MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2010) **10**, 378–384 doi: 10.1111/j.1755-0998.2009.02733.x

COMPUTER PROGRAM NOTE

INTROGRESS: a software package for mapping components of isolation in hybrids

ZACHARIAH GOMPERT and C. ALEX BUERKLE
Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY 82071, USA

Abstract

A new software package (INTROGRESS) provides functions for analysing introgression of genotypes between divergent, hybridizing lineages, including estimating genomic clines from multi-locus genotype data and testing for deviations from neutral expectations. The software works with co-dominant, dominant and haploid marker data, and does not require fixed allelic differences between parental populations for the sampled genetic markers. Permutation and parametric procedures generate neutral expectations for introgression and provide a basis for significance tests of observed genomic clines. The software also implements maximum likelihood estimates of hybrid index from genotypic data and a number of graphical analyses. The package is an extension of the R statistical software, is written in the R language and is freely available through the Comprehensive R Archive Network (CRAN; http://cran.r-project.org/). In this study, we describe INTROGRESS and demonstrate its use with a sample data set.

Keywords: admixture, genetic mapping, genomic clines, hybrid zones, reproductive isolation

Received 6 March 2009; revision received 17 April 2009; accepted 5 May 2009

Considerable interest exists in identifying the genetic architecture of reproductive isolation and adaptive evolution (Howard & Berlocher 1998; Coyne & Orr 2004; Butlin & Ritchie 2009). Knowledge regarding the number and location of genes that contribute to reproductive isolation between divergent populations or species may allow for discrimination among competing models of speciation and identification of the types of genetic changes that facilitate speciation (Rieseberg *et al.* 1995; Wu 2001; Coyne & Orr 2004; Qvarnstrom & Bailey 2009). Natural hybrid zones generate recombinant individuals that provide a unique opportunity for mapping the genetic architecture of reproductive isolation (Rieseberg *et al.* 1999; Buerkle & Lexer 2008; Gompert & Buerkle 2009). In recombinant, admixed individuals, the fitness effects of genotype combinations determine the extent of introgression at individual loci. Contrasts among loci allow for the identification of loci that lower the fitness of hybrids and thus contribute to reproductive isolation, or that increase fitness and facilitate adaptive introgression.

Gompert & Buerkle (2009) proposed the genomic clines method for mapping components of reproductive isolation in admixed populations (i.e. hybrid zones). This method estimates the change in frequency of marker genotypes along a genome-wide admixture gradient. These estimates are referred to as genomic clines and differ from geographical clines, which estimate changes in the population frequency of characters (alleles or phenotypes) along geographical gradients. Specifically, the genomic clines method uses multinomial regression to predict the probability of a given genotype for a marker as a function of genome-wide admixture (e.g. hybrid index between a pair of species or divergent populations; Buerkle 2005). Expected genomic clines can then be generated based on a null model of neutral introgression and compared with genomic clines from observed marker data to identify molecular markers with patterns of introgression inconsistent with neutral expectations (see details below and Gompert & Buerkle 2009). These non-neutral markers are probably linked to genes that contribute to reproductive isolation (markers with reduced introgression or a deficit of heterozygotes) or hybrid vigour (markers with increased introgression or an excess of heterozygotes). For detailed information

Correspondence: Alex Buerkle, Fax: 307 766 2851;
E-mail: buerkle@uwyo.edu

on the genomic clines method, see Gompert & Buerkle (2009).

We have implemented the analytical approach described in Gompert & Buerkle (2009) in a new software package (INTROGRESS, version 1.1) that is written in the statistical language R. INTROGRESS is freely available from the Comprehensive R Archive Network (CRAN; http://cran.r-project.org/), as is R (R Development Core Team 2008). R is a powerful software environment for statistical computing and is available for all common computer operating systems. The INTROGRESS software is an add-on package for R and contains functions for processing genotype data for analysis, estimating genomic clines, testing genomic clines for deviations from neutral expectations and graphical analysis. In addition, INTROGRESS provides maximum likelihood estimates of hybrid index (equivalent to those in HINDEX; Buerkle 2005). The INTROGRESS package also includes documentation of all functions, sample data sets and examples of analyses. The software is written and is available as R code, and consequently, the internals of the functions can be examined and utilized by other R programmers under the GNU General Public License. Additional information and files that are not part of the R package distributed through CRAN are maintained at http://www.uwyo.edu/buerkle/software.

The functions in INTROGRESS work with co-dominant (e.g. SNP, SSR), dominant (e.g. AFLP, RAPD) and haploid (e.g. organellar DNA) marker data and do not require markers to exhibit fixed differences between the parental populations. Ideally, studies that utilize the software will be based on a sufficient number of informative markers to cover the genetic map at a reasonable density and will utilize potentially admixed individuals that cover the full range of variation between parental species. In this study, we describe the functionality of INTROGRESS and illustrate its use with a sample data set.

Software usage and description

The starting point for analysis with INTROGRESS is a set of text files containing genotype data for one or more admixed populations (i.e. hybrid zones) and for the parental populations. Researchers typically manage genotype data for individuals in spreadsheet software and then save the data as a text file. Text files are read into R using standard functions read.table or read.csv; R Development Core Team 2008). Genotype data should be arranged in a matrix with data for individuals in columns and genetic loci in rows. The first two rows of the file with genotype data for the admixed individuals can be used to provide the sampling locality and individual identification (this is optional, but can be desirable). Alleles for co-dominant data should be separated by a forward-slash (e.g. A/A or 152/152 for a homozygote, and A/C or 146/152 for a heterozygote) and dominant data should be coded as 0 or 1 (where 0 denotes the recessive allele). Missing data for a locus are coded as NA. Alternatively, the alleles for a locus can be recorded in two adjacent rows or columns (similar to input files for *structure*; Pritchard *et al.* 2000). In addition to genotype data, a matrix with information on marker loci is required. At a minimum, the locus information should not only include the locus name and type (co-dominant, *C*; dominant, *D*; or haploid, *H*), but it may also provide map information (linkage group and location) that can be used for plots (see Table 1 for an example).

Once the input data have been read into R, the resulting data objects with genotypes for potentially admixed individuals, for individuals from each parental population, and the locus information are provided to prepare.data (a function in the INTROGRESS package). This first function calculates parental allele frequencies and counts alleles derived from each of the parental populations for each individual from the admixed populations (i.e. 0, 1 or 2 for co-dominant data; 0 or 1 for dominant or haploid data). For each locus, allele frequency differentials between the parental populations are calculated (δ) and if there are more than two alleles, alleles are combined into two allelic classes with frequency differentials between populations equal to the observed differential based on all alleles (Gregorius & Roberds 1986; Zhu *et al.* 2005). This procedure creates a biallelic classification, without loss of information or distortion of the relationship between the parental populations. If the parental populations are fixed for different alleles for all loci, the data for the admixed populations can be coded simply with characters for alleles inherited from each of the parental populations (e.g. *m* vs. *d*, or *P1* vs. *P2*) and the genotype data for parental populations are not needed. The prepare.data function returns a data object (a list) that includes individual identifications (if these data were provided), the matrix of allele counts, allele frequencies for the allelic classes in parental populations, the original genotype data matrix and other data generated associated with allele frequency

Table 1 Example format for loci data object. The first row is a header that provides the column names

Locus	Type	lg	Marker pos.
msat1	C	1	1.13
msat2	C	1	1.35
aflp1	D	1	1.78
msat3	C	2	2.03
aflp2	D	2	2.51
msat4	C	2	2.88
msat5	C	3	3.22
msat6	C	3	3.77

differentials and allelic classes. The resulting data object can be used as the input for genomic cline analysis. Alternatively, the genotype data for the admixed population can be pre-coded in a 0, 1, 2 format and the prepare.data function is not necessary.

The genomic clines method requires an estimate of hybrid index (i.e. genome-wide admixture) for each of the admixed individuals. These estimates can be obtained within R with the est.h function from INTROGRESS or with separate software: HINDEX (Buerkle 2005) or Bayesian admixture proportions from *structure* (estimates of Q with $k = 2$; Pritchard *et al.* 2000). The est.h function uses as its input the list output from prepare.data and the data object with loci information (as described earlier). The est.h function returns a data object (a data frame) containing maximum likelihood hybrid index estimates with 95% confidence intervals.

The genomic.clines function from INTROGRESS fits genomic clines for the admixed population using logistic regression. The regressions use the observed data to estimate probabilities of observing homozygous and heterozygous genotypes as a function of hybrid index (Gompert & Buerkle 2009). The genomic.clines function can not only simply estimate the genomic clines for observed data, but can also perform significance testing for departures from neutral expectations for genomic clines. Two alternative methods exist for generating neutral expectations: a permutation procedure and parametric simulations. The permutation procedure, which permutes an individual's allele counts among loci, should generally be preferred as it is less likely to lead to false positives. The parametric procedure simulates individual genotypes on the basis of parameters of the parental and admixed populations (allele frequency differentials between parental populations, hybrid index estimates and deviations from expected heterozygosity). The parametric procedure for dominant markers has been altered slightly from that described by Gompert & Buerkle (2009); parental allele frequencies are now estimated from the frequency of the null homozygote assuming Hardy–Weinberg equilibrium and then used for generating neutral expectations. The parametric approach should be used when parental populations do not exhibit fixed (or nearly fixed and largely equal) differences for alleles at all of the assayed molecular markers. Also, the permutation procedure cannot be used if genomic clines are being fit simultaneously for multiple types of marker data (i.e. some combination of co-dominant, dominant and haploid markers). See Gompert & Buerkle (2009) for more details on these procedures.

Both approaches to significance testing utilize stochastic samples (permutations or parametric simulations) and the genomic.clines function has an argument (a variable) with which the number of samples can be modified (the default is 1000 permutations or simulations, which should generally be sufficient). Estimating hybrid indexes and generating the neutral expectations for genomic clines are the most computationally intensive steps of the analysis. For a data set consisting of 100 admixed individuals scored for 110 markers, a complete genomic clines analysis including constructing a matrix of allele counts, estimating hybrid indexes, fitting genomic clines and performing significance tests based on 1000 permutations took approximately 25 min on a Macintosh computer with two 2.8 GHz Quad-Core Intel Xeon processors (OS X, ver. 10.5.6). Significance testing is a memory intensive process, with the above example requiring 730 MB of

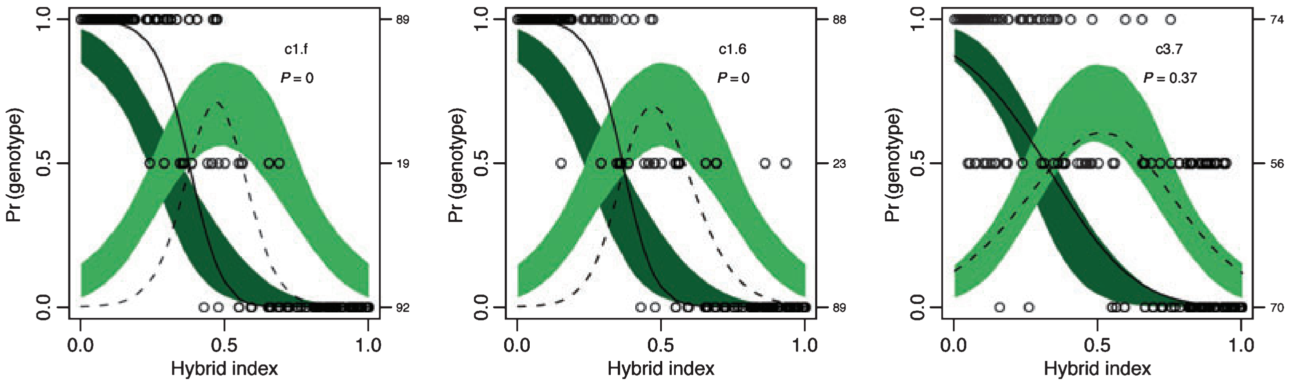

Fig. 1 Genomic cline plots for three markers from a simulated data set. The name of each locus and *P*-value from the analysis are given in each plot. Locus c1.f was under selection, c1.6 was 10 cM from this marker, and c3.7 was on a different linkage group. Solid coloured regions represent the 95% confidence intervals for the P1/P1 (dark green) and P1/P2 (light green) genotypes. The solid line and dashed lines denote the genomic clines for the P1/P1 and P1/P2 genotypes respectively. Circles indicate the raw genotype data (P1/P1 on the top line, P1/P2 in the middle and P2/P2 on the bottom line), with counts of individuals on the right vertical axis. The hybrid index quantifies the fraction of alleles derived from population 2.

Table 2 A complete genomic clines analysis using functions in INTROGRESS and R. Lines beginning with ## are comments. The code in this table can be copied from the electronic version of this manuscript and pasted directly into R. Text files with sample data sets should be downloaded from http://www.uwyo.edu/buerkle/software and placed in the directory that is being used for as the 'working directory' for R

```
## We assume that introgress and additional packages have been installed from CRAN

## load the introgress library
library(introgress)

## set the working directory in R to an existing directory where you will work
## and where the downloaded data files have been saved
## On Mac OS and other UNIX systems:
##     setwd("~/Documents/introgress/")
## On Windows ... replace your_username with appropriate name:
##     setwd("C:/Documents and Settings/your_username/Documents/introgress")

## read in data for individuals from admixed population
AdmixDataSim1 <- read.csv(file="AdmixDataSim1.txt", header=FALSE)
## alternatively use the built-in data set included with introgress
## data(AdmixDataSim1)

## read in marker information
LociDataSim1 <- read.csv(file="LociDataSim1.txt", header=TRUE)
## alternatively use the built-in data set included with introgress
## data(LociDataSim1)

## look at help page that describes the data set
help(AdmixDataSim1)

## code to convert genotype data into a matrix of allele counts,
## the results are saved to the list data object count.matrix
count.matrix <- prepare.data(admix.gen=AdmixDataSim1, loci.data=LociDataSim1,
                             parental1="P1", parental2="P2", pop.id=FALSE,
                             ind.id=FALSE, fixed=TRUE)

## estimate hybrid index values and save the results to the data frame hi.index.sim
hi.index.sim <- est.h(introgress.data=count.matrix, loci.data=LociDataSim1,
                      fixed=TRUE, p1.allele="P1", p2.allele="P2")
## write the hi.index.sim data frame to a text file
write.table(hi.index.sim, file="hindex.txt", quote=FALSE, sep=",")

## make plot to visualize patterns of introgression
## this saves the plot to a pdf in the current directory for R
mk.image(introgress.data=count.matrix, loci.data=LociDataSim1,
         marker.order=NULL, hi.index=hi.index.sim, ylab.image="Individuals",
         xlab.h="population 2 ancestry", pdf=TRUE, out.file="image.pdf")

## conduct the genomic clines analysis and save the results to a
## data object (list) called gen.out
## this uses the permutation procedure with 1000 permutations
gen.out <- genomic.clines(introgress.data=count.matrix,
                          hi.index=hi.index.sim, loci.data=LociDataSim1,
                          sig.test=TRUE, method="permutation")
## view summary of analysis in gen.out; could use write.table() to write results to file
gen.out$Summary.data

## make plots to visualize the genomic clines
## this saves the plots to a pdf in the current directory for R
clines.plot(cline.data=gen.out, rplots=3, cplots=3, pdf=TRUE, out.file="clines.pdf")
```

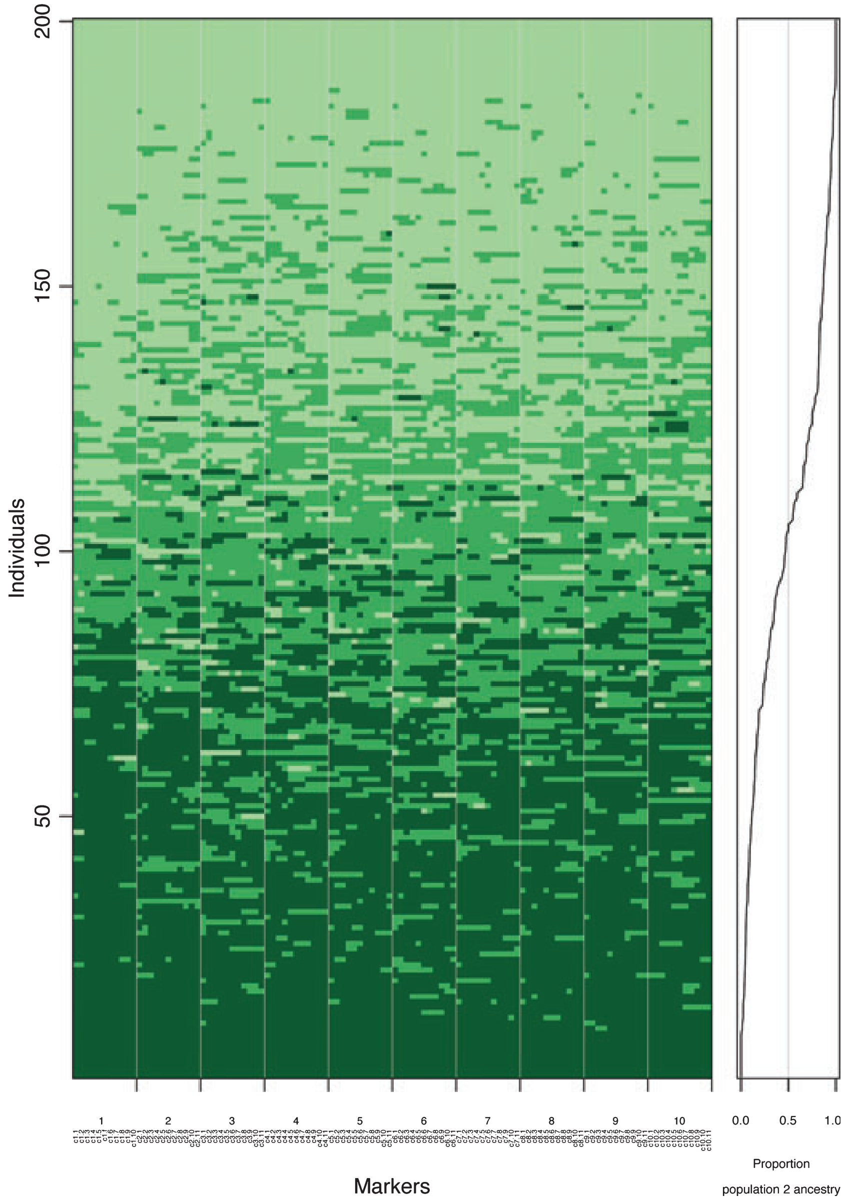

Fig. 2 Overview plot of patterns of introgression for all markers and individuals in an admixed population. Markers are ordered based on map locations. Each rectangle denotes an individual's genotype at a given locus: dark green (*P1*/*P1*), green (*P1*/*P2*) and light green (*P2*/*P2*). On the right is a plot of the fraction the genome inherited from population 2 for each individual (hybrid index or admixture coefficient). In this case, individuals are sorted by increasing hybrid index. The plot was produced with the `mk.image` function in INTROGRESS.

memory. Windows users may need to increase the allocation of memory to R using the memory.limit utilities function.

The genomic.clines function produces a list data object as output. The first element in this list gives the log likelihood ratio test statistic for each marker and the *P*-value for the significance test. This is followed by additional elements that give fitted values for the genomic clines estimated from the observed data (this includes one element for each genotype). In addition this list object contains elements specifying the upper and lower limits of the 95% confidence interval for genomic clines under neutral expectations (from permutations or simulations). The list object contains input data for the clines.plot function, which produces graphical output for the genomic clines analysis (Fig. 1).

Most of the functions included in INTROGRESS have arguments that allow the user to conduct analyses using a subset of the individuals or markers and to alter the order of markers for plotting purposes. Arguments also exist for specifying various graphical parameters for plots. Several accessory functions are included in the

INTROGRESS package, such as calc.intersp.het and compare.clines. The former estimates admixed individuals' mean observed inter-class heterozygosities, while the latter contrasts marker specific patterns of introgression between two admixed populations derived from the same parental populations (see the *Helianthus* example from Gompert & Buerkle 2009). The package also contains additional functions that produce graphics for visualizing patterns of admixture (e.g. mk.image and triangle.plot) or are internal functions that are called by the high-level functions described above. The INTROGRESS R documentation provides complete information on each function, including the arguments, return values and examples for each function.

Analysis of a sample data set

As an illustration of genomic clines analysis using INTROGRESS, we simulated admixture in a population to generate a sample data set and used the functions in the R package for analysis. We generated an admixed population with 500 diploid hermaphroditic individuals with 10 pairs of chromosomes, each one Morgan in length. The admixed population resulted from an initial hybridization event between two parental populations (combined in equal proportions) and mating continued for five additional generations. All matings occurred within the admixed population (i.e. there was no additional migration from the parental populations). We applied viability selection against heterozygotes at a single locus in the centre of linkage group 1. The fitness of heterozygotes at this locus was 0.1, whereas homozygotes had full fitness. At the completion of the simulation, we sampled 200 individuals from the admixed population and scored each for 110 co-dominant markers spaced evenly across all 10 linkage groups (10 cM intervals). One of these markers (c1.f) was at the locus under selection. The parental populations were assumed to be fixed for different alleles at all sampled markers (for more information on the simulation, see Gompert & Buerkle 2009). This and two other simulated data sets are distributed with the R package and are also available as text files from the project website.

The specific commands for the complete genomic clines analysis are given in Table 2. For this data set, genotypes for the simulated admixed individuals were saved in a text file (AdmixDataSim1.txt) and coded as *P1*/*P1*, *P1*/*P2* and *P2*/*P2* for alleles from parental populations 1 and 2. We also created a text file with information on marker loci (similar to Table 1), including the genetic map position of the markers. These data objects served as input for the prepare.data function, but we did not provide parental genotype data objects to the function, as parental populations exhibited fixed differences for all markers. Instead, in the arguments for prepare.data, we only specified the characters that designated alleles from each species (Table 2).

The next step of the analysis was to estimate hybrid indexes for the admixed individuals using est.h (Table 2). To provide a clear visualization of variation in patterns of introgression and ancestry across markers and for each of the admixed individuals, we used mk.image (Fig. 2). We then carried out the genomic clines analysis for the 110 markers using the genomic.clines function. The results of the genomic clines analysis were saved in a data object (gen.out), which was used as input for the clines.plot function to generate plots of the clines for each locus. Sample plots for the marker under selection, a neighbouring marker on the same linkage group and a marker on a different linkage group are in Fig. 1. Of the 11 markers on the linkage group that had the locus under selection at its centre, eight markers showed patterns of introgression inconsistent with neutral expectations ($P \leq 0.01$). Conversely, only 5 of the 99 markers on other linkage groups showed patterns of introgression inconsistent with neutral expectations. For a detailed analysis of power and false-positives for the genomic clines method, see Gompert & Buerkle (2009).

Analysis with INTROGRESS is achieved with a few high-level functions in R and should be usable by those without prior experience with R. The interaction with the software involves textual commands that can be copied and pasted into R, rather than the use of a graphical interface. One of the advantages of this approach is that the commands used for analysis can be saved in a text file (such as Table 2), so that analyses can easily be repeated and modified. The INTROGRESS package builds on well-tested and powerful functions in R and places the results in an environment that is suitable for additional analysis.

Acknowledgements

We thank the following for comments on an earlier version of this manuscript and testing a trial version of INTROGRESS: Monia Haselhorst, Jarrett Johnson, Tom Parchman, Christian Lexer, Arne Nolte and Subject Editor Vincent Castric. This work was funded by NSF DBI award 0701757 to CAB and a NSF Pre-doctoral Graduate Research Fellowship to ZG.

References

Buerkle CA (2005) Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes*, **5**, 684–687.

Buerkle CA, Lexer C (2008) Admixture as the basis for genetic mapping. *Trends in Ecology & Evolution*, **23**, 686–694.

Butlin RK, Ritchie MG (2009) Genetics of speciation. *Heredity*, **102**, 1–3.

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

Gompert Z, Buerkle CA (2009) A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, **18**, 1207–1224.

Gregorius HR, Roberds JH (1986) Measurement of genetical differentiation among subpopulations. *Theoretical and Applied Genetics*, **71**, 826–834.

Howard DA, Berlocher SH (1998) *Endless Forms: Species and Speciation*. Oxford University Press, New York.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Qvarnstrom A, Bailey RI (2009) Speciation through evolution of sex-linked genes. *Heredity*, **102**, 4–15.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rieseberg LH, Linder CR, Seiler GJ (1995) Chromosomal and genic barriers to introgression in Helianthus. *Genetics*, **141**, 1163–1171.

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.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Zhu X, Luke A, Cooper RS *et al.* (2005) Admixture mapping for hypertension loci with genome-scan markers. *Nature Genetics*, **37**, 177–181.