Ecography

ECOG-00462
Balkenhol, N., Holbrook, J. D., Onorato, D., Zager, P., White, C. and Waits, L. P. 2014. A multi-method approach for analyzing hierarchical genetic structures: a case study with cougars *Puma concolor*. – Ecography 37: xxx–xxx.

Supplementary material

Appendix 1
Description of cougar sampling

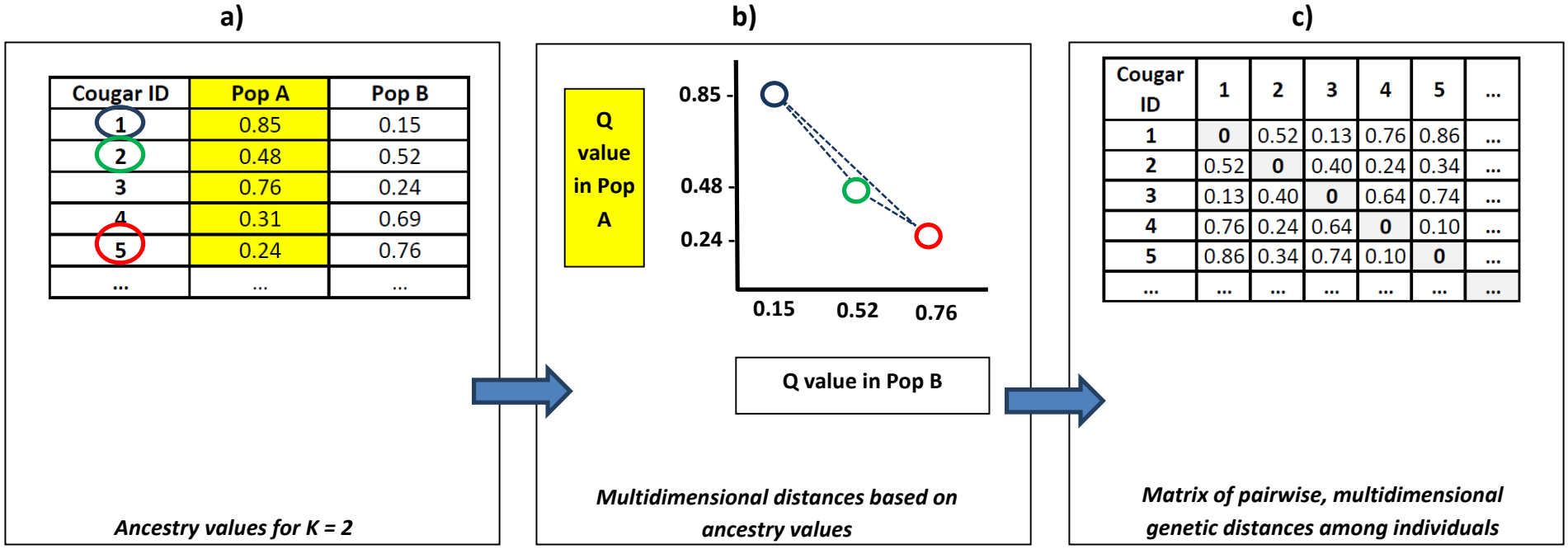
In Montana, cougars were captured during winter with the assistance of trailing hounds. We collected 106 tissue samples for genetic analyses between 1997-2005, including ear tissue (n = 91), blood (n = 13), and hair (n = 2; >10 follicles/sample). Samples were collected as part of a study of cougar mating system (Onorato et al. 2011) and we subsampled the dataset (n = 39) so that only one individual from each family group was included in the dataset. We stored tissue samples at -80° C until we performed DNA extraction.

In Idaho, hunters are required to register their legally harvested cougar with Idaho Department of Fish and Game (IDFG) biologists within 10 days of harvest. During the registration process, biologists collected a sample of tissue and recorded biological information, including harvest location, for cougars harvested during 2007-2008. Samples were placed in a sterile vial with desiccant beads, and stored until DNA extraction. When possible, exact harvest locations were verified by contacting the hunter. Otherwise, we used ARCIEW 3.2 GIS software to draw buffers around the reported locations. We drew reasonably sized buffers (i.e., 800 – 4,800 meter radius) around each general location based on our knowledge of trails, topography and accessibility for hunters. We then used GIS to randomly generate a point within the buffer and recorded the coordinate. In some instances reported general locations were major rivers that extended long distances (e.g., >16 km). In these instances, we generated a coordinate in riparian habitat at the approximate midpoint of the river within the reported Game Management Unit. We also collected tissue samples from treed cougars in Idaho by firing biopsy darts from an air pistol (Zager and White 2003). For these samples, we determined the sampling location with a handheld GPS unit.

Appendix 2

Graphical, Step-by-step illustration for calculating hierarchical genetic distances from ancestry values

Step 1: Calculating genetic distances for the first level of the hierarchy



At the first level of the hierarchy, *Structure* suggested 2 subpopulations ($K = 2$), so that each individual received two ancestry or Q-values (A). These Q-values were then used as coordinates on two axes (one for each subpopulation), to define a location for each individual in two-dimensional space (B). The Euclidean distances among these locations (shown as blue-dotted lines) are then used as a genetic distance among individuals (for illustration, only three individuals are shown). The final result is a pairwise matrix (C) showing genetic distances among all individuals; this distance captures the similarity among individuals with respect to the genetic ancestry values given by software *Structure*.

Step 2: Calculating genetic distances for the second level of the hierarchy

a)

	Pop A	
Cougar ID	Pop A1	Pop A2
1	0.47	0.53
3	0.51	0.49
...	...	...

	Pop B	
Cougar ID	Pop B1	Pop B2
2	0.71	0.29
4	0.63	0.37
5	0.24	0.78
...	...	...

1st level of hierarchy (see 1a) ***2nd level of hierarchy***

b)

Cougar ID	Pop A	Pop B	Pop A1	Pop A2	Pop B1	Pop B2
1	0.85	0.15	0.47	0.53		
2	0.48	0.52			0.71	0.29
3	0.76	0.24	0.51	0.49		
4	0.31	0.69			0.63	0.37
5	0.24	0.76			0.24	0.78
...	...	...	...	...	...	...

c)

Cougar ID	1	2	3	4	5	...
1	0	0.90	0.17	1.32	1.49	...
2	0.90	0	0.69	0.33	0.93	...
3	0.17	0.69	0	1.10	1.27	...
4	1.32	0.33	1.10	0	0.70	...
5	1.49	0.93	1.27	0.70	0	...
...	...	...	...	...	...	...

Genetic distances for 1st and 2nd level combined

~

Cougar ID	1	2	3	4	5	...
1	0	0.52	0.13	0.76	0.86	...
2	0.52	0	0.40	0.24	0.34	...
3	0.13	0.40	0	0.64	0.74	...
4	0.76	0.24	0.64	0	0.10	...
5	0.86	0.34	0.74	0.10	0	...
...	...	...	...	...	...	...

Genetic distances for 1st level

=

Cougar ID	1	2	3	4	5	...
1	0	-0.06	-0.23	0.02	0.05	...
2	-0.06	0	-0.10	-0.24	0.23	...
3	-0.23	-0.10	0	-0.02	0.01	...
4	0.02	-0.24	-0.02	0	0.01	...
5	0.05	0.23	0.01	0.01	0	...
...	...	...	...	...	...	...

Residual genetic distances for 2nd level

At the second level of the hierarchy, we analyzed the two populations detected at the first level (Population A and B) separately in *Structure*. Both populations split up into two subpopulations, again leading to two *Q*-values for all individuals (a). The ancestry values for the second level were then combined into a single data table with the ancestry values of the first level (b). Based on this table, we again calculated Euclidean distances among individuals. Note that these distances can no longer be shown graphically, because they are based on coordinates in six-dimensional space (i.e., two axes for the 1st level of the hierarchy, plus four axes for the 2nd level). The resulting multivariate genetic distances include the combined information from the 1st and 2nd level of the hierarchy. We then regressed these distances against the distances obtained at the first level (see previous page) to obtain the residual genetic distances for the 2nd level of the hierarchy (c). These genetic distances describe the genetic structure at the 2nd level, while accounting for the structure already contained in the 1st level.

Step 3: Calculating genetic distances for the third level of the hierarchy

a)

	Pop B	
	Pop B1	
Cougar ID	Pop B1a	Pop B1b
2	0.46	0.54
4	0.62	0.28
...	...	...

b)

1st level of hierarchy | *2nd level of hierarchy* | *3rd level of hierarchy*

Cougar ID	Pop A	Pop B	Pop A1	Pop A2	Pop B1	Pop B2	Pop B1a	Pop B1b
1	0.85	0.15	0.47	0.53				
2	0.48	0.52			0.71	0.29	0.46	0.54
3	0.76	0.24	0.51	0.49				
4	0.31	0.69			0.63	0.37	0.62	0.28
5	0.24	0.76			0.24	0.78		
...	...	...	...	...	...	...	...	...

c)

Cougar ID	1	2	3	4	5	...
1	0	1.05	0.20	1.53	1.73	...
2	1.05	0	0.79	0.47	1.07	...
3	0.20	0.79	0	1.27	1.47	...
4	1.53	0.47	1.27	0	0.81	...
5	1.73	1.07	1.47	0.81	0	...
...	...	...	...	...	...	...

Genetic distances for 1st, 2nd, and 3rd level combined

~

Cougar ID	1	2	3	4	5	...
1	0	0.52	0.13	0.76	0.86	...
2	0.52	0	0.40	0.24	0.34	...
3	0.13	0.40	0	0.64	0.74	...
4	0.76	0.24	0.64	0	0.10	...
5	0.86	0.34	0.74	0.10	0	...
...	...	...	...	...	...	...

Genetic distances for 1st level

+

Cougar ID	1	2	3	4	5	...
1	0	0.90	0.17	1.32	1.49	...
2	0.90	0	0.69	0.33	0.93	...
3	0.17	0.69	0	1.10	1.27	...
4	1.32	0.33	1.10	0	0.70	...
5	1.49	0.93	1.27	0.70	0	...
...	...	...	...	...	...	...

Genetic distances for 2nd level

=

Cougar ID	1	2	3	4	5	...
1	0	-0.012	-0.037	0.001	0.007	...
2	-0.012	0	-0.019	0.058	0.003	...
3	-0.037	-0.019	0	-0.006	-0.0003	...
4	0.001	0.058	-0.006	0	0.004	...
5	0.007	0.003	-0.0003	0.004	0	...
...	...	...	...	...	...	...

Residual genetic distances for 3rd level

At the third level of the hierarchy, we analyzed the four populations detected at the second level (Populations A1, A2, B1 and B2) separately in *Structure*. Population B1 split up into two subpopulations, leading to two *Q*-values for individuals previously assign to population B1 (a). The ancestry values for the third level were then combined into a single data table with the ancestry values of the first and second level (from Step 1 and 2; b). Based on this table, we again calculated Euclidean distances among individuals. Note that these distances are now based on eight-dimensional space. The resulting genetic distances include the combined information from the 1st, 2nd, and 3rd level of the hierarchy. We then regressed these distances against the distances obtained at the first and second level to obtain the residual genetic distances for the 3nd level of the hierarchy (c). These genetic distances quantify genetic structure contained in the 3rd level that is not already explained by the 1st and 2nd level of the hierarchy.

Appendix 3

R-Code for calculation of hierarchical genetic distances.

General Information

You can download a zipped folder “hier.dist.zip” at http://goo.gl/rv0nCg. In this zip-file, the file "R-function hierarchical genetic distances.R" contains a small R-script for calculating hierarchical genetic distances from ancestry values through the function "hier.dist". Ancestry values can be obtained through different genetic assignment methods, and it is recommended to average ancestry values across multiple runs, for example using software CLUMPP (Jakobsson & Rosenberg 2007).

Data format

The function requires a data structure similar to the one provided in files “example1.txt” and “example2.csv”. These files hold the ancestry- or Q-values of individuals in the different genetic clusters inferred at each level of the genetic hierarchy. The first column lists the ID of the individuals, and the following columns hold the ancestry-values. For example, in file “example1.txt”, three genetic levels were detected, with clusters A & B constituting the first genetic level. These clusters were further subdivided into subclusters A1/A2, and B1/B2, respectively. Finally, subcluster B2 was further subdivided into subclusters B2a and B2b. Note that individuals not assigned to a certain sub-cluster receive an ancestry value of zero (0) in that subcluster. For example, individual 1 in file “example1.txt” was assigned to clusters B and B1, so that it gets ancestry values of zero in clusters A1/A2 and B2a/B2b.

In the file “example2.csv”, three genetic clusters were detected at the first level of the hierarchy (1, 2, 3) and only cluster 1 was further subdivided into subclusters 1a and 1b.

Using the function

The function requires two arguments: 1) the input data (‘input’), and 2) the number of clusters detected at each level of the hierarchy (‘levels’). For example, in the file “example1.txt”, we have 2, 4, and 2 clusters at the three genetic levels. Thus, levels = c(2,4,2), which means that there are

three genetic levels, with 2 clusters at the first level, 4 subclusters at the second level, and 2 subclusters at the third level. For file “example2.csv”, levels = c(2,2), because we have 2 levels with 2 clusters each.

Running the function produces an R matrix that lists the hierarchical genetic distances for each level in a separate column, without headers. The distances are derived from the symmetric (i.e., pairwise) distance matrix, but stored in a column-order lower triangular matrix (i.e., only the lower half of the full matrix is used).

Cougar ID	1	2	3	4	5	…
1	0	0.52	0.13	0.76	0.86	…
2	0.52	0	0.40	0.24	0.34	…
3	0.13	0.40	0	0.64	0.74	…
4	0.76	0.24	0.64	0	0.10	…
5	0.86	0.34	0.74	0.10	0	…
…	…	…	…	…	…	…

Full, symmetric matrix of genetic distances

0.52
0.13
0.76
0.86
0.40
0.24
0.34
0.64
0.74
0.10

Column-order lower triangular matrix (= output format)

This object is also automatically saved in .txt and .csv format in the working directory, with the name “hierarchical-genetic-distances”.

An example

First, set the working directory. You can either do this through the GUI, or through the command “setwd()”. For example, if your data is stored on the C drive in folder “WORK”, you can use:

```
setwd('C:\WORK')
```

To use the function, you have to load the data into R, for example by reading in a .txt or .csv file with a data format as described above. For file “example1.txt”, the command would be:

```
q.values<-read.table('example1.txt', header=TRUE)
```

Similarly, for file “example2.csv”, the command would be:

```
ancestry.values<-read.csv('example2.csv')
```

Next, you have to load the function ‘hier.dist’ provided in the file “R-function hierarchical genetic distances.R” into R. There are several ways to do this, for example, you can simply open the file in a text-editor and just copy and paste the function into R. Or, you can go to “File”, “Read in R Code…”. Once the function is loaded, you can use it with the two example files, e.g.,

```
example.1<-hier.dist(q.values, levels=c(2,4,2))
```

In addition to the R object “example.1”, this creates two output files in the working directory: “hierarchical-genetic-distances.csv” and “hierarchical-genetic-distances.txt”. These files are identical, except that one is comma-separated (.csv) and one is tab-delimited (.txt). Note that the files have three columns, because there are three hierarchical levels in the input file.

If you run the function with another data set, e.g.,

```
example.2<-hier.dist(ancestry.values, levels=c(2,2))
```

In addition to creating an object “example.2”, this will overwrite the two output files in the working directory: “hierarchical-genetic-distances.csv” and “hierarchical-genetic-distances.txt”, which now only have two columns for the two levels of the hierarchy.

If you encounter problems or have questions, please email Niko Balkenhol: nbalken@gwdg.de

Appendix 4
Simulations

To assess the utility of calculating genetic distances from ancestry values, we used Easypop (Balloux 2002) to simulate a scenario showing hierarchical genetic structures. The scenario consisted of three populations (Populations **1**, **2**, **3** in Fig. S1), each of which was further subdivided into two subpopulations (**a** and **b**). The hierarchical genetic structure was caused by two types of landscape filters to gene flow: The first level of genetic structure was caused by a relatively strong filter allowing a migration rate of 0.05 for males, and 0.01 for females. This filter is shown through dark grey lines with white dashes in Fig. A1. The second level of the hierarchy was caused by a weaker filter (shown as light grey lines in Fig. A1) that allowed a migration rate of 0.09 for males and 0.05 for females. Such a scenario could, for example, arise when two types of roads that show high versus low traffic volumes (e.g., highways vs. country roads) affect a study organism in an area. In each subpopulation, we simulated 30 males and 30 females, leading to an overall population size of 360. We ran simulations for 50 generations and sampled individuals at 15 microsatellite loci. The simulations lead to a population showing low overall genetic structure (i.e., F_{ST} = 0.028). We then analyzed the genetic data using software Structure (Pritchard et al. 2000) and the Evanno method (Evanno et al. 2005). As expected, the Evanno method first detected the clusters **1**, **2**, and **3**. When analyzing these three clusters separately in Structure, each was further divided into subpopulations **a** and **b**. We then used the ancestry values for individuals to calculate hierarchical genetic distances at the two hierarchical levels (see main manuscript), and tested for significant influences of the two types of roads using multiple regression on distance matrices (MRDM). We also calculated the same MRDM models using a standard, non-hierarchical genetic distance (proportion of shared alleles, PSA). Results

are shown in Table A1. Using PSA, both types of roads have a significant effect on genetic structures. However, the R^2 of the model is low ($R^2 = 0.037$), and the coefficients of the two variables are quite similar (0.082 for highways vs. 0.122 for country roads) and do not accurately reflect the relative strength of two types of roads (i.e., higher coefficient for the weaker barrier). Thus, the information content of this analysis is very limited and could even lead to incorrect inferences. In contrast, the hierarchical genetic distances are able to clearly separate the relative influence of the two types of roads on the two levels of the hierarchy. For the first hierarchical level only highways are significant (coefficient = 0.718, $p > 0.001$), and the model R^2 of 0.416 is more than a magnitude higher than for PSA indicating better model fit. Similarly, only the variable 'country roads' is significant at the second level of the hierarchy (coefficient = 0.414, $p > 0.001$; model $R^2 = 0.150$). Overall, our simulations indicate that the hierarchical genetic distances based on ancestry values lead to much more meaningful results than standard PSA.

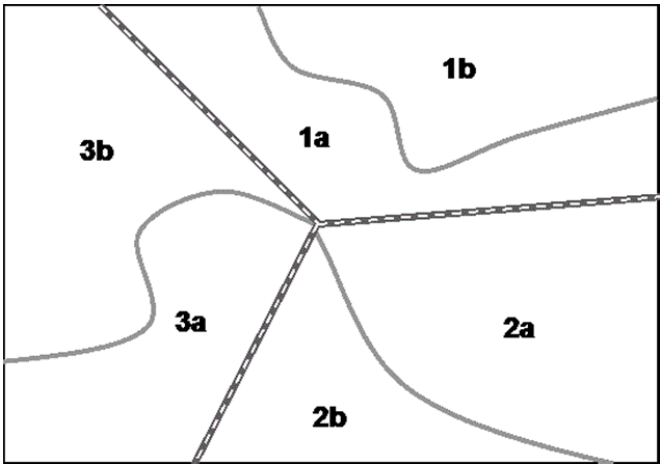
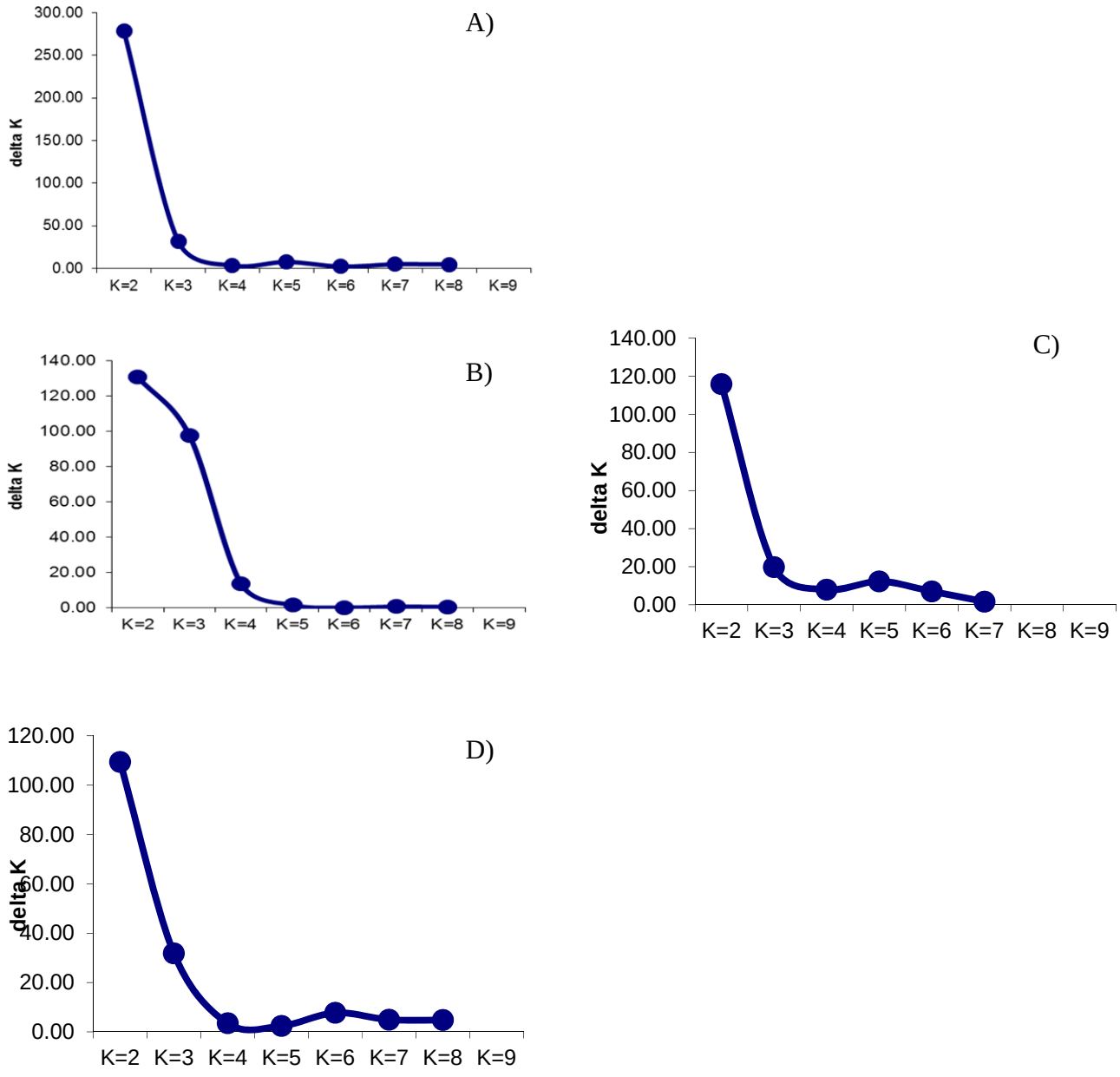


Fig. A1: Simulated scenario showing hierarchical genetic structures

Table A1: Results of multiple regression based on distance matrices for simulated scenario.

Genetic Distance	Model R^2 (p-value)	Variable	Coefficient	p-value
PSA	0.037 (p>0.001)	**highways**	**0.082**	**<0.001**
		country roads	**0.122**	**<0.001**
HGD level 1	0.416 (p<0.001)	**highways**	**0.718**	**<0.001**
		country roads	-0.026	0.100
HGD level 2	0.150 (p>0.001)	Highways	-0.021	0.509
		country roads	**0.414**	**<0.001**

Fig. A2: Plots of delta K statistics for clustering results obtained from program Structure for the cougar genetic data. A) complete data set; B) cluster A; C) cluster B; and D) cluster B1. See also Fig. 3 in main manuscript.



Appendix 5
Assessing the utility of hierarchical genetic distances for understanding landscape influences on overall genetic structure

We wanted to ensure that using the novel hierarchical genetic distances used in our analyses (see main text) led to results that were meaningful for understanding landscape effects on the overall gradient structure of the population. We used the proportion of shared alleles (PSA) as our standard, non-hierarchical genetic distance. We calculated a cumulative effective distance that incorporated the information on landscape influences on genetic structure across hierarchical levels. Specifically, we calculated the cumulative effective distance (CED) as:

$$CED = \sum_{\substack{i=1 \\ j=1}}^{\substack{n \\ k}} w_i * rc_{j,i} * Var_j$$

Where w_i is a weight that that equals 0 if two individual cougars are in the same genetic cluster at hierarchical level *i*, and 1 otherwise, $rc_{j\ i}$ is the regression coefficient for variable $_j$ (Var_j) at level $_i$ of the hierarchy. Only significant variables are considered, and summation is across all variables (1 to n) and all hierarchical levels (1 to k). Thus, for the cougar data set with three hierarchical levels, the formula becomes

$$CED = w_1 * rc_{1(SRP)} * SRP + w_1 * rc_{1(GEO)} * GEO + w_2 * rc_{2(Forest)} * FOR + w_2 * rc_{2(GEO)} * GEO + w_3 * rc_{3(GEO)} * GEO$$

Using coefficients estimated for all samples (see Table 1 in main manuscript), we get:

$$CED = w_1 * 0.629 * SRP + w_1 * 0.040 * GEO + w_2 * 0.108 * FOR + w_2 * 0.142 * GEO + w_3 * 0.170 * GEO$$

The cumulative effective distance weighs the simple effective distances by their relative importance for explaining a certain hierarchical level of genetic structure. To assess whether CED explained overall gradient genetic structures, we correlated this distance with PSA using simple and partial Mantel tests. Mantel tests were conducted in R 2.12.1 using the package ecodist (Goslee and Urban 2007) and 9,999 permutations to assess significance.

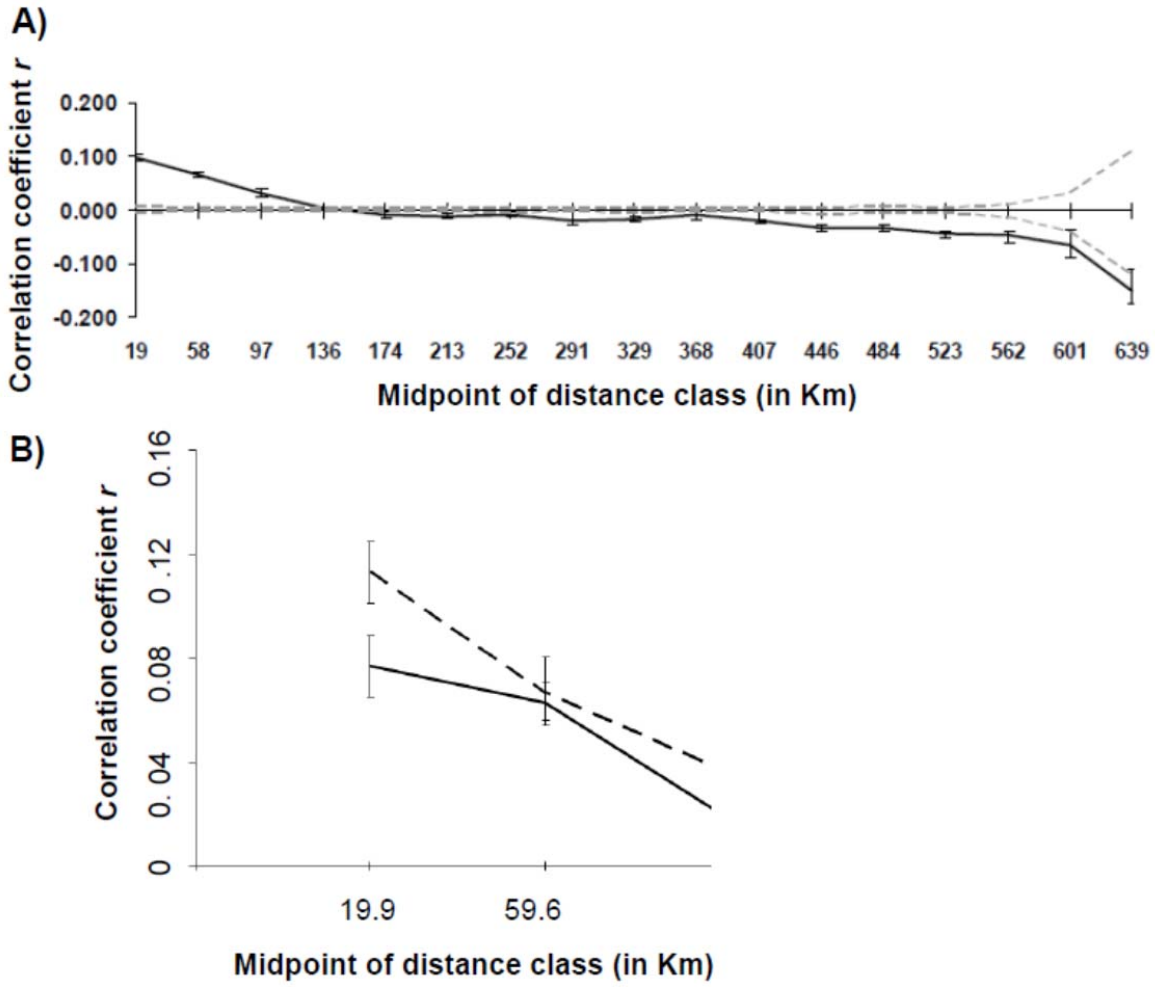
The cumulative effective distance based on the hierarchical analysis led to a highly significant correlation coefficient of 0.426 ($p = 0.0001$) with the PSA genetic distance. When partialling out the effect of geographic distance, the partial correlation coefficient was 0.339, which was still highly significant ($p = 0.0001$). Results were consistent when analysing males and females separately, but correlations were higher for males than for females. Correlation and partial correlation coefficients were 0.459 and 0.0.381 for males, and 0.396 and 0.294 for females, respectively ($p < 0.001$ for all tests).

Appendix 6
Sex-specific spatial-autocorrelation analysis

We compared the pattern and strength of fine-scale spatial genetic structure of males versus females within each distance class using the method of Smouse et al. (2008) as implemented in software GenALEx 6.5 (Peakall and Smouse 2012). We used Sturge's rule (Sturges 1926) to determine the optimal number of spatial lags and significance of spatial autocorrelation was assessed via confidence intervals obtained from 999 permutations.

The autocorrelation analysis suggested that significant positive autocorrelation existed up to a distance of ~136 km when using all samples, and that the spatial autocorrelation in the first distance class (up to ~40 km) was significantly higher in females than in males (Fig. A2).

Fig. A3: Spatial genetic correlograms for A**)** all samples combined and **B)** close-up of females (dashed line) versus males (solid line) in the first two distance classes. Error bars show 95% confidence intervals based on 1,000 bootstraps.



References

Balloux, F. 2001. EASYPOP (version 1.7):A computer program for the simulation of population genetics. — J. Heredity 92: 301–302.

Evanno, G. et al. 2005. Detecting the number of clusters of inidividuals using the software Structure: a simulation study. — Mol. Ecol. 14: 2611–2620.

Goslee, S. C. and Urban, D. L. 2007. The ecodist package for dissimilarity-based analysis of ecological data. — J. Stat. Softw. 22: 1–19.

Jakobsson, M. and Rosenberg, M.A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure — Bioinformatics 23: 1801-1806.

Onorato, D. et al. 2011. Genetic assessment of paternity and relatedness in a managed population of cougars. — J. Wildl. Manage. 75: 378–384.

Peakall, R. and Smouse P.E. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. — Bioinf. 28: 2537–2539.

Pritchard, J. K. et al. 2000. Inference of population structure using multilocus genotype data. — Genetics 155: 945–959.

Smouse P.E. et al. 2008. A heterogeneity test for fine-scale genetic structure. — Mol. Ecol. 17: 3389-3400.

Sturges H. 1926. The choice of a class-interval — J. Am. Stat. Assoc. 21: 65-66.

Zager, P. and White, C. 2003. Study IV. Factors influencing elk calf recruitment. Job numbers 2-3. Calf mortality causes and rates. Predation effects on elk calf recruitment. —Idaho

Department of Fish and Game Federal Aid in Wildlife Restoration Job Progress Report, Boise, ID, USA.