MOLECULAR ECOLOGY

Supplemental Information for:

Community metabarcoding reveals the relative role of environmental filtering and dispersal in metacommunity dynamics of soil microarthropods across a mosaic of montane forests

Víctor Noguerales, Emmanouil Meramveliotakis, Adrián Castro-Insua,
Carmelo Andújar, Paula Arribas, Thomas J. Creedy, Isaac Overcast,
Hélène Morlon, Brent C. Emerson, Alfried P. Vogler and Anna Papadopoulou

Table of Contents:

TABLE S1. Geographical location and elevation of the 44 sampling points across the main five forest habitat types of the Troodos mountain range (Cyprus). Geographic distances between sampling points ranged from 45 m to 65 km. Two different samples were collected from each locality corresponding to the leaf litter and the deep soil layers, for a total of 88 samples. In order to avoid biases due to recent natural or human-driven disturbances, we selected pure forest sites that appeared intact (within a 2 km radius) based on historical orthophotos of the past 60 years, available from the Department of Land and Surveys, Ministry of Interior, Republic of Cyprus (*https://eservices.dls.moi.gov.cy/*).

Forest habitat	Habitat code	Locality code	Latitude	Longitude	Elevation (m)
Pinus brutia	*Pb*	Pb01NWE	35.05890	32.62383	960
Pinus brutia	*Pb*	Pb02PSI	35.01363	32.65182	1100
Pinus brutia	*Pb*	Pb03TRI	35.01081	32.67200	1182
Pinus brutia	*Pb*	Pb04OIK	35.03740	32.83374	451
Pinus brutia	*Pb*	Pb05THE	35.05109	32.93183	472
Pinus brutia	*Pb*	Pb06SAI	34.88597	32.91084	857
Pinus brutia	*Pb*	Pb07LAG	34.98339	33.03513	814
Pinus brutia	*Pb*	Pb08MIT	35.02863	33.10117	550
Pinus brutia	*Pb*	Pb09KAP	34.95303	33.22019	693
Pinus brutia	*Pb*	Pb10LIT	34.93824	33.32448	426
Quercus alnifolia	*Qa*	Qa01NWE	35.05032	32.62390	964
Quercus alnifolia	*Qa*	Qa01TEST	34.95216	32.74701	890
Quercus alnifolia	*Qa*	Qa02PSI	35.02238	32.64947	1050
Quercus alnifolia	*Qa*	Qa03TRI	34.99646	32.67845	1393
Quercus alnifolia	*Qa*	Qa04STR	34.98886	32.68803	1137
Quercus alnifolia	*Qa*	Qa05MIL	34.92007	32.74375	945
Quercus alnifolia	*Qa*	Qa06MAD	34.95435	32.99693	1572
Quercus alnifolia	*Qa*	Qa07ALO	34.92540	33.04498	1279
Quercus alnifolia	*Qa*	Qa08PAL	34.89243	33.07495	1351
Quercus alnifolia	*Qa*	Qa09KIO	34.92168	33.19400	1347
Quercus alnifolia	*Qa*	Qa10VAV	34.90582	33.21685	1147
Cedrus brevifolia	*Cb*	Cb01PSI	35.02256	32.64991	1027
Cedrus brevifolia	*Cb*	Cb02TRI	35.00095	32.67910	1341
Cedrus brevifolia	*Cb*	Cb03TRI	35.00021	32.68005	1359
Cedrus brevifolia	*Cb*	Cb04TRI	34.99837	32.67926	1366
Cedrus brevifolia	*Cb*	Cb05TRI	34.99722	32.67742	1372
Cedrus brevifolia	*Cb*	Cb06TRI	34.99460	32.68099	1400
Cedrus brevifolia	*Cb*	Cb07STR	35.02256	32.64991	1124
Pinus nigra	*Pn*	Pn01PRO	34.95261	32.83689	1440
Pinus nigra	*Pn*	Pn02CHI	34.95487	32.85660	1696
Pinus nigra	*Pn*	Pn03CHI	34.94499	32.86504	1818
Pinus nigra	*Pn*	Pn04CHI	34.93476	32.86170	1905
Pinus nigra	*Pn*	Pn05CHI	34.93015	32.87531	1766
Pinus nigra	*Pn*	Pn06CHI	34.93917	32.89204	1712

Pinus nigra	*Pn*	Pn07TRO	34.91629	32.88539	1727
Pinus nigra	*Pn*	Pn08AMI	34.92676	32.89821	1600
Juniperus foetidissima	*Jn*	Jn01CHI	34.95515	32.85732	1671
Juniperus foetidissima	*Jn*	Jn02CHI	34.94367	32.86582	1798
Juniperus foetidissima	*Jn*	Jn03CHI	34.93476	32.86017	1896
Juniperus foetidissima	*Jn*	Jn04CHI	34.93439	32.87003	1833
Juniperus foetidissima	*Jn*	Jn05TRO	34.91908	32.86950	1716
Juniperus foetidissima	*Jn*	Jn06CHI	34.93451	32.88286	1619
Juniperus foetidissima	*Jn*	Jn07AMI	34.92804	32.89826	1564
Juniperus foetidissima	*Jn*	Jn08CHI	34.93525	32.86301	1921

TABLE S2. Details on the software, arguments and parameter values used in each step of the read processing and filtering pipeline.

Step	Software and version	Arguments and parameter values	Task
Primer removal	CUTADAPT v.2.10	*min_overlap=20 min_overlap=26* *--discard_untrimmed --error-rate=0*	Filter out those reads that do not contain an adapter and/or exhibit variation in primer length and primer composition in either forward or reverse raw sequences
Merging	PEAR v.0.9.11	*-q 26 -v 100*	Merge forward and reverse reads by setting the minimum overlap size to 100 bp and trim the low-quality part of a read when two consecutive bases have a quality score below 26
Quality filtering	VSEARCH v.2.9.1	*--fastx_filter --fastq_maxee 1*	Discard reads with more than 1 expected error
Dereplication I	VSEARCH v.2.9.1	*--derep_fulllength –sizeout --minuniquesize 2*	Pool strictly identical reads and calculate abundance value per library. Discard reads with a post-dereplicated abundance value smaller than 2
Length filtering	VSEARCH v.2.9.1	*--fastx_filter --fastq_minlen 416* *--fastq_maxlen 420*	Retain only reads with length between 416 bp and 420 bp
Denoising	VSEARCH v.2.9.1	*--cluster_unoise --minsize 2 --unoise_alpha 2*	Denoise reads using the UNOISE3 algorithm with a value of 2 for the α parameter (default) and discard reads with an abundance value smaller than 2
Chimera filtering	VSEARCH v.2.9.1	*--uchime3_denovo --abskew 16*	Remove chimeras using an abundance skew value of 16 (default)
Dereplication II	VSEARCH v.2.9.1	*--derep_fulllength --sizeout –sizein --minuniquesize 1*	Pool strictly identical reads and calculate abundance value across all libraries using previous library-based abundance information
BLAST search	NCBI-BLAST v.2.8.1	*-outfmt 5 -evalue 0.001 -max_target_seqs 100*	BLAST search against a database using an e-value of 0.001 and output the first 100 matches per query. Format result output file to XML format suitable for MEGAN software
Community table generation	VSEARCH v.2.9.1	*--search_exact --otutabout*	Search for 100% exact and full-length matches of the query sequences in the database of target sequences. Summarise sequence abundance per site in a species x site formatted table
Read-abundance filtering	METAMATE v.0.1b18	*find.py --refmatchpercent 100 --refmatchlength 350* *--dbmatchpercent 100 --dbmatchlength 400* *--expectedlength 418 --percentvariation 0 -s 5*	Remove putative NUMTs and other types of low-frequency erroneous sequences using stringent values to assign sequences as verified authentic and verified non-authentic haplotypes

TABLE S3. PCA loadings of environmental variables for the two first principal components (PC). Variables with a loading value >0.4 for a given PC are highlighted in bold. The fraction of variance explained by each PC is reported at the bottom of the table.

Variables	Code	Data source[1][2][3]	PC1	PC2
Elevation	ALT	NASA STRM	**0.981**	0.020
Annual mean temperature	BIO1	WorldClim	**-0.989**	-0.021
Maximum temperature of warmest month	BIO5	WorldClim	**-0.883**	0.032
Minimum temperature of coldest month	BIO6	WorldClim	**-0.955**	-0.078
Annual precipitation	BIO12	WorldClim	**0.984**	-0.010
Precipitation of wettest quarter	BIO16	WorldClim	**0.968**	-0.042
Precipitation of driest quarter	BIO17	WorldClim	**0.952**	0.079
Climatic moisture index	CMI	ENVIREM	**0.990**	0.025
Thornthwaite aridity index	TAI	ENVIREM	**-0.649**	**-0.430**
Topographic wetness index	TWE	ENVIREM	**-0.538**	**0.772**
Explained variance (%)			81.312	7.965

[1] NASA STRM: *https://srtm.csi.cgiar.org*

[2] WorldClim: *https://www.worldclim.org/data/worldclim21.html*

[3] ENVIREM: *https://envirem.github.io*

TABLE S4. Results of model selection testing for the relationship between average richness (α diversity, RICH) or community uniqueness (local contribution to β diversity - LCBD) and topoclimatic variation (ENV_{PC1} and ENV_{PC2}; see Table S3). Latitude (Lat) and longitude (Lon) were included as covariates in the models. Response variables were calculated at ASV and OTU level. Generalized linear mixed models (GLMMs) were used to fit forest habitat type as random effect in order to account for non-independence among samples from the same forest habitat. Generalized linear models (GLMs) were applied when random effect variance was zero or nearly zero indicating a poor model fit and no influence of random effect on response variation. For each model we indicate K, number of parameters in the model; AIC_C, sample-size adjusted Akaike's information criterion (AIC) value; ΔAIC_C, difference in AIC_C value from that of the most supported model; ω_i, AIC_C weight. Only best ranked equivalent models ($\Delta AIC_C \leq 2$) are shown.

ASVs (haplotypes)						OTUs (3% lineages)					
Model	Parameters	K	AIC_C	ΔAIC_C	ω_i	Model	Parameters	K	AIC_C	ΔAIC_C	ω_i
$RICH_{ASV}$						$RICH_{OTU}$					
1	Lon	1	320.40	0.00	0.48	1	ENV_{PC2} + Lon	2	267.20	0.00	0.67
2	Lon + ENV_{PC2}	2	321.10	0.70	0.34	2	ENV_{PC2}	1	268.57	1.37	0.33
3	Lon + Lat	2	322.32	1.92	0.18						
$LCBD_{ASV}$						$LCBD_{OTU}$					
1	ENV_{PC1} + ENV_{PC2}	2	-493.40	0.00	1.00	1	ENV_{PC1}	1	-435.40	0.00	1.00

TABLE S5. Results of PCNM axes selection using dbRDA and a forward selection procedure as implemented in the 'ordistep' function of the *vegan* R package (Oksanen et al., 2020). Significance of PCNM axes was evaluated against community dissimilarity matrices as estimated using the Simpson dissimilarity index (β_{SIM}) at ASV and OTU levels. The adjusted coefficient of determination (R^2_{ADJ}) for each model is also provided. The *Pinus nigra* (*Pn*) and *Juniperus foetidissima* (*Jn*) sampling sites were analysed both separately and jointly, according to the results of NMDS-based ordinations (Figure 3). Predictor information in those models containing no significant variables (null) is replaced by dashes.

Model	ASVs (haplotypes)				OTUs (3% lineages)			
	Predictors	F	p-value	R^2_{ADJ}	Predictors	F	p-value	R^2_{ADJ}
Across habitats	SPA_{PCNM1}	5.332	<0.001	0.155	SPA_{PCNM1}	4.586	<0.001	0.148
	SPA_{PCNM2}	2.392	<0.001		SPA_{PCNM2}	2.369	<0.001	
	SPA_{PCNM5}	1.629	0.007		SPA_{PCNM4}	1.456	0.027	
	SPA_{PCNM10}	1.367	0.032		SPA_{PCNM5}	1.714	0.002	
	SPA_{PCNM17}	1.637	0.007		SPA_{PCNM15}	1.334	0.042	
					SPA_{PCNM17}	1.426	0.029	
Pinus brutia (*Pb*)	Null	-	-	-	Null	-	-	-
Quercus alnifolia (*Qa*)	SPA_{PCNM1}	1.543	0.006	0.051	SPA_{PCNM1}	2.892	0.005	0.231
					SPA_{PCNM2}	1.850	0.007	
Cedrus brevifolia (*Cb*)	SPA_{PCNM2}	2.603	0.045	0.210	SPA_{PCNM1}	2.387	0.024	0.188
Pinus nigra (*Pn*)	Null	-	-	-	Null	-	-	-
Juniperus foetidissima (*Jn*)	Null	-	-	-	Null	-	-	-
Pinus nigra (*Pn*) + *Juniperus foetidissima* (*Jn*)	Null	-	-	-	SPA_{PCNM3}	1.755	0.008	0.048

TABLE S6. Results from multivariate generalized linear models (mvGLMs) testing for the relationship between community composition as response variable and forest habitat type (HAB), spatial (SPA_{PCNMi}) and topoclimatic (ENV_{PCi}) variables as predictors. Analyses were performed at ASV and OTU levels, either with all sites (across habitats) or separately for each of the forest habitat types. Variable significance was first assessed by single-term models and those predictors with a significant effect were used to build a full model following a backward stepwise selection with all terms in the final model being statistically significant. Results of all significant single-term models are shown and variables included in the final model are highlighted in bold. We provide the correlation coefficient (Hopper´s R^2) of each significant variable as calculated using the 'best.r.sq' function from the *mvabund* R package (Wang, Naumann, Wright, & Warton, 2012). The *Pinus nigra* (*Pn*) and *Juniperus foetidissima* (*Jn*) sampling sites were analysed both separately and jointly, according to the results of NMDS-based ordinations (Figure 3).

Model	ASVs (haplotypes)				OTUs (3% lineages)			
	Predictors	Deviance	*p*-value	R^2	Predictors	Deviance	*p*-value	R^2
Across habitats	**HAB**	4220	0.001	0.121	**HAB**	2197	0.001	0.138
	SPA_{PCNM1}	2530	0.001	0.040	**SPA_{PCNM1}**	1110	0.002	0.044
	SPA_{PCNM2}	1677	0.001	0.030	**SPA_{PCNM2}**	813.7	0.002	0.037
	SPA_{PCNM3}	1080	0.011	0.025	**SPA_{PCNM3}**	501.5	0.016	0.028
	SPA_{PCNM4}	1442	0.001	0.029	**SPA_{PCNM4}**	645.3	0.002	0.032
	SPA_{PCNM5}	1084	0.034	0.023	**SPA_{PCNM5}**	503.9	0.044	0.027
	SPA_{PCNM7}	1040	0.006	0.028	**SPA_{PCNM7}**	449.5	0.024	0.028
	SPA_{PCNM10}	1067	0.019	0.024	**SPA_{PCNM10}**	455.1	0.036	0.025
	SPA_{PCNM11}	844.3	0.016	0.032	**SPA_{PCNM11}**	462.1	0.008	0.044
	SPA_{PCNM17}	1224	0.004	0.030	**SPA_{PCNM15}**	380.8	0.030	0.036
	SPA_{PCNM20}	940.9	0.015	0.029	**SPA_{PCNM17}**	478.6	0.029	0.026
	SPA_{PCNM25}	996.4	0.019	0.030	**SPA_{PCNM20}**	427.2	0.026	0.028
	ENV_{PC1}	1620	0.001	0.030	**ENV_{PC1}**	866.7	0.001	0.040
	ENV_{PC2}	1772	0.001	0.032	**ENV_{PC2}**	897.5	0.001	0.041
Pinus brutia (*Pb*)	**SPA_{PCNM1}**	872.3	0.006	0.128	**SPA_{PCNM1}**	437.5	0.015	0.131
	ENV_{PC1}	770.9	0.023	0.130	**ENV_{PC1}**	424.0	0.022	0.138

Quercus alnifolia (*Qa*)	**SPA$_{PCNM1}$**	684.1	0.032	0.107	**SPA$_{PCNM1}$**	392.4	0.010	0.136
	SPA$_{PCNM6}$	842.8	0.001	0.119	**SPA$_{PCNM6}$**	349.6	0.025	0.108
	SPA$_{PCNM7}$	609.5	0.029	0.115			
	ENV$_{PC2}$	652.0	0.039	0.113			
Cedrus brevifolia (*Cb*)	**ENV$_{PC1}$**	586.6	0.045	0.281	**ENV$_{PC1}$**	275.5	0.044	0.259
Pinus nigra (*Pn*)	Null	-	-	-	Null	-	-	-
Juniperus foetidissima (*Jn*)	Null	-	-	-	Null	-	-	-
Pinus nigra (*Pn*) + *Juniperus foetidissima* (*Jn*)	**SPA$_{PCNM2}$**	551.9	0.028	0.072	**SPA$_{PCNM3}$**	257.4	0.013	0.089
	SPA$_{PCNM3}$	490.0	0.033	0.080	**ENV$_{PC1}$**	254.4	0.019	0.086
	SPA$_{PCNM4}$	516.8	0.040	0.076			
	SPA$_{PCNM5}$	512.2	0.024	0.076			
	SPA$_{PCNM6}$	515.5	0.018	0.070			
	ENV$_{PC1}$	508.0	0.026	0.082			
	ENV$_{PC2}$	438.6	0.032	0.087			

TABLE S7. Univariate matrix regressions with randomization (MRR) testing for the relationship between dissimilarity in community composition (β_{SIM}) across *Quercus alnifolia* (*Qa*) sampling sites as response variable and forest fragmentation (FRA$_{IBR}$) as explanatory variable. Community dissimilarity was estimated using the Simpson dissimilarity index (β_{SIM}) at ASV and OTU levels. The alternative FRA$_{IBR}$ distance matrices were calculated under an isolation-by-resistance (IBR) scenario based on the spatial distribution of *Quercus alnifolia* patches and assuming a range of increasing resistance values (5 - 1,000,000) for the non-*Quercus* habitat cells in order to identify the value that best explained the observed estimates of community dissimilarity (β_{SIM}). We also provide the univariate results for the remaining distance matrices based on topoclimate (ENV$_{PC1-2}$), weighted topographic distances (SPA$_{TWD}$), topographic complexity (TRI$_{IBR}$) and a 'flat' scenario (NULL$_{IBR}$), with their respective coefficient of determination (R^2) and *p*-value.

Terms	Resistance value of non-*Quercus* habitat cells	*ASVs (haplotypes)*		*OTUs (3% lineages)*	
		p-value	R^2	*p*-value	R^2
ENV$_{PC1-2}$	-	0.002	0.224	0.005	0.178
SPA$_{TWD}$	-	0.037	0.083	0.002	0.242
TRI$_{IBR}$	-	0.027	0.110	0.006	0.229
NULL$_{IBR}$	-	0.049	0.072	0.003	0.232
FRA$_{IBR}$	5	0.021	0.108	0.002	0.265
FRA$_{IBR}$	10	0.020	0.115	0.001	0.266
FRA$_{IBR}$	50	0.010	0.124	0.002	0.264
FRA$_{IBR}$	100	0.001	0.125	0.004	0.263
FRA$_{IBR}$	250	0.009	0.124	0.002	0.261
FRA$_{IBR}$	500	0.010	0.122	0.002	0.260
FRA$_{IBR}$	1000	0.018	0.121	0.003	0.259
FRA$_{IBR}$	10000	0.022	0.119	0.002	0.257
FRA$_{IBR}$	100000	0.016	0.119	0.003	0.256
FRA$_{IBR}$	1000000	0.020	0.118	0.006	0.256

FIGURE S1. Position in the topoclimatic space (first two principal components of a PCA based on 10 environmental variables, see Table S3) of the 44 sampling sites. The inset plot shows the position of the 44 sampling points (red circles) in relation to the 500 randomly distributed points throughout Cyprus (grey circles), all used to perform PCA. For a biological interpretation of each PC, see Table S3.
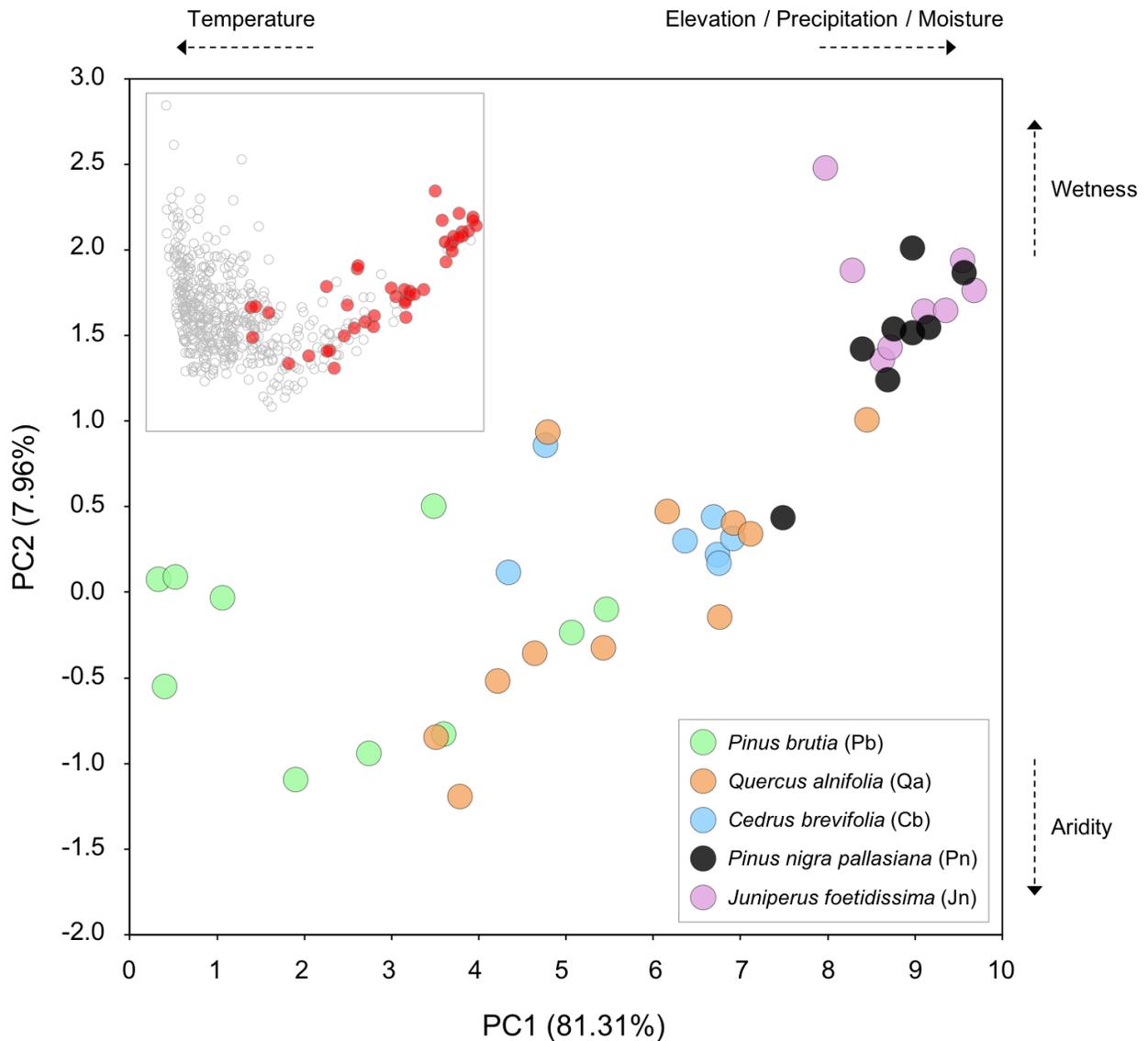
**FIGURE S2.** Average richness (α diversity, top panels) and community uniqueness (local contribution to β diversity - LCBD, bottom panels) per soil layer (deep soil and leaf litter) within each forest habitat, at ASV (left panels) and OTU (right panels) levels. ANOVA results assessing the difference in mean values between the two layers are provided on each respective panel. Inset graphs show the difference in variation of α diversity (ASV, $F_{1,86}$ = 152.4, p-value <0.001; OTU, $F_{1,86}$ = 145.6, p-value <0.001) and LCBD (ASV, $F_{1,86}$ = 13.11, p-value <0.001; OTU, $F_{1,86}$ = 16.17, p-value <0.001) between soil layers when combining all forest habitats: *Pb, Pinus brutia* (light green); *Qa, Quercus alnifolia* (orange); *Cb, Cedrus brevifolia* (blue); *Pn, Pinus nigra* (black); *Jn, Juniperus foetidissima* (purple).
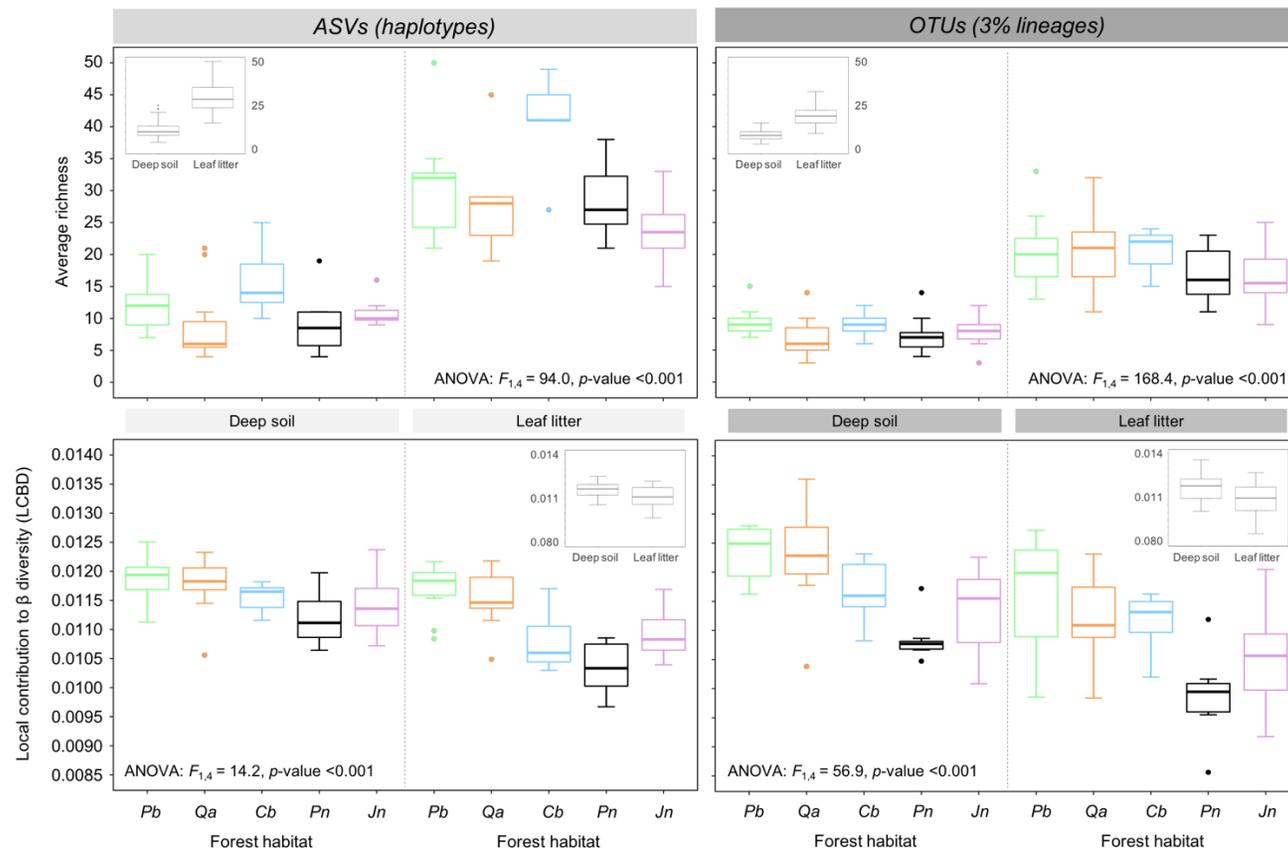
FIGURE S3. Relationships of average richness (α diversity, top panels) and community uniqueness (local contribution to β diversity - LCBD, bottom panels) with their most explicative predictors according to GLMM/GLM analysis (Table 1; Table S4), at ASV (left panels) and OTU (right panels) levels. In the main four scatter plots, circle size represents the altitude of each sampling site. Additionally, inset graphs show the relationship between the respective response variable and elevation using the LOESS function as implemented in the *ggplot2* (Wickham, 2016) R package, for displaying purposes. Regression lines and 95% confidence intervals are represented in dark and light grey, respectively. Circle colour represents the different forest habitats: *Pb*, *Pinus brutia* (light green); *Qa*, *Quercus alnifolia* (orange); *Cb*, *Cedrus brevifolia* (blue); *Pn*, *Pinus nigra* (black); *Jn*, *Juniperus foetidissima* (purple).
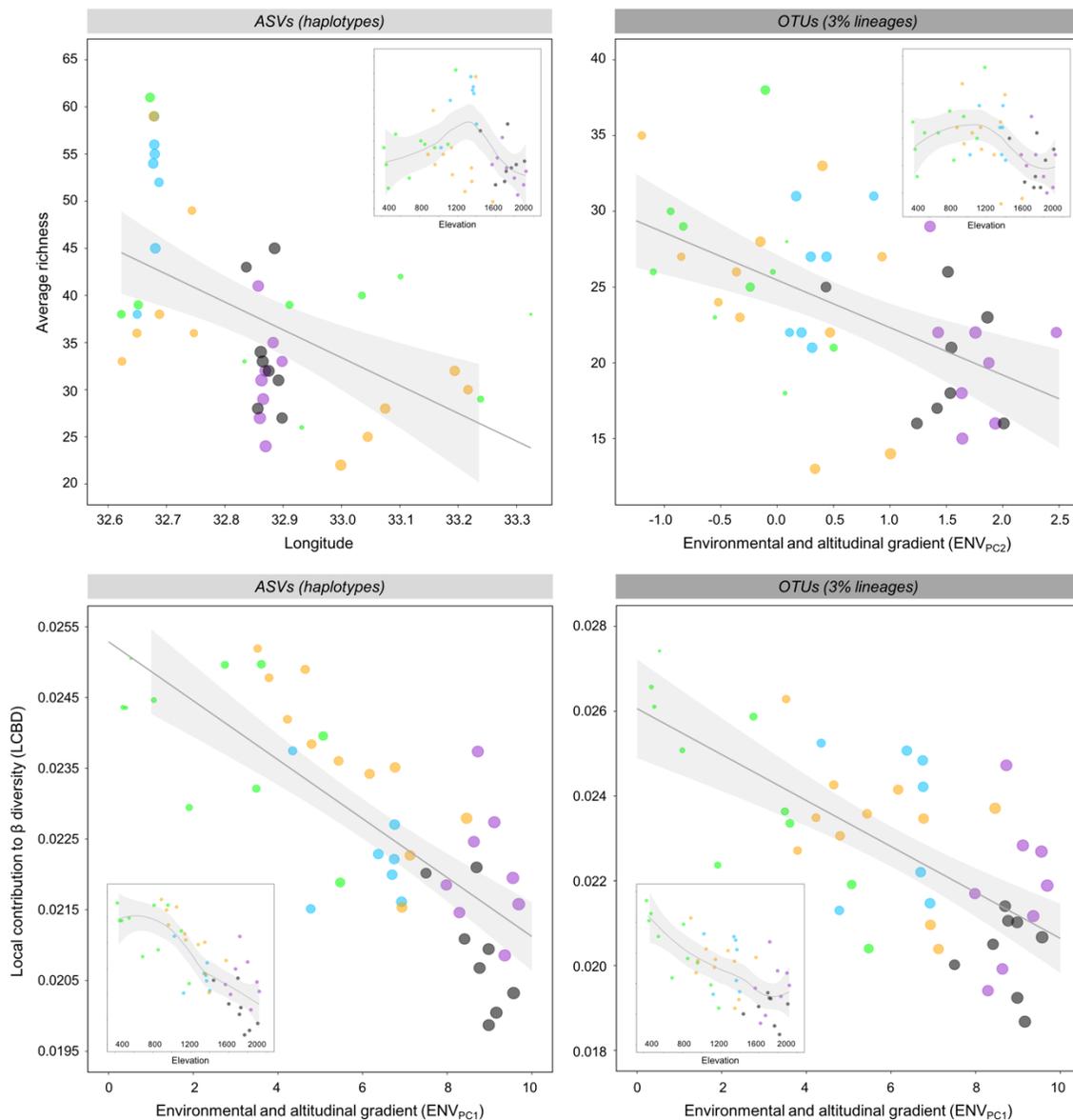
APPENDIX I - SUPPLEMENTAL METHODS

Read-abundance filtering in METAMATE

The average number of raw reads per sample after demultiplexing was 158,262 (SD = 25,865). Once raw reads were quality filtered and dereplicated in VSEARCH v.2.9.1 (Rognes, Flouri, Nichols, Quince, & Mahé, 2016) (see Table S2), we retained reads of ±2 bp with respect to amplicon length as this variation (416-420 bp) is required for the application of posterior filterings using the METAMATE v.0.1b18 software (Andújar et al., 2021). After denoising and chimera filtering (see Table S2), each sample contained on average 435.8 unique sequences (SD = 176.7) resulting in a total of 35,776 ASVs of which 28,499 were taxonomically assigned to Acari (6,388), Collembola (4,710) and Coleoptera (17,401). These ASVs were further filtered to remove putative nuclear copies of mitochondrial DNA (NUMTs or pseudogenes) and other erroneous sequences by applying the METAMATE software. This software allows the application of multiple read-abundance filtering strategies and posterior evaluation of their effects on the prevalence of known authentic mitochondrial haplotypes and presumed non-mitochondrial copies (*e.g.*, those violating the reading frame or expected length, as expected for NUMTs and erroneous sequences) in the final filtered dataset. We designated ASVs as either known mitochondrial haplotypes or presumed non-authentic sequences on the basis of a reference database composed of the NCBI *nt* collection (accessed November 2020) and a curated reference catalogue including 561 previously available sequences corresponding to soil lineages of Acari, Collembola and Coleoptera from the Iberian Peninsula (Arribas, Andújar, Salces-Castellano, Emerson, & Vogler, 2021; Arribas, Andújar, Hopkins, Shepherd, & Vogler, 2016) plus 344 Sanger sequences of the 'voucher' specimens from Cyprus (generated by this study).

We ran *find.py* Python script in METAMATE with the above-described reference database and set the reference-matching arguments as follows: *--refmatchpercent 100*, *--refmatchlength 350, --dbmatchpercent 100, --dbmatchlength 400*. The length-based arguments were set by applying the next options: *--expectedlength 418* and *--percentvariation 0.* For the translation-based filtering, *i.e.*, amino acid translation to identify erroneous sequences, we applied the invertebrate mitochondrial genetic code (translation-based argument; *--table 5*), letting the reading frame to be automatically detected. When not specified, the parameters were set as default (see *https://github.com/tjcreedy/metamate* for more details about the different options).

We assessed a total of 7,872 different filtering strategies at both across- and within-sample levels, and also within clades. ASVs of each taxonomic group (Acari, Collembola and Coleoptera) were separately analysed. Specifically, we evaluated a broad range of threshold values for a comprehensive suite of filtering procedures based on (a) absolute haplotype abundance across the whole dataset [range: 3 - 500], (b) relative haplotype abundance across the whole dataset [range: 0.000005 - 4%], (c) relative haplotype abundance across the whole

dataset by phylogenetic clade [range: 0.0000025 - 4%], (d) absolute haplotype abundance by library [range: 3 - 500], (e) relative haplotype abundance by library [range: 0.000005 - 4%] and (f) relative haplotype abundance by library and phylogenetic clade [range: 0.0000025 - 4%]. We also evaluated the effect of combinatory filtering resulting from jointly considering (a)+(b), (a)+(c), (b)+(c) and (d)+(e), (d)+(f), (e)+(f) criteria, assuming a similar range of threshold values as described above. We excluded a given ASV if it did not pass the specified threshold in all libraries where it was found. Following Andújar et al. (2021), clades were defined assuming a 20% divergence threshold on a UPGMA tree reconstructed with F84 model-corrected distances (Felsenstein & Churchill, 1996). Distances were calculated on an alignment which was obtained in MAFFT (FFT-NS-i method) with 1000 cycles (Katoh & Standley, 2013).

For each filtering strategy, we calculated the ratio of retained estimated non-authentic ASVs to the total number of estimated retained ASVs (estimated retained non-authentic ASVs plus estimated retained authentic ASVs). For each taxonomic group (Acari, Collembola and Coleoptera), we selected the filtering solution that yielded the lowest ratio value (always 0%) while maximizing the total number of actual ASVs in the final dataset (Andújar et al., 2021), in order to ensure the removal of most erroneous sequences. The best-fit filtering strategy for each taxonomic group included the following criteria and threshold values: Acari, (b) relative haplotype abundance across the whole dataset [value = 0.0005%] and (c) relative haplotype abundance across the whole dataset by phylogenetic clade [value = 0.065%]; Collembola, (d) absolute haplotype abundance by library [value = 15] and (f) relative haplotype abundance by library and phylogenetic clade [value = 0.03%]; and Coleoptera, (d) absolute haplotype abundance by library [value = 300] and (f) relative haplotype abundance by library and phylogenetic clade [value = 0.15%]. Finally, the *dump.py* Python script was used to generate the sequence file containing only the METAMATE-filtered haplotypes that passed the abundance thresholds, as specified by the best-fit filtering solutions. After examining these sequence files, we found no presence of stop codons or unexpected amplicon lengths.

Once the METAMATE filtering was completed, we used VSEARCH (Rognes et al., 2016) to generate a read-count community table of the METAMATE-filtered haplotypes by matching them against the raw read dataset (before dereplicating, length filtering and denoising; see Table S2). The read-count community table was further filtered by removing those ASVs showing abundances of 2 or fewer reads and also those whose contribution to the total number of reads per taxonomic group and library was lower than 1%. After completing the read processing and filtering pipeline, there were recalcitrant sequences remaining in the negative controls, which were absent or presented very low-abundance in any given actual sample, and were thus removed completely from the dataset (see Taberlet, Bonin, Zinger, & Coissac, 2018). Filtered read-count community tables were converted to presence/absence tables.

Digital elevation model (DEM) and weighted topographic distances

We obtained a digital elevation model (DEM) at 90 meters resolution in GeoTIFF format from the CGIAR-CSI GeoPortal (Jarvis, Reuter, Nelson, & Guevara, 2008; *https://srtm.csi.cgiar.org*). This DEM is derived from the NASA´s Shuttle Radar Topographic Mission (SRTM) data distributed by the United States Geological Survey (USGS), which has been processed to fill data voids using interpolation methods in order to produce a continuous topographic surface.

Given the high topographic complexity of the study region, we did not rely on Euclidean geographic distances among sampling sites, but we calculated the weighted topographic distance between each pair of sampling sites using the 'topoWeightedDist' function from the package *topoDistance* (Wang, 2020) in ʀ (ʀ Core Team, 2020). Topographic distances account for the additional overland distance covered by an organism due to elevation changes imposed by topographic relief. The coordinates of the sampling sites and the previously described DEM were used as inputs. We calculated the weighted topographic distances using a linear function to weight aspect changes (*hFunction* parameter) and an exponential function to weight the slope (*vFunction* parameter), as recommended by Wang (2020). This assumes that the energetic cost to traverse a slope varies exponentially with the change in angle.

WorldClim and ENVIREM variables

We downloaded near current (1970-2000) climate data at 30-arcseconds (~1 km) resolution from WorldClim v.2.1 (Fick & Hijmans, 2017; *https://www.worldclim.org/*). We obtained climate data for monthly average, minimum, and maximum temperature and monthly precipitation, and the 19 standard WorldClim bioclimatic variables. Additionally, we downloaded the 30-arcseconds DEM that was specifically used to produce the WorldClim dataset.

We interpolated climate raster variables to downscale the WorldClim data to a finer resolution (90 meters), which allowed us to distinguish the closest sampling sites in terms of their environmental profiles since the WorldClim data are not spatially detailed enough to account for the spatial heterogeneity of such a mountainous terrain (Poggio, Simonetti, & Gimona, 2018). We used the available elevation data at 90 m resolution as an independent variable to interpolate temperature and precipitation rasters. The interpolation models utilized here were similar to those used by the authors of the WorldClim data set, that include elevation as an independent variable to fit thin-plate spline interpolations of temperature and precipitation variables (Fick & Hijmans, 2017; see Figure S1 in their supplementary material). Firstly, we fitted thin-plate spline surfaces to the original 30-arcsecond resolution raster data, with the elevation from the 30-arcsecond DEM of WorldClim as an independent covariate, using the 'fastTps' function from the *fields* ʀ package (Nychka, Furrer, Paige, & Sain, 2017). This function uses a compactly supported Wendland covariance, and we set its θ scale parameter to

5 geographical miles (about 5 arcminutes). This parameter is the tapering range that is passed to the Wendland compactly supported covariance, and larger values make this model closer to a standard thin-plate spline model. We used the 'fastTpsMLE' function to estimate the smoothing parameter lambda (λ) from the data. We selected the 'fastTps' function instead of the standard 'Tps' function because the former was computationally faster and produced qualitatively similar results to the standard one in preliminary runs. Secondly, we generated 90 meters resolution raster data from these thin-plate spline surfaces and the 90 m resolution DEM with the 'interpolate' function in the *raster* R package (Hijmans, 2020). We did this for all monthly temperature and precipitation variables and for six selected bioclimatic variables: BIO1 (annual mean temperature), BIO5 (maximum temperature of warmest month), BIO6 (minimum temperature of coldest month), BIO12 (annual precipitation), BIO16 (precipitation of wettest quarter), and BIO17 (precipitation of driest quarter).

We used the 90 m resolution DEM and the interpolated monthly temperature and precipitation data to calculate three environmental and one topographic variable of the ENVIREM dataset (Title & Bemmels, 2018, and references therein) at the required resolution: climatic moisture index (CMI), Thornthwaite aridity index (TAI), topographic wetness index (TWI), and terrain roughness index (TRI). CMI is a metric of relative wetness and aridity, computed as the ratio of annual precipitation to annual potential evapotranspiration (Vörösmarty, Douglas, Green, & Revenga, 2005; Willmott & Feddema, 1992). TAI is a measure of the degree of water deficit below water need (Thornthwaite, 1948). TWI quantifies water flow over the topography and is correlated with soil attributes such as organic matter and horizon depth (Böehner et al., 2002; Conrad et al., 2005; Title & Bemmels, 2018). TWI provides information which is not correlated with elevation and can be used to characterize sites that are located in the same area, and thus with similar regional climate, but differ in soil conditions owing to their position within a watershed (Title & Bemmels, 2018). TRI is an index of terrain variability summarizing local variation to surface morphology (Wilson, O'Connell, Brown, Guinan, & Grehan, 2017), being used as a surrogate for topographic complexity, habitat heterogeneity and microsite availability (Title & Bemmels, 2018).

We used the 90 m resolution DEM to estimate the TWI from the DEM using SAGA v.2.1.4 (Conrad et al., 2015), and the TRI using the 'terrain' function from the *raster* R package. In order to calculate the climatic moisture index (CMI) and Thornthwaite aridity index (TAI) at the required resolution (90 meters), we first estimated annual potential evapotranspiration (PET) using the 'monthlyPET' function from the *envirem* R package (Title & Bemmels, 2018). This function calculates the potential evapotranspiration from the monthly mean temperature, temperature range and the extraterrestrial solar radiation. The extraterrestrial solar radiation can be generated with the 'ETsolradRasters' function from the *envirem* R package. Following Title and Bemmels (2018), the solar radiation was estimated for the year 1990 and the annual PET was calculated as the sum of the monthly values. The Thornthwaite aridity index was generated using the 'aridityIndexThornthwaite' function from the monthly precipitation and

monthly PET data. The CMI was calculated using the 'climaticMoistureIndex' function from the annual PET and the annual precipitation (sum of the monthly values), both functions implemented in the *envirem* R package (Title & Bemmels, 2018, see references therein).

Isolation by resistance (IBR) scenarios of landscape connectivity

Given that the Golden Oak - *Quercus alnifolia* (*Qa*) is an endemic species to Cyprus (also known as the national tree of the country) and its habitat (EU priority habitat type 9390) is protected by European legislation (listed in the Annex I of Habitat Directive, 92/43/EEC), the exact distribution and limits of its highly fragmented forest patches along the Troodos mountain range have been mapped at a fine scale by the local Department of Forests (Ministry of Agriculture, Rural Development and Environment, Republic of Cyprus). This available information on the *Quercus alnifolia* (*Qa*) forest patches allowed us to implement an isolation-by-resistance (IBR) approach to assess the effect of habitat fragmentation on β diversity patterns of the soil microarthropod communities. To this end, the digital cartography delimiting the geographic distribution of *Quercus alnifolia* (*Qa*) habitat fragments was obtained upon request from the Department of Forests (*https://www.data.gov.cy/*). The shapefile was carefully examined for inconsistencies using Google satellite imagery in ARCGIS v.10.3 (ESRI, Redlands, CA, USA). After some minor corrections regarding the edge location of specific small patches located in well-known areas by the authors, the shapefile was converted to a raster layer using the 'rasterize' function in the *raster* R package (Hijmans, 2020) and a cell size of 90 m. This raster file was further examined for potential errors in patch edges as a result of the rasterizing process.

For the IBR analyses, cells classified as *Quercus alnifolia* (*Qa*) were assigned a resistance value of 1. In order to assign an appropriate resistance value to non-*Quercus* cells, we explored a range of values and selected the one that best explained the variation of our response variables (for a similar approach, see Noguerales, Cordero, Knowles, & Ortego, 2021). Specifically, we generated 10 alternative models by assigning to non-*Quercus* cells increasing resistance values (from 5 to 1,000,000) using the 'reclassify' function in the *raster* R package. Then, we used univariate matrix regression with randomization (MRR; Wang, 2013) to evaluate the significance of each alternative model and the one showing the highest coefficient of determination (R^2) was selected for downstream analyses. Community dissimilarity matrices based on the Simpson dissimilarity index (β_{SIM}) at ASV and OTU levels were considered as response variables in these landscape sensitivity analyses.

Additionally, we constructed two alternative IBR scenarios considering (i) the topographic complexity of the study area as calculated using the topographic roughness index (TRI_{IBR}), and (ii) a "flat landscape" with all cells assigned an equal resistance (=1) value ($NULL_{IBR}$). The terrain roughness index (TRI) per cell was estimated based on the 90 m resolution DEM using the 'terrain' function from the *raster* R package, as explained above. The "flat landscape"

scenario (NULL$_{IBR}$) is considered as a proxy of classical Euclidean distances between geographic points that has proven to be appropriate for comparisons with competing IBR models using a circuit theory framework (Jha & Kremem, 2013; Noguerales, Ortego, & Cordero, 2017; Velo-Antón, Parra Parra-Olea, & Zamudio, 2013).

Finally, we calculated pairwise resistance distances between *Quercus alnifolia* (*Qa*) sampling points (*n* = 11) based on each IBR scenario using CIRCUITSCAPE v.4.0.5 (McRae, 2006; McRae & Beier, 2007), an eight-neighbor cell connection scheme, and average resistance for connections between cells. All raster data used to generate the above IBR scenarios were at 90 meters of resolution.

Statistical analyses

Principal coordinates of neighbour matrices (PCNM)

The topographic weighted distance matrix was transformed into spatial predictors using principal coordinates of neighbour matrices (PCNM), an approach allowing to transform distances to rectangular data that are suitable for constrained ordination or regression (Borcard & Legendre, 2002; Dray, Legendre, & Peres-Neto, 2006). Resulting PCNM eigenvectors are known to be robust predictors to describe both simple and complex spatial patterns across different geographic scales (Jones et al., 2008). In order to avoid an overrepresentation of spatial variables in posterior analyses owing to the high number of PCNM eigenvectors retrieved (up to 24 axes in our case), we ran a preliminary dbRDA (Legendre & Anderson, 1999) using the PCNM eigenvectors as explanatory terms and applied a forward selection procedure using the 'ordistep' function in the *vegan* R package (for a similar approach, see Peres-Neto & Legendre, 2010; Zinger et al., 2019). PCNM eigenvectors were independently selected for each spatial (across-habitats and within-habitat) and genetic similarity (ASVs and OTUs) scale using the respective community dissimilarity matrix (β_{SIM}). Only the PCNM eigenvectors remaining in the final model were considered for posterior analyses (dbRDA and mvGLMs).

Generalized linear mixed models (GLMMs)

We used generalized linear mixed models (GLMMs) and an information-theoretic model selection approach (Burnham & Anderson, 2002) to analyse the relationship between average richness per site (α diversity, RICH) or local contribution to β diversity (LCBD) per site and the topoclimatic variables (ENV$_{PC1}$ and ENV$_{PC2}$) as predictors, with latitude and longitude as covariates. GLMMs were fitted using a Gaussian error distribution and an identity link function. Forest habitat type (categorical predictor, HAB) was initially fitted in all models as a random effect in order to account for the potential non-independence among samples from the same forest habitat.

We fitted GLMMs using the maximum likelihood (ML) method, instead of using the restricted maximum likelihood (REML) approach, as recommended when a model selection approach is followed (Bolker et al., 2009; Bunnefeld & Phillimore, 2012; Harrison et al., 2018). We assessed the variance of the random effect in all GLMMs, and when its value was zero or nearly zero as indicative of a poor model fit and no influence of random effect on the variation of the response variable, we applied generalized linear models (GLMs) using a Gaussian error distribution and an identity link function. We analyzed the collinearity among predictors by means of variance inflation factor (VIF) and ensured none of them showed VIF >10.

The goodness of fit of the models was evaluated using AICc (Akaike's information criterion, adjusted for small samples). Best ranked equivalent models (ΔAICC ≤2) were full-averaged and the unconditional 95% confidence intervals (CI) of their estimators were estimated. The effect of variables was considered as consistent and significant if the 95 % confidence interval (CI) of its estimator excluded the value 0 (Burnham & Anderson, 2002). We calculated the marginal (R^2_m) and conditional (R^2_c) coefficients of determination for each final model including only predictors considered significant. These statistics represent the variance explained by the fixed effects (R^2_m) and by the entire model including both fixed and random effects (R^2_c). The R^2_c estimator was not calculated in those cases where model fitting was conducted using GLMs, as random effects had no influence on the response variable. GLMMs/GLMs were built using 'lmer' and 'glm' functions in the *lme4* R package (Bates, Maechler, Bolker, & Walker, 2015). Model selection and averaging were performed using 'dredge', 'model.sel', 'model.avg', 'par.avg' and 'r.squaredGLMM' functions in the *MuMIn* R package (Barton, 2015).

REFERENCES

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., … Emerson, B.C. (2021). Validated removal of nuclear pseudogenes and sequencing artifacts from mitochondrial metabarcode data. *Molecular Ecology, in press.* doi: 10.1111/1755-0998.13337

Arribas, P., Andújar, C., Hopkins, K., Shepherd, M., & Vogler, A. P. (2016). Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution*, *7*, 1071-1081.

Arribas, P., Andújar, C., Salces-Castellano. A., Emerson, B. C., & Vogler, A. P. (2021). The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. *Molecular Ecology*, *30*, 48-61.

Barton, R. J. (2020). *MuMIn*: multimodel inference. R package version 1.43.17. Available at *https://CRAN.R-project.org/package=MuMIn*

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, *67*, 1–48.

Böehner, J., Köethe, R., Conrad, O., Gross, J., Ringeler, A., & Selige, T. (2002). Soil regionalization by means of terrain analysis and process parameterization. In Micheli, E., Nachtergaele, F., & Montanarella, L. (Eds.), *Soil Classification 2001 European Soil Bureau, Research Report No. 7.* (pp. 213-222). Luxembourg.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*, 127-135.

Borcard, D., & Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, *153*, 51-68.

Bunnefeld, N., & Phillimore, A. B. (2012). Island, archipelago and taxon effects: mixed models as a means of dealing with the imperfect design of nature's experiments. *Ecography*, *35,* 15-22.

Burham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference. A practical information-theoretic approa*ch (2nd edition). New York: Springer-Verlag.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., … & Boehner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v.2.1.4. *Geoscientific Model Development*, *8*, 1991-2007.

Dray, S., Legendre, P., & Peres-Neto, P. R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological modelling*, *196*, 483-493.

Felsenstein, J., & Churchill, G. A. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, *13*, 93-104.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*, 4302-4315.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., ... & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6,* e4794.

Hijmans, R. J. (2020). *raster*: geographic data analysis and modeling. ʀ package version 3.4-5. Available at *https://CRAN.R-project.org/package=raster*

Jha, S., & Kremen, C. (2013). Urban land use limits regional bumble bee gene flow. *Molecular Ecology, 22*, 2483-2495.

Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. (2008). Hole-filled seamless SRTM data V4. International Centre for Tropical Agriculture (CIAT). Available at *https://srtm.csi.cgiar.org*

Jones, M. M., Tuomisto, H., Borcard, D., Legendre, P., Clark, D. B., & Olivas, P. C. (2008). Explaining variation in tropical plant community composition: influence of environmental and spatial data quality. *Oecologia*, *155,* 593-604.

Katoh, K., & Standley D. M. (2013). ᴍᴀғғᴛ - multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30,* 772-780.

Legendre, P., & Anderson, M. J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, *69,* 1-24.

McRae, B. H. (2006). Isolation by resistance. *Evolution*, *60*, 1551–1561.

McRae, B. H., & Beier, P. (2007). Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 19885-19890.

Noguerales, V., Cordero, P. J., Knowles, L. L, & Ortego, J. (2021). Genomic insights into the origin of trans-Mediterranean disjunct distributions. *Journal of Biogeography*, *48*, 440-452.

Noguerales, V., Cordero, P. J., & Ortego, J. (2017.) Testing the role of ancient and contemporary landscapes on structuring genetic variation in a specialist grasshopper. *Ecology and Evolution*, *7*, 3110-3122.

Nychka, D., Furrer, R., Paige, J., & Sain, S. (2017). *fields*: tools for spatial data. ʀ package version 11.6. Available at *https://github.com/NCAR/Fields*

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2020). *vegan*: community ecology package. ʀ package version 2.5-7. Available at *https://CRAN.R-project.org/package=vegan*

Peres-Neto, P. R., & Legendre, P. (2010). Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography*, *19*, 174-184.

Poggio, L., Simonetti, E., & Gimona, A. (2018). Enhancing the WorldClim data set for national and regional applications. *Science of the Total Environment, 625*, 1628-1643.

ʀ Core Team. (2020). ʀ: a language and environment for statistical computing. ʀ Foundation for Statistical Computing, Vienna, Austria. Available at *https://www.r-project.org/*

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahe, F. (2016). ᴠsᴇᴀʀᴄʜ: a versatile open source tool for metagenomics. *PeerJ, 4*, e2584.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: for biodiversity research and monitoring.* UK: Oxford University Press.

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review, 38*, 55-94.

Title, P. O., & Bemmels, J. B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, *41*, 291-307.

Velo-Antón, G., Parra, J. L., Parra-Olea, G., & Zamudio, K. R. (2013). Tracking climate change in a dispersal-limited species: reduced spatial and genetic connectivity in a montane salamander. *Molecular Ecology*, *22*, 3261-3278.

Vörösmarty, C. J., Douglas, E. M., Green, P. A., & Revenga, C. (2005). Geospatial indicators of emerging water stress: an application to Africa. *Ambio*, *3*, 230-236.

Wang, I. J. (2013). Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution*, *67*, 3403-3411.

Wang, I. J. (2020). Topographic path analysis for modelling dispersal and functional connectivity: calculating topographic distances using the *topoDistance* R package. *Methods in Ecology and Evolution*, *11*, 265–272.

Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). *mvabund* - an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution, 3*, 471-474.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag.

Willmott, C., & Feddema, J. (1992). A more rational climatic moisture index. *The Professional Geographer*, *44*, 84-88.

Wilson, M. F. J., O'Connell, B., Brown, C., Guinan, J. C., & Grehan, A. J. (2007). Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Marine Geodesy*, *30*, 3-35.

Zinger, L., Taberlet, P., Schimann, H., Bonin, A., Boyer, F., De Barba, M., ... Chave, J. (2019). Body size determines soil community assembly in a tropical forest. *Molecular Ecology*, *28,* 528–543.