Analysing landscape effects on dispersal networks and gene flow with genetic graphs
Supporting information

Savary, Paul*[1, 2, 3], Foltête, Jean-Christophe[2], Moal, Hervé[1], Vuidel, Gilles[2], and Garnier, Stéphane[3]

[1] *ARP-Astrance, 9 Avenue Percier, 75008 Paris, France*
[2] *ThéMA, UMR 6049 CNRS, Université Bourgogne-Franche-Comté, 32 Rue Mégevand, 25030 Besançon Cedex, France*
[3] *Biogéosciences, UMR 6282 CNRS, Université Bourgogne-Franche-Comté, 6 Boulevard Gabriel 21000 Dijon, France*

Contents

1 Supplementary figures **2**

2 Glossary of acronyms **8**

3 Mathematical background : independence graphs **8**

4 Computation of the DPS genetic distance **13**

*Corresponding author: `paul.savary univ-fcomte.fr`

1 Supplementary figures

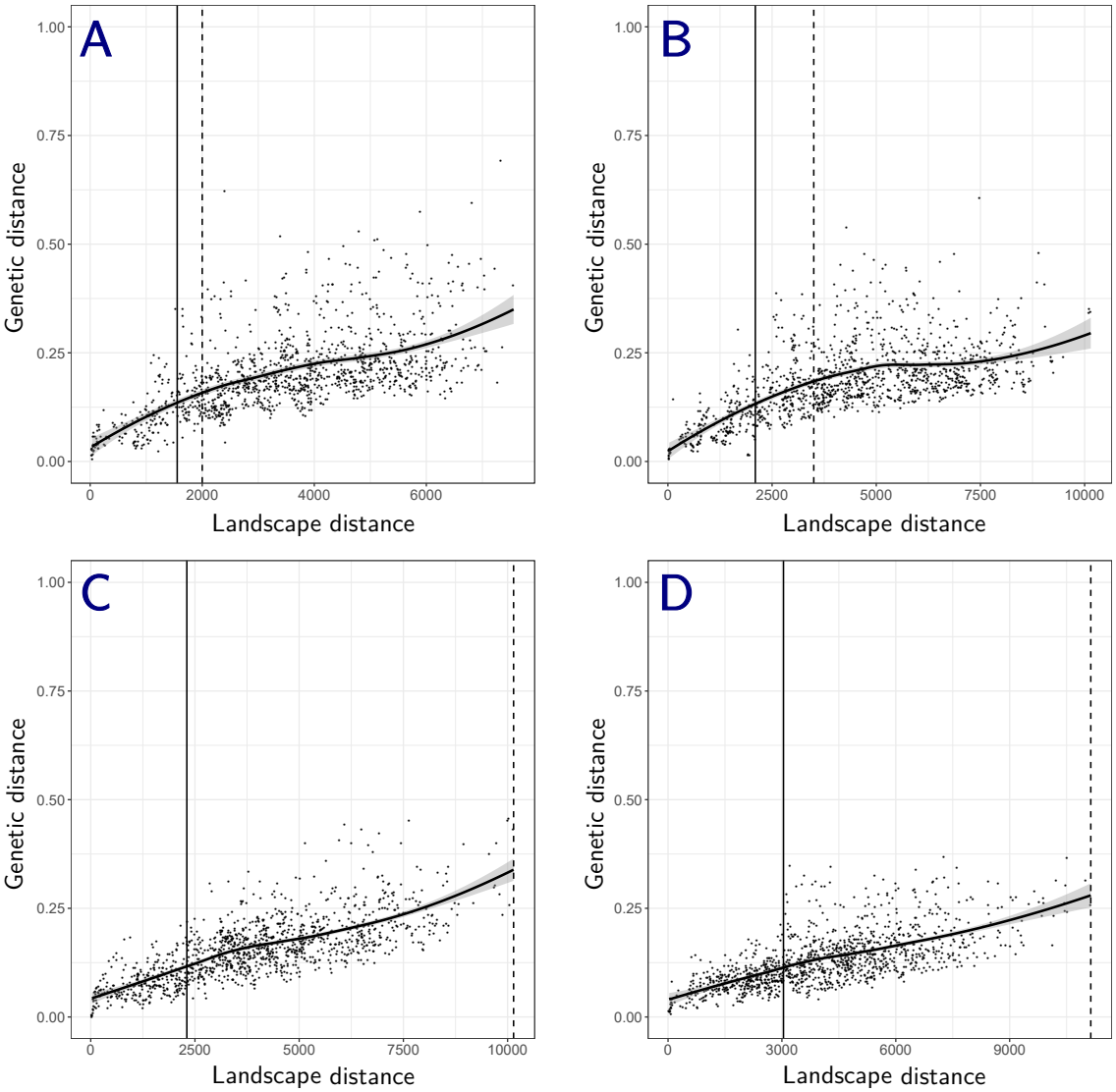

Figure S1: Scatter plots of the genetic distance (F_{ST}) plotted against cost-distance at generation 50. Cases A to D illustrate the gradient of IBD patterns (from type-IV to type-I) along the first component of the PCA. Solid vertical lines indicate the maximum dispersal distance, dashed lines indicate the DMC.

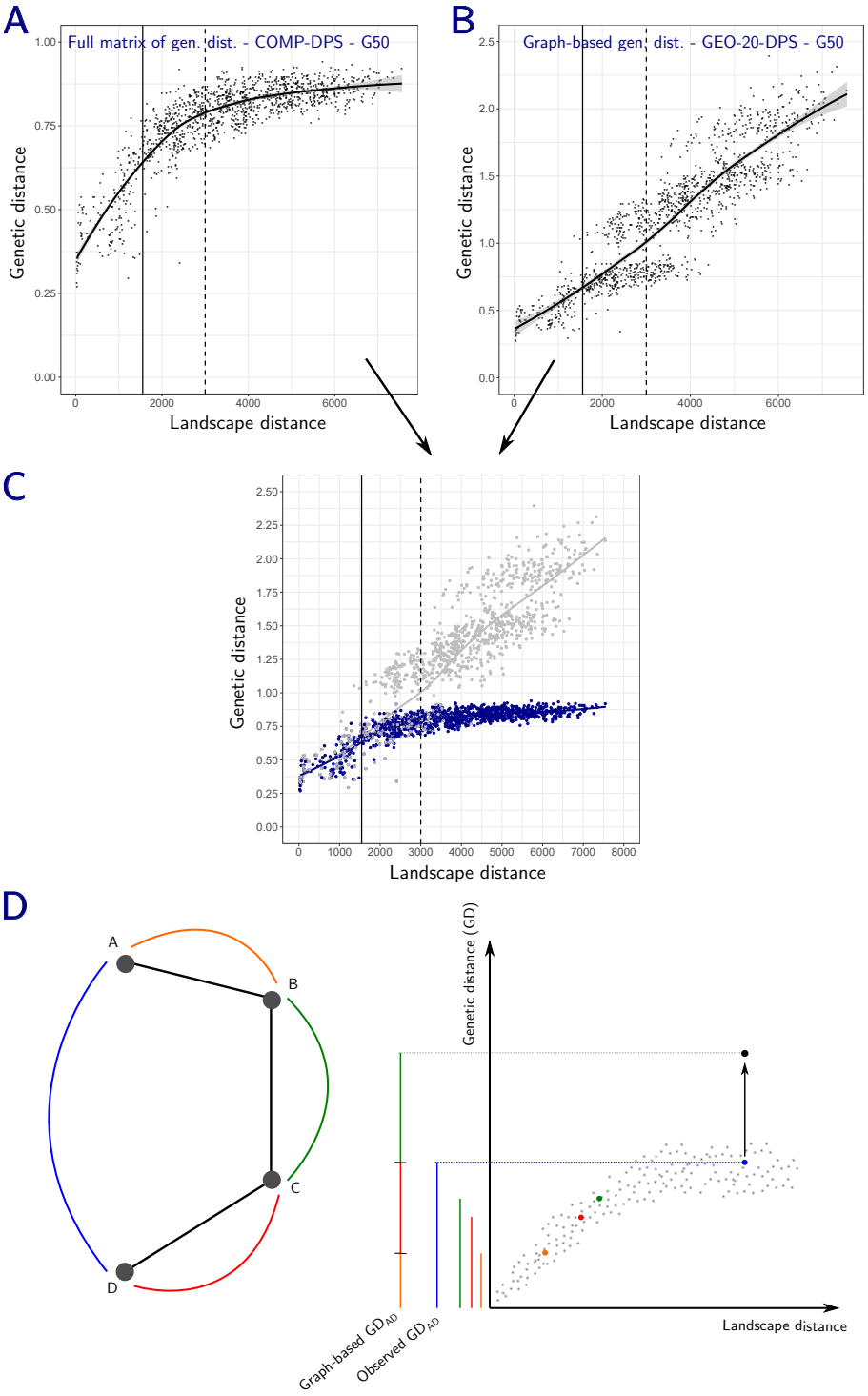

Figure S2: Illustration of the potential bias induced by the use of the graph-based genetic distance. Scatter plots of the complete matrix of D_{PS} (A) or the graph-based D_{PS} from the graph GEO-20-DPS (B) against the CD. Both scatter plots are combined in panel C to display genetic distance value differences induced by the sum of genetic distances. Panel D illsutrates the mechanism behind this. Populations A and D cannot directly exchange propagules while dispersal occurs in a stepwise way between population pairs A-B, B-C and C-D. Although the genetic distance between A and D is high, it is not directly proportional to the distance between and A and D because the relationship between genetic and landscape distance flattens out, as expected under a case-IV pattern of IBLR. Considering that the genetic distance GD_{AD} between A and D is equal to $GD_{AB} + GD_{BC} + GD_{CD}$ therefore over-estimates it. Example data come from the simulation A (generation 50). Solid vertical lines indicate the maximum dispersal distance, dashed lines indicate the DMC computed with the D_{PS} at generation 50.

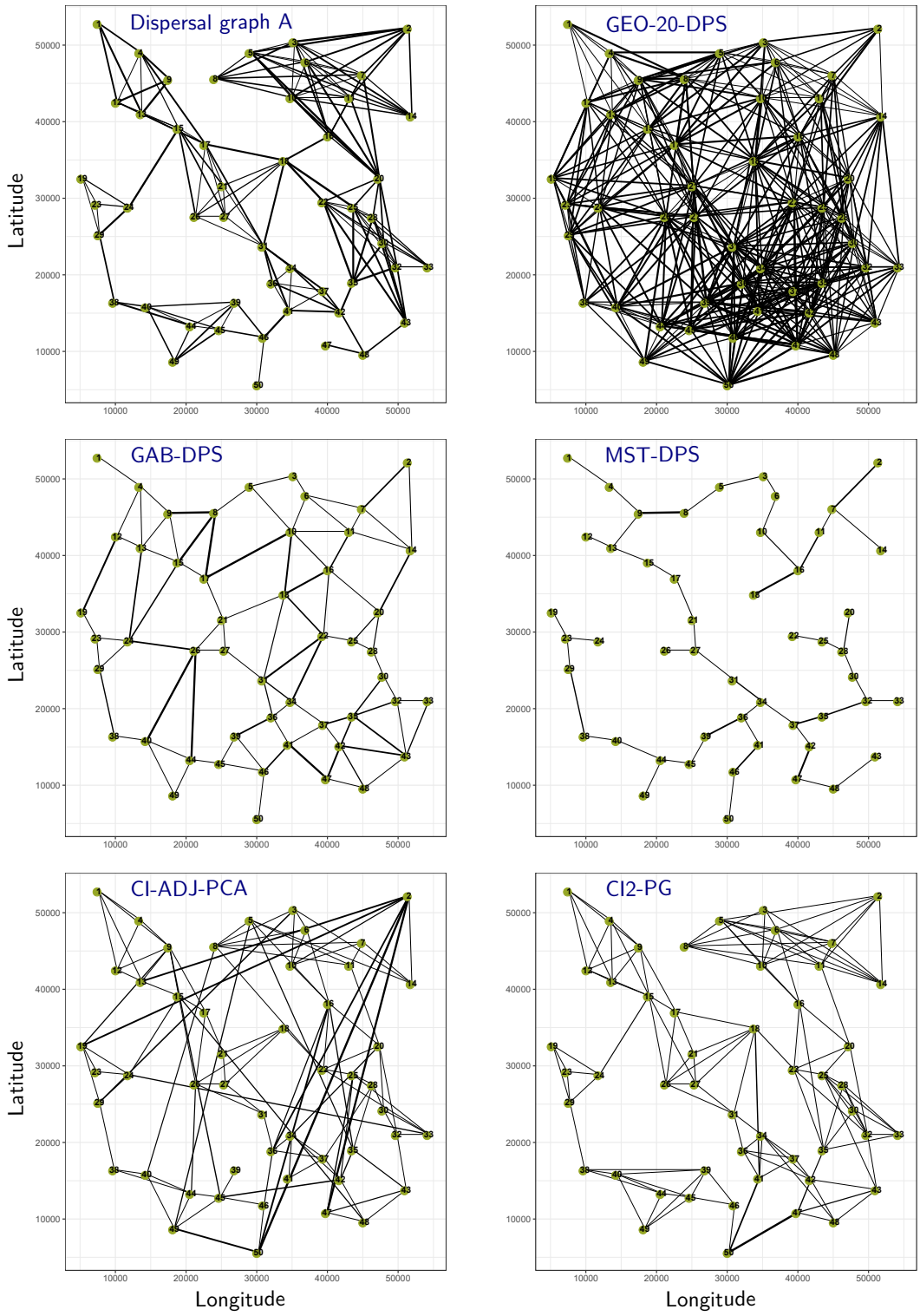

Figure S3: 6 examples of graphs (from left to right and top to bottom): a dispersal graph, a graph pruned at a geographical distance of 15 km, a Gabriel graph, a Minimum Spanning Tree and two independence graphs (see table 1 for the graph names). Example data come from a run of the simulations performed for configuration A (generation 50). Link width is proportional to genetic distance values.

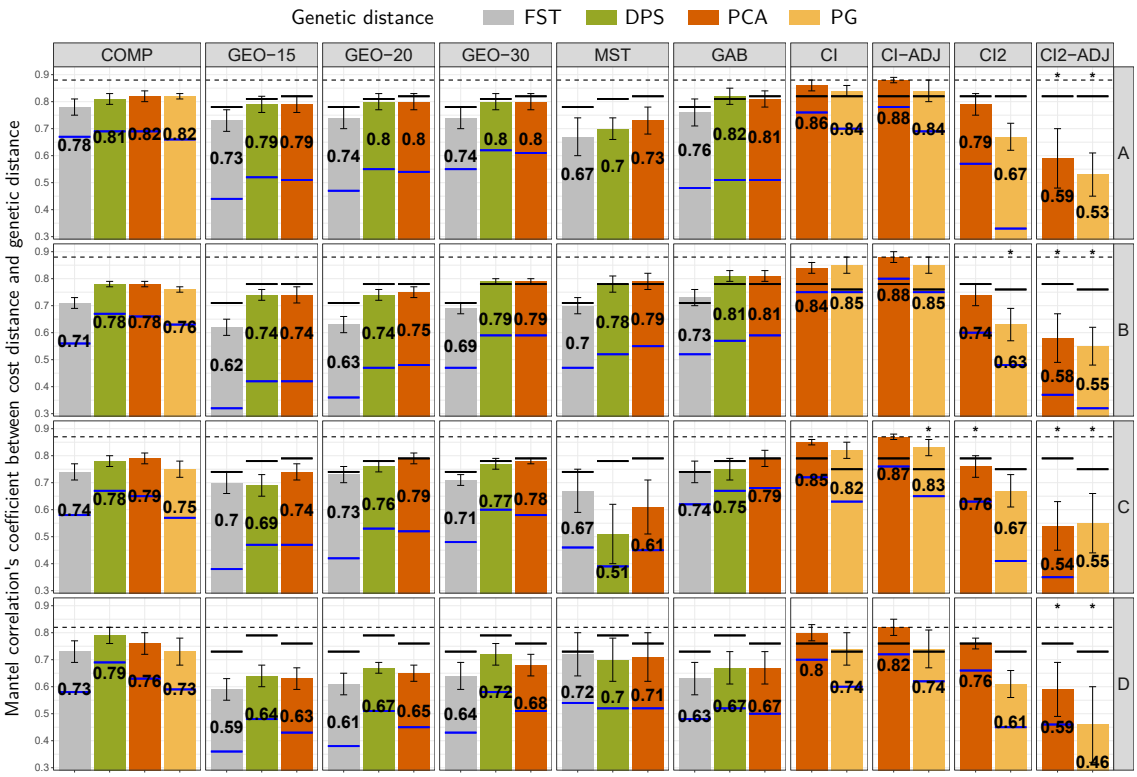

Figure S4: Mantel correlation between genetic distances and cost-distances separating nodes directly connected on the genetic graphs, according to the type of genetic distance and the pruning method at generation 500 (see table 1 for the graph names). Mean $\pm$ SD values were computed for the 10 runs simulated in each scenario. Blue bars refer to the correlation coefficient between genetic distance and geographical distance, when it is above 0.3. Black bars refer to the correlation coefficient obtained using every population pair to compute the correlation. When black and blue bars overlap, the bar is black. Stars indicate graphs counting several components. The dashed line indicates the maximum r value obtained for each configuration.

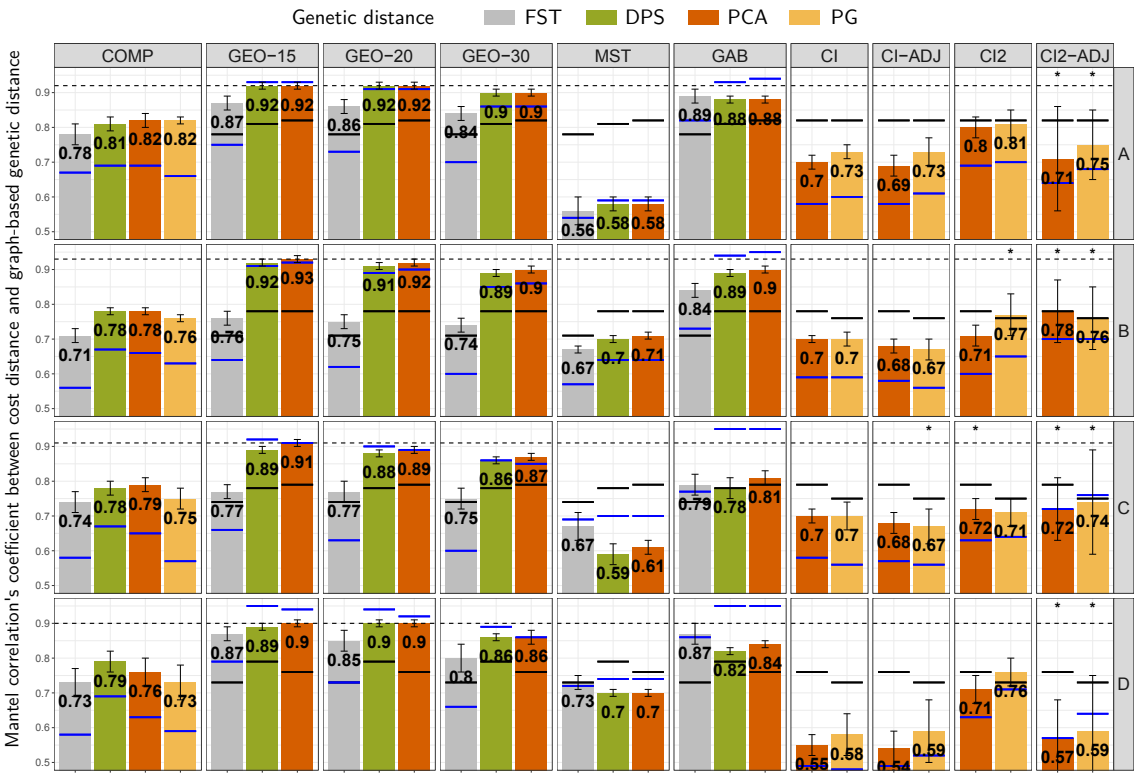

Figure S5: Mantel correlation between graph-based genetic distances and cost-distances separating nodes on the genetic graphs, according to the type of genetic distance and the pruning method at generation 500 (see table 1 for the graph names). Mean $\pm SD$ values were computed for the 10 runs simulated in each scenario. Blue bars refer to the correlation coefficient between genetic distance and geographical distance, when it is above 0.5. Black bars refer to the correlation coefficient obtained using every population pair to compute the correlation. When black and blue bars overlap, the bar is black. Stars indicate graphs counting regularly several components. The dashed line indicates the maximum r value obtained for each configuration.

2 Glossary of acronyms

ADJ Designates a genetic graph pruned using the conditional independence principle and p-value Adjustment

CD Cost Distance

cGD Conditional Genetic Distance

CI Designates a genetic graph pruned using the Conditional Independence principle and computing the covariance using squared genetic distances between populations

CI2 Designates a genetic graph pruned using the Conditional Independence principle and computing the covariance using squared genetic distances between populations

COMP Designates a Complete genetic graph (not pruned)

DMC Distance of Maximum Correlation. Landscape distance threshold below which the subset of population pairs maximize the linear correlation between genetic and landscape distances

G50 (G500) Generation 50 of the simulation (Generation 500)

GAB Designates a genetic graph with the topology of a Gabriel graph

GEO Designates a genetic graph pruned using a geographical distance threshold

IBD Isolation By Distance

IBLR Isolation By Landscape Resistance

MST Designates a genetic graph with the topology of a Minimum Spanning Tree

PCA Principal Component Analysis. Also designates a genetic graph whose links are weighted using a Euclidean genetic distance computed after the Principal Component Analysis of the population allelic frequencies.

PG Designates a genetic graph whose links are weighted using a Euclidean genetic distance computed using the same formula as that used in `popgraph` package.

3 Mathematical background : independence graphs

Two events or independent variables are conditionally independent if they are statistically independent after accounting for a third event or variable (Magwene, 2001). An independence graph is a graph that summarizes conditional independence relationships between a set of variables (Magwene, 2001). Genetic independence graphs were first used by Dyer and Nason (2004) in population genetics. In this case, the "variables" are populations and the series of values of each "variable" are allelic frequencies. Creating a genetic independence graph is tantamount to identifying pairs of populations that can be considered independent once all relationships with other populations have been taken into account.

Let $\mathbf{Y}$ be a set of p variables following a normal multivariate distribution: $\mathbf{Y} = \{y_1, y_2, \ldots, y_p\}$. The three following assumptions are equivalent (Krzanowski and Marriott, 1995, in Magwene, 2001):

- y_1 and y_2 variables are independent, conditionally to $\mathbf{Y}_K$, with $\mathbf{Y}_K$ every subset of $\mathbf{Y}$ excluding y_1 and y_2.
- Partial correlation between y_1 and y_2 is null : $\rho_{ij.\{K\}} = 0$
- If $\mathbf{C}$ is the covariance matrix of the set of variables $\mathbf{Y}$, then the element π_{ij} of the inverse covariance matrix $\mathbf{\Pi} = \mathbf{C}^{-1}$ (precision matrix), is null.

Therefore, to assess conditional independence between a set of populations, partial correlation matrix or precision matrix have first to be calculated from genetic data. In population genetics, the multilocus genotypes of individuals from populations are frequently coded as a matrix with alleles as columns and individuals as rows. The absence of an allele is coded as a 0. The presence of 1 or 2 copies of an allele in the genotype of an individual are coded respectively as a 0.5 or a 1. If these data are coded with 0, 1 and 2 values, as in Fortuna et al. (2009) and Smouse and Peakall (1999), it does not affect the calculation.

First, mean allelic frequencies in each population are computed. These frequencies are elements of a matrix $\mathbf{F}$ counting as many columns as alleles and as many rows as populations. The allele frequencies are the series of values characterizing each population, considered as variables in the construction of the genetic independence graph. The next step consists in computing the covariance between populations (betweens rows of $\mathbf{F}$). Dyer and Nason (2004) calculates this covariance by first calculating a matrix of Euclidean genetic distance between populations and then following Gower (1966), who demonstrated the duality between distance and covariance.

To that purpose, the matrix $\mathbf{F}$ of mean allelic frequencies by population has to be centered both by rows and by columns for the calculation of covariance from genetic distance in subsequent steps to be correct. However, in this particular case, this step is not mandatory given 1) the row sums are all equal to the number of loci because the allelic frequencies sum to 1 for each locus, and 2) the centering by columns does not affect the Euclidean distance between populations (rows). Without the double-centering, the between populations covariance matrix calculated from the genetic distances is however equivalent to the matrix of covariance between the columns of the transpose $\mathbf{X}$ of the double-centered matrix $\mathbf{F}$ of allelic frequencies. We demonstrate why thereafter. We also demonstrate why the covariance has to be calculated from the squared distances and not from distances, from a strict mathematical point of view, following Everitt and Hothorn (2011)(page 107), Gower (1966) and Smouse and Peakall (1999)(equation 13).

The Euclidean genetic distance d_{ij} between populations i and j is calculated from the transpose matrix $\mathbf{X}$ of the matrix $\mathbf{F}$ of allelic frequencies. $\mathbf{X}$ is of dimension $n \times p$, with n the number of alleles and p the number of populations. The genetic distance is computed with the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ki} - x_{kj})^2} \tag{1}$$

The sample covariance c_{ij} between variables/populations i and j is:

$$c_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}) \tag{2}$$

As $\mathbf{F}$ has been centred both by rows and by columns, $\bar{x_i} = \bar{x_j} = 0$. Then, the sample

covariance between variables/populations i and j is simply:

$$c_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} \tag{3}$$

And consequently:

$$\begin{aligned} c_{ii} &= \frac{1}{n}\sum_{k=1}^{n} x_{ki}^2 \\ c_{jj} &= \frac{1}{n}\sum_{k=1}^{n} x_{kj}^2 \end{aligned} \tag{4}$$

Then, the covariance matrix $\mathbf{C}$ is:

$$\mathbf{C} = \frac{1}{n}\mathbf{X^T X} \tag{5}$$

such that $\mathbf{X^T}$ is of size $p \times n$, $\mathbf{X}$ of size $n \times p$ and $\mathbf{C}$ of size $p \times p$.

The sum of the elements of each row of $\mathbf{C}$ is:

$$\begin{aligned} \sum_{j=1}^{p} c_{ij} &= \sum_{j=1}^{p} \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} \\ &= \frac{1}{n}\left[(\sum_{k=1}^{n} x_{ki}x_{k1}) + (\sum_{k=1}^{n} x_{ki}x_{k2}) + \ldots + (\sum_{k=1}^{n} x_{ki}x_{kp})\right] \\ &= \frac{1}{n}\left[(x_{1i}x_{11} + x_{2i}x_{21} + \ldots + x_{ni}x_{n1}) + \ldots + (x_{1i}x_{1p} + x_{2i}x_{2p} + \ldots + x_{ni}x_{np})\right] \\ &= \frac{1}{n}\left[x_{1i} \times (\sum_{j=1}^{p} x_{1j}) + x_{2i} \times (\sum_{j=1}^{p} x_{2j}) + \ldots + x_{ni} \times (\sum_{j=1}^{p} x_{nj})\right] \\ &= \frac{1}{n}\left[x_{1i} \times 0 + x_{2i} \times 0 + \ldots + x_{ni} \times 0\right] \\ &= 0 \end{aligned} \tag{6}$$

as the row sums of $\mathbf{X}$ are null since $\mathbf{F}$ was centered by rows and by columns.

The trace T of $\mathbf{C}$ is:

$$T = \sum_{i=1}^{p} c_{ii} \tag{7}$$

Let express d_{ij}^2 in function of the elements of $\mathbf{C}$:

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{n} (x_{ki} - x_{kj})^2 \\ &= \sum_{k=1}^{n} (x_{ki}^2 - 2x_{ki}x_{kj} + x_{kj}^2) \\ &= \sum_{k=1}^{n} x_{ki}^2 + \sum_{k=1}^{n} x_{kj}^2 - 2\sum_{k=1}^{n} x_{ki}x_{kj} \\ &= n \times (c_{ii} + c_{jj} - 2c_{ij}) \end{aligned} \tag{8}$$

We then have:

$$\begin{aligned}\sum_{i=1}^{p} d_{ij}^2 &= \sum_{i=1}^{p} n \times (c_{ii} + c_{jj} - 2c_{ij}) \\ &= n \times (\sum_{i=1}^{p} c_{ii} + \sum_{i=1}^{p} c_{jj} - 2\sum_{i=1}^{p} c_{ij})\end{aligned} \tag{9}$$

As $\sum_{j=1}^{p} c_{ij} = 0$ and $\mathbf{C}$ is a symmetric matrix, $\sum_{i=1}^{p} c_{ij} = 0$. We then have:

$$\begin{aligned}\sum_{i=1}^{p} d_{ij}^2 &= n \times (T + pc_{jj} - 2 \times 0) \\ &= n \times (T + pc_{jj}) \\ \sum_{j=1}^{p} d_{ij}^2 &= n \times (T + pc_{ii})\end{aligned} \tag{10}$$

We calculate $\sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij}^2$:

$$\begin{aligned}\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 &= \sum_{i=1}^{p}\sum_{j=1}^{p} n \times (c_{ii} + c_{jj} - 2c_{ij}) \\ &= n \times (\sum_{i=1}^{p}\sum_{j=1}^{p} c_{ii} + \sum_{i=1}^{p}\sum_{j=1}^{p} c_{jj} - 2\sum_{i=1}^{p}\sum_{j=1}^{p} c_{ij}) \\ &= n \times (pT + pT - 2 \times 0) \\ &= n \times 2pT\end{aligned} \tag{11}$$

We then calculate $d_{i\bullet}^2$, $d_{\bullet j}^2$ and $d_{\bullet\bullet}^2$:

$$\begin{aligned}d_{i\bullet}^2 &= \frac{1}{p}\sum_{j=1}^{p} d_{ij}^2 \\ &= \frac{1}{p} \times n \times (T + pc_{ii}) \\ &= n \times (\frac{T}{p} + c_{ii}) \\ &= n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{ii}) \\ d_{\bullet j}^2 &= n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{jj}) \\ d_{\bullet\bullet}^2 &= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 \\ &= \frac{n}{p^2} \times 2pT \\ &= 2 \times \frac{n}{p}\sum_{i=1}^{p} c_{ii}\end{aligned} \tag{12}$$

Because of the formula used to calculate the Euclidean distance, we have :

$$
\begin{aligned}
d_{ij}^2 &= n \times (c_{ii} + c_{jj} - 2c_{ij}) \\
c_{ij} &= -\frac{1}{2}(\frac{d_{ij}^2}{n} - c_{ii} - c_{jj}) \\
&= -\frac{1}{2n}(d_{ij}^2 - n \times c_{ii} - n \times c_{jj}) \\
&= -\frac{1}{2n}(d_{ij}^2 - n \times c_{ii} - n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii} - n \times c_{jj} - n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii} + 2n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii}) \\
&= -\frac{1}{2n}\left[d_{ij}^2 - n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{ii}) - n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{jj}) + 2n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii}\right] \\
&= -\frac{1}{2n}(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2)
\end{aligned}
\tag{13}
$$

Hence, to conform with the covariance definition, c_{ij} has to be calculated from squared distances although Dyer and Nason (2004) in `popgraph` package use the following formula:

$$c_{ij} = -\frac{1}{2}(d_{ij} - d_{i\bullet} - d_{\bullet j} + d_{\bullet\bullet}) \tag{14}$$

The division by n in (13) does not have any influence on subsequent computation steps, given the covariance matrix $\mathbf{C}$ is then standardised into a correlation matrix $\mathbf{R}$. This correlation matrix is inverted into the inverse correlation matrix $\mathbf{\Omega}$, which is also standardised. The non-diagonal elements ω_{ij} of $\mathbf{\Omega}$ are multiplied by -1 to obtain the partial correlation matrix $\mathbf{P}$ such that (Magwene, 2001):

$$\rho_{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \tag{15}$$

Finally, to determine if populations i and j are independent conditionally to all other populations, we have to test if each element ρ_{ij} is significantly different from 0. To that purpose, the Edge Exclusion Deviance criterion (EED) is calculated following Whittaker (2009) as:

$$EED = -N \ln(1 - \rho_{ij}^2) \tag{16}$$

with N the total number of observations (total number of individuals, as implemented by Dyer and Nason (2004)).

We assumed that an independence genetic graph should have links between populations positively correlated if it is to represent direct gene flow between populations. Therefore, we converted negative elements of $\mathbf{P}$ into 0 before the calculation of EED, although it was not the case in the original method of Dyer and Nason (2004).

EED has an asymptotic χ^2 distribution with one degree of freedom (Whittaker, 2009). This property allows to test the significance of every EED value and thereby to test the hypothesis $H_0 : \rho_{ij} = 0$ against $H_1 : \rho_{ij} \neq 0$. When H_0 is rejected, there is a link between populations i and j in the resulting graph.

The 0.05 level is commonly used to test the significance of EED, without p-values adjustment in the original method. However, we adjusted p-values using Holm (1979) method to limit the risk of type-I error because $\frac{p(p-1)}{2}$ tests are carried out to build a graph.

4 Computation of the DPS genetic distance

The D_{PS} is a genetic distance that relies upon the dissimilarities between the allele pools of different populations. It was initially developed as an inter-individual genetic distance (Bowcock et al., 1994). An "inter-population version" exists and has been used repeatedly in landscape genetics (Murphy et al., 2015).

To compute it, we used the formula used in MSA software:

$$D_{PS} = 1 - \frac{\sum_d^D \sum_k^K min(f_{a_{kd,i}}, f_{a_{kd,j}})}{D}$$

such as a_{kd} is the allele k at locus d, $f_{a_{kd,i}}$ is the frequency of a_{kd} in population i, D is the total number of loci and K is the allele number at each locus.

This genetic distance can be computed in R with the function `mat_gen_dist()` in `graph4lg` package.

References

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *nature*, 368(6470):455–457.

Dyer, R. J. and Nason, J. D. (2004). Population graphs: the graph theoretic shape of genetic structure. *Molecular ecology*, 13(7):1713–1727.

Everitt, B. and Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.

Fortuna, M. A., Albaladejo, R. G., Fernández, L., Aparicio, A., and Bascompte, J. (2009). Networks of spatial genetic variation across species. *Proceedings of the National Academy of Sciences*, 106(45):19044–19049.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Krzanowski, W. and Marriott, F. (1995). *Multivariate Analysis vol. 2: Classification, Covariance Structures, and Repeated Measurements*. London: Arnold.

Magwene, P. M. (2001). New tools for studying integration and modularity. *Evolution*, 55(9):1734–1745.

Murphy, M., Dyer, R., and Cushman, S. A. (2015). Graph theory and network models in landscape genetics. In Balkenhol, N., Cushman, S., Storfer, A., and Waits, L., editors, *Landscape genetics: Concepts, methods, applications*, pages 165–180. John Wiley & Sons, 1 edition.

Smouse, P. E. and Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, 82(5):561–573.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.