# MEMGENE package for R: Tutorials

Paul Galpern[1,2] and Pedro Peres-Neto[3]

[1] *Faculty of Environmental Design, University of Calgary*
[2] *Natural Resources Institute, University of Manitoba*
[3] *Département des sciences biologiques, Université du Québec à Montréal*

## Contents

# 1 Introduction

MEMGENE is a tool for spatial pattern detection in genetic distance data. It uses a multivariate regression approach and Moran's Eigenvector Maps (MEM) to identify the spatial component of genetic variation. MEMGENE variables are the output, and can be used in visualizations or in subsequent inference about ecological or movement processes that underly genetic pattern. Please see the publication associated with the MEMGENE package (Galpern et al., 2014) for more information.

Three tutorials are presented here. The first shows a MEMGENE analysis of a simulated data set produced for the publication associated with the MEMGENE package (Galpern et al., 2014) and a second demonstrates an analysis for field-collected caribou data contained in the same paper. The third tutorial demonstrates how to use MEMGENE in the context of landscape genetic analysis also using simulated data.

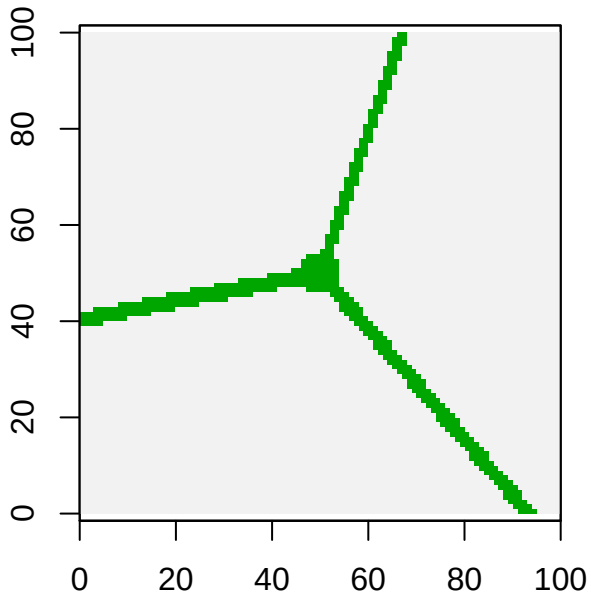
## 2 Tutorial: Spatial genetic patterns in simulated data

This tutorial demonstrates how to use MEMGENE when the primary objective is to identify spatial genetic patterns. It is possible to reproduce these examples directly in R. The tutorial focuses on the radial data set, which is also provided with the package.

### 2.1 The data set

A full description of how the radial spatial genetic data were simulated is available in the publication associated with this package (Galpern et al., 2014). Briefly, we simulated the moving and mating of 1000 individuals over 300 non-overlapping generations. Movement across the arms of the radial structure (see figure below) was less likely than within the three regions of the landscape, due to landscape resistance to movement imposed on the simulated individuals. This makes the radial structure into a semi-permeable barrier, reducing dispersal and therefore gene flow. Given a sufficient number of generations for genetic drift under reduced gene flow, we expect a spatial genetic pattern that reflects the landscape resistance pattern in this Figure (below).

The data set provided with the package (`radial.csv` installed in the `extdata` folder) represents a spatially stratified sampling of 200 individuals at generation 300 of this simulation. It includes 200 rows, one for each individual, two columns giving coordinates at which the individual was "sampled", and 30 paired columns giving the alleles at 15 codominant loci.




Above: The figure shows the radial resistance surface used to generate the spatial genetic data set used in this tutorial.

## 2.2 MEMGENE analysis

**Step 1** Produce a genetic distance matrix

MEMGENE requires a genetic distance matrix giving the pairwise genetic distances among individual genotypes. Any genetic distance metric can be used. In principle the method will also work with a population genetic distance matrix (e.g. pairwise Fst).

In this first step we find the genetic distance matrix using the proportion of shared alleles among individuals (Bowcock et al., 1994) as the metric. We use a convenience function included in the package to produce this that wraps functions in the `adegenet` package (Jombart, 2008).

```
## Load the radial genetic data
radialData <- read.csv(system.file("extdata/radial.csv",
    package="memgene"))
```

```
## Create objects for positional information and genotypes
radialXY <- radialData[ ,1:2]
radialGen <- radialData[, 3:ncol(radialData)]
```

```
## Produce a proportion of shared alleles genetic distance matrix
## using the convenience wrapper function provided with the package
radialDM <- codomToPropShared(radialGen)
```

**Step 2** Extract MEMGENE variables

In this second step we extract the MEMGENE variables, using the typical interface to the MEMGENE package (the `mgQuick` function). The analysis framework is discussed in detail in the publication associated with this package.

The `mgQuick` function does the following: (1) Finds the MEM eigenvectors given the sampling locations of the individuals (`mgMEM` function); (2) Uses these eigenvectors to identify significant spatial genetic patterns (`mgForward` and `mgRDA` functions); (3) Returns MEMGENE variables that describe these significant patterns on a reduced set of axes (`mgRDA` function). For additional detail on these functions, and for more control over the MEMGENE analysis see the R help files.

```
## Run the MEMGENE analysis
## May take several minutes
if (!exists("radialAnalysis"))
    radialAnalysis <- mgQuick(radialDM, radialXY)
```

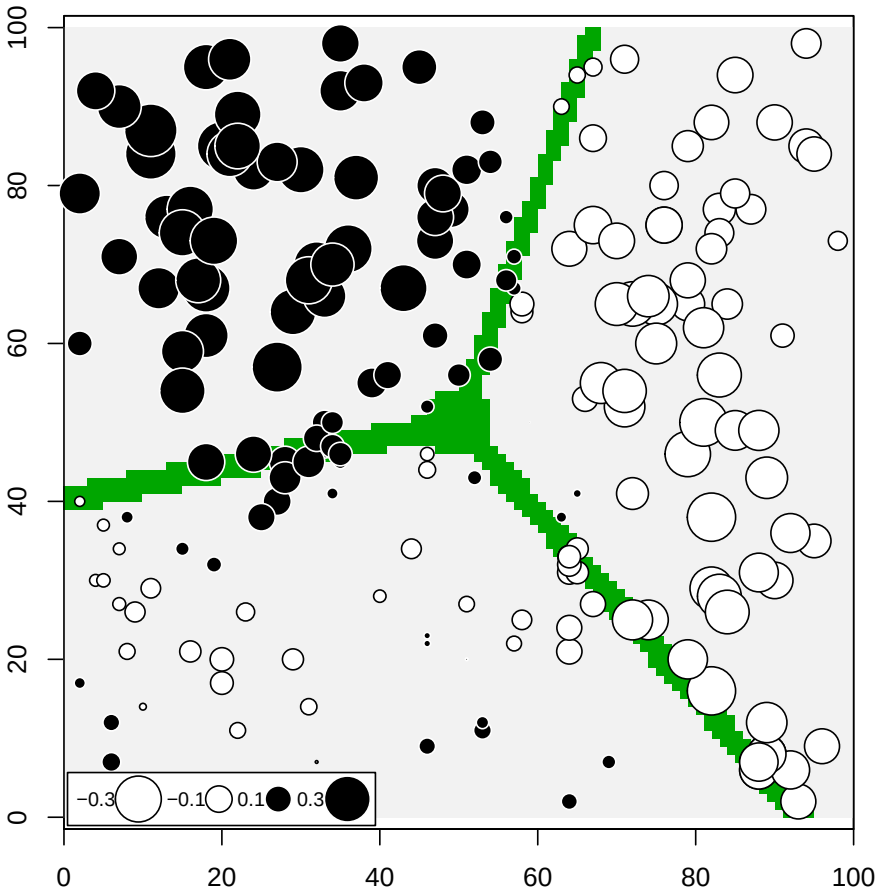
**Step 3** Visualize MEMGENE variables

The MEMGENE variables represent orthonormal patterns of significant spatial genetic variation, and are ordered in terms of the amount of variation they explain from most to least. Typically, much of the variation is summarized in the first two variables, so it can often be convenient to visualize these two initially.

```
## Visualize the first two MEMGENE variables
## by providing only the first two columns of the memgene matrix
mgMap(radialXY, radialAnalysis$memgene[, 1:2])
```

However, it is often more interesting to visualize the MEMGENE variables superimposed over some map or other. In the figure below we superimpose the first MEMGENE variable (MEMGENE1) over the resistance surface used to create the spatial genetic data. This can be done using the `add.plot=TRUE` parameter.

```
library(raster)
radialRas <- raster(system.file("extdata/radial.asc", package="memgene"))
plot(radialRas, legend=FALSE)
mgMap(radialXY, radialAnalysis$memgene[, 1], add.plot=TRUE, legend=TRUE)
```




Above: The scores of individuals on the MEMGENE1 axis superimposed on the resistance surface used to create the spatial genetic data. Circles of similar size and colour

represent individuals with similar scores on this axis. Note how the pattern of spatial genetic variation in MEMGENE1 (spatial genetic neighbourhoods) reflects the structure of the landscape used to create it.

Although visualization may often be an end in itself, the MEMGENE variables can also be used singly or in combination to test hypotheses about the creation of the spatial genetic neighbourhoods they describe.

# 3 Tutorial: Spatial genetic patterns in wildlife data

This tutorial demonstrates the use of MEMGENE to identify spatial genetic patterns in a data set for boreal woodland caribou, a North American ungulate.

## 3.1 The data set

A full description of how these spatial genetic data were collected and genotyped can be found in the publication associated with this package (Galpern et al., 2014). Briefly, these are genotypes for 87 caribou sampled on both sides of the Mackenzie River (Northwest Territories, Canada). The Mackenzie is a major North American river that varies between 1 and 4.5 km through the study area. Caribou have occasionally been reported crossing the river.

Boreal woodland caribou are a threatened species under Canada's Species at Risk Act. For this reason the caribou data included with the package have obfuscated sampling locations produced by reprojecting them in a way that maintains the Euclidean distance matrix among the points, but is not easily assignable to a precise location on the Earth's surface.

## 3.2 MEMGENE analysis

**Step 1** Produce a genetic distance matrix

```
## Load the caribou genetic data
caribouData <- read.csv(system.file("extdata/caribou.csv",
    package="memgene"))
```

```
## Create objects for positional information and genotypes
caribouXY <- caribouData[ ,1:2]
caribouGen <- caribouData[, 3:ncol(caribouData)]
```

```
## Produce a proportion of shared alleles genetic distance matrix
## using the convenience wrapper function provided with the package
caribouDM <- codomToPropShared(caribouGen)
```

**Step 2** Extract MEMGENE variables

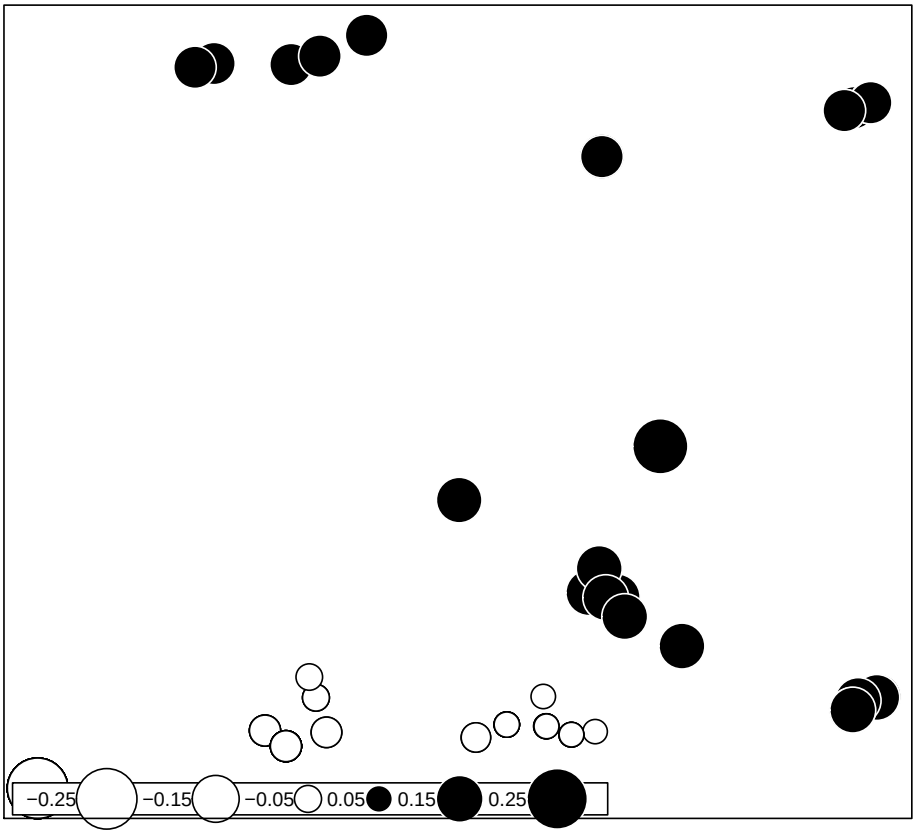
```
## Run the MEMGENE analysis
## May take several minutes
if (!exists("caribouAnalysis"))
    caribouAnalysis <- mgQuick(caribouDM, caribouXY)
```

**Step 3** Visualize MEMGENE variables

The results of the visualization of MEMGENE1 is shown in Figure below. This figure also appears in the publication associated with this package, sumperimposed over a map of the region.

```
plot(caribouXY, type="n", xlab="", ylab="", axes=FALSE)
mgMap(caribouXY, caribouAnalysis$memgene[, 1], add.plot=TRUE,
    legend=TRUE)
```

```
box()
```




Above: The scores of individual caribou on the MEMGENE1 axis. The Mackenzie River separates the white and black circles diagonally through the lower half of the map (not shown). For the full presentation of these results see the publication associated with this package.

**Step 4** Additional interpretation

Finding the adjusted R-squared (i.e. the genetic variation explained by spatial pattern) is just a matter of referencing the list element in the `caribouAnalysis` object as follows:

```
caribouAnalysis$RsqAdj
```

```
## [1] 0.02905906
```

Note that this low value should be interpreted not as an inadequacy of the regression to explain variation, but rather that there is only a small proportion of all genetic variation

that can be attributed to spatial patterns; or more specifically, to the `N-1` (where `N` is the number of sampling locations) MEM spatial eigenfunctions that were extracted. It is important to note, however, that adjustments to how the MEM eigenfunctions are extracted have the potential to subtly change which spatial patterns are captured, as well as increase R squared. Further work is required to explore the effects of these modelling decisions.

Then determining the proportion of the this variation that is explained by each of the MEMGENE variables is also straightforward:

```
## Find the proportional variation explained by each MEMGENE variable
caribouMEMGENEProp <- caribouAnalysis$sdev/sum(caribouAnalysis$sdev)
```

```
## Neatly print proportions for the first three MEMGENE variables
format(signif(caribouMEMGENEProp, 3)[1:3], scientific=FALSE)
```

```
##       MEMGENE1       MEMGENE2       MEMGENE3
## "0.7220000000" "0.2780000000" "0.0000000153"
```

It is clear that there are only two distinctive patterns in these data, and the dominant pattern is that created by the Mackenzie River (i.e. MEMGENE1)

# 4 Tutorial: Landscape genetics using simulated data

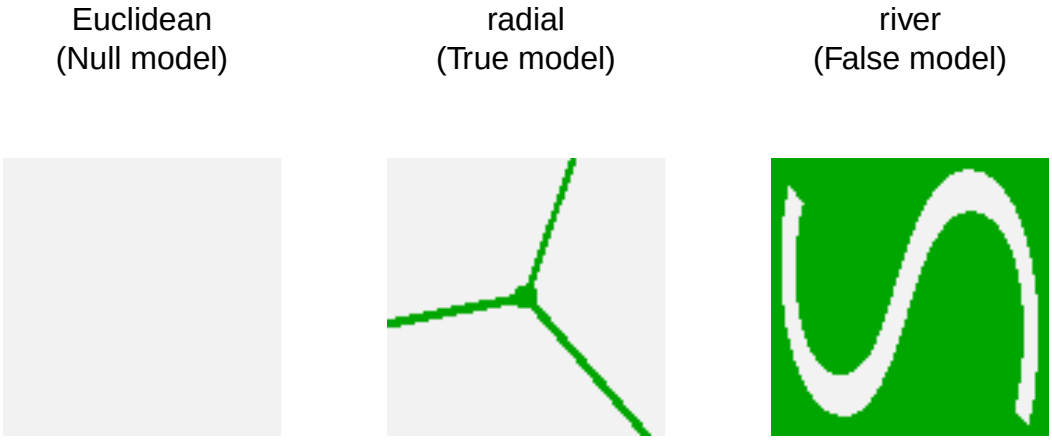
This tutorial demonstrates the use of MEMGENE to explore a common research question in landscape genetics. Given genetic samples collected from across a broad spatial extent researchers may be interested in determining whether landscape or spatially-variable environmental conditions have influenced organism dispersal and, by extension, gene flow. With a sufficient number of generations (e.g. for genetic drift) and insufficient gene flow to homogenize genetic differences, evidence of reduced or biased dispersal may be read from spatial patterns in neutral genetic markers (Segelbacher et al., 2010; Storfer et al., 2010).

A common approach to testing the influence of landscape features on gene flow has been to create a landscape resistance surface; a hypothesis in the form of a map about the degree to which geographical, ecological and anthropogenic features such as mountains, land cover and roads may reduce dispersal and gene flow for an organism of interest. This resistance surface is then used to predict genetic patterns observed among individuals or populations. Typically the predictive performance is also compared to a null model, a Euclidean surface where any differentiation among genetic samples across space is considered to be a function of distance (known as isolation-by distance; IBD) (Wright, 1943). By contrast, isolation-by-resistance (IBR) is inferred where effective distances (e.g. of least-cost paths among samples on the resistances surface) better explain genetic patterns than the straight-line or Euclidean distances implied by the null model.

Here we use MEMGENE to compare the predictive performance of Euclidean and resistance surface hypotheses for genetic pattern. In previous tutorials MEM eigenvectors were extracted from a truncated Euclidean distance matrix among sampling locations which provided a set of spatial patterns. Note that truncation is necessary because a fully Euclidean matrix produces only two eigenvectors (or spatial patterns) exactly representing the original variables. A subset of these spatial patterns was then identified using forward selection and used for visualization. In the following analysis, resistance surfaces are included by instead finding the MEM eigenvectors from a truncated matrix containing least-cost path distances among sampling locations. This produces a different set of spatial patterns to test using forward selection. We can then compare the proportion of genetic variation explained by the selected MEM eigenvectors from a Euclidean surface as well as resistance surface models. Using a variation partitioning procedure we partition genetic variation among two sources of spatial variation: (1) the selected eigenvectors from a given surface; and, (2) the spatial coordinates of sampling.

## 4.1 Landscape genetic data sets

The radial data set was used where spatial genetic patterns were simulated to reflect resistance created by three semi-permeable linear features (see first tutorial). Three different landscape models are tested for these data: (1) A null Euclidean surface; (2) the true radial resistance surface used to generate the data; and (3) a false river resistance surface. The three surfaces are shown below.




## 4.2 MEMGENE analysis of multiple landscape models

**Step 1** Prepare landscape resistance models

First we assemble the two resistance surfaces we are going to test into a `RasterStack` object (a set of rasters with identical extent, resolution and coordinate reference system). These rasters represent the radial model (used to generate the genetic data) and the river model (not used to generate the data).

```
resistanceMaps <- stack(
           raster(system.file("extdata/radial.asc", package="memgene")),
           raster(system.file("extdata/river.asc", package="memgene")))
```

**Step 2** Prepare the genetic data

We next prepare the radial genetic data for analysis in the same way as we prepared genetic data for previous tutorials. Again, we are using the proportion of shared alleles among pairs of individuals (Bowcock et al., 1994) as the genetic distance metric, but this is not a requirement of the analysis.

```
radialData <- read.csv(system.file("extdata/radial.csv",
    package="memgene"))
radialGen <- radialData[, -c(1,2)]
radialXY <- radialData[, 1:2]
radialDM <- codomToPropShared(radialGen)
```

**Step 3** Compare three landscape models

Finally, we use the `mgLandscape` function to compare the proportion of spatial genetic variation explained by each these two resistance surface models and a Euclidean model.

```
## Note permutations are set high for greater accuracy
## Reduce to 100 in each case for a faster run (note
## results may differ slightly because forward selection of
## spatial patterns differs)
if (!exists("compareThree")) {
compareThree <- mgLandscape(resistanceMaps,
                            radialDM, radialXY, euclid=TRUE,
                            forwardPerm=500, finalPerm=1000)
}
```

```
## Analyzing Euclidean surface (landscape model 1 of 3)
## Extracting Moran's eigenvectors from Euclidean distance matrix
## Forward selections of positive Moran's eigenvectors
## ----Selected: 1, 2, 3, 5, 6, 7, 8, 9, 13, 21
## Forward selections of negative Moran's eigenvectors
## ----Selected: None
## Partitioning spatial genetic variation
##
## Analyzing resistance surface (landscape model 2 of 3) [radial]
## Calculating least-cost path distance matrix
## Extracting Moran's eigenvectors from least-cost path distance matrix
## Forward selections of positive Moran's eigenvectors
## ----Selected: 1, 2, 3, 4, 6, 10, 13, 14, 44
## Forward selections of negative Moran's eigenvectors
## ----Selected: None
## Partitioning spatial genetic variation
##
## Analyzing resistance surface (landscape model 3 of 3) [river]
## Calculating least-cost path distance matrix
## Extracting Moran's eigenvectors from least-cost path distance matrix
## Forward selections of positive Moran's eigenvectors
## ----Selected: 3, 4, 5, 6, 9, 11, 13, 23
## Forward selections of negative Moran's eigenvectors
## ----Selected: None
## Partitioning spatial genetic variation
```

**Step 4** Interpretation

```
print(compareThree)
```

```
## mgLandscape Analysis
##      model [abc] P[abc]    [a]  P[a]     [c]  P[c]    [b]   [d]
##  Euclidean 0.135  0.001 0.0462 0.001 0.01017 0.001 0.0788 0.865
##     radial 0.156  0.001 0.0672 0.001 0.00547 0.026 0.0835 0.844
##      river 0.118  0.001 0.0291 0.001 0.05595 0.001 0.0330 0.882
##
## Interpretation:
## Proportion of variation in genetic distance that is... (RsqAdj)
## [abc] explained by spatial predictors
## [a]   spatial and explained by selected patterns in the model
## [c]   spatial and explained by coordinates not patterns in the model
## [b]   spatial and confounded between the model and coordinates
## [d]   residual and not explained by spatial predictors
```

There are several important results to interpret from the above table:

1. Comparing the `[abc]` fraction for all three models indicates that incorporating the MEM eigenvectors derived from spatial patterns in the radial resistance surface explains the highest proportion of spatial genetic variation.
2. In the radial case the `[a]` fraction is the highest and the `[c]` fraction is the lowest, indicating that the majority of spatial genetic variation has been partitioned to the selected MEM eigenvectors and not to the coordinates (which describe linear sources of spatial genetic variation not described by the Moran's eigenvectors).
3. By contrast, in the river model, the `[c]` fraction is higher than the `[a]` fraction showing that the selected eigenvectors in this case are relatively poor at capturing spatial genetic pattern, compared to the coordinates. In other words, the linear pattern of genetic differentiation (i.e. coordinates) is more important than a non-linear one (Moran's eigenvectors).
4. The similar performance of the Euclidean and the radial model underlines the fact that the radial resistance surface is highly correlated with the Euclidean surface (i.e. individuals that fall within a given third of the landscape are expected to differ as a function of Euclidean distance).

In summary, the radial resistance surface, the "true" model that was used to generate the simulated genetic data, uniquely explains the highest proportion of spatial genetic pattern (i.e. in its `[a]` fraction). But because of its similarity to a Euclidean surface, it is not easily distinguished from this latter model. This may imply that that a truncated Euclidean distance matrix for which no a priori hypothesis is imposed is capable of describing complex patterns closer to the true spatial constraints underlying genetic variability in landscapes.

## References


Bowcock AM, Ruizlinares A, Tomfohrde J, et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368, 455-457.

Galpern, P., Peres-Neto, P., Polfus, J., and Manseau, M. (2014) MEMGENE: Spatial pattern detection in genetic distance data. *Submitted*.

Jombart T. (2008) adegenet: a R package for the multivariate analysis of genetic markers *Bioinformatics* 24, 1403-1405.

Segelbacher, G., Cushman, S.A., Epperson, B.K., Fortin, M.-J., Francois, O., Hardy, O.J., Holderegger, R., Taberlet, P., Waits, L.P. and Manel, S. (2010) Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics*, 11, 375-385.

Storfer, A., Murphy, M.A., Spear, S.F., Holderegger, R. and Waits, L.P. (2010) Landscape genetics: where are we now? *Molecular Ecology*, 19, 3496-3514.

Wright, S. (1943) Isolation by distance. *Genetics*, 28, 114-138.