ResDisMapper v1.0 User Manual

Qian Tang and Tak Fung

Oct 2018

Introduction

Generally, this *R* package is used to map environmental resistance to dispersal on the basis of individual-based genetic distance. The package uses the principle of Isolation by distance (IBD), which is widely used in population genetic studies to quantify natural dispersal. Specifically, the package models IBD as the expected relationship between genetic and geographical distances for individuals in a population or set of geographically continuous populations, and deviations from this general trend (IBD residual, Keis et al., 2013) are used to map resistance to dispersal. Positive IBD residuals accumulate in areas which have a relatively high resistance to dispersal and vice versa.

There are four functions in this *R* package:

rdm_IBD: Calculates IBD residuals for each pair of individuals in a population.

rdm_residual: Creates and plots IBD residuals for pairs of individuals in a population, represented as line segments over a defined landscape.

rdm_resistance: Produces environmental resistance to dispersal over a landscape, for individuals in a population.

rdm_mapper: Visualizes the environmental resistance to dispersal over a landscape, for individuals in a population.

Who may find it useful

This *R* package is designed to study the effects of landscape or other environmental features on the dispersal of individuals within a relatively small area. All studied individuals should come from a single population or a set of geographically continuous populations. Because the *R* package uses IBD residuals to indicate the resistance to dispersal, inputting multiple isolated (highly diverged) populations or species can confuse the IBD modelling as different populations may have different IBD trends. Therefore, please check PCA plots to make sure that there are no significantly diverged clusters before using the *R* package.

To guarantee the accuracy of the IBD model and subsequent analyses, we recommend that users have ensured that there is a continuous distribution of geographical distance among samples. Therefore, a sampling scheme of even distribution across the study area is highly recommended.

How to cite

Qian, T., Fung, T., Rheindt, F.E. (In review). *ResDisMapper*: An *R* package for fine-scale mapping of environmental resistance to dispersal.

Contributors

Tak Fung <tfung2000@gmail.com>

Tang Qian <tangqiannus@gmail.com>

Installation

The *R* package is available on GitHub (https://github.com/takfung/ResDisMapper).

ResDisMapper can be installed from GitHub by typing the following line in *R*, which uses the *R* package devtools:

```
devtools::install_github("takfung/ResDisMapper")
```

Input data

The *R* package is designed for multilocus data, such as SNPs, AFLPs and microsatellites. The *R* package intake is in the form of two files, a genetic data file in GENEPOP format and a geographical data file containing tab-delimited coordinates of sampling localities.

Description of the functions

rdm_IBD

Arguments

Gen_raw: A string specifying the path to a file of genotype data in GENEPOP format, with alleles specified using three digits. The GENEPOP file must have the extension *.gen*.

Geo_raw: A string specifying the path to a text file showing the name (1st column) and geographical x and y coordinates (2nd and 3rd columns) of each individual (each row). The first row needs to specify the column headings, and all entries need to be tab-delimited.

Dist_method: An integer from 1 to 6 specifying the method used to calculate genetic distance among individuals. Integers 1 to 6 correspond to the methods incorporated into the following six functions from the poppr *R* package, respectively: diss.dist, nei.dist, rogers.dist, reynolds.dist, edwards.dist, and provesti.dist. The default value is 1.

IBD_method: An integer specifying the method used to calculate the IBD residual for each pair of individuals. A value of 1 means that the residuals are calculated by fitting a straight line ($y = ax + b$, where y is the genetic distance; x is the geographical distance; and a and b are the fitted parameters) to all pairs of genetic and geographic distances, and measuring the distance from each pair to the fitted line. The fitted straight line represents the combinations of genetic and geographic distances that are expected from IBD. A value of 2 means that a non-linear curve of the form $y = a + b\ (1\text{-}\exp(\text{-}\exp(c)\ x))$ is fitted instead of a straight line, where y is the genetic distance; x is the geographical distance; and a, b and c are the fitted parameters. The default value is 1. The function rdm_IBD automatically plots the data points together with the fitted line or curve.

Output

After running rdm_IBD, an object of class dist, containing a matrix showing the IBD residuals for pairs of individuals, is output. Also, a scatter plot of the IBD model will be presented.

Example

```
IBD.res <- rdm_IBD(Gen_raw = 'files/genotypes.gen', Geo_raw = 'files/coordinates.txt', Dist_method = 1, IBD_method = 1)
```

rdm_residual

Arguments

IBD.res: A distance matrix of pairwise IBD residuals output from rdm_IBD.

Geo_raw: A string specifying the path to a text file showing the name (1st column) and geographical x and y coordinates (2nd and 3rd columns) of each individual (each row). The first row needs to specify the column headings, and all entries need to be tab-delimited.

min.dist: IBD residuals are only calculated for pairs of individuals that are separated by a distance greater than min.dist. The default value is 1.

max.dist: IBD residuals are only calculated for pairs of individuals that are separated by a distance less than max.dist. The default value is Inf.

n_resolution: Specifies the number of cells that the landscape is divided into, for the purposes of creating the 3-D plot. The landscape is divided into n_resolution cells along both coordinate axes. Individuals that appear in the same cell are represented by a single point on the plot. The default value is 50.

proj: The coordinate system that is used for plotting. The default is EPSG 4326.

Output

After running rdm_residual, an object of class SpatialLinesDataFrame, containing coordinates of the line segments joining each pair of individuals and the corresponding IBD residuals, is output. Also, a 3D plot of the spatial lines data frame will be presented.

Example

```
IBD.res <- rdm_IBD(Gen_raw = 'files/genotypes.gen', Geo_raw = 'files/coordinates.txt', Dist_method = 1, IBD_method = 1)

Res_SLDF <- rdm_residual(IBD.res = IBD.res, Geo_raw = 'files/coordinates.txt', min.dist = 1, max.dist = Inf, n_resolution = 50, proj = sp::CRS('+init=epsg:4326'))
```

rdm_resistance

Arguments

Res_SLDF: An object of class SpatialLinesDataFrame, containing the IBD residuals for each pair of individuals in a population and the coordinates of line segments connecting the pairs of individuals over the defined landscape. This object is produced by the function rdm_residual.

nrows: Number of grid cells in each row of the raster map of environmental resistance to dispersal. The default value is 30.

ncols: Number of grid cells in each column of the raster map of environmental resistance to dispersal. The default value is 30.

conf_intervals: The coverage of the confidence interval generated for each resistance value in a grid cell, expressed as a proportion. This interval measures the uncertainty in the (observed) resistance in each cell. If a cell has no intersecting line segments or only one, then a confidence interval cannot be calculated and the cell is no longer considered for further calculation. If the confidence interval does not overlap 0, then the resistance is statistically different from 0. The default value is 0.95 (corresponding to 95% intervals).

random_rep: Number of random resamples of the IBD residuals used to construct the null distribution of resistances in each grid cell. For a grid cell with *n* intersecting line segments and *n* corresponding IBD residuals, a random resample consists of randomly sampling *n* IBD residuals from the set of all IBD residuals over the entire landscape, without replacement. If the observed resistance is above a threshold percentile of the null distribution, then the cell is inferred to be a genetic barrier. If instead the observed resistance is below another threshold percentile of the null distribution, then the cell is inferred to be a genetic corridor. The default value is 1,000.

outputfile: Name of the .csv file that is produced, containing resistance information in the output data frame. The default is "resistance_map.csv".

Output

After running rdm_resistance, a data frame with resistance information for each grid cell in the landscape with more than one intersecting line segment is output. In the data frame, each row corresponds to one cell, and the eight different columns refer to the resistance, the number of intersecting line segments, the x and y coordinates specifying the location of the mid-point of the cell, the lower and upper limits of the confidence interval for the resistance, the sign of the product of the lower and upper limits of the confidence interval, and the percentile of the null distribution of resistances corresponding to the observed resistance. When calculating this data, the function rdm_resistance prints out messages showing the current stage of calculation. There are four stages: Calculating (1) the resistances, (2) the numbers of intersecting line segments, (3) the lower limits of the confidence intervals for the resistances, and (4) the upper limits of the confidence intervals for the resistances.

Example

```
IBD.res <- rdm_IBD(Gen_raw = 'files/genotypes.gen', Geo_raw = 'files/coordinates.txt', Dist_method = 1, IBD_method = 1)

Res_SLDF <- rdm_residual(IBD.res = IBD.res, Geo_raw = 'files/coordinates.txt', min.dist = 1, max.dist = Inf, n_resolution = 50, proj = sp::CRS('+init=epsg:4326'))

F.df <- rdm_resistance(Res_SLDF = Res_SLDF, nrows = 30, ncols =30, conf_intervals = 0.95, random_rep = 1000, outputfile = 'resistance_map.csv')
```

rdm_mapper

Arguments

F.df: A data frame containing eight columns with resistance information for each grid cell in a landscape. Each row corresponds to one cell, and the eight different columns refer to the resistance,

the number of intersecting line segments, the x and y coordinates specifying the location of the mid-point of the cell, the lower and upper limits of the confidence interval for the resistance, the sign of the product of the lower and upper limits of the confidence interval, and the percentile of the null distribution of resistances corresponding to the observed resistance. A .csv file with this information is produced by the function rdm_resistance.

Geo_raw: A string specifying the path to a text file showing the name (1st column) and geographical x and y coordinates (2nd and 3rd columns) of each individual (each row). The first row needs to specify the column headings, and all entries need to be tab-delimited.

r_size: Specifies the size of each grid cell in the plotted raster map of resistances, which can be adjusted to match the size of the plotted landscape. Default value of 5.

p_signf: Specifies the percentiles of the null distribution of resistances used to define genetic barriers and corridors. A genetic barrier is a grid cell with a resistance that is above the 100*(1-p_signf)% percentile, whereas a genetic corridor is a grid cell with a resistance that is below the 100*(p_signf)% percentile. Default value of 0.05.

p_size: Specifies the size of the sampling points in the plotted raster map of resistances. Default value of 2.

Output

After running rdm_mapper, a map of the distribution of resistance will be presented. There will be contour lines to highlight areas of statistical significance and certainty of resistance values.

Example

```
F.df <- read.csv('files/resistance_map.csv')

pdf('resistance_map.pdf', width = 5, height = 5)

rdm_mapper(F.df = F.df, Geo_raw = 'files/coordinates.txt', r_size = 5, p_signf = 0.05, p_size = 2)

dev.off()
```

Missing data

The *R* package can handle missing genetic data. Missing data generate NAs in the genetic distance matrix and hence in the IBD residuals. In the *R* package, all NAs in the calculated IBD residuals are removed in the rdm_residual function. For other ways of handling missing genetic data, please refer to the documentation of poppr *R* package (Kamvar et al., 2014). We caution that a high level of missing genetic data will substantially affect the accuracy of the results.

Workflow

The entire flow of analyses will typically take about one to a few hours, but the actual time depends on the amount of data and the proposed resolution of the resistance map.

Step 1. Prepare your data according to the required input format.

Step 2. Run rdm_IBD to check the distribution of genetic and geographic distances calculated from the data, and choose a method to calculate the genetic distance and a method to model IBD.

We recommend that users run rdm_IBD with all available methods of genetic distance calculation. For example, using the pigeon dataset that can be downloaded at https://github.com/takfung/ResDisMapper/tree/master/Documentation (for description of dataset, see Tang et al., 2018), we run rdm_IBD six times with the six different methods (Fig. 1).

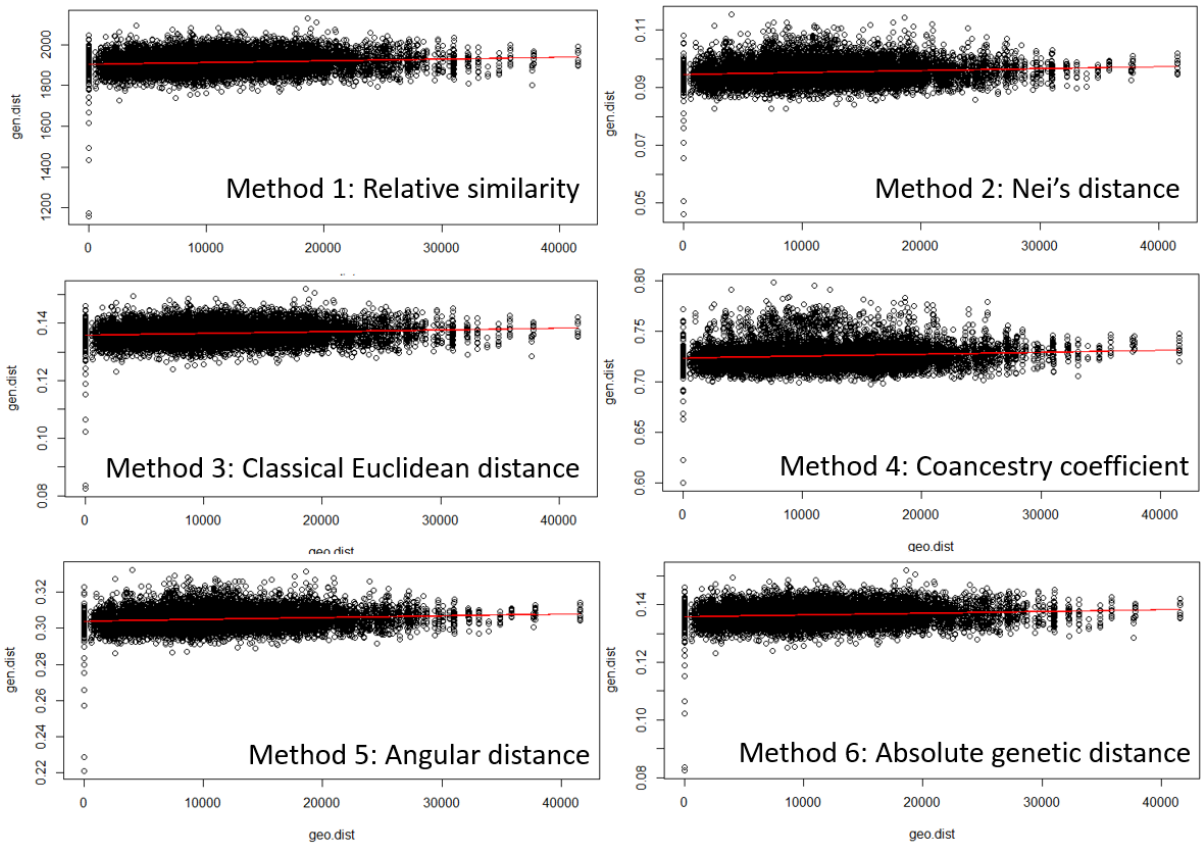

Fig. 1. Plots of genetic distance against geographic distance, produced by running rdm_IBD with the example pigeon dataset and each of the six available methods of calculating genetic distance. The black dots refer to the data points.

After running rdm_IBD with the six different methods of calculating genetic distance, one method can be chosen. Using the chosen method of calculating genetic distance, we recommend running

rdm_IBD with the two methods of modelling IBD, using a linear or non-linear model. For example, using Method 4 for calculating genetic distance, we run rdm_IBD using the pigeon dataset and the two methods of modelling IBD (Fig. 2).

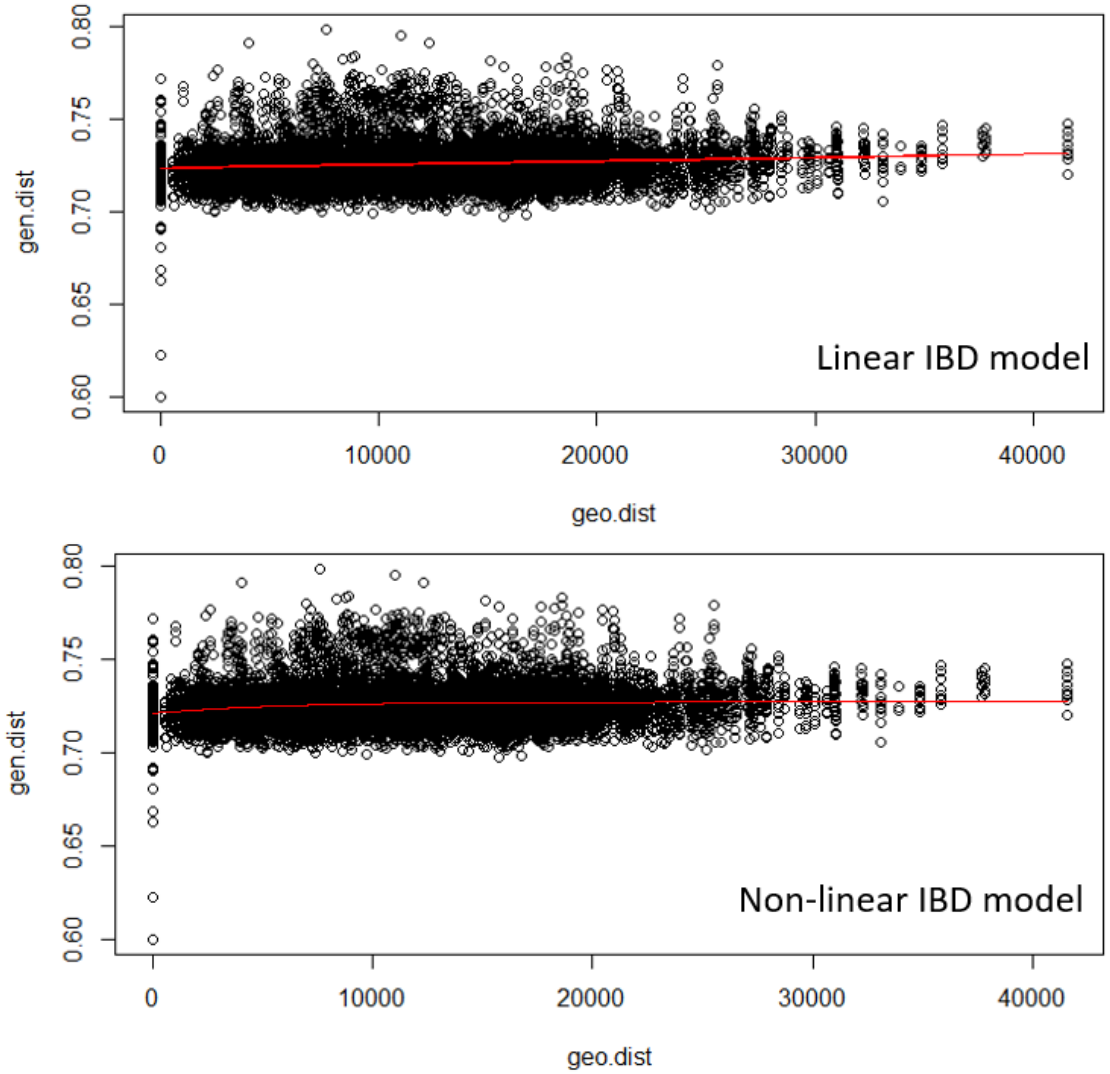

Fig. 2. Plots of genetic distance against geographic distance, produced by running rdm_IBD with the example pigeon dataset, Method 4 for calculating the genetic distance, and the two different methods of modelling IBD (linear and non-linear). The black dots refer to the data points, whereas the red lines refer to the fitted IBD model.

After running rdm_IBD with a particular method of calculating genetic distance and the two different methods of modelling IBD, one method for modelling IBD can be chosen. Using the chosen method of calculating genetic distance and the chosen method of modelling IBD, rdm_IBD is run for the last time. For example, using our pigeon dataset, we ran rdm_IBD using Method 4 for calculating genetic distance and a linear model for IBD:

```
IBD.res <- rdm_IBD(Gen_raw = 'Pigeon.gen', Geo_raw = 'Pigeon.coordinates', Dist_method = 4, IBD_method = 2)
```

Please take note of the scale of geographical distance for the dataset and if there is any discrete pattern of genetic distance along the geographical distance. This will help you to manipulate min.dist and max.dist in the next step to examine dispersal patterns at a certain geographical scale.

Step 3. Run rdm_residual to generate an object of class SpatialLinesDataFrame and visualize the distribution of IBD residuals in the form of line segments. You will get a rotatable 3D plot to check IBD residuals in the form of line segments.

If the plot of genetic distance against geographic distance in Step 2 shows a discrete pattern, or if the dispersal range is known for the population considered, users can specify the geographical distance range (min.dist and max.dist) within which line segments are considered for calculation of resistance in Step 4 (Fig. 3).

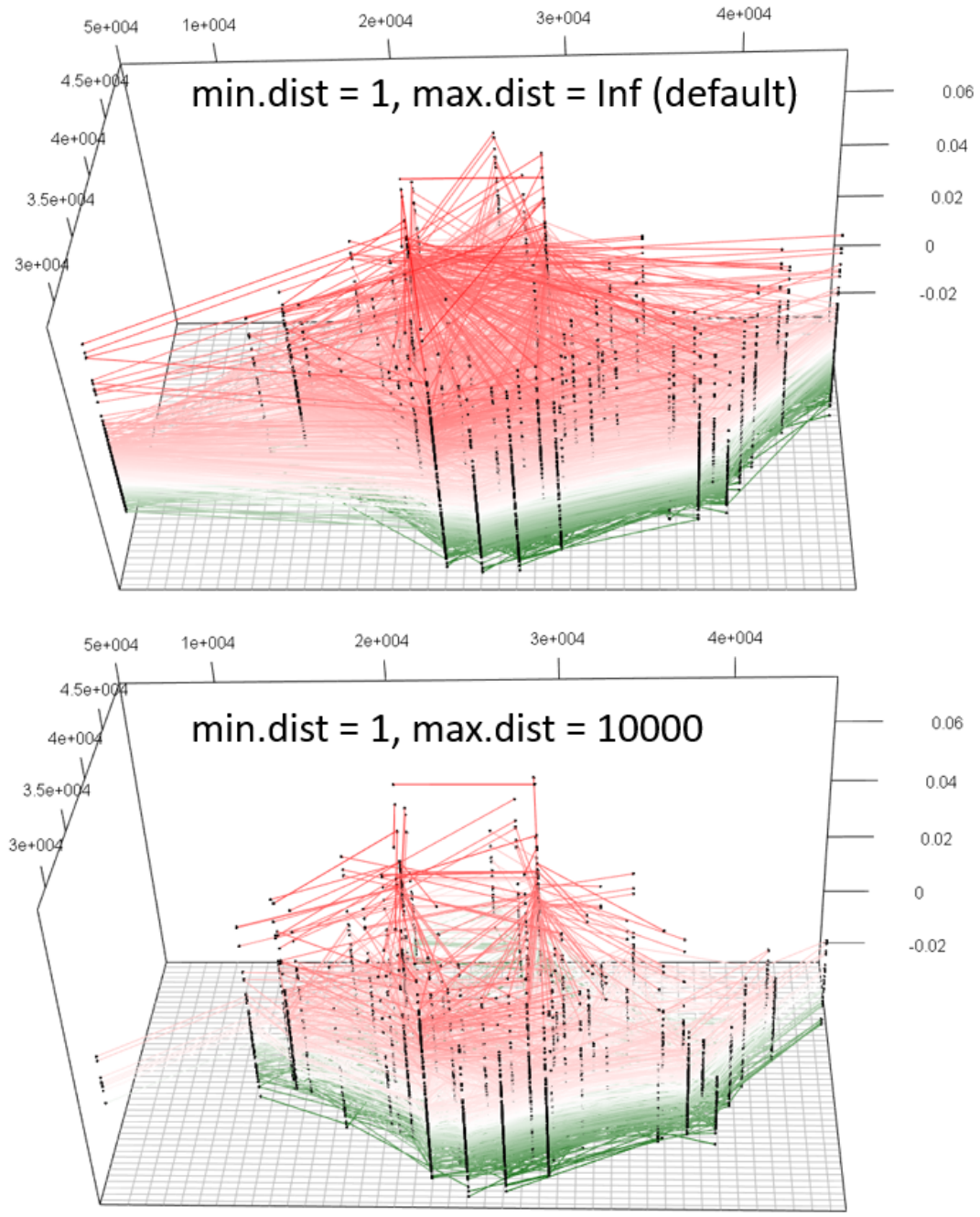

Fig. 3. 3D plot visualization of IBD residuals as line segments, produced by running rdm_residual with the example pigeon dataset and with different geographical distance ranges. Only line segments within the geographical distance range specified are plotted. Line segments with different colors from dark green to dark red correspond to IBD residuals that are increasingly positive. The vertical axis refers to the values of the residuals.

The 3D plot of IBD residuals can also be used to decide what resolution the grid representing the landscape considered should be, which is used in calculating resistance in Step 4. Users need to make

sure that most of the grid cells have enough intersecting line segments to provide a continuous resistance map in Step 5 (Fig. 4).

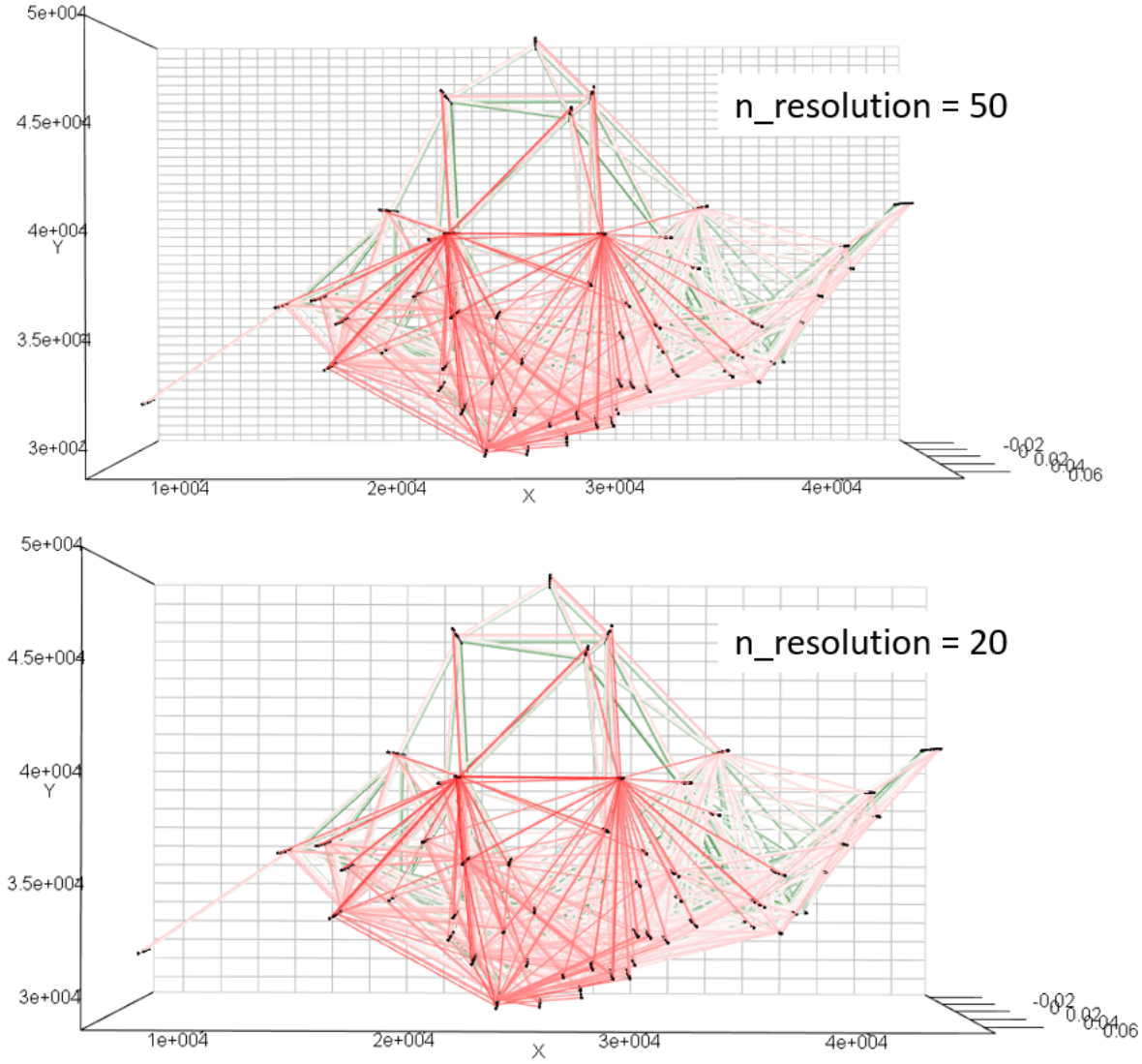

Fig. 4. 3D plot visualization of IBD residuals as line segments, produced by running rdm_residual with the example pigeon dataset and with different grid size resolution (number of cells in each row or column). Line segments with different colors from dark green to dark red correspond to IBD residuals that are increasingly positive. The *z* axis refers to the values of the residuals.

After deciding what geographical distance range and grid size resolution to use, run rdm_residual for the last time with this range and resolution. For example in our pigeon case, we ran:

```
Res_SLDF <- rdm_residual(IBD.res = IBD.res, Geo_raw = “Pigeon.coordinates”, min.dist = 1, max.dist = Inf, n_resolution = 50, proj = CRS("+init=epsg:4326"))
```

Step 4. Run rdm_resistance to calculate the resistance over the landscape considered, together with whether the resistance values are statistically different from null values and are statistically different from zero.

The function rdm_resistance takes as input an object of class SpatialLinesDataFrame, which is the output of Step 3. The resolution of the resistance raster map produced (number of cells in a row and number of cells in a column) has to be no smaller than the resolution of the SpatialLinesDataFrame object produced in Step 3. We would recommend you choose the resolution according to the density and distribution of your samples. The function also requires a name for the output file (in .csv format), including all useful information of the resistance raster map. The other options in this function (coverage of confidence intervals and number of random resamples for the generation of null distributions of resistances) can generally be left as the default values unless you have a particular reason to use different values. For example, in our pigeon case, we ran:

```
F.df <- rdm_resistance(Res_SLDF = Res_SLDF, nrows = 25, ncols = 50, conf_intervals = 0.95, random_rep = 1000, outputfile = 'resistance_map.csv')
```

Running rdm_resistance typically takes about an hour to a few hours, so you can monitor progress in the R console (Fig. 5).

```
[1] "Calculating resistance (1/4)"
  |======================================================================================================================================================| 100%
[1] "Counting intersects (2/4)"
  |======================================================================================================================================================| 100%
[1] "Calculating lower bound of confidence interval (3/4)"
  |======================================================================================================================================================| 100%
[1] "Calculating upper bound of confidence interval (4/4)"
  |==========                                                                                                                                            |  10%
```

Fig. 5. Screen shot of R console while processing a call of rdm_resistance, using the example pigeon dataset.

Step 5. Run rdm_mapper to visualize the resistance map.

To visualize the resistance map, you need to input the data frame returned from Step 4. The input data frame can be directly set as the output of the function rdm_resistance in Step 4, or imported from the .csv file produced by the function. In the map, you will see cells of different colors ranging from red to green. The colors indicate different levels of resistance, with red referring to high resistance and green referring to low resistance. Red and green contour lines are drawn on the map, delineating genetic barriers and corridors, respectively. These correspond to areas with resistance values that are higher or lower than those from a null distribution with high probability. In addition, blue contour lines are drawn to delineate areas with resistances that are statistically different from 0. For the interpretation of the map, please see the annotated example below (Fig. 6). If a contour line

encounters a grid cell with no data (no intersecting line segments) for calculating a resistance value, then the line terminates without forming a loop.

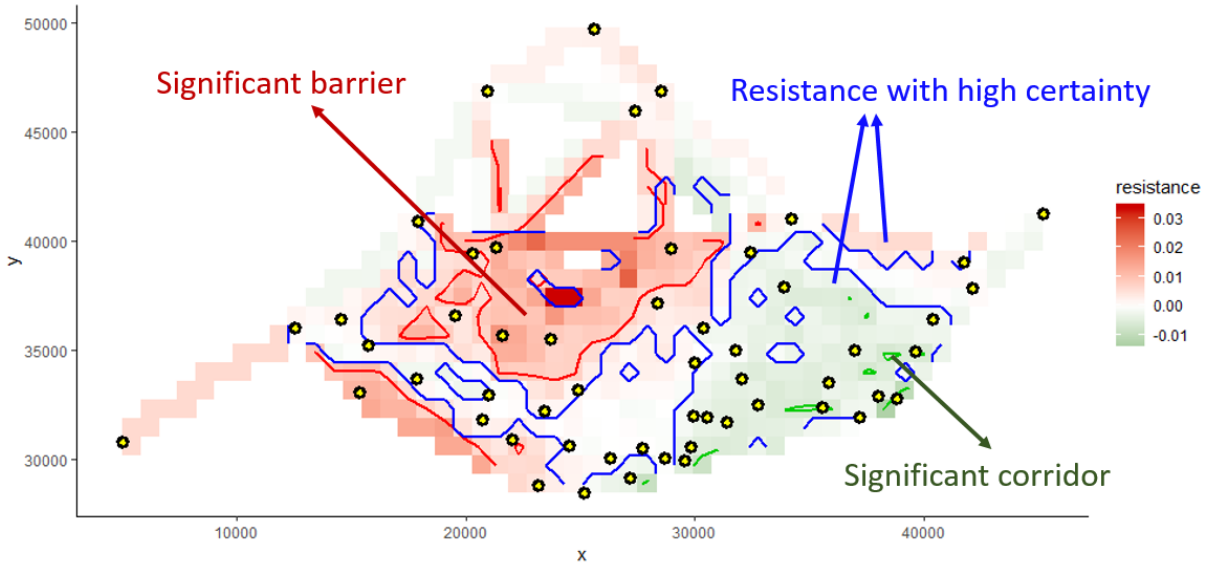

Fig. 6. Resistance map produced by the function rdm_mapper with the example pigeon dataset, with annotations describing the meaning of the different colors and contours. A significant barrier/corridor refers to areas with resistance values that are higher/lower than those from a null distribution with high probability, and lie within the red/green contours. Areas that lie inside the blue contours have resistance values with high probability being positive or negative (high "certainty"). The yellow circles indicate sampling points.

We note that in our example using the pigeon dataset, using different resolutions changes the pattern of resistance quantitatively but not qualitatively (Fig. 7).

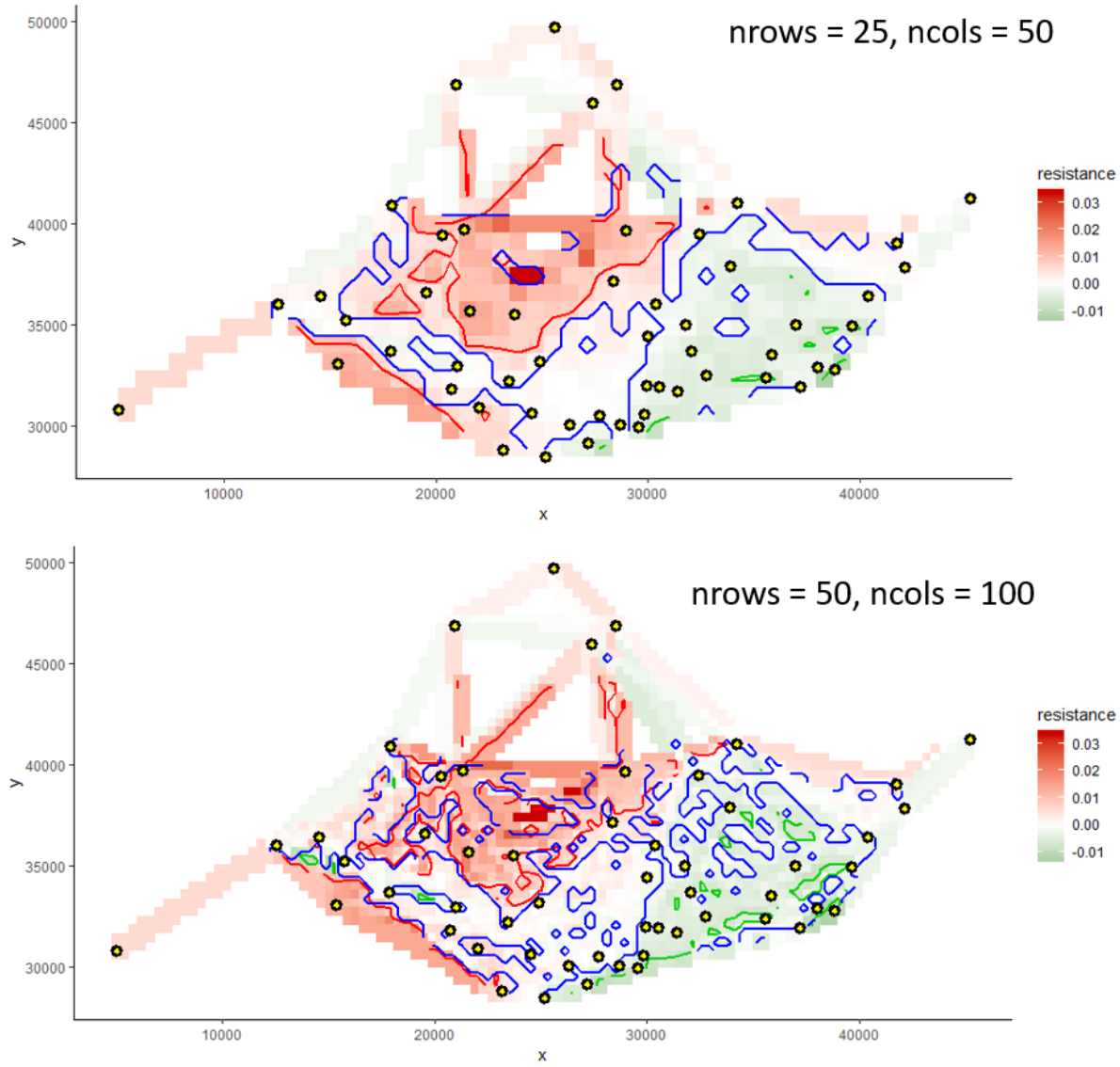

Fig. 7. Resistance maps produced by the function rdm_mapper with the example pigeon dataset, with different resolutions. The yellow circles indicate sampling points.

If you wish to further customize the resistance map, for example to change the colors, we recommend you write your own function with use of ggplot2.

Citations

Kamvar ZN, Tabima JF, Grünwald NJ. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:e281.

Keis, M., Remm, J., Ho, S. Y., Davison, J., Tammeleht, E., Tumanov, I. L., ... & Margus, T. (2013). Complete mitochondrial genomes and a novel spatial genetic method reveal cryptic

phylogeographical structure and migration patterns among brown bears in north-western Eurasia. *Journal of Biogeography*, 40(5), 915-927.

Qian, T., Low, G. W., Lim, J. Y., Gwee, C. Y., Rheindt, F. E. (2018). Human activities and landscape features interact to closely define the distribution and dispersal of an urban commensal. *Evolutionary Applications*, doi: 10.1111/eva.12650.