

## COMPUTER PROGRAM NOTE

**CDPOP: A spatially explicit cost distance population genetics program**

ERIN L. LANDGUTH\*† and S. A. CUSHMAN†

*\*University of Montana, Mathematics Building, Missoula, MT 59812, USA, †USDA Forest Service, Rocky Mountain Research Station, 800 E Beckwith, Missoula, MT 59801, USA***Abstract**

**Spatially explicit simulation of gene flow in complex landscapes is essential to explain observed population responses and provide a foundation for landscape genetics. To address this need, we wrote a spatially explicit, individual-based population genetics model (CDPOP). The model implements individual-based population modelling with Mendelian inheritance and  $k$ -allele mutation on a resistant landscape. The model simulates changes in population and genotypes through time as functions of individual based movement, reproduction, mortality and dispersal on a continuous cost surface. This model will be a valuable tool for the study of landscape genetics by increasing our understanding about the effects of life history, vagility and differential models of landscape resistance on the genetic structure of populations in complex landscapes.**

*Keywords:* distance matrices, individual-based model, landscape resistance, population genetics

*Received 12 November 2008; revision accepted 24 March 2009*

A primary goal for landscape genetics is predicting how landscape features influence genetic structure and gene flow patterns in natural populations (Manel *et al.* 2003; Holderegger & Wagner 2006). Classic population genetic theory is founded on several critical simplifying assumptions, including discretely bounded and panmictic populations, or simple isolation by distance (Wright 1943). However, real populations occupy spatially complex landscapes in which populations are rarely discretely bounded or internally panmictic, and in which spatial heterogeneity fundamentally alters simple predictions of isolation by Euclidean distance or geographical barriers. The mathematics of population genetics in spatially complex environments is difficult and closed form solutions akin to those of classical population genetics may be intractable. Recent studies have investigated the effects of landscape structure and seascape structure on gene flow patterns (Antolin *et al.* 2006; Cushman *et al.* 2006; Neville *et al.* 2006; Hansen & Hemmer-Hansen 2007), but simulation modelling might be the best means to help establish mechanistic explanations and provide a foundation for spatial population genetic theory.

We have developed a simulation program (CDPOP) to predict the influences of landscape structure on emergence of spatial patterns in population genetic data as functions of individual-based movement, breeding and dispersal. The model represents landscape structure flexibly as resistance surfaces whose value represents the stepwise cost of crossing each location. The model simulates mate selection and dispersal as probabilistic functions of cumulative cost across these resistance surfaces. Breeding is simulated with Mendelian inheritance and  $k$ -allele mutation, a commonly used mutation model (Balloux 2001; Manel *et al.* 2007). The user specifies the locations and genotypes of the initial population and the model simulates spatially explicit population genetic change through time as functions of individual based movement (mate choice and dispersal), mating, mutation and mortality on a continuous cost surface. The model is specifically designed to enable explicit quantification of how landscape resistance affects gene flow patterns. Simulations with different resistance grids allow quantification of the effects of different landscape conditions on genetic connectivity and the time required for the spatial patterns of genetic relatedness to equilibrate. CDPOP provides functionality to simulate different movement and mating functions, direct calculation of  $H_E$ ,  $H_O$ ,  $F$ ,  $F_{IS}$ ,  $F_{ST}$ ,

Correspondence: Erin L. Landguth, Fax: 406 210 9332; E-mail: erin.landguth@umontana.edu

and  $F_{IT}$  for each simulated time step, and Monte Carlo and batch capability to enable simulation of many runs of many different parameter sets from a single input file. CDPOP provides a number of functions not available in other published spatial genetic simulation models. It is the only individual-based model published to date that simulates genotypes of each individual in a spatially structured population as functions of mating, dispersal and mutation on flexible resistance landscapes. There are very basic differences between CDPOP and SPLATCHE (Currat *et al.* 2004) to the extent that we feel there is no real redundancy among them in terms of functionality or questions to which they are most suited. SPLATCHE combines SIMCOAL (Excoffier *et al.* 2000) simulation of molecular genetic diversity in a set of discrete demes using backward coalescent approaches with spatial migration. SPLATCHE simulations first simulate migration as a function of environmental friction and local carrying capacity, and then simulate genetic diversity under a coalescent framework. This differs fundamentally in several ways from CDPOP: (i) individual based rather than deme based; (ii) a forward simulation of individual genotypes of all individuals in spatially structured population; and (iii) dispersal and mating in CDPOP are explicitly cost functions across heterogeneous landscapes. CDPOP is partly redundant with EASYPop (Balloux 2001) when used to simulate panmictic demes separated by full or partial barriers (e.g. demes with zero or nonzero migration rates among them). However, CDPOP greatly extends functionality in simulating spatial population genetics by being able to incorporate virtually unlimited models of landscape effect on gene flow. The flexibility of CDPOP to represent gene flow as any resistance surface allows comparison of isolation by barriers, Euclidean distance and multiple landscape resistance hypotheses within a single simulation framework. This is unique among the tools available at present.

### Program description

CDPOP simulates population dynamics and genetic exchange for a given cost distance matrix and  $n - (x,y)$  located individuals. Examples of types of cost distances from each individual to every other individual include least accumulative cost distance (i.e. Dijkstra's algorithm) or Euclidean distance. The model can flexibly incorporate absolute or partial barriers, and panmixia within this cost distance modelling framework. This is critical to provide a consistent and comparable framework to evaluate the pattern-process relationships resulting from isolation by landscape resistance, isolation by Euclidean distance and isolation by barriers (Cushman *et al.* 2006). The user must provide an existing cost distance matrix recording the cumulative movement cost among all pairs of genotyped

individuals as input to the model. Common approaches to calculating cost distance matrices in landscape genetics include the `COSTDISTANCE` function in ArcGIS (ESRI Corp.), the `PATHMATRIX` program (Ray 2005) and the `CIRCUITSCAPE` program (McRae & Beier 2007).

The simulation assumes constant population size and density over time. Individuals are assumed to occupy a fixed grid on the landscape that is user defined by the  $n - (x,y)$  located individuals. The density of the grid may be uniform, or a function of resistance level or habitat suitability. Fixing the population on a constant grid serves several purposes, with associated tradeoffs. The major advantage of simulating constant population is that it enables explicit separation of the effects of landscape resistance on gene flow from those caused by spatially dynamic population size. This separation is important for those seeking to quantify the independent effects of landscape patterns on population structure, differentiation and gene flow. The second advantage is computational efficiency. Initializing the model to a fixed grid of occupied locations greatly increases efficiency as recalculation of cost distances following movement or dispersal is obviated. The major drawback is that by fixing population size the model cannot explore interactions between population dynamics, gene flow and landscape structure, which is a potentially illuminating area. Future versions of CDPOP will allow for dynamic fluctuation of population size and density.

Figure S1 shows the model logic and sequence of tasks implemented in the simulation. The first step in implementing the model is reading in a table of input parameters specifying movement, mating, birth, death and genetic processes and other run parameters (Table 1). Next, the model reads each  $n - (x,y)$  individual location, initial genotype, age and sex. The genotype of each locus for each individual can be initialized by randomly choosing from a file containing allele frequencies for each locus, or by reading in a file containing the initial multilocus genotypes of all the individuals. The initial age structure of the population is specified by an input file specifying the initial age frequency. The sex of each individual is randomly assigned.

There are four movement functions that define how individuals choose a mate and disperse to a free grid location: linear, inverse square, nearest neighbour and random mixing. The probability of movement for each individual is a function of the cost distance from that individual to every other individual, based on one of these four movement functions. With the nearest neighbour movement function, an individual moves to the available grid location nearest to its initial location. Random mixing moves an individual to a grid location that is randomly chosen from the  $n$  grids in the population. In linear and inverse square movement functions,

**Table 1** Program input and output

	Input	Description
Program files	cdmatrix.csv	A $[N \times N]$ cost distance matrix, where $N \geq n$ and $n$ is the number of grid values (or individuals) on the landscape
	xycdmatrix.csv	The $N - (x,y)$ coordinate values
	agedistribution.csv	The distribution that is used to initialize the model's $n$ individuals' age structure
	allelefrequency.csv	The allele frequency distribution for each locus, used to initialize the model's $n$ individual's genotype
Model initialize	xygrid.csv	The $n - (x,y)$ grid location values
	Batch lines	The repeated number of input lines to implement multiple runs with differing parameter inputs
	Monte Carlo Loop time	The repeated number of simulations to be conducted for the Monte Carlo method Simulation run time [year or generation]
Mating process	Reproduction answer	Y for sexual reproduction and N for asexual reproduction
Movement process	Function answer	Movement function answer [1 = Linear, 2 = Inverse Square, 3 = Nearest Neighbour, 4 = Random Mixing]
	Movement threshold	A threshold option (in cost distance units) for how far an individual can search for a mate and an offspring can disperse
	Birth process	Reproduction age Litter size
Death process	Female Born	Per cent number of female born in each litter
	Adult mortality	Per cent mortality in the adult population
	Offspring mortality	Per cent mortality in the offspring population
Genetic process	Initialization	Initializing the genotype for each $n - (x,y)$ [Random = maximized the genetic diversity, File = draws from allele frequency distribution file, Known = reads genotype information from a given known file]
	Gene swap year	The time that genetic information is exchanged
	Loci	The number of loci
	Alleles	The number of alleles per locus
	Mutation rate	The $k$ -allele model rate for mutation
	Output	Description
Output files	Grid(time).csv	The genotype for each $n - (x,y)$ across specified time units
	Cdmatrix(n).csv	The cost distance matrix used for the $n$ grid locations
	Edmatrix(n).csv	The Euclidean distance matrix used for the $n$ grid locations
	Gdmatrix(n_time).csv	The genetic distance matrix used for the $n$ grid locations for specified year of the simulation run [braycurtis = the Bray-Curtis distance measure, Dps = proportion of shared alleles, Da = Nei's] following
	Output.csv	A file of the various data measures at each year: (Population, Emigrants, Deaths, Births, Immigrants, Total Females, Total Males, Total Breeding Age Females, Total Breeding Age Males, Alleles, $H_E$ , $H_O$ , $F$ , $F_{ST}$ , $F_{IT}$ , $F_{IS}$ , Alleles Mutated, Average Mate Distance, Average Offspring Dispersal Distance)

individuals move a distance from their initial location based on a draw from a probability distribution inversely proportional to a linear or inverse square distance function. The user specifies the maximum dispersal distance. In the linear distance function, this maximum dispersal distance is where the probability function goes to zero. The probability is maximum at no distance from the original location and goes to zero at the maximum dispersal distance, with a constantly decreasing probability at intermediate distances, such that the area under the line from no distance to the maximum dispersal distances integrates to 1, corresponding to 100% probability that

individuals will move a distance between zero and the maximum dispersal distance.

Reproduction is defined by the user as either asexual or sexual. With asexual reproduction, all  $n$  individuals mate and bear offspring, with mates selected according to the movement function choice and without regard to any gender or mating type. It is important to note that this 'asexual' reproduction is functionally the sexual pairing of hermaphroditic individuals; thus it is asexual in the sense that there are no distinct sexes, but is sexual in the sense that individuals mate with other individuals, exchanging genes in Mendelian reproduction. In sexual

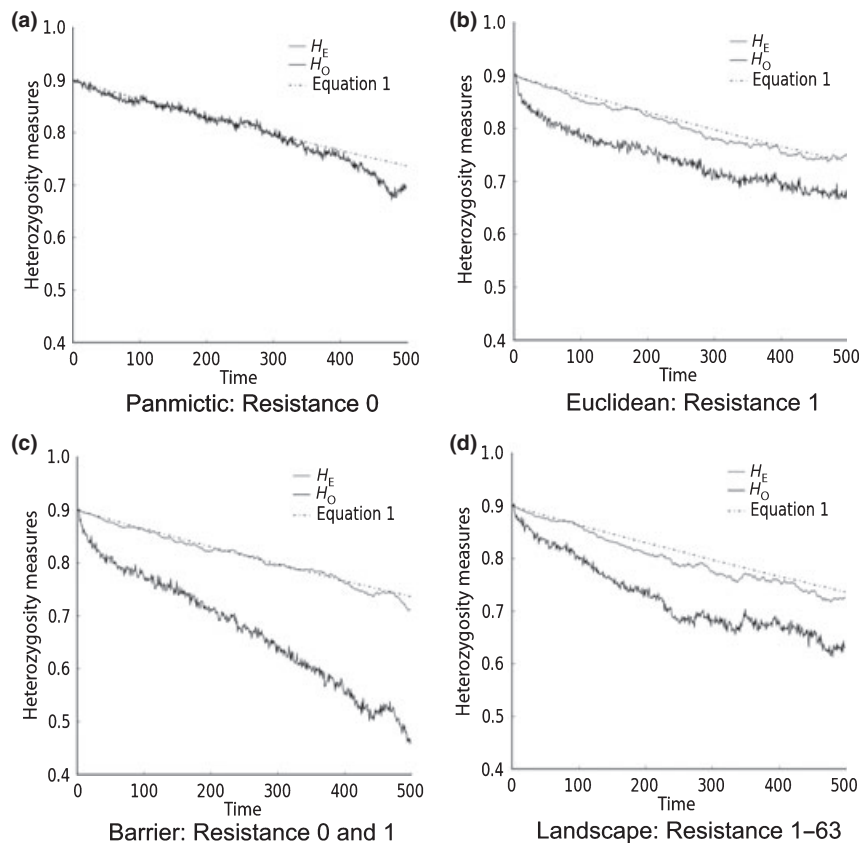
reproduction, mated pairs are one male to possibly many or no females, and the end of the mating process occurs when all females have mated. Each mated pair can have a number of offspring that is a bounded random draw, a Poisson draw with specified mean, or a constant number. Mendelian inheritance with  $k$ -allele mutation (rate chosen by the user) is used to generate the offspring's genotype and the sex assignment is random.

Dispersal of offspring occurs from the mother's  $(x,y)$  location according to the selected movement function. The vital rates (birth and death) define whether or not the population will have emigrants or immigrants. If mortality of the adults creates a situation in which the number of free grid cells is less than the number of offspring, then the remaining offspring that did not disperse to the free cells emigrate out of the population (i.e. are removed). If mortality of the adults creates a situation in which the number of free cells is greater than the number of offspring, then the offspring will disperse to the free cells and the remaining free cells are filled with immigrants. Immigration can also occur if there are free grid

spots that are beyond the movement threshold of all offspring, specified in units of the cost distance. The immigrants have age 0, random assigned sex and genotype drawn randomly from the current populations' allele frequency distribution.

There are five model outputs (Table 1). The first is a series of comma delimited files specifying the locations, genotypes, age and sex of each individual at specified simulation years. The second and third outputs are the cost distance matrix and Euclidean distance matrix for the  $n - (x,y)$  grid locations. The fourth is the genetic distance matrix at specified years among individuals calculated using either the Bray-Curtis percentage dissimilarity measure (Legendre & Legendre 1998), proportion of shared alleles (Bowcock *et al.* 1994), or Nei's genetic distance (Nei *et al.* 1983). The final output is a file containing yearly values for various population parameters, formatted to facilitate graphical display (Table 1).

As the model simulates stochastic processes, most applications will necessarily involve simulating multiple runs and quantifying mean and variability of genetic



**Fig. 1** Expected heterozygosity (grey line) and observed heterozygosity (black line), compared with theoretical expectation (black-dashed line, see eqns 1–2) for the example of the four landscape resistance scenarios: (a) panmictic, (b) isolation by Euclidean distance, (c) isolation by barrier and (d) isolation by a simple model of landscape resistance (Note that this is only an example of one simulation run for each of the scenarios.).

structure across many runs. Thus, we have added Monte Carlo option, which allows the user to choose the number of runs to simulate given a single set of input parameters. In addition, a user may also frequently wish to launch several runs with different parameter values simultaneously. To address this, we have included a batch capability, in which each line of the input model parameter file specifies a separate analysis, each of which could include any number of Monte Carlo runs. The user specifies the number of Monte Carlo runs in the input parameter file line, with one folder created for each batch line containing subfolders created for each Monte Carlo run containing all five outputs described above.

### Example simulation

An example provides a basic understanding of model functionality and performance. The example simulates genetic exchange without mutation in a population of 1248 individuals under four landscape resistance scenarios: (i) panmictic, (ii) isolation by Euclidean distance, (iii) isolation by barrier, and (iv) isolation by a model of landscape resistance taken from Cushman *et al.* (2006). We provide basic graphical output from the model, including comparison of simulation results to theoretical predic-

tions of loss of heterozygosity over time (Fig. 1) and basic analysis of the genetic distances over time, plots of Mantel's  $r$  correlating genetic distance and cost distance over time (Fig. 2). Mantel's  $r$  was produced using the Ecodist library in R (rproject.org).

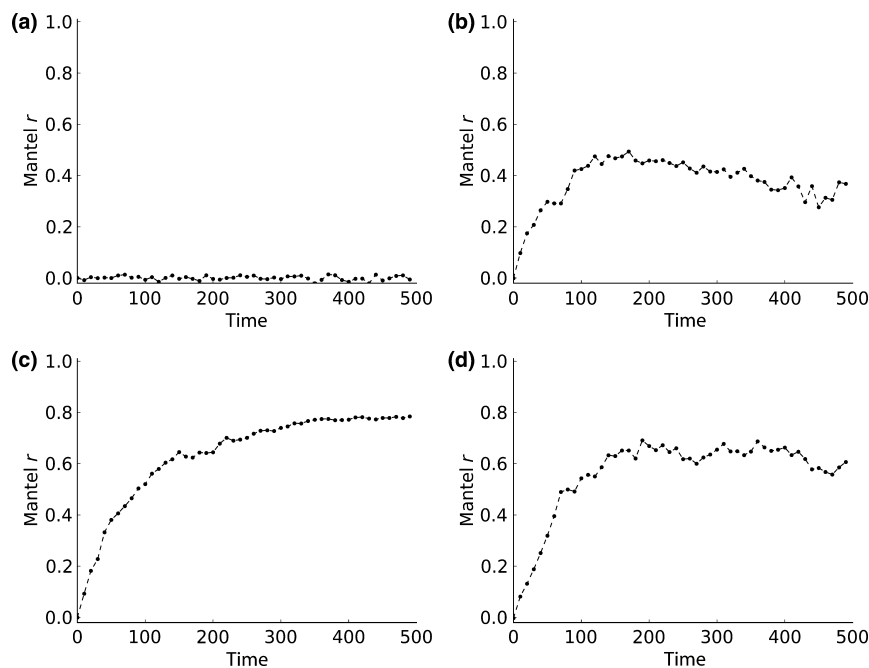
$H_E$  and  $H_O$  (the expected and observed heterozygosity respectively for the  $n$  individuals) are shown in Fig. 1, compared with curves for decay of heterozygosity produced according to the equation:

$$H_t = \left(1 - \frac{1}{2N_e + 1}\right)^t H_O(0), \quad (\text{eqn1})$$

where

$$N_e = \frac{4N_M N_F}{N_M + N_F}, \quad (\text{eqn2})$$

and  $H_t$  is the theoretical rate of decay after  $t$  generations,  $H_O(0)$  is initial heterozygosity,  $N_M$  is the total number of males and  $N_F$  is the total number females.  $H_E$  is calculated at each simulation time as the estimated fraction of all individuals who would be heterozygous based on allele frequency in an ideal Hardy-Weinberg population.  $H_O$  is the proportion of individuals that are heterozygous at each simulation year.



**Fig. 2** Mantel tests on a sample of the 1248 total points (875 points within the interior and away from the edge points) with 1999 permutations to test for significance of correlation for the example of the four landscape resistance scenarios: (a) panmictic (Mantel  $r$  between genetic distance and Euclidean distance; eight of 101 values were significant at  $\alpha = 0.05$ , which is approximately the expectation of experiment-wise error because of randomness), (b) isolation by Euclidean distance (Mantel  $r$  between genetic distance and Euclidean distance; all values were significant), (c) isolation by barrier (Mantel  $r$  between genetic distance and barrier model distance; all values were significant) and (d) isolation by a simple model of landscape resistance (Mantel  $r$  between genetic distance and landscape model distance; all values were significant).

The model shows Wahlund effects for the three spatial resistance scenarios such that a deficit of heterozygotes develops, as expected from population genetic theory whenever there is substructure or barriers (Fig. 1; Allendorf & Luikart 2007). The simulations show rapid creation of genetic substructure in the three spatial scenarios, with no substructure developing in the panmictic scenario, as expected (Fig. 2).

The program is written in PYTHON 2.5.2 and provided with installation instructions for most platforms, along with sample input files, which may be used as templates for constructing new input files according to the format described above. CDPOP has been debugged as carefully as possible by testing simulation results against analytical expectations in all cases where they exist. This program can be obtained free of charge <http://LandguthResearch.dbs.umt.edu/software/>, by emailing a request for electronic delivery to [erin.landguth@umontana.edu](mailto:erin.landguth@umontana.edu).

## References

- Allendorf FW, Luikart G (2007) *Conservation and the Genetics of Populations*. Blackwell, Malden, MA.
- Antolin MF, Savage LT, Eisen RJ (2006) Landscape features influence genetic structure of Black-Tailed Prairie Dogs (*Cynomys ludovicianus*). *Landscape Ecology*, **21**, 867–875.
- Balloux F (2001) EASYPOP (Version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.
- Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368**, 455–457.
- Curat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, **4**, 139–142.
- Cushman SA, McKelvey KS, Hayden J, Schwartz MK (2006) Gene flow in complex landscapes: testing multiple hypotheses with casual modeling. *The American Naturalist*, **168**, 486–499.
- Excoffier L, Novembre J, Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity*, **91**, 506–510.
- Hansen M, Hemmer-Hansen J (2007) Landscape genetics goes to sea. *Journal of Biology*, **6**.
- Holderegger R, Wagner HH (2006) A brief guide to Landscape Genetics. *Landscape Ecology*, **21**, 793–796.
- Legendre P, Legendre L (1998) *Numerical Ecology*, 2nd English edn. Elsevier, Amsterdam.
- Manel S, Schwartz M, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, **18**, 189–197.
- Manel S, Berthoud F, Bellemain E *et al.* (2007) A new individual-based spatial approach for identifying genetic discontinuities in natural populations. *Molecular Ecology Resources*, **16**, 2031–2043.
- McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences, USA*, **104**, 19885–19890.
- Nei M, Tajima F, Tatenno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, **19**, 153–170.
- Neville HM, Dunham JB, Peacock MM (2006) Landscape attributes and life history variability shape genetic structure of trout populations in a stream network. *Landscape Ecology*, **21**, 901–916.
- Ray N (2005) PATHMATRIX: a GIS tool to compute effective distances among samples. *Molecular Ecology Notes*, **5**, 177–180.
- Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

## Acknowledgements

We are grateful to Gordon Luikart for helpful discussions regarding population genetic theory, genetic distance measures and mutation models. Kevin McKelvey and Joe Glassy provided helpful discussion of simulation approaches in the initial conceptual design phase of the project. This research is supported by the National Science Foundation grant #DGE-0504628.

## Supporting Information

Additional supporting information may be found in the online version of this article.

**Fig. S1** CDPOP schematic of program steps.

**Fig. S2** An example of a population of 1248 individuals and four landscape resistance scenarios: (a) panmictic, (b) isolation by Euclidean distance, (c) isolation by barrier and (d) isolation by a simple model of landscape resistance. In the panmictic case resistance values across the full landscape are set to 0, effectively making all mating and dispersal spatially random. In the Euclidean case all resistance values are set to 1, making all mating and dispersal a direct function of Euclidean distance. In the barrier case there is resistance of 0 in all cells except for the barrier, which can have any positive non-zero value. This results in random mating and dispersal on either side of the barrier, and no breeding or dispersal across the barrier. In the landscape case resistance is defined by the user. In this example we use the resistance map identified by Cushman *et al.* (2006) as the strongest predictor of gene flow in an American black bear population.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.