1]Equal first authors.

geographic transects. Cline shape can be modelled by combining three equations (Szymura & Barton 1986, 1991) that describe a sigmoid shape at the centre of a cline and two exponential decay curves on either side of the central cline. Estimated parameters can then be used to determine concordance and coincidence of clines from different traits as well as to infer the strength of selection against hybrids.

Hybrid zone analysis for R (HZAR) is a package for the R programming environment (R-Development-Core-Team 2008), which fits molecular (allele frequency) and morphological data to equilibrium geographic cline models (Szymura & Barton 1986, 1991; Barton & Gale 1993; Gay *et al.* 2008) using the Metropolis–Hastings Markov chain Monte Carlo (MCMC) algorithm (Metropolis *et al.* 1953; Hastings 1970). HZAR is licensed under the GNU General Public Licence and leverages existing R libraries, MCMCpack (Martin *et al.* 2011), which provides the Metropolis–Hastings algorithm, and foreach (Revolution Analytics 2012), which distributes computing power.

HZAR was written to extend the capabilities of existing cline fitting software, to use the graphing and high-performance computing utilities of R and to provide an open-source package for the development of new methods by end-users. There are several software platforms in which to fit geographic cline models, including Analyse (Barton & Baird 1995), ClineFit (Porter *et al.* 1997) and Cfit (Gay *et al.* 2008). These programs are powerful, but are platform limited. Given the strong research focus on hybrid zones in evolutionary biology, and the shift towards implementing evolutionary models in R language (e.g. 'ape' (Paradis 2006), 'GEIGER' (Harmon *et al.* 2008), 'LASER' (Rabosky 2006), 'apTreeshape' (Bortolussi *et al.* 2006), 'INTROGRESS' (Gompert & Buerkle 2010), 'HIest' (Fitzpatrick 2013)), HZAR provides a powerful statistical tool for hybrid zone studies. HZAR has several significant strengths. The modular model set-up allows models of varying complexity to be fit to the observed data and automatic model selection functions search for the best model. This automation-friendly approach obviates user-implemented iterative model fitting. HZAR allows different, appropriate likelihood functions for genetic and morphological data. Finally, the R platform allows HZAR to leverage existing highly developed tools (such as R's advanced plotting functions), to be leveraged by other R libraries and to be extended to incorporate new methods. HZAR does not include all possible methods. We mention limitations where they apply in the description of the software below as well as suggest additional functions that could be added by open-source users. Users need to have some experience with R, particularly with the creation and handling of data frames, in order to use HZAR successfully. Here, we describe the current functionality of HZAR and illustrate its use by re-analysing a published data set for an avian hybrid zone (Brumfield *et al.* 2001, 2003).

Software usage and description

The functions contained in HZAR can be grouped into four analysis steps: data set-up, model description, cline model fitting and postprocessing. All functions are documented, and examples provided in the help file associated with the HZAR package (also see Data S3, Supporting Information). Briefly, we describe the types of functions associated with each of the analysis steps.

Current functionality of HZAR requires data collected along one-dimensional transects of natural hybrid zones. One-dimensional sampling of the hybrid zone assumes minimal variation perpendicular to the cline. Setting a one-dimensional transect in the wrong direction can introduce error (Macholán *et al.* 2008), and reducing sampling to a line of localities can result in loss of power and potential misinferences (Dufková *et al.* 2011). Additional sampling should be carried out perpendicular to the cline transect to ascertain the validity of using one-dimensional data on a case-by-case basis. A very useful future contribution to the package would be analysis of multidimensional geographic data.

HZAR includes a function to aid users in processing georeferenced data. This function generates a list of distances from a user-specified locality and accounts for the non-Euclidean geometry of geographic coordinates.

HZAR assumes genetic equilibrium within loci and equal relatedness across localities. The user should test their data beforehand for any departure from Hardy–Weinberg equilibrium (HWE). Localities with departure from HWE will have too much influence on fitting because the estimates of allele frequency at those localities will be overconfident (Phillips *et al.* 2004; Macholán *et al.* 2007, 2008). The user thus must account for departure from HWE by correcting effective sample sizes (Szymura & Barton 1986) before running these data in HZAR.

Data set-up

Data set-up functions assist in loading, formatting and processing input data. Analysis begins with an input file containing the genotypic or phenotypic data for a set of admixed localities and for the parental localities. To input genetic data, you will need one file. Input files for genetic data should include information on locality distance, allele frequency (between 0 and 1) and number of sampled alleles. (Note: users can run data in which the

coordinates of each individual are unique. You can generate a data object with one individual at each locality. Assuming that each individual has two alleles then, the observed frequencies would be 0, ½ or 1 with a sample size of two. For haploid samples, the observed frequencies would be 0 or 1 with a sample size of one.). If all loci share the same information for distance and number of sampled alleles, then these two columns of information only need to be included once. However, if loci have different numbers of sampled alleles, then one should include a separate column of number of alleles sampled for each locus. Header row labels must begin with a letter and include only letters, numbers and a period (.). The table should include one additional row for each locality that has been sampled. Column order is not set, but we recommend including as the first column the names of the sample sites, the second column the distance information, and each subsequent pair of columns the frequency and the sample size of each individual allele. These data can be formatted in spreadsheet software and then saved as a text file that can be read into R using standard functions, such as `read.table` or `read.csv` (R-Development-Core-Team 2008). Read the help files for `read.table` or `read.csv` to insure correct formatting. Running the example scripts provided in supplemental material will output example data sets from the HZAR package to provide models for formatting input files.

Once the allele frequency data are imported in table format, it is necessary to create a data object (named `hzar.obsData`) for each allele for each locus. This object is created using the function `hzar.doMolecularData1DPops`. If you are creating more than one of these objects for molecular data, then it is useful to collect all of these objects in a list. This list of objects can be created using the `list` method in R. The arguments are the data objects (see example file below). It is recommended that you name each entry in the list. For information on how to do this, see the help file for `name` in R (Table 1).

Quantitative trait data (such as body weight) can be imported in the format it was measured. For example, body weight observations could be imported in grams. HZAR requires that quantitative traits be reasonably approximated by normal distributions. If the data are far from normal, then the user can transform the data according to allometric considerations (e.g. a log transform). If the transformed data pass a test for normalcy, then the transformed data can be used for quantitative trait cline fitting (Barton & Gale 1993). To input quantitative trait data, you will need one file. Input files should include information on locality distance, observed trait means, observed trait variances and number of sampled individuals. If all traits share the same information for distance and number of sampled individuals, then these two columns of information only need to be included once. However, if traits have different numbers of sampled individuals, then one should include a separate column of number of individuals sampled for each trait. Header row labels must begin with a letter and include only letters, numbers and a period (.). The table should include one additional row for each locality that has been sampled. Column order is not set, but we recommend including as the first column the names of the sample sites, the second column the distance information and each subsequent set of columns per trait the observed trait means, observed trait variances and the number of individuals sampled. These data can be formatted in spreadsheet software and then saved as a text file that can be read into R using standard functions, such as `read.table` or `read.csv` (R-Development-Core-Team 2008). Read the help files for `read.table` or `read.csv` to insure correct formatting. See Table 2 and example data set 'manakinQuantitative' in the HZAR package.

Table 1 Allele frequencies and number of samples for the hybrid zone between *Manacus candei* and *M. vitellinus* in northwestern Panama (Brumfield *et al*. 2001). These frequencies are taken from the example data file distributed with the software package

Locality_ID	Locality	Distance_(km)	*Ada*[a]	*Ada*[b]	*Ada*_nSamples
A	Costa_Rica	0.00	0.10	0.90	10
B	Rio_Sixaola	138.25	0.36	0.64	14
C	Rio_Teribe	151.75	0.20	0.80	40
D	Rio_Changuinola	159.50	0.45	0.55	40
E	Rio_Oeste	182.25	0.31	0.69	42
F	Quebrada_Pastores	188.25	NA	NA	NA
G	Tierra_Oscura	198.50	0.55	0.46	22
H	Rio_Uyama	201.25	0.57	0.43	44
I	Rio_Robalo	210.00	0.44	0.56	52
J	Chiriqui_Grande	230.75	0.63	0.38	40
K	Valiente_Peninsula	319.50	0.43	0.58	40
L	Soberania	569.50	0.68	0.33	40

Table 2 Showing data for beard length (mm) to demonstrate input format appropriate for HZAR. Locality ID corresponds to the Locality ID for molecular data (along with Locality name) and nSamples refers to sample size (number of alleles)

Locality_ID	Locality	Distance_(km)	Observed_mean	Observed_variance	nSamples
A	Costa_Rica	0.00	12.43	0.80	21
B	Rio_Sixaola	138.25	11.75	0.92	4
C	Rio_Teribe	151.75	11.94	0.65	9
D	Rio_Changuinola	159.50	12.22	0.69	9
E	Rio_Oeste	182.25	13.17	0.95	15
F	Quebrada_Pastores	188.25	12.50	0.25	5
G	Tierra_Oscura	198.50	12.63	3.25	11
H	Rio_Uyama	201.25	12.18	1.48	20
I	Rio_Robalo	210.00	16.33	5.59	20
J	Chiriqui_Grande	230.75	17.40	1.60	10
K	Valiente_Peninsula	319.50	18.80	3.70	5
L	Soberania	569.50	18.68	2.58	22

Once the quantitative trait data are imported in table form, it is necessary to create a data object (named `hzar.obsData`) for each observed trait. This object is created using the function `hzar.doNormalData1DPops`. If you are creating more than one of these objects, then it is useful to collect all of these objects in a list. This list of objects can be created using the `list` method in R. The arguments are the data objects (see example file below). It is recommended that you name each entry in the list. For information on how to do this, see the help file for `name` in R.

If you want to fit clines to all of the genetic and quantitative trait data objects at once, then you need to make a combined list of these objects. You can join a list of quantitative trait data objects to a list of genetic data objects using the `c(list1, list2)` function in R. If either `list1` or `list2` was not named before joining, then the combined list will not be named (see help file for `name` in R).

Model description

Functions for model description are used to create objects that will be passed to functions for fitting cline models. This is a two-step process for both genetic and quantitative trait (e.g. morphological) data. You first create a cline model object (object name, clineMetaModel). For genetic data, you pass the `hzar.obsData` object and the arguments describing the scaling and tails of the cline model that you want to fit to `hzar.makeCline1DFreq`. For quantitative trait data, you pass the `hzar.obsData` object and the arguments describing the tails of the cline model to `hzar.makeCline1DNormal`. For frequency based clines, HZAR automatically selects ascending or descending direction as informed by the observed data. For quantitative trait data, the direction of the cline does not matter. If you want to use a list of `hzar.obsData` objects, you can use the R function `lapply`. For molecular data, fifteen different model variants can be described in `hzar.makeCline1DFreq`. These fifteen models represent three possible combinations of trait interval [pMin, pMax] (fixed to 0 and 1; observed values; estimated values) and five possible combinations of fitting tails (none fitted; left only; right only; mirror tails; both tails estimated separately). There is also a sixteenth cline model, which is a null model. This last model is only used in the postprocessing stage (see below). All sixteen models can be fit to molecular data. For quantitative trait data, five different types of models can be described in `hzar.makeCline1DNormal`. All models estimate trait mean and variance on the left and right and additional variance in the centre, as well as centre and width. The models vary in fitting exponential tails (none fitted; left only; right only; mirror tails; both tails estimated separately). To test hypotheses, parameter values can be fixed at a specific value while other parameters are estimated. One can also edit clineMetaModel objects to place limits on parameters. For example, one can place limits on the width of the cline or on the centre of the cline to limit the amount of parameter space explored. This increases the speed and efficiency of the MCMC process.

The second step of the process is to create an `hzar.fitRequest` object using the function `hzar.first.fitRequest.old.ML`. This function takes the `clineMetaModel` object and the `hzar.obsData` object to create the `hzar.fitRequest` object. An additional parameter can be set to report diagnostics data. Some modifications can be made to the `hzar.fitRequest` object; you can set the chain length, the burn-in, the thin (generations subsampled) and the random number generator seed used in the MCMC process. The default settings for these

parameters of the MCMC operator are adequate in most cases, but the user can make modifications if desired.

Cline model fitting

To begin the process of fitting the desired cline model, the user passes the `hzar.fitRequest` object to the function `hzar.doFit`. This function is used to fit the cline model. The output of this function is a second `hzar.fitRequest` object that contains the information from the first object plus the results of the fit. The user then takes this second object and passes it to the function `hzar.nextFitRequest`. This function uses the results of the previous fit to optimize the covariance matrix used to drive the MCMC process but does not fit a cline model. This function updates the seed channel of the random number generator so that the random number process for the second fit request is independent of the first fit request. The default setting for the second fit request has the same values for the chain length, the burn-in and the thinning, but the user can change these settings if desired. The output of `hzar.nextFitRequest` is a third `hzar.fitRequest` object that can then be passed back to `hzar.doFit`. The user can repeat this process iteratively to create new runs of the same MCMC chain. It is important to note that these runs within the same chain are not independent of one another.

To check whether runs are stable and converging, one can plot the raw data from the MCMC process using a standard plot function of the MCMC raw entry (*foo* `$mcmcRaw`) from an `hzar.fitRequest` object returned from `hzar.doFit`. Plotting the raw MCMC data returns a plot with two columns: the first column is the trace (value of the parameter versus the number of generations), and the second column is the density distribution (relative density vs. parameter value). The number of rows returned is equal to the number of free parameters. For example, plotting the raw MCMC data for a model in which only width and centre vary will return plots of width and centre. Fig. 1 illustrates what these data look like.

To create an independent chain, the user must pass the first `hzar.fitRequest` object directly to the

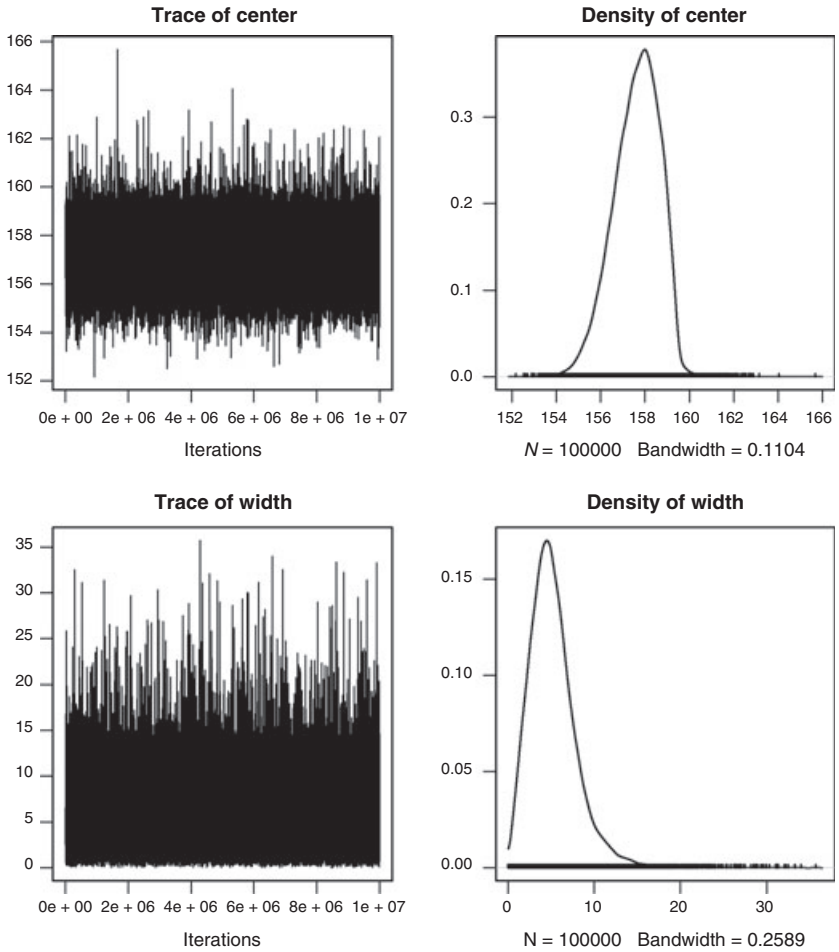

Fig. 1 A plot of the raw MCMC data for a model in which only width and centre vary. The first column is the trace (value of the parameter vs. the number of generations), and the second column is the density distribution (relative density vs. parameter value). Y-axis in left panels and X-axis in right panels are the values of estimated parameter (km). Y-axis in right panel is the density distribution.

function `hzar.nextFitRequest` to create a new second `hzar.fitRequest` object. This step does not fit the cline model, but it does create an independent object. The user then passes this new second `hzar.fitRequest` object to `hzar.doFit`. The output `hzar.fitRequest` object is then passed again to `hzar.nextFitRequest` to create a new `hzar.fitRequest` object. As described above, the user can repeat this process iteratively to create new runs for this second, independent MCMC chain. The user can repeat this process to create as many independent chains with as many dependent runs as desired.

There is a function that automates the within-chain process described above called `hzar.chain.doSeq`. This function iteratively passes `hzar.fitRequest` objects to `hzar.doFit` and then to `hzar.nextFitRequest` as many times as the user specifies. This function then returns a list of all the `hzar.fitRequest` objects that contain results. If the argument `collapse` of `hzar.chain.doSeq` is set to `TRUE`, then the function will return a single `hzar.fitRequest` object with all of the results concatenated. Postprocessing functions can use either output format.

If you have a list of `hzar.fitRequest` objects representing independent chains, then you can use the R function `lapply` with `hzar.chain.doSeq` to run all of the independent chains automatically. The return will be a list of all of the results. Users familiar with R can use functions (e.g. `lapply` or `mclapply`) to automate this process for fitting multiple models. Functions exist within postprocessing to handle complex lists of results as long as all of the models share the same `hzar.obsData` objects.

To provide a benchmark of run-time, if one runs the example script in this manuscript, which fits three different cline models to one molecular locus and one morphological trait, the total run-time is 15.5 min (on a Linux (64bit, v2.6.38) 2.3 GHz Intel Core 2 CPU). Each of the fittings includes ten runs of 100,000 generations each.

Postprocessing

Once clines are fit, HZAR provides a set of postprocessing functions. These functions can aggregate multiple fits to allow for model comparison and model selection using AIC. They can also generate summary statistics, confidence intervals and plots of results. We have also included limited ability to generate likelihood profiles of single parameters. The R platform allows users to extend these functions as needed.

To perform model selection, it is necessary to aggregate the results from multiple cline fits if models were fit individually, or to split results if models were fit using batch processing. The function `hzar.dataGroup.add` aggregates results from the same model for the same `hzar.obsData` object. The output of this function is an `hzar.dataGroup` object. To concatenate the results for the same `hzar.obsData` object and automatically split different models, the user can use the function `hzar.make.obsDataGroup`. The output of this function is an `hzar.obsDataGroup` object. The function `hzar.dataGroup.null` represents the null model hypothesis that there is no cline in the sampled region. This function generates an `hzar.dataGroup` object that can be compiled into the `hzar.obsDataGroup` object as a point of comparison in model selection. These objects can be passed to model selection functions.

HZAR compares model performance using Akaike information criterion (AIC) or the AIC score corrected for small sample size (AICc). For a given trait, the model with the lowest AIC (AICc) score is the model that best fits the data. The function `hzar.AIC.default` (`hzar.AICc.default`) calculates the AIC (AICc) for the given likelihood, number of parameters and number of observations for a given object. The function `hzar.AIC.hzar.obsDataGroup` (`hzar.AICc.hzar.obsDataGroup`) returns a table of AIC (AICc) values for all of the models contained in an `hzar.obsDataGroup` object. The package also provides the function `hzar.get.ML.cline` for extracting the maximum-likelihood cline from an `hzar.dataGroup` object. Summary statistics for model fitting can be extracted from an `hzar.dataGroup` object using `hzar.getLLCutParam`. This function returns the range of parameter values that are within two log-likelihood units of the maximum likelihood for a provided character vector of parameters. Using a 2-unit support envelope around a cline allows one to visualize uncertainty in model fit (Devitt *et al.* 2011; Macholan *et al.* 2011).

A likelihood profile is a representation of the maximum likelihood of a model as a function of a single fixed parameter value, such as centre (Phillips *et al.* 2004; Alexandrino *et al.* 2005; Bimova *et al.* 2011; Devitt *et al.* 2011). The independent variable is a range of parameter values, and the dependent variable is the maximum likelihood for the given parameter value. For example, the independent variable may be a range of centre values (such as 10, 20 or 30 km), and the dependent variable is the maximum likelihood of the model given each of those fixed centre values. Users may want to examine likelihood profiles in order to gauge the quality of the model fitting or to compare cline shape across multiple genetic markers and/or quantitative traits. Profiling reduces the dimensionality of

MCMC searches, thus reducing overconfidence in estimates by increasing the chance that the search finds all solutions within 2 units of support. Profiling can also facilitate nested hypothesis tests, for example, for concordance and coincidence. A straightforward method to create the likelihood profile involves estimating the maximum likelihood at a series of fixed parameter values. Users can generate likelihood profiles manually by choosing one model as the base (typically the best model identified during model selection) and deriving secondary models where the specific parameter is fixed to each value. In HZAR, this process is automated to a certain extent. The workflow is such that users give the function `hzar.profile.dataGroup` an `hzar.dataGroup` object of the desired model to profile and the desired parameter to profile. This function will extract the model from the data group and use either additional parameters given to the method or estimates extracted from the data group. The result is a list of `hzar.fitRequest` objects. Each of those objects reuses the original model associated with the data group but fixes the specified parameter to each of a series of values. The series of values are derived by default based upon the information in the `hzar.dataGroup` object, but can be specified explicitly by the user. The list of `hzar.fitRequest` objects then needs to be fitted using methods described above in cline model fitting. There is a set of specialized methods to automate this process, namely `hzar.doChain.multi` and `hzar.doFit.multi`. Please see help files for details on the functions `hzar.profile.dataGroup`, `hzar.multiFitRequest`, `hzar.doChain.multi` and `hzar.doFit.multi`. As the output from the fitted objects is a series of models with the same observational data, the trace information can be compiled for postprocessing using the method `hzar.make.obsDataGroup`. Extracting the estimated maximum-likelihood value for each of the parameter values is not yet automated, but can be performed manually to produce a graph of the likelihood profile.

Finally, HZAR provides a series of plotting functions for plotting results. These plotting functions can be used in conjunction with generic R plotting functions to format plots for manuscript preparation. Here, we describe the three most important plotting functions provided in HZAR. The function `hzar.plot.obsData` plots observation data (i.e. mean frequencies for molecular clines and mean values for morphological clines). This function can be used to check whether data were entered correctly. The function `hzar.plot.cline` plots the cline from an `hzar.cline` object, the maximum-likelihood cline (and, by default, the observation data) from an `hzar.dataGroup` results object or the maximum-likelihood cline for each model on top of the observation data from a `hzar.obsDataGroup` object.

HZAR also allows the user to produce fuzzy cline plots, which allow the user to view the distribution of uncertainty in a model prediction. A fuzzy cline plot allows the user to view the distribution of the estimated true local mean given the selected model. The function `hzar.plot.fzCline`, using an `hzar.dataGroup` object, plots the maximum-likelihood cline and observed frequency data over the associated fuzzy cline region. This function provides two separate approaches to constructing the fuzzy cline region. The first method extracts a subset of the cline model distribution generated from the MCMC trace, either using a subset which is 95% credible or the subset that is within two log-likelihood units of the maximum likelihood. The approach takes a set of distances (e.g. 20 evenly spaced points) across the observed sample localities and calculates the maximum and minimum estimated allele frequency (or mean trait value) at each distance for all of the clines in the model subset. The fuzzy cline region is the region enclosed by these maximum and minimum values across the transect. The second method constructs the fuzzy cline plot (using either the entire MCMC trace of clines or a randomly chosen subset of clines) by calculating the estimated allele frequency (or mean trait value) for each of those clines and then uses either the 95% confidence interval or 95% credible interval of that distribution for each locality. The default region for the function is the region enclosed by the maximum and minimum values of the 95% credible subset of the cline model distribution. Please see Fig. 2 for an example of such a plot. Use of this function depends on a stable, convergent MCMC trace. The two log-likelihood unit methods and the 95% confidence interval method are both dependent on the MCMC traces being well sampled. The second method for constructing the fuzzy cline plots is more computationally intensive than the first method.

Analysis of a sample data set: Manakin Hybrid Zone

Brumfield *et al.* (2001, 2003) fit cline models to a series of molecular and morphological data sets from a natural hybrid zone between two species of manakins (*Manacus candei* and *M. vitellinus*). This analysis was originally conducted using the software program Analyse v1.3 (Barton & Baird 1999). Here, we repeat this analysis using the same data sets but on the HZAR platform in R. We use the same sample sizes as Brumfield *et al.* (2001) to make our results comparable. Overall, our results are similar to those reported in Brumfield *et al.* (2001), demonstrating the functionality of HZAR. To demonstrate

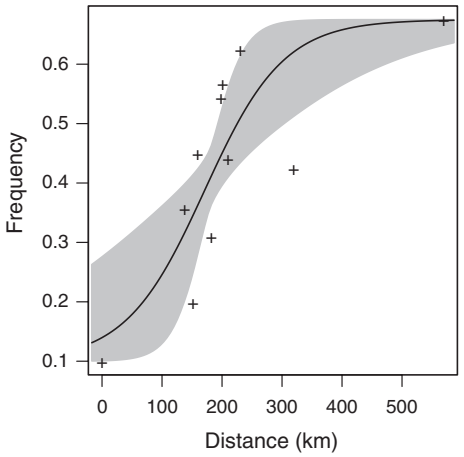

Fig. 2 A plot of the maximum-likelihood cline and observed frequency data over the associated fuzzy cline region (95% credible cline region) as returned by the function `hzar.plot.fzCline`.

automated model fitting, we fit all possible models to these data sets. Below, we provide a detailed comparison of these analyses.

Molecular data

Brumfield *et al.* (2001) fit three cline models to seven genetic loci using the program Analyse v1.3 (Barton & Baird 1999). The three models were as follows: model I: p_{min}/p_{max} set to observed values with no exponential decay curves (tails) fitted, model II: p_{min}/p_{max} estimated with no tails fitted and model III: p_{min}/p_{max} estimated and both tails fitted. We fit these same models to the same genetic loci using HZAR (e.g. script and output fitting these three models to one of the loci, please see Data S1 and Data S2, Supporting Information), and compared our results to those generated by Analyse v1.3. One important difference between the two software platforms is the parameters used to estimate tails. Analyse v1.3 fits tails using the parameters β and Θ, whereas HZAR uses δ and τ, which represent the parameters d and t described by Gay *et al.* (2008). The latter parameters can be transformed to the former parameters. Because of this difference in parameter usage, we expected quantitative but not qualitative differences in our results for model III.

We also fit the 15 possible models to the same seven genetic loci to demonstrate the full functionality of HZAR. Users must select cline models to compare based on the biological relevance of those cline models to their organism.

Results. Of the three models, both programs selected the same model as the best-fit model for each of the seven genetic loci (Table 3). The log-likelihood scores were nearly identical across software platforms (Table 3). For the three models, we also compared estimated parameters for the best-fit model for each of the genetic loci. Values for all estimated parameters were identical or nearly identical for all seven genetic loci (Table 4). We plotted the maximum-likelihood clines for the best-fit model for each of the seven genetic traits to facilitate comparison with the results in Brumfield *et al.* (2001) (see Fig. 3).

Table 3 Log-likelihood scores for fitted clines under different models using Analyse v1.3 (nonshaded columns) and HZAR (shaded columns)

Locus	ln L Model I	ln L & AICc Model I	ln L Model 2	ln L & AICc Model 2	ln L Model 3	ln L & AICc Model 3
Ada[a]	−9.6*	−9.7	−8.3	−8.2	−7.9	−8.3
		23.3*		24.6		32.9
Ak-2[a]	−43.9	−43.9	−6.2*	−6.3	−6.2	−5.7
		91.8		20.6*		27.8
Pgm−2[b]	−17.7	−17.7	−6.7*	−6.7	−6.7	−5.9
		39.4		21.5*		28.1
Gsr[b]	−15.7	−15.7	−5.8*	−5.8	−5.0	−6.5
		35.4		19.8*		29.3
λ5[b]	−36.9	−34.6	−12.2	−11.6	−6.6*	−6.4
		73.2		31.2		29.2*
pSCN3[b]	−17.9	−19.1	−2.6*	−2.6	−1.2	−1.4
		42.3		13.3*		19.1
mtDNA[b]	−19.1	−21.3	−1.8*	−1.9	−1.8	−1.9
		46.7		11.9*		20.5

*Denotes the model that provides the best fit with the fewest number of estimated parameters.
Brumfield *et al.* (2001) compared models using goodness-of-fit tests, and we compared models using AICc values (lowest value indicates best fit). Table is modified from Brumfield *et al.* (2001) Table 3.

Table 4 Parameter estimates for the genetic clines using Analyse v1.3 (nonshaded columns) and HZAR (shaded columns)

Locus	w	c	p_{min}	p_{max}
Ada[a]	262.5 (110.0–856.4)	171.0 (108.7–200.1)	0.1 (fixed)	0.675 (fixed)
	263.1 (96.2–630.0)	170.9 (109.9–201.2)	0.1 (fixed)	0.675 (fixed)
Ak-2[a]	9.9 (0.9–15.3)	208.6 (207.3–210.2)	0.0 (0.0–0.1)	1.0 (1.0–1.0)
	9.5 (0.1–15.7)	208.7 (207.2–210.2)	0.0 (0.0–0.1)	1.0 (0.9–1.0)
Pgm-2[b]	7.7 (0.5–12.0)	206.5 (201.5–209.5)	0.0 (0.0–0.0)	0.8 (0.7–0.8)
	7.7 (0.07–12.2)	206.4 (201.3–209.9)	0.0 (0.0–0.0)	0.8 (0.7–0.8)
Gsr[b]	2.9 (0.2–37.9)	209.9 (202.1–223.1)	0.0 (0.0–0.1)	0.3 (0.2–0.4)
	5.7 (0.009–44.3)	209.9 (201.7–224.5)	0.0 (0.0–0.1)	0.3 (0.2–0.4)
λ5[b]	10.4 (7.2–14.4)	208.2 (206.2–208.4)	0.1 (0.1–0.1)	1.0 (1.0–1.0)
	8.1 (2.9–18.4)	208.1 (204.7–210.0)	0.1 (0.1–0.1)	1.0 (1.0–1.0)
pSCN3[b]	2.8 (0.6–15.4)	209.4 (206.7–209.9)	0.2 (0.1–0.2)	1.0 (0.9–1.0)
	1.4 (0.1–15.5)	209.7 (206.6–210.0)	0.2 (0.1–0.2)	1.0 (0.9–1.0)
mtDNA[b]	11.1 (6.9–19.0)	208.3 (206.4–210.3)	0.0 (0.0–0.0)	0.9 (0.9–1.0)
	11.2 (6.8–19.5)	208 (206.0–210.8)	0.0 (0.0–0.0)	1.0 (0.9–1.0)
Locus	β_c/δ_L	Θ_c/τ_L	B_v/δ_R	Θ_v/τ_R
λ5[b]	1535.9 (1520–1556)	1.0 (0.2–1.0)	3.1 (2.6–13.7)	0.1 (0.0–0.1)
	204.0 (4.0–628.9)	0.3 (0.0–1.0)	1.2 (0.0–7.8)	0.1 (0.0–0.6)

Two log-likelihood unit support limits are presented in parentheses. Cline width is presented as 1/maximum slope (w). Parameter c is the cline centre measured in distance (km) from locality 1, p_{min} is the minimum estimated frequency at the west end of the cline, and p_{max} is the maximum estimated frequency at the eastern end. The exponential decay curves (tails) shape parameters are given as β and Θ for Analyse v1.3 results and as δ and τ for HZAR results. Table modified from Brumfield *et al.* (2001) Table 4.

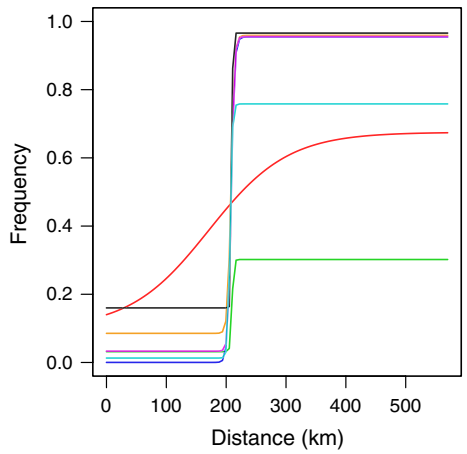

Fig. 3 Plot of allele frequency vs. distance for seven genetic loci describing the *Manacus* hybrid zone in Panama. Each line is the maximum-likelihood cline for the best-fit model from the original three models compared in Brumfield *et al.* (2001). Legend: *Ada*[a] (red); *Ak-2*[a] (purple); *Gsr*[b] (green); *Pgm-2*[b] (cyan); *λ5*[b] (orange); *pSCN3*[b] (black); *mtDNA*[b] (blue).

We next compared the fit of all 15 models for the seven genetic loci. In some cases, we found that the best-fit model was different from the original model found by Brumfield *et al.* (2001), although in most cases the original model provided the best or an equal fit (Table 5). If we adopt the convention of selecting as the best model one that has an AICc value more than two points less than the AICc of the next best model, then we found a best-fit model for two of the seven loci (*λ5*[b] and *pSCN3*[b]) different from Brumfield *et al.* (2001). For *λ5*[b], the original best-fit model was one that estimated p_{min}/p_{max} and fits both tails separately. The new best-fit model for this locus is one that estimated p_{min}/p_{max} and fits only a right tail. For *pSCN3*[b], the original best-fit model was one that estimated p_{min}/p_{max} and fits no tails. The new best-fit model for this locus is one that used the observed values of p_{min}/p_{max} and fit mirror tails (Table 5).

Table 5 The AICc value for the best-fit model from the Brumfield *et al.* (2001) three-model comparison compared with the AICc value for the best-fit model from the 15-model comparison

Locus	Brumfield best-fit model	15-model best-fit model
Ada[a]	23.3 (p_{min}/p_{max} observed, no tails)	22.9 (p_{min}/p_{max} fixed, right tail)
Ak-2[a]	20.6 (p_{min}/p_{max} estimated, no tails)	19.5 (p_{min}/p_{max} fixed, mirror tails)
Pgm-2[b]	21.5 (p_{min}/p_{max} estimated, no tails)	21.5 (p_{min}/p_{max} estimated, no tails)
Gsr[b]	19.8 (p_{min}/p_{max} estimated, no tails)	19.7 (p_{min}/p_{max} estimated, no tails)
λ5[b]	29.2 (p_{min}/p_{max} estimated, both tails)	25.1 (p_{min}/p_{max} estimated, right tail)
pSCN3[b]	13.3 (p_{min}/p_{max} estimated, no tails)	9.3 (p_{min}/p_{max} observed, mirror tails)
mtDNA[b]	11.9 (p_{min}/p_{max} estimated, no tails)	10.6 (p_{min}/p_{max} fixed, right tail)

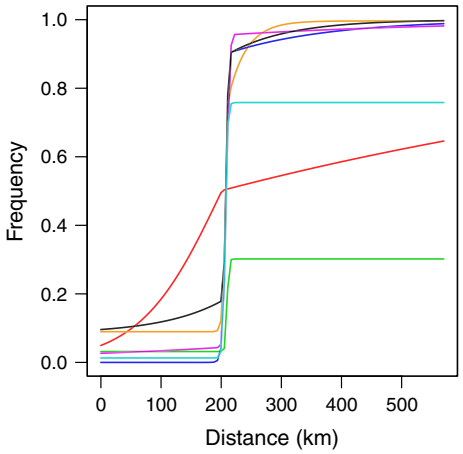

Fig. 4 Plot of allele frequency vs. distance for seven genetic loci describing the *Manacus* hybrid zone in Panama. Each line is the maximum-likelihood cline for the best-fit model from the 15-model comparison. Legend: Ada^a (red); $Ak\text{-}2^a$ (purple); Gsr^b (green); $Pgm\text{-}2^b$ (cyan); $\lambda 5^b$ (orange); $pSCN3^b$ (black); $mtDNA^b$ (blue).

Table 6 AICc values for five quantitative trait model comparisons for each of the morphological traits

Models	Collar colour	Beard length	Epaulette width	Belly colour
model.none	921.2	544.2	603.7	1088.6
model.left	925.8	544.3	598.4*	1092.4
model.right	843.5*	548.2	605.1	1062.5*
model.mirror	867.1	542.8*	607.6	1090.4
model.both	848.0	550.2	604.9	1080.0

The models vary in fitting exponential tails (model.none = none fitted; model.left = left only; model.right = right only; model.mirror = mirror tails; model.both = both tails estimated separately).
*indicates the best-fit model.

We also plotted the maximum-likelihood cline for the best-fit model for each trait from the 15-model comparison (Fig. 4). One thing to note from the figure is that although the best-fit model for Ada^a did not have an AICc value much lower than that of the original best-fit model, the maximum-likelihood cline of this model (illustrated in red in Fig. 4) now indicates a cline centre coincident with the other seven genetic loci, in contrast to the displaced centre found by Brumfield *et al.* (2001).

Morphological data

Brumfield *et al.* (2001) took 11 morphological measurements on specimens from the *Manacus* hybrid zone. Four of these morphometric characters (mass, wing length, tail length and tarsus length) were collapsed into a principle

Table 7 Parameter estimates for the morphological clines using Analyse v1.3 (nonshaded columns) and HZAR (shaded columns)

Locus	w	c
Collar colour	4.4	157.6
	1.2 (0.1–5.6)	155.4 (152.3–159.3)
Beard length	10.3	208.8
	9.9 (4.4–15.4)	208.5 (207.7–209.8)
Epaulette width	65.2	200.0
	30.0 (15.0–34.0)	201.1 (198.6–203.5)
Belly colour	3.0	157.8
	4.3 (0.2–7.6)	159.2 (153.7–159.8)

Locus	β_c/δ_L	Θ_c/τ_L	B_v/δ_R	Θ_v/τ_R
Collar colour	96.4	0.4	54.2	0.0
	no tail	no tail	0.4 (0.0–2.4)	0.01 (0.0–0.05)
Beard length	1.3	0.8	0.9	0.2
	7.3 (2.7–8.9)	0.08 (0.05–0.27)	7.3 (2.7–8.9)	0.08 (0.05–0.27)
Epaulette width	7500.0	0.9	12.3	0.1
	0.6 (0.2–2.9)	0.18 (0.09–0.26)	no tail	no tail
Belly colour	1041.2	0.2	8.2	0.0
	no tail	no tail	0.1 (0.0–0.9)	0.06 (0.0–0.11)

Two log-likelihood unit support limits are presented in parentheses. Cline width is presented as 1/maximum slope (w). Parameter c is the cline centre measured in distance (km) from locality 1. The exponential decay curves (tails) shape parameters are given as β and Θ for Analyse v1.3 results and as δ and τ for HZAR results. Table modified from Brumfield *et al.* (2001) Table 4.

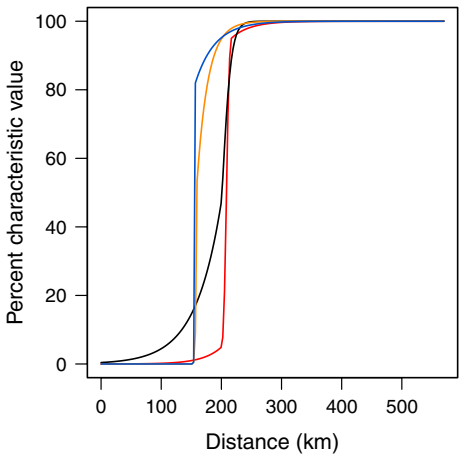

Fig. 5 Plot of the per cent characteristic value vs. distance for four morphological traits describing the *Manacus* hybrid zone in Panama. To facilitate cline comparison, results are plotted as ascending clines, which for quantitative data necessitates plotting the per cent characteristic value from the unscaled cline function. Each line is the maximum-likelihood fundamental cline for the best-fit model from the five-model comparison. Legend: beard length (red); belly colour (orange); epaulette width (black); collar colour (blue).

component vector (PC1) capturing size variation. The four characters describing plumage variation – belly colour, collar colour, beard length and epaulette width – were fitted separately. Brumfield *et al.* (2001) fit the most general model of the three fitted to the genetic loci (model III). Model III estimates p_{min}/p_{max} and fits both tails. Before clines were fit to the morphological character locality means, individual measurements were scaled to values between zero and one. The clines for PC1 and beard length were similar in position to the genetic clines, and the cline for epaulette width was broader and slightly northwest of these clines' centres but not dramatically displaced. The two plumage colour clines, however, transitioned steeply about 50 km northwest of the other cline centres and coincident with the Changuinola River.

Using HZAR, we fit the five different quantitative trait models to the four characters describing plumage variation: beard length, collar colour, belly colour and epaulette width (e.g. script and output fitting these models to one of the traits, please see Data S1 and Data S2, Supporting Information). HZAR requires that quantitative traits be reasonably approximated by normal distributions, and past studies suggest that quantitative trait likelihood functions can be fit as long as the parental populations can be reasonably approximated by normal distributions (Barton & Gale 1993). This is the case for the manakin parental populations for these four quantitative traits. We selected the best-fit model based on comparison of AICc values.

Results. A model with a right tail to the cline provided the best fit to collar colour and belly colour, whereas a model with a left tail provided the best fit to epaulette width (Table 6). For beard length, a mirrored left tail and right tail provided the best fit (Table 6).

Because we did not rescale the plumage characters as in Brumfield *et al.* (2001), we can perform a qualitative if not quantitative comparison. In Table 7, we provide the estimated value and the two log-likelihood unit support limits for each estimated parameter for each trait for the best model. A comparison of the Analyse v1.3 and HZAR results found a good agreement between the estimated centre values. The width estimates fell within two log-likelihood units for collar colour, belly colour, and beard length, but not for epaulette width.

We also plotted the maximum-likelihood cline for the best-fit model for each trait from the five-model comparison (Fig. 5). Note that similar to the results in Brumfield *et al.* (2001), we found that the centres of the beard length and epaulette width clines are similar to those of the seven genetic loci, whereas the collar colour and belly colour clines are shifted northwest by approximately 50 km, coincident with the Changuinola River.

Acknowledgements

We thank S. Baird, M. Blum, O. Gaggiotti, A. Porter and an anonymous reviewer for helpful comments on earlier drafts of the manuscript and software package. We are also grateful to individuals who beta-tested the software package, including S. Lipshutz, S. Shaak, C. Duffie, S. Singhal. This work was funded by NSF Grant DEB-0841729 to RTB and BoR Pfund 167 to RTB and EPD.

References

Alexandrino J, Baird SJ, Lawson L *et al.* (2005) Strong selection against hybrids at a hybrid zone in the *Ensatina* ring species complex and its evolutionary implications. *Evolution*, **59**, 1334–1347.

Barton NH (1979) Gene flow past a cline. *Heredity*, **43**, 333–339.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551–568.

Barton NH, Baird SJE (1995) *Analyse: An Application for Analysing Hybrid Zones*. Freeware, Edinburgh, UK.

Barton NH, Baird SJE (1999) Analyse: software for the analysis of geographic variation and hybrid zones. Ver.1.03 for Macintosh. University of Edinburgh, Edinburgh.

Barton NH, Gale KS (1993) Genetic analysis of hybrid zones. In: *Hybrid Zones and the Evolutionary Process*(ed. Harrison RG), pp. 13–42. Oxford University Press, Oxford.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Barton NH, Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.

Bimova BV, Macholan M, Baird SJE *et al.* (2011) Reinforcement selection acting on the European house mouse hybrid zone. *Molecular Ecology*, **11**, 2403–2424.

Bortolussi N, Durand E, Blum M, François O (2006) apTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*, **22**, 363–364.

Brumfield RT, Jernigan RW, McDonald DB, Braun MJ (2001) Evolutionary implications of divergent clines in an avian (*Manacus*: Aves) hybrid zone. *Evolution*, **55**, 2070–2087.

Brumfield RT, Jernigan RW, McDonald DB, Braun MJ (2003) Erratum correcting "Evolutionary implications of divergent clines in a manakin (Manacus; Aves) hybrid zone. *Evolution*, **57**, 2919.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Sunderland, MA.

Devitt TJ, Baird SJE, Moritz C (2011) Asymmetric reproductive isolation between terminal forms of the salamander ring species *Ensatina eschscholtzii* revealed by fine-scale genetic analysis of a hybrid zone. *BMC Evolutionary Biology*, **11**, 245.

Dufková P, Macholán M, Piálek J (2011) Inference of selection and stochastic effects in the house mouse hybrid zone. *Evolution*, **65**, 993–1010.

Ender JA (1977) *Geographic Variation, Speciation and Clines*. Princeton University Press, Princeton, NJ.

Fitzpatrick BM (2013) Alternative forms for genomic clines. *Ecology and Evolution*, **3**, 1951–1966.

Gay L, Crochet P-A, Bell DA, Lenormand T (2008) Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution*, **62**, 2789–2806.

Gompert A, Buerkle CA (2010) introgress: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, **10**, 378–384.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.

Harrison RG (1990) Hybrid zones: windows on the evolutionary process. In: *Oxford Surveys in Evolutionary Biology* (eds Futuyma D, Antonovics J), pp. 69–128. Oxford University Press, Oxford, UK.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Macholan M, Baird SJE, Dufkova P *et al.* (2011) Assessing multilocus introgression patterns: a case study on the mouse X-chromosome in Central Europe. *Evolution*, **65**, 1428–1446.

Macholán M, Munclinger P, Sugerková M *et al.* (2007) Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. *Evolution*, **61**, 746–771.

Macholán M, Baird SJE, Munclinger P *et al.* (2008) Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? *BMC Evolutionary Biology*, **8**, 271.

Martin AD, Quinn KM, Park JH (2011) MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, **42**, 1–21.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Paradis E (2006) *Analysis of Phylogenetics and Evolution with R*. Springer, New York, NY.

Phillips BL, Baird SJE, Moritz C (2004) When vicars meet: a narrow contact zone between morphologically cryptic phylogeographic lineages of the rainforest skink, *Carlia rubrigularis*. *Evolution*, **58**, 1536–1548.

Porter AH, Wenger R, Geiger HJ, Scholl A, Shapiro AM (1997) The *Pontia daplidice-edusa* hybrid zone in northwestern Italy. *Evolution*, **52**, 1561–1573.

Rabosky DL (2006) LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evolutionary Bioinformatics*, **2**, 247–250.

R-Development-Core-Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.

Revolution Analytics (2012) foreach: Foreach looping construct for R. R package version 1.4.1.

Rieseberg LH, Whitton J, Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.

Szymura J, Barton NH (1986) Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution*, **40**, 1141–1159.

Szymura J, Barton NH (1991) The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution*, **45**, 237–261.

R.T.B. conceived the study and provided the manakin molecular and morphological data. G.E.D. conceived, designed and implemented the HZAR package, with input from all authors. E.P.D., G.E.D. and J.M.M. conducted analyses; E.P.D. and G.E.D. produced figures and tables. E.P.D. prepared and edited the manuscript, with input from all authors.

Data Accessibility

All data are publicly available. Analyses are available in R files; raw data files are part of the HZAR package on CRAN.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1 hzarExampleScript.R and ExampleScriptOutput.pdf. Provides script and output files for fitting 3 cline models to one genetic locus.

Data S2 hzarExampleScriptQT.R and ExampleScriptOutputQT.pdf. Provides script and output files for fitting 3 quantitative cline models to one quantitative trait.

Data S3 Model equations.