



Digital supplementary material to

CENTANNI, J., KAUFMANN, B., BLATRIX, R., BLIGHT, O., DUMET, A., JAY-ROBERT, P. & VERGNES, A. 2022: High resolution mapping in Southern France reveals that distributions of supercolonial and monodomous species in the *Tapinoma nigerrimum* complex (Hymenoptera: Formicidae) are related to sensitivity to urbanization. – Myrmecological News 32: 41-50.

The content of this digital supplementary material was subject to the same scientific editorial processing as the article it accompanies. However, the authors are responsible for copyediting and layout.

Fig. S1: Square grid mesh of 1km<sup>2</sup> on the study area.

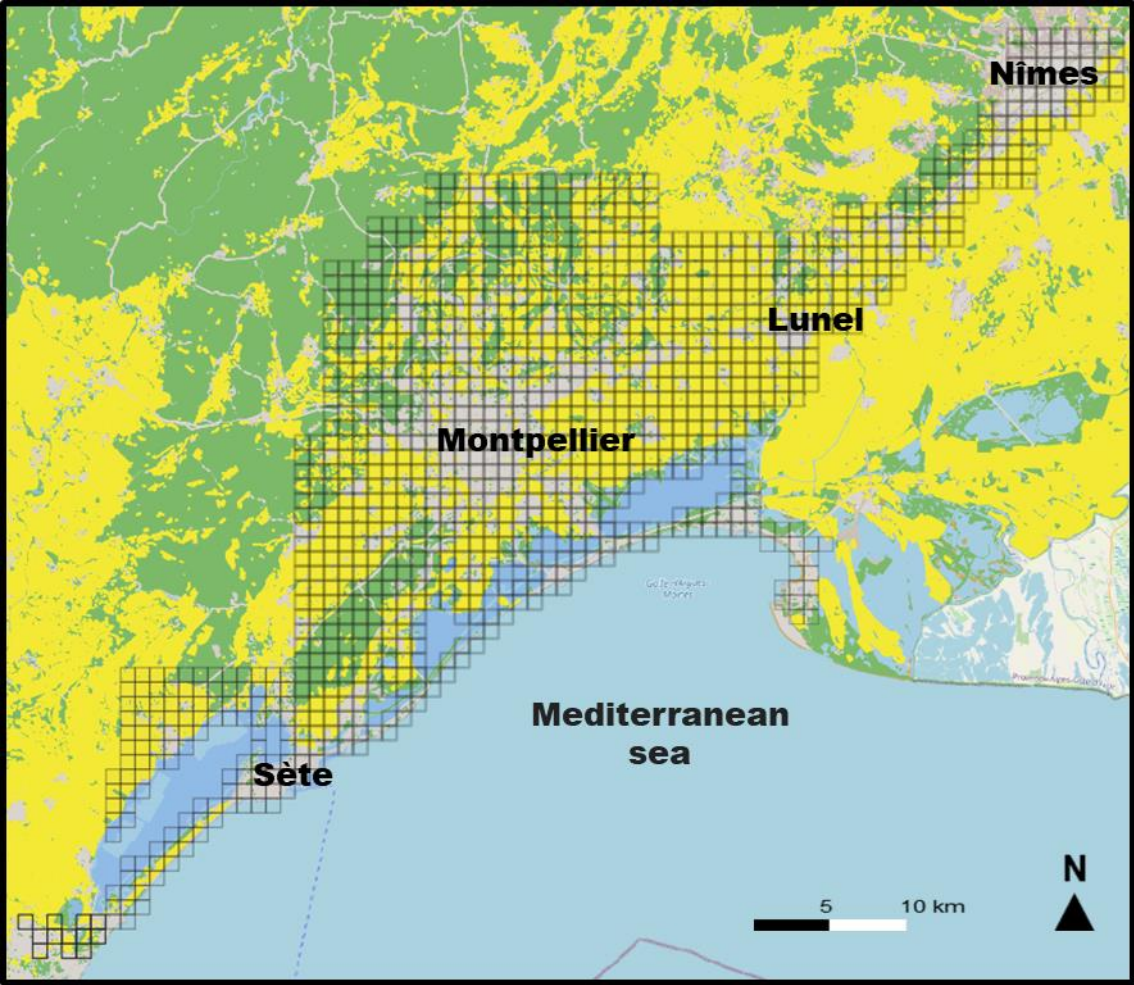


Fig. S2: Correlogram between the variables studied: “percentage of agricultural areas”, “percentage of urbanization”, “percentage of natural areas”, and the variable distance to the coast. The key in red corresponds to a negative correlation and the key in blue corresponds to a positive correlation according to the Spearman correlation test.

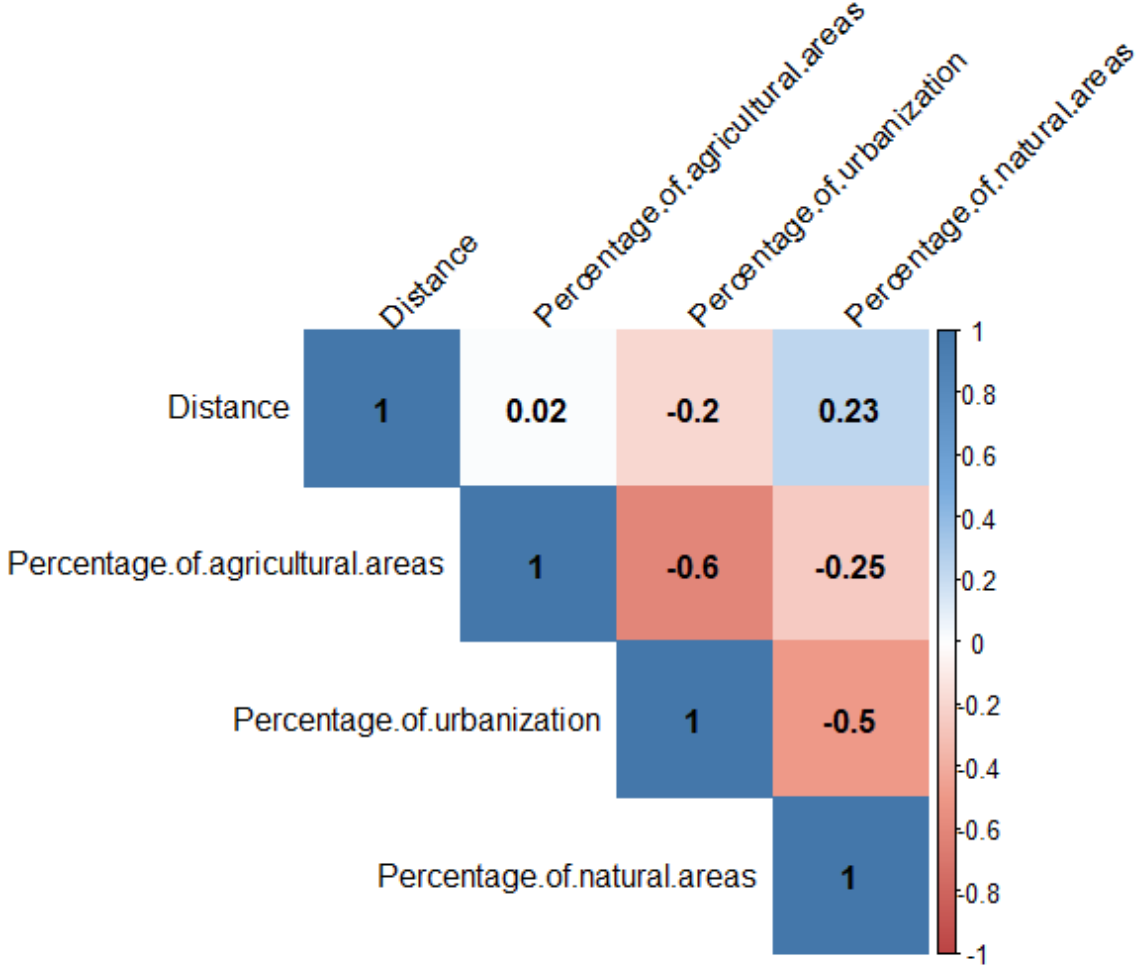


Fig. S3: Probabilities of occurrence for (A) *T. nigerrimum* (quadratic model, pvalue <0.001) and (B) *T. darioi* (linear model, pvalue <0.001), as a function of percentage of area for agricultural use in a 250m radius circular buffer around sampling sites (% agri). The light blue area represents the standard error. The rug plot at the bottom of the graph shows the number of sampling points corresponding to % agri values.

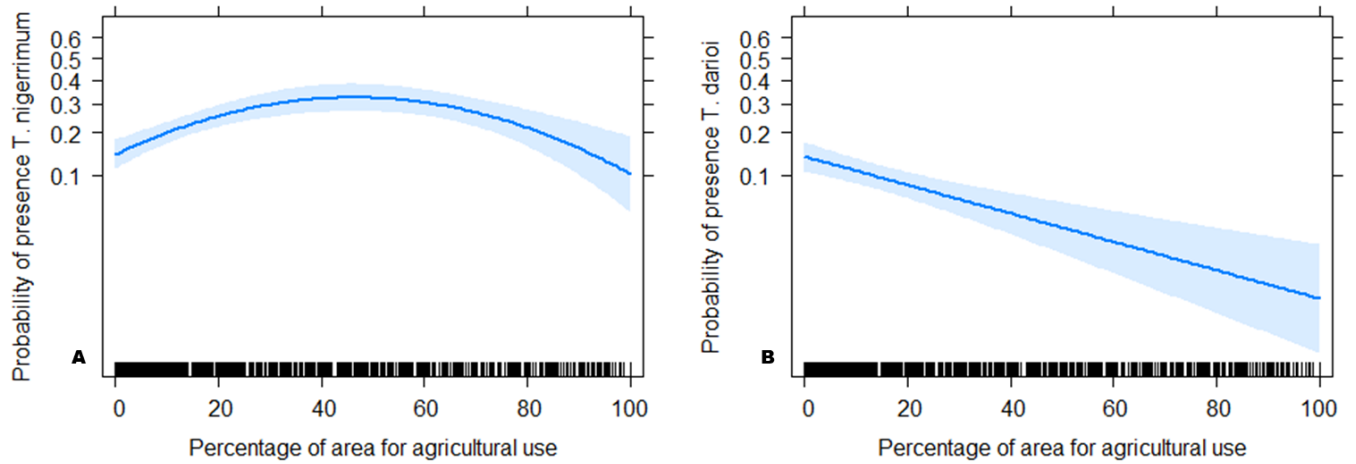
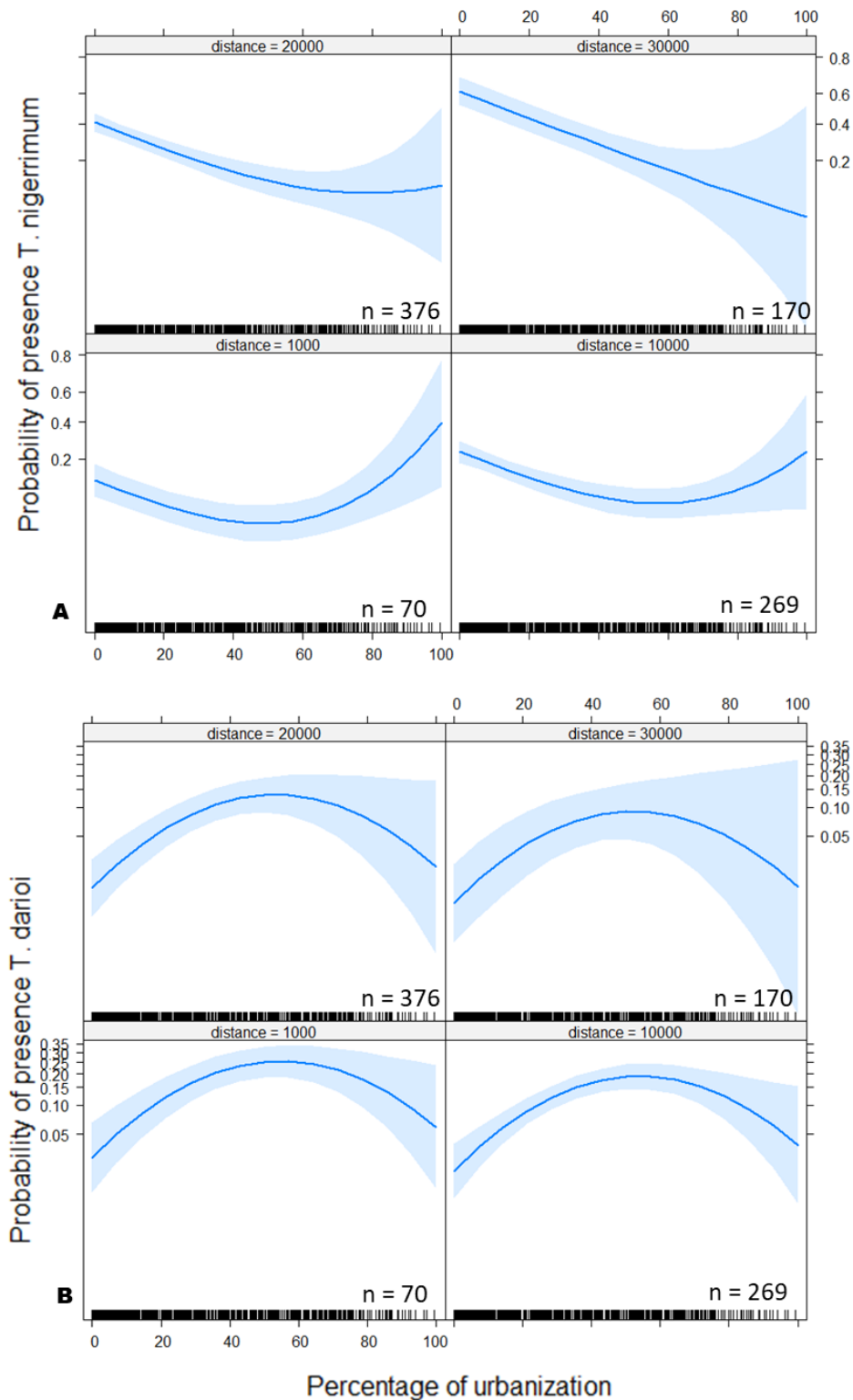


Fig. S4: Probabilities of occurrence for (A) *T. nigerrimum* (pvalue <0.01) and (B) *T. darioi*, (pvalue = 0.87) according to the interaction of urbanization and distance to the coast. The light blue area represents the standard error. n: number of points per class. The rug plot at the bottom of the graph shows the number of sampling points corresponding to % urban values. The direction in which the graphics are read begins at the bottom, from left to right and then at the top from left to right.



## Appendix S1. Genetic identification by microsatellites

Three other Appendix files are available:

- *Appendix S2.docx*, characteristics of microsatellite markers
- *Appendix S3.xlsx*, genotypes, geographic coordinates, PCA, DAPC and GC2 result table, is available as a separate file from the journal's web pages.
- *Appendix S4.docx*, PCA, DAPC and STRUCTURE figures.

SEIFERT & al. (2017) used NUMOBAT morphological methods to define species within the *Tapinoma nigerrimum* species complex and to identify individuals to the species level. They also used mitochondrial DNA (Cytochrome Oxidase gene) sequencing, which supported the NUMOBAT taxonomy. However, a number of disagreements between NUMOBAT and mtDNA appeared in *T. magnum* and *T. darioi* from Italy and from France. The identification of workers (but not of males) by NUMOBAT also led to some ambivalent results in *T. darioi* and *T. magnum*.

NUMOBAT identification methods are extremely time consuming (KLARICA & al. 2011) and mtDNA markers do not give exact identifications in species complexes where hybridization or incomplete lineage sorting occur (as is the case in *T. nigerrimum* complex, SEIFERT & al. 2017). Following the methods developed in CORDONNIER & al. (2019), we therefore used 15 microsatellite markers designed for *Tapinoma magnum* to genotype one individual per colony collected. The details for microsatellite development, sequences and primers can be found in Appendix S1. The rationale behind identification by genotyping is that mutations that accumulate in microsatellite sequences should result in strong differentiation over time, especially in species which are reproductively isolated. Microsatellites are, however, known to converge by homoplasy. To overcome this obstacle, several markers have to be analyzed together. In CORDONNIER & al. (2019), five species of the ant genus *Tetramorium* (MAYR, 1855) were identified with 17 highly polymorphic microsatellite markers, which also revealed numerous interspecific hybrids between two species (CORDONNIER & al. 2018). In the present paper, the same approach was used, with one major difference: we used the DNA of individuals that were subjected to both NUMOBAT and to mtDNA identification in SEIFERT & al. (2017) to assign species names to genotype groups defined by microsatellite analysis.

### ***Microsatellite development***

Fifteen markers were developed from *Tapinoma magnum* individuals: The total genomic DNA was isolated from the ants collected in Lyon in 2014. The DNA was extracted following crushing (3 min at 30Hz in a Retsch MM200 ball mill), standard digestion (proteinase K [75 µg] + 200 µl of TNES buffer [0.05 M Tris, 0.1 M NaCl, 0.01 M EDTA, 0.5% SDS]) and salt-chloroform purification. The total genomic DNA was sent to GenoScreen, Lille, France ([www.genoscreen.fr](http://www.genoscreen.fr)), where 1µg was used for the development of microsatellite libraries through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries as described in MALAUSA & al. (2011). The total DNA was mechanically fragmented and enriched for AG,

AC, AAC, AAG, AGG, ACG, ACAT and ATCT repeated patterns. The enriched fragments were subsequently amplified. The PCR products were purified, quantified and the GsFLX libraries were then carried out following the manufacturer's protocols and sequenced on a GsFLX PTP. Details for microsatellite development, sequences and primers can be found in Appendix S1. Primer characteristics and marker accession numbers can be found in Appendix S2 Table 1.

### ***DNA extraction***

The DNA from one individual per colony collected in the field was extracted by grinding each ant in proteinase K (10 $\mu$ L at 15mg/ml) and adding Chelex 7% (150 $\mu$ L), followed by a 2-hour incubation at 55°C and an additional 15 min at 90°C to inactivate the proteinase K.

### ***Genotyping***

Fifteen microsatellite markers were organized into three genotyping mixes (Appendix S2 Table 1) and run through a 3730xl DNA Analyzer (Applied Biosystems) by a service provider (GENTYANE, Clermont-Ferrand).

Forward primers were end-labelled with universal tails following BLACKET & al. (2012); four different tails were used with the following fluorescent dyes - Tail\_A\_6FAM [6FAM]GCCTCCCTCGCGCCA, Tail\_B\_HEX [HEX]GCCTTGCCAGCCCGC, Tail\_C\_ATTO550 [ATTO550]CAGGACCAGGCTACCGTG, Tail\_D\_ATTO565 [ATTO565]CGGAGAGCCGAGAGGTG. The primer and tail correspondence are indicated in Appendix S2 Table 1. All three PCR mixes had a total volume of 12  $\mu$ l with 1X Master Mix Type-it microsatellite PCR kit (QIAGEN, 206246) and 2  $\mu$ l extracted DNA with 0.02-0.1  $\mu$ M of primer Forward, 0.04-0.2  $\mu$ M of primer reverse and 0.1-0.2  $\mu$ M of tail. All PCRs were run with the same cycling program: 5 min. annealing at 95°C, 36 cycles with denaturation (30 s) at 95°C, annealing (3 min) at 60°C, extension (30 s) at 72°C, and a final 30 min extension at 60°C. Electrophoregrams were read and interpreted with Genemarker 1.95 (Softgenetics).

### ***Genetic analyzes***

Powerful Bayesian methods such as STRUCTURE (PRITCHARD & al. 2000) require data that conform to several conditions, in particular the Hardy-Weinberg equilibrium within a cluster, the sample size equilibrium of clusters and the allelic diversity equilibrium between clusters (PUECHMAILLE 2016). As basic population genetics (i.e. allelic richness, heterozygosities, F indices, null allele frequency) could not be computed before identification of species, none of these three conditions could be verified before clustering. Species identification was therefore made by using three methods concurrently: simple PCA, discriminant PCA (dACP, JOMBART 2008, JOMBART & al. 2010) and GENECLASS 2 (PIRY & al. 2004) which is based on Monte-Carlo resampling algorithms. We then checked the consensus results using STRUCTURE with or without priors.

In addition to samples belonging to the *Tapinoma nigerrimum* species complex, our samples might also include individuals from distant clades such as the *T. erraticum* and *T. simrothi* species groups. The *T. erraticum* species group is comprised of three species in France: *T. erraticum*, *T. madeirense* and *T. subboreale*. The precise taxonomy of *T. simrothi* is not stable, with a possible subspecies (*T. simrothi phoenicaeum*) found in the Western

Mediterranean, and disjoint distribution in Spain, Italy and North-Africa hinting at further complexity ([www.antmaps.com](http://www.antmaps.com)). No *T. simrothi* colony has recently been found in France, but older records exist which makes its presence possible.

Species from the three groups cannot always be told apart in the field or identified reliably in the laboratory without using NUMOBAT methods.

We therefore built up a reference dataset from the colonies that were identified using NUMOBAT by B. Seifert, most of which were part of SEIFERT & al. (2017). For *Tapinoma madeirense* where too few NUMOBAT colonies were available, we used samples identified by mtDNA sequencing in another study. 9-16 individuals per species made up the reference dataset (Appendix S2 Table 3), except for the *T. simrothi* group, which might comprise more than one species, and where we chose to use 44 individuals that cover the geographical distribution of the species group. Individuals sampled for the present study constitute the test dataset, except for four samples which are part of the reference dataset because they were first identified by mtDNA sequencing and NUMOBAT by B. Seifert.

PCA and DAPC were run in the R package *adegenet* (JOMBART & al. 2008) with reference and test datasets simultaneously. For PCA, we manually grouped samples that formed coherent clusters. Discriminant Analysis of Principal Components (DAPC) is a multivariate method designed to identify clusters of genetically related individuals (JOMBART & al. 2010). It uses sequential K-means and model selection to infer genetic clusters. Probabilities of assignment to each cluster is computed for all samples. In both methods, we used the reference dataset to identify the species contained in each cluster. In DAPC, individuals within a species can be split between separate clusters if there is a strong structure within the species, which can also occur with Bayesian methods such as STRUCTURE (e.g. CORDONNIER & al. 2019). The probability of assignment to species is therefore computed by addition of the probabilities into clusters belonging to the same species, which are identified thanks to the reference dataset.

GeneClass2 (GC2) is a program used to assign genotypes to populations from a reference dataset (PIRY & al. 2004). For each sample, the result of the analysis is a probability of assignment to each population, in the present case, to each species. As the reference dataset showed hierarchical clustering into species groups first, which confused assignment within species groups, GC2 was first run on reference and tested individuals of all species; subsequently, GC2 was run separately for each species group. Assignment probabilities of the second GC2 runs were used for species identification. The highest probability score in each reference group was retained for species identifications, except when the probabilities were below 0.01 for each species. Assignment tests were run with the simulation algorithm of PAETKAU & al. (2004), with criteria as per PAETKAU & al. (1995) and 1000 simulated individuals.

The identification or assignment results from all three methods were compared for each sample to reach a consensus. All samples with at least 2 identical species identifications, and a non-contradictory third, were assigned to the consensus species. When two methods contradicted a third, the assignment and clustering probabilities were checked, and if the contradicting method was ambiguous, it was assigned to the majority. For instance, individual N\_e05 was identified as *madeirense* by PCA and GC2, but as *subboreale* by DAPC; however, DAPC probabilities for *subboreale* were 0.516 vs 0.484 for *madeirense*; it was therefore considered as *madeirense*.



To check the identification results and help the reader who is more used to seeing STRUCTURE outputs, we ran STRUCTURE (PRITCHARD & al. 2000) with two different strategies, both with 10 runs of 200,000 burn-in iterations followed 200,000 iterations: i) the default STRUCTURE strategy is admixture, no priors; ii) an informed strategy with no admixture, priors (POPINFO flag) for the reference dataset, and with simulations updated only with information from the reference dataset. We ran simulations for the *nigerrimum* (with K=4 and K=5) and *erraticum* (with K=3 and K=4) groups separately, using the group information from the consensus between PCA, DAPC and GC2. We chose to run simulations with numbers of clusters corresponding to the number of species in each group, and also with that number + 1, as it is expected that that the species with the larger number of samples could be split between clusters (PUECHMAILLE 2016b). STRUCTURE results were analyzed with CLUMPAK (KOPELMAN & al. 2015), to control for clustering modes. All STRUCTURE figures show CLUMPAK major modes.

### ***Results of species identification-by-genotyping***

147 individuals belonging to the reference dataset and 413 from the test dataset. were genotyped. All microsatellite markers produced legible genotypes for all species. Allele numbers and sizes by species and markers can be found in Appendix S2 Table 3. Result tables from all three methods, as well as raw genotype data and geographical coordinates can be found in Appendix S3.

**For PCA**, ten axes were retained. All species could be separated by using the coordinates of the reference dataset genotypes (Appendix S4, Fig. 1-10).

**DAPC** revealed 15 clusters (Appendix S4, Fig. 11). For the reference individuals, clustering was clear:

- *T. darioi* samples were split into two clusters, with split probabilities (minimum 0.115) in each cluster, but with  $p=0$  in all other 13 clusters.
- All *T. ibericum* samples were all found in one single cluster with  $p=1$ .
- *T. magnum* samples were split into two clusters, but with exclusive  $p = 1$  in one or the other cluster.
- In *T. nigerrimum*, the situation was similar to *T. darioi* but over 3 clusters.
- *T. simrothi* samples were split into two exclusive clusters (samples from the Eastern Mediterranean in one cluster, Western Mediterranean samples in the other).
- All *T. erraticum* samples were all in one single cluster with  $p=1$ .
- In *T. madeirense* and *T. subboreale*, two closely related species the situation was more complex. Only nine *T. madeirense* were available for inclusion into the reference dataset vs 14 for *T. subboreale*. *T. madeirense* were split over three clusters with split probabilities which combined amount to 0.756 to 1; one individual had  $p=0.24$  in the *T. subboreale* cluster. *T. subboreale* were mostly found in one cluster, with probabilities above 0.75 for 10 individuals; 3 individuals had higher combined probabilities of belonging to one or more *T. madeirense* clusters, 1 individual was clearly placed within one of the *T. madeirense* clusters.

With GC2, the first run (Appendix S4, Fig. 12) resulted in 101 individuals assigned to the *Tapinoma erraticum* group, and 310 to the *T. nigerrimum* group. No *T. simrothi* could be detected. In the second hierarchical run, GC2 was able to assign all individuals from the *T. erraticum* group to species with  $>0.001$  probability (Appendix S4, Fig. 13). In contrast, 11 of the *T. nigerrimum* group could not be assigned to species with probabilities  $>0.001$  (Appendix S4, Fig. 14).

All the individuals could be assigned to a species. For 4 individuals, a third method was contradictory to the other two. Two individuals were identified as *Tapinoma madeirense* with PCA and with GC2 ( $p=0.17$  vs  $p=0$ , and  $p=0.02$  vs  $p=0$ ) but *T. subboreale* with DAPC (resp.  $p=0.516$  vs  $p=0.484$  and  $p=0.709$  vs  $p=0.291$ ); they were considered *T. madeirense* for the subsequent analyzes. Two other individuals were assigned to *T. darioi* by PCA and DAPC, but to *T. ibericum* by GC2. Both individuals were  $p=0$  for *T. ibericum* in DAPC ( $p=1$  for *T. darioi*), and  $p=0.005$  and  $p=0.019$  for *T. ibericum* in DAPC ( $p=0.001$  and  $p=0$  for *T. darioi*). As the probabilities were very low in GC2, but very high in DAPC, both individuals were therefore assigned to *T. darioi* for the subsequent analyzes.

STRUCTURE results with the informed strategy gave exactly the same identifications as the consensus method (Fig. 13 and 15) except for the *erraticum* group, where the single *subboreale* sample was identified as *subboreale* for  $K=3$ , but as *madeirense* for  $K=4$ . It would seem that when an extra cluster is added, *madeirense* samples from the test dataset cluster together against *madeirense* from the *Reference* dataset. The default strategy results (Fig. 16) mixed *madeirense* and *subboreale* at  $K=3$ , grouping all the *madeirense* from the local sampling (i.e. the three in the *Reference* dataset as well as those from the *test* dataset) versus other European *madeirense* as well as all *subboreale*. With  $K=4$ , *madeirense* was split in two (local vs European), but *subboreale* was clearly separated from the two *madeirense* clusters; the *subboreale* sample from the test dataset as *subboreale* was recovered. It is difficult to conclude on this individual which could either truly belong to *subboreale* or to an additional *madeirense* cluster. The default strategy results for the *nigerrimum* group (Fig. 14) were as expected biased by the more numerous and diverse *nigerrimum* sensu stricto: at  $K=4$  and for one mode (5/10) at  $K=5$ , *magnum* and *ibericum* clustered together while *nigerrimum* was split into 2 or 3 clusters. The second majority mode (5/10) however, recovered species as expected, with 2 *nigerrimum* clusters, but *ibericum* and *magnum* well separated.

Final identification results are therefore: 91, *Tapinoma madeirense*, 1 *T. subboreale* (dubious), 78 *T. darioi*, 6 *T. magnum* and 193 *T. nigerrimum* (+3 in the *Reference* dataset).

## References

- BLACKET, M.J., ROBIN, C., GOOD, R.T., LEE, S.F. & MILLER, A.D. 2012: Universal primers for fluorescent labelling of PCR fragments-an efficient and cost-effective approach to genotyping by fluorescence. – *Molecular Ecology Resources* 12: 456-463.

- CORDONNIER, M., GIBERT, C., BELLEC, A., KAUFMANN, B. & ESCARGUEL, G. 2019: Multi-scale impacts of urbanization on species distribution within the genus *Tetramorium*. – *Landscape Ecology* 34: 1937-1948.
- CORDONNIER, M., BELLEC, A., DUMET, A., ESCARGUEL, G. & KAUFMANN, B. 2018: Range limits in sympatric cryptic species: a case study in *Tetramorium* pavement ants (Hymenoptera: Formicidae) across a biogeographical boundary. – *Insect Conservation and Diversity* 12:109–120.
- JOMBART, T. 2008: Adegnet: A R package for the multivariate analysis of genetic markers. – *Bioinformatics* 24: 1403-1405.
- JOMBART, T., DEVILLARD, S. & BALLOUX, F. 2010: Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. – *BMC Genetics* 11: 94.
- KLARICA, J, BITTNER, L, PALLUA, J, PEZZEI C, HUCK-PEZZEI V, DOWELL F, SCHIED J, BONN GK, HUCK C, SCHLICK-STEINER BC, STEINER FM (2011) NEAR-INFRARED IMAGING SPECTROSCOPY AS A TOOL TO DISCRIMINATE TWO CRYPTIC *TETRAMORIUM* ANT SPECIES. *JOURNAL OF CHEMICAL ECOLOGY*, 37:549–552.
- KOPELMAN, N.M., MAYZEL, J., JAKOBSSON, M., ROSENBERG, N.A. & MAYROSE, I. 2015: Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. – *Molecular Ecology Resources* 15: 1179-1191.
- PAETKAU, D., CALVERT, W., STIRLING, I. & STROBECK, C. 1995: Microsatellite analysis of population structure in Canadian polar bears. – *Molecular Ecology* 4: 347-354.
- PAETKAU, D., SLADE, R., BURDEN, M. & ESTOUP, A. 2004: Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. – *Molecular Ecology* 13: 55-65.
- PIRY, S., ALAPETITE, A., CORNUET, J.M., PAETKAU, D., BAUDOUIN, L. & ESTOUP, A. 2004: GENECLASS2: A software for genetic assignment and first-generation migrant detection. – *Journal of Heredity* 95: 536-539.
- PRITCHARD, J.K., STEPHENS, M. & DONNELLY, P. 2000: Inference of population structure using multilocus genotype data. – *Genetics* 155: 945-959.
- PUECHMAILLE, S.J. 2016: The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. – *Molecular Ecology Resources* 16: 608-627.
- SEIFERT, B., D'EUSTACCHIO, D., KAUFMANN, B., CENTORAME, M., LORITE, P. & MODICA, M.V. 2017: Four species within the supercolonial ants of the *Tapinoma nigerrimum* complex revealed by integrative taxonomy (Hymenoptera: Formicidae). – *Myrmecological News* 24: 123-144.

**Appendix S2. Characteristics of microsatellite markers (Table 1), Fluorochrome tails sequence (Table 2) and population genetics characteristics of microsatellite markers per species (Table 3).**

**Appendix S2 Table 1:** Characteristics of the microsatellite markers developed and used for the study.

Locus name	Accession number	Primer sequence 5'_3' F	Primer sequence 5'_3' R	Repeat motif	PCR Mix	Fluorochrome tails
Tma-F08		<b>GCCTCCCTCGCGCCAACAAATACGCGCAGATTCGGCTC</b>	<b>ATACAGCCGGTCTTATCGTGTCTG</b>	(GA) <sub>11</sub>	A	tail a
Tma-H05		<b>GCCTTGCCAGCCC GCGACTCTACAAATCAGCGAGCGCAC</b>	<b>GATACCCGGCAGATATCGAGCCGT</b>	(AC) <sub>10</sub>	A	tail b
Tma-B01		<b>GCCTTGCCAGCCC GCGCCGTAACCTAGATCCCTACGT</b>	<b>GTACAATAAAGAACGCGTGGAACG</b>	(TC) <sub>11</sub>	A	tail b
Tma-H09		<b>CAGGACCAGGCTACCGTGTTAACCGCGTAGAGCAAACTCTG</b>	<b>ATTGCAGCGAAGATTAATCAAGAC</b>	(CT) <sub>10</sub>	A	tail c
Tma-D05		<b>CGGAGAGCCGAGAGGTGCCGCGACTCGATCTTGCGAAACTG</b>	<b>CATCACGTGCGATATGAGAACACG</b>	(AG) <sub>13</sub>	A	tail d
Tma-F10		<b>GCCTCCCTCGCGCCAATCGCACGGGATAATGAAAGATGC</b>	<b>TTCACTCGGGTCGTATTGTATCGC</b>	(TG) <sub>12</sub>	B	tail a
Tma-A11		<b>GCCTCCCTCGCGCCATTGTCGGAGATTATTAGGCGGCCG</b>	<b>TAAGAACGAAATGCTTCATGGTCC</b>	(AG) <sub>12</sub>	B	tail a
Tma-F12		<b>GCCTTGCCAGCCC GCTTCGAGATGTGCTATTCTCGGTAC</b>	<b>GATAATAGCGGTTAAAAGACCTGC</b>	(GA) <sub>11</sub>	B	tail b
Tma-B10		<b>GCCTTGCCAGCCC GCGCCGTTAATGACTATACCC</b>	<b>TACGCCACTCTCAATTCTCGAAAG</b>	(TC) <sub>11</sub>	B	tail b
Tma-C03		<b>CAGGACCAGGCTACCGTGATTGAAAATCTTGAAAAGTCTTG</b>	<b>GTCGAAACATCTTGCGGCTCTTGT</b>	(TG) <sub>16</sub>	B	tail c
Tma-A09		<b>GCCTCCCTCGCGCCACTTTGTACGTCCGCGGATTATTAC</b>	<b>TTTGTGTTATAAACACGCGCTGGC</b>	(TC) <sub>14</sub>	C	tail a
Tma-A10		<b>GCCTTGCCAGCCC GCTATACACCGTCCGATGACGAATGC</b>	<b>ACGAAGGTTTCGTTATCGCTTTTCG</b>	(AC) <sub>13</sub>	C	tail b
Tma-B03		<b>CAGGACCAGGCTACCGTGAGGGATCTGTGTATCCCCGAGTAC</b>	<b>TGACACCACATGAGAGTGGGAAGG</b>	(CT) <sub>11</sub>	C	tail c
Tma-A03		<b>CAGGACCAGGCTACCGTGCACGCGTGCGTACATGACTCAACC</b>	<b>TCCGCGCAATTTATTGCTCCCTTC</b>	(CA) <sub>11</sub>	C	tail c
Tma-E11		<b>CGGAGAGCCGAGAGGTGACAAAGAAGAAACGGAGCTTGCCG</b>	<b>CAGAGGAGGAGAGGTGTATCGGTG</b>	(TC) <sub>11</sub>	C	tail d

**Appendix S2 Table 2:** Fluorochrome tails sequences.

Tail	Fluorochrome	Tail sequence
Tail_A	6FAM	GCCTCCCTCGCGCCA
Tail_B	HEX	GCCTTGCCAGCCCGC
Tail_C	ATTO550	CAGGACCAGGCTACCGTG
Tail_D	ATTO565	CGGAGAGCCGAGAGGTG

**Appendix S2 Table 3:** Microsatellite markers characteristics for each species genotyped, based on both reference and test datasets. Classical indices were not relevant as both reference and test datasets comprise more than one population. NR = number of reference individuals, NT = number of test individuals, AR = number of alleles in reference samples, AT = number of allele in test samples, Span = sizes of the smallest and largest allele (reference and test samples together).

	<i>darioi</i>		<i>ibericum</i>		<i>magnum</i>		<i>nigerrimum</i>		<i>erraticum</i>		<i>madeirense</i>		<i>subboreale</i>		<i>Gr. simrothi</i>	
NR / NT	16	78	14	1	16	6	16	193	14	0	9	91	14	1	48	0
Marker	AR/AT	Span	AR/AT	Span	AR/AT	Span	AR/AT	Span	AR/AT	Span	AR/AT	Span	AR/AT	Span	AR/AT	Span
Tma-F08	12/12	126-172	8/0	124-160	10/6	122-158	5/8	128-152	14/0	132-182	4/3	122-128	2/1	122-124	7/0	118-132
Tma-H05	1/3	122-126	4/1	122-128	6/4	124-166	2/3	122-136	1/0	126	4/12	132-166	3/2	130-138	25/0	120-192
Tma-B01	4/3	223-235	5/2	225-233	7/5	227-245	3/5	233-241	16/0	231-285	3/5	231-239	2/2	231-235	9/0	215-233
Tma-H09	4/9	142-164	9/1	140-172	12/6	138-174	3/3	146-150	1/0	160	11/18	168-220	12/1	180-204	19/0	136-180
Tma-D05	3/2	116-124	3/1	116-136	6/3	122-132	3/2	118-122	2/0	126-132	2/2	128-136	2/1	130-136	4/0	120-126
Tma-F10	3/3	130-134	4/0	126-134	6/3	134-144	1/1	130	1/0	128	1/1	124	1/1	124	3/0	128-142
Tma-A11	2/2	187-189	5/0	185-193	5/4	187-195	2/2	187-191	8/0	199-221	3/3	183-187	2/1	183-185	2/0	175-177
Tma-F12	14/15	162-206	12/1	158-192	9/6	154-204	7/15	166-198	4/0	160-168	4/5	152-164	1/1	160	3/0	158-162
Tma-B10	1/1	235	2/1	231-243	9/7	247-277	1/5	231-283	14/0	281-317	17/57	231-480	25/1	325-459	3/0	239-245
Tma-C03	3/3	176-182	6/2	174-188	9/3	172-220	1/3	176-180	13/0	186-212	8/12	170-242	5/1	168-176	2/0	164-182
Tma-A09	2/2	108-110	4/1	110-116	8/7	112-144	2/1	108-110	14/0	104-140	10/22	102-206	21/2	106-210	13/0	106-164
Tma-A10	3/4	131-139	5/1	131-139	6/2	125-145	5/6	133-145	2/0	131-133	3/6	129-141	4/1	135-141	4/0	125-129
Tma-B03	6/4	110-140	6/1	120-138	6/6	114-140	5/6	124-138	10/0	140-160	3/3	124-138	1/1	124	15/0	120-164
Tma-A03	1/3	246-250	3/0	246-250	7/3	244-296	1/1	246-250	7/0	256-298	2/3	254-262	4/2	256-260	11/0	242-264
Tma-E11	3/5	169-239	15/1	165-223	5/4	159-191	11/21	173-217	13/0	167-203	11/20	161-243	16/2	167-229	15/0	153-185

**Appendix S3. Excel file with genotypes, geographic coordinates, PCA, DAPC and GC2 result table. This appendix is available as separate file from the journal's webpages.**

## Appendix S4. Graphs of PCA, DAPC and STRUCTURE results.

- **Appendix S4.1 PCA results:** Figures 1-10 show PCA Axis 1 vs axes 2-10, with Fig. 1. showing Reference dataset only for Axes 1-2
- **Appendix S4.2 DAPC results:** Figure 11 upper panel shows DAPC probability of membership graphs for the Reference dataset, Figure 11 lower panel for the Test dataset; the two datasets were analyzed in a single DAPC run
- **Appendix S4.3 STRUCTURE results:** Figures 13-14 show STRUCTURE-CLUMPAK results for Reference and Test datasets in *T. nigerrimum* group. Figures 15-16 show results for both Reference and Test datasets for the *T. erraticum* group. In all figures, each bar represents a single individual, for which membership in each cluster (C1 to C5) is indicated by colors

### Appendix S4.1 PCA results

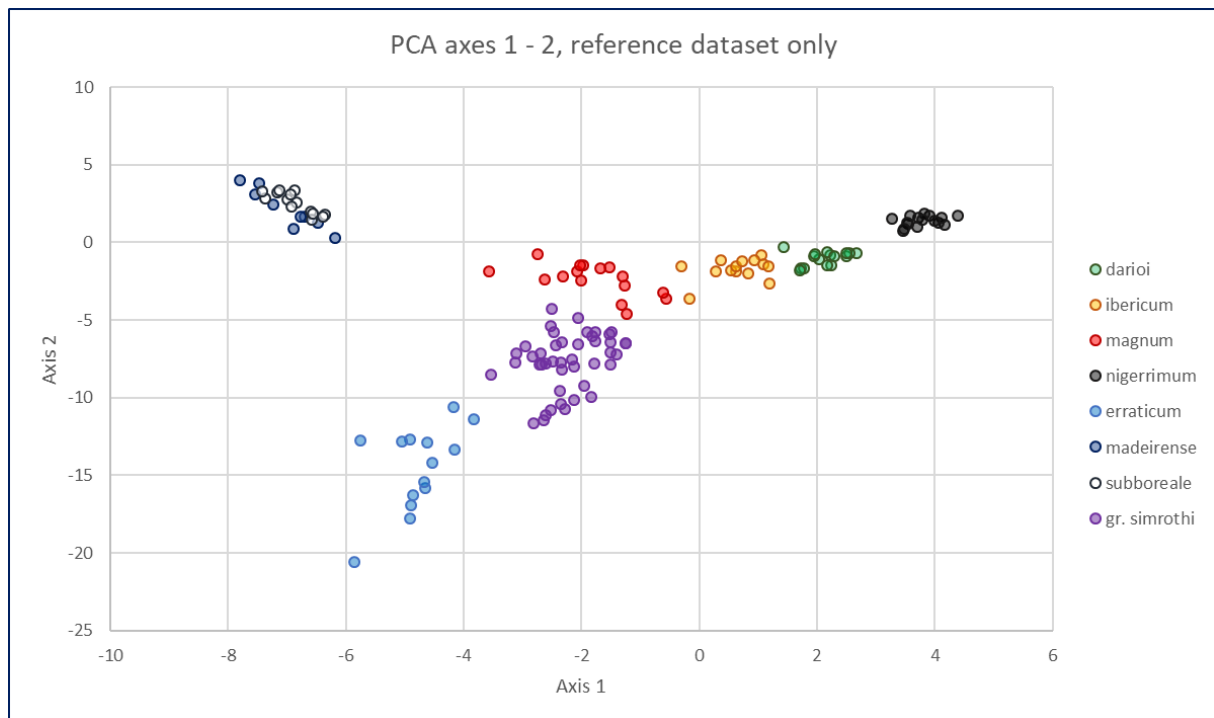


Fig. 1: Reference dataset PCA axes 1 and 2. All taxa except *madeirensis* and *subboreale* are well separated. Two *ibericum* and two *magnum* samples are close on axis 1 and 2.

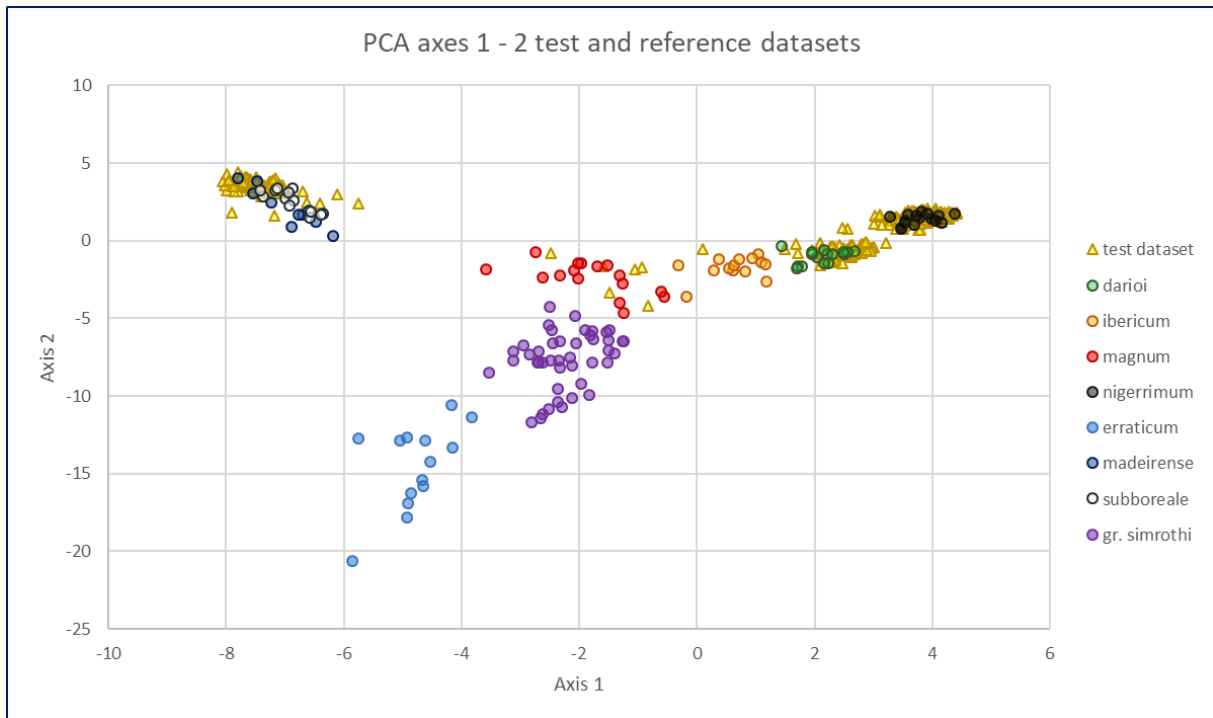


Fig. 2: Reference and test datasets PCA axes 1 and 2. No sample from the test dataset cluster with reference *T. simrothi* or *T. erraticum*.

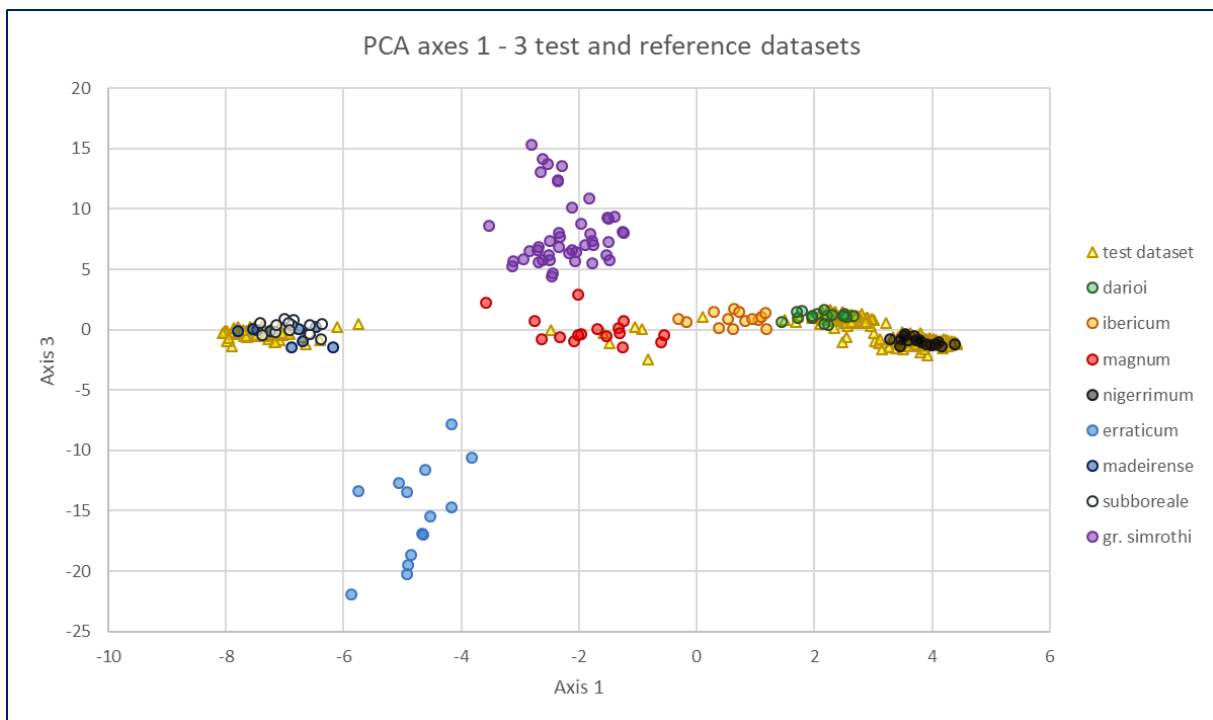


Fig. 3: Reference and test datasets PCA axes 1 and 3. Axis 3 make the *erraticum* and *simrothi* clusters more distinct.



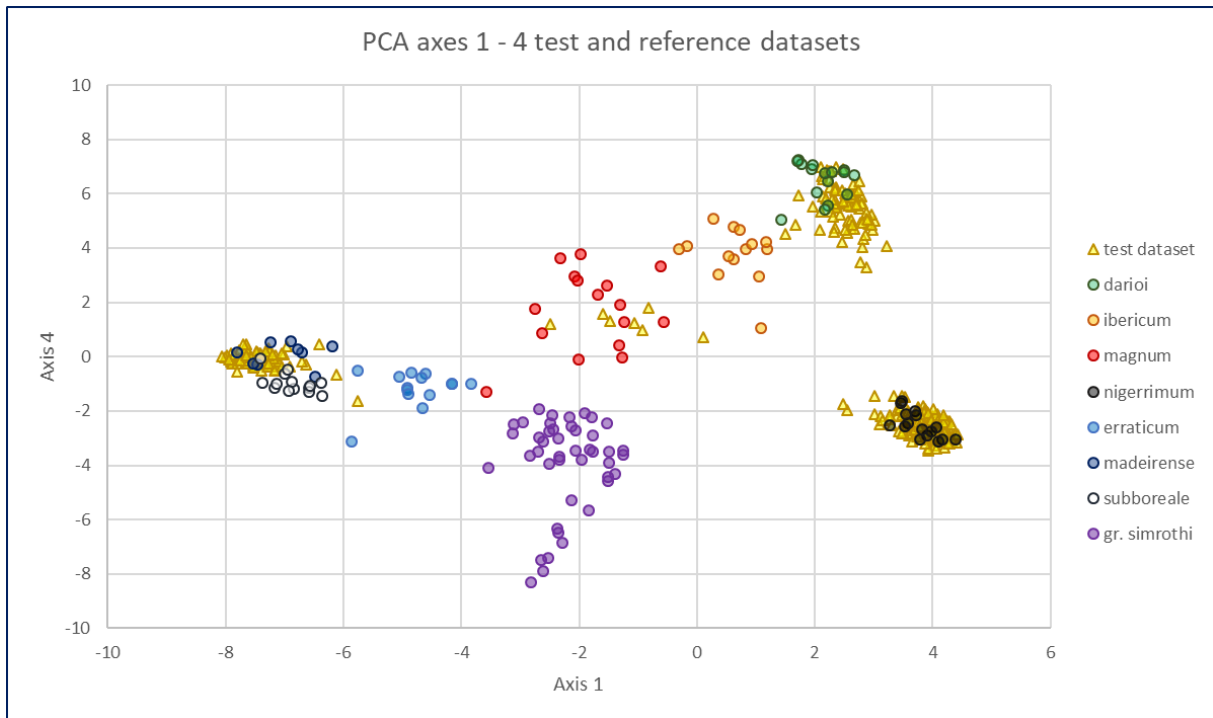


Fig. 4: Reference and test datasets PCA axes 1 and 4. Axis 4 clearly separate *darioi* from *nigerrimum*.

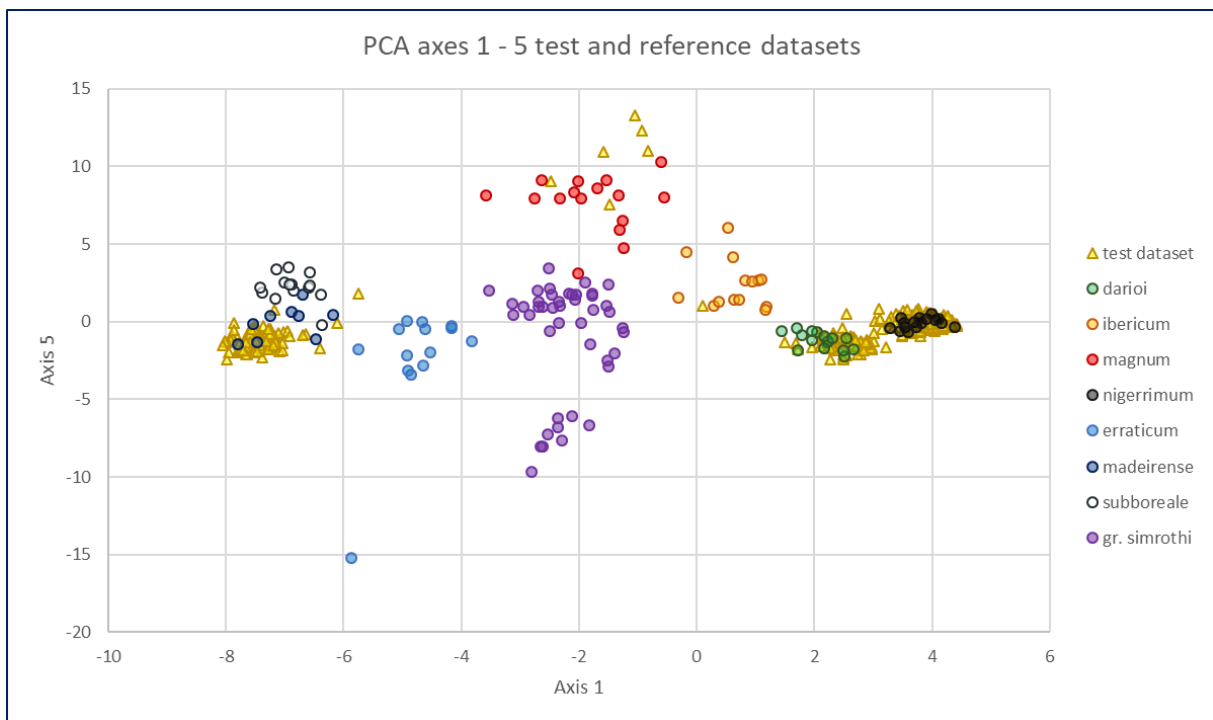


Fig. 5: Reference and test datasets PCA axes 1 and 5. Axis 5 separates *magnum* from *ibericum*, *ibericum* from *darioi*, and partly *madeirense* from *subboreale*.

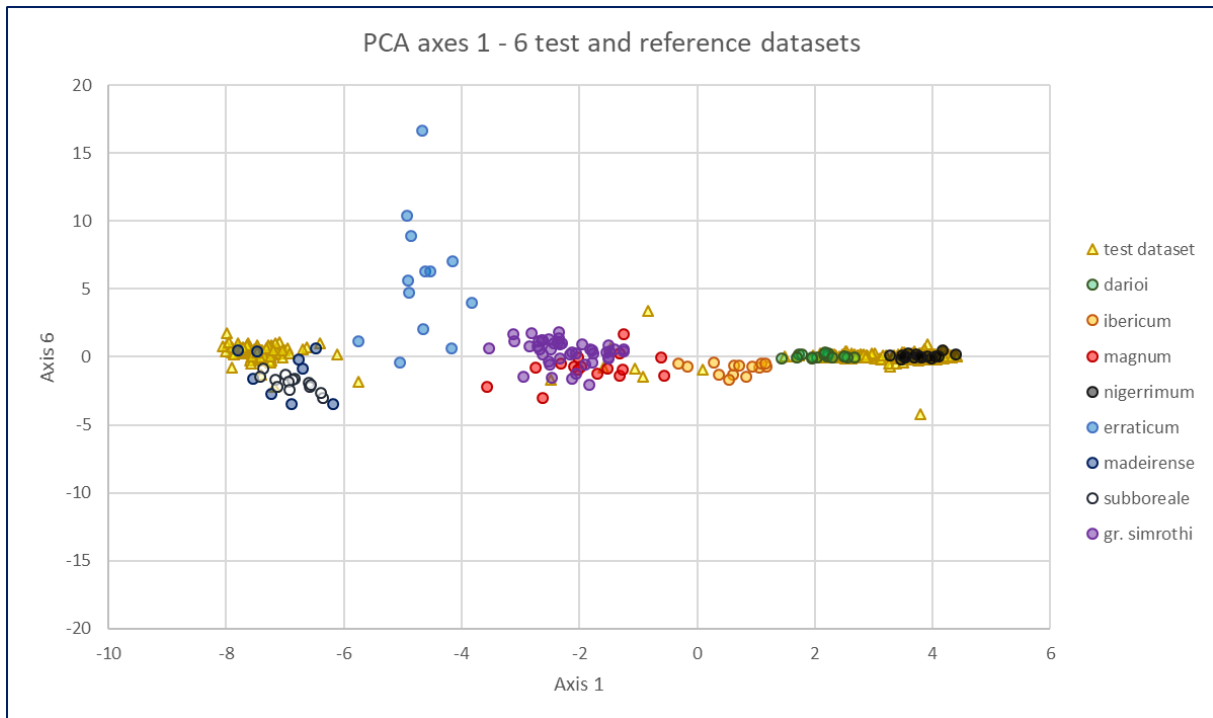


Fig. 6: Reference and test datasets PCA axes 1 and 6. Axis 6 shows diversity within reference *T. erraticum* and one differing test *T. nigerrimum*.

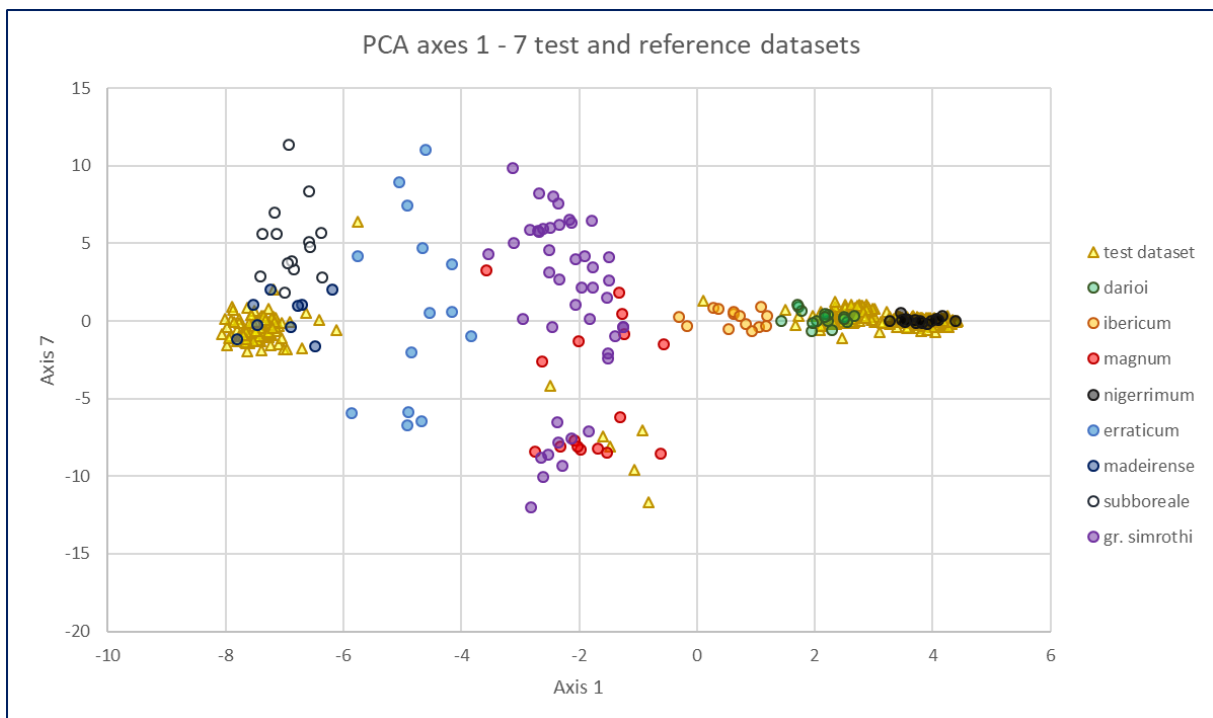


Fig. 7: Reference and test datasets PCA axes 1 and 7. Axis 7 shows diversity within reference *T. erraticum*, *T. simrothi* and *T. magnum*. Test dataset *T. magnum* samples are separated from all other species from the *T. nigerrimum* group. It also separates *T. madeirense* from *T. subboreale*, showing individual M\_e09 clustering with *T. subboreale*.

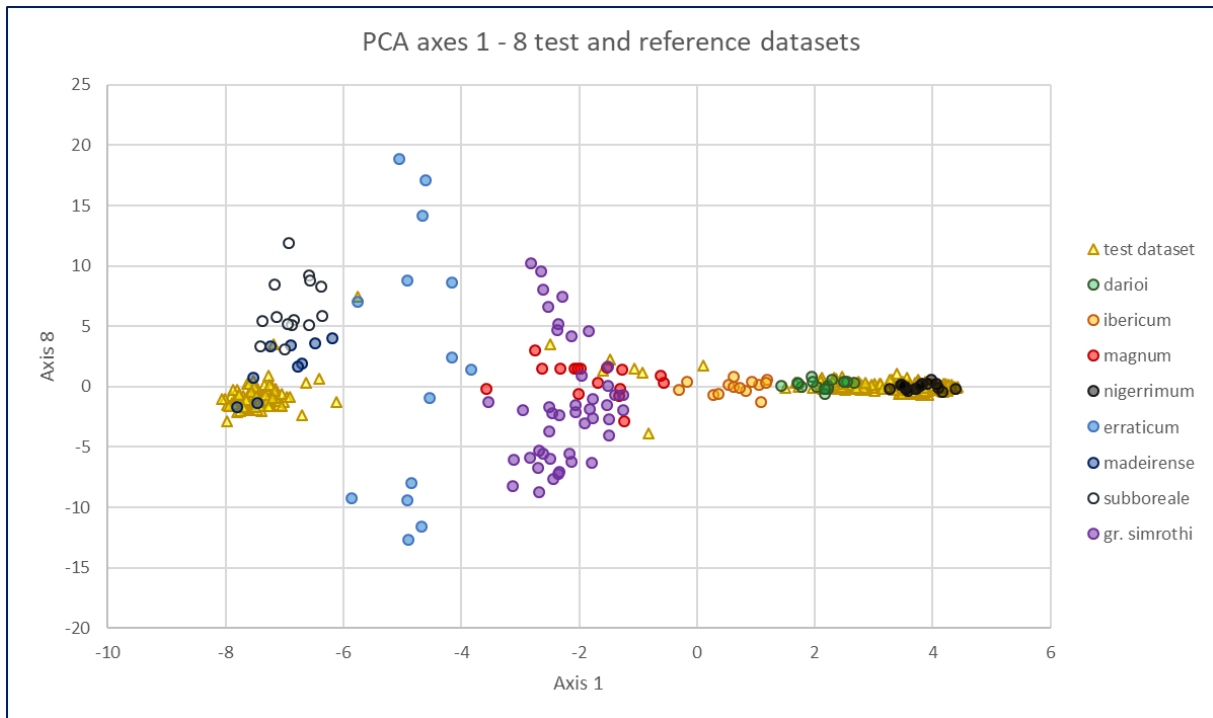


Fig. 8: Reference and test datasets PCA axes 1 and 8. Axis 8 again separate *T. subboreale* and *T. madeirense*, and shows diversity within *T. erraticum*.

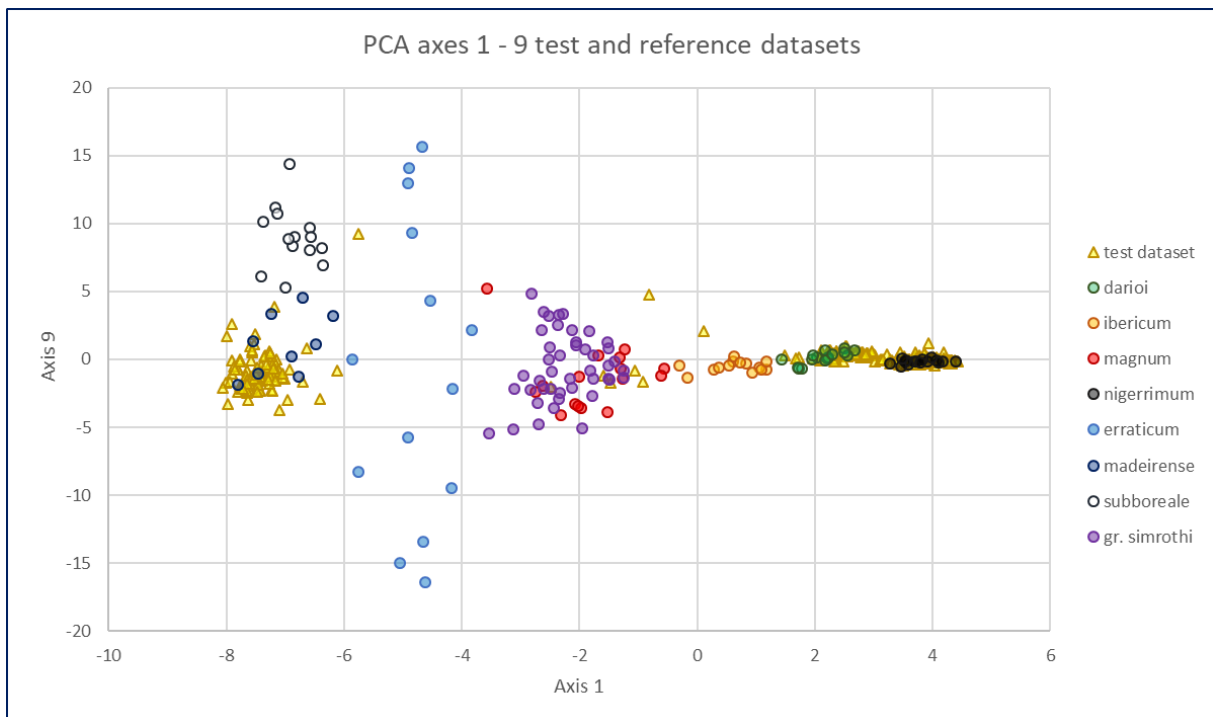


Fig. 9: Reference and test datasets PCA axes 1 and 9. Axis 9 again separate *T. subboreale* and *T. madeirense* (sample M\_e09 clusters with *T. subboreale* in this graph), and shows diversity within *T. erraticum*.

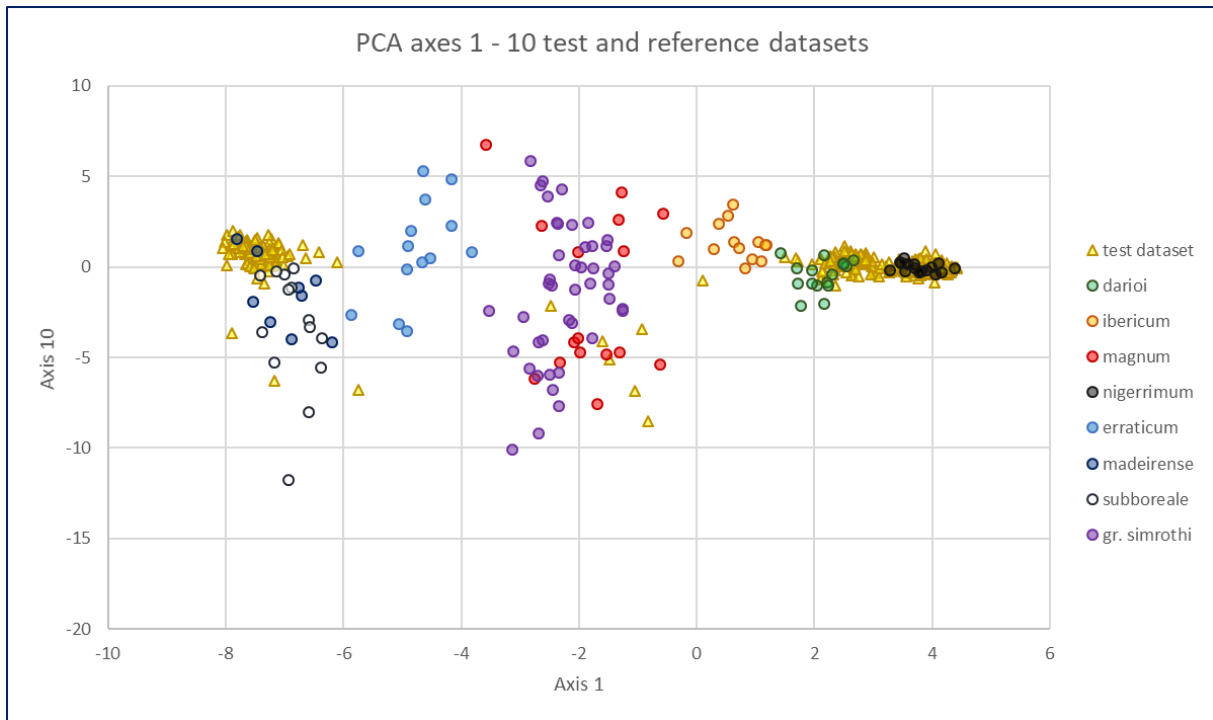


Fig. 10: Reference and test datasets PCA axes 1 and 10. Axis 10 again partly separate *T. subboreale* and *T. madeirense*, and shows diversity within *T. erraticum*, *T. magnum* and *T. simrothi*.

## Appendix S4.2. DAPC results

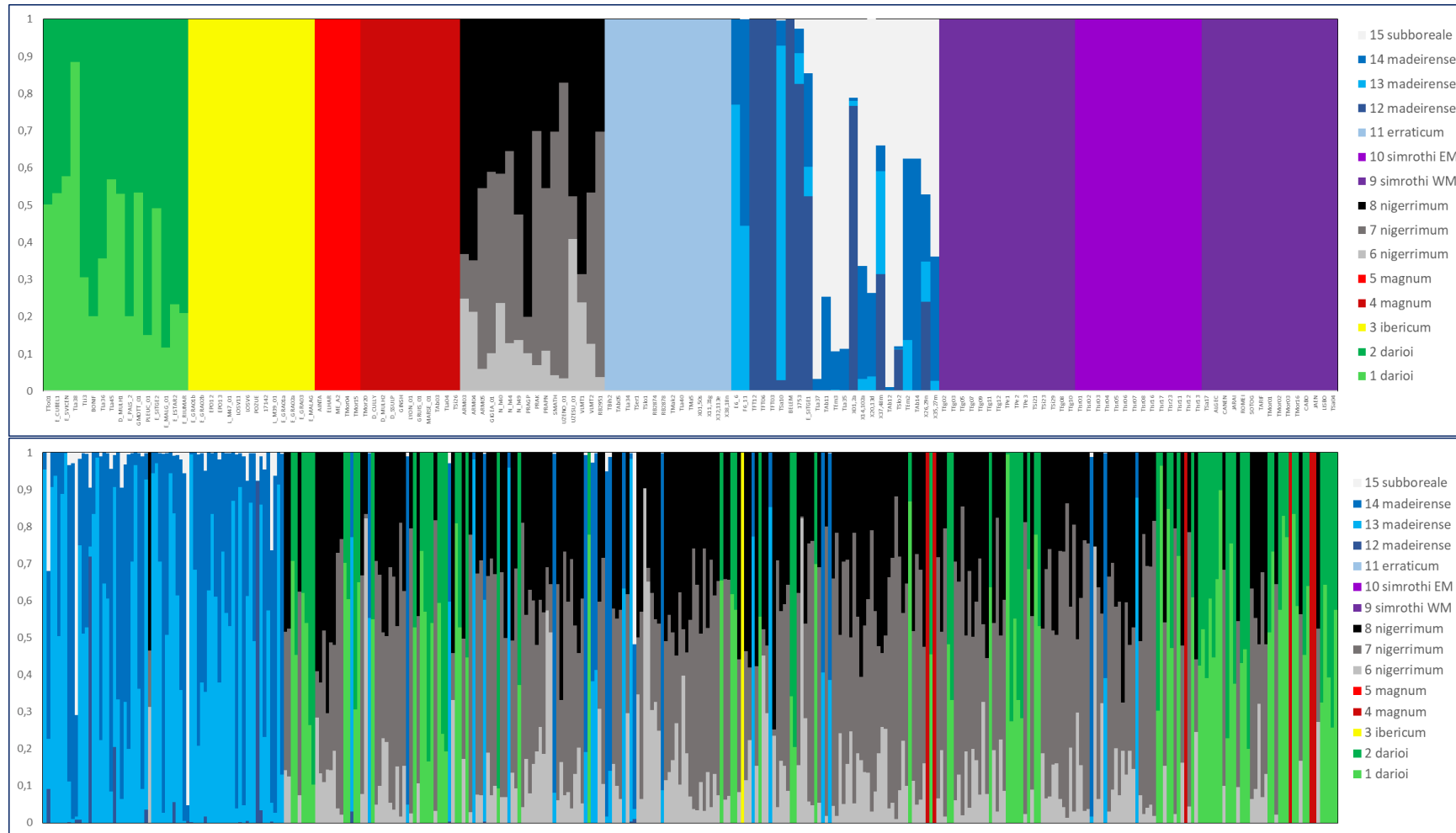


Fig. 11: Upper panel shows DAPC probability of membership for the Reference dataset, lower panel for the Test dataset; the two datasets were analyzed in a single DAPC run. Colors correspond to DAPC clusters, not to species, which can be split over more than one cluster, and in two different ways. The first way (e.g. *darioi*, *nigerrimum* or *madeirense* in the reference dataset), individual samples are split over the two or three clusters. The second way (e.g. *magnum* and *gr. simrothi* in the reference dataset), individuals have 100% membership in one of the two clusters; this suggests underlying genetic structure, as is the case here for Western and Eastern Mediterranean *T. simrothi* taxa; *T. magnum* could have a similar substructuring.

**Appendix S4.3 STRUCTURE results:** Figure 13-14 show STRUCTURE-CLUMPAK results for Reference and Test datasets in *T. nigerrimum* group. Figure 15-16 show results for both Reference and Test datasets for the *T. erraticum* group. In all figures, each bar represents a single individual, for which membership in each cluster (C1 to C5) is indicated by colors

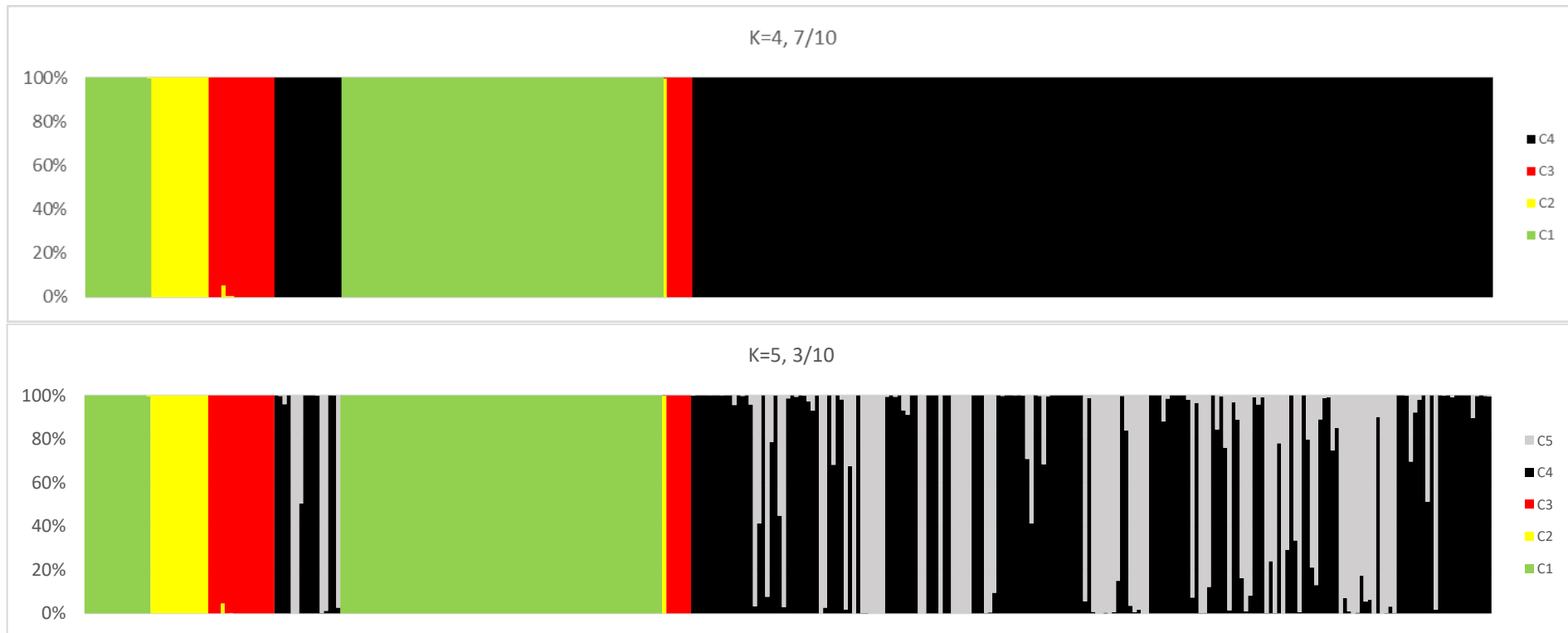


Fig. 13: Structure runs with priors for *T. nigerrimum* group. Parameters for Structure Runs: 10 runs, 200k Burn-in Iterations, 200k Iterations, model updated only with Reference dataset (POPFLAG on), no admixture. Upper panel shows results for K=4, major clustering model from Clumpak (7 runs/10). Lower panel shows results for K=5, major clustering model (there were 5 clustering models). Order of the individuals is from left to right: Reference dataset (*darioi*, *ibericum*, *magnum*, *nigerrimum*), Test dataset in order of species obtained from Consensus between PCA, DAPC and GENECLASS 2 (*darioi*, *ibericum*, *magnum*, *nigerrimum*), then in alphabetical order of sample name. Here in the upper panel Cluster 1 = *darioi*, Cluster 2 = *ibericum*, Cluster 3 = *magnum*, Cluster 4 = *nigerrimum*. In the lower panel, the additional Cluster 5 = *nigerrimum*.

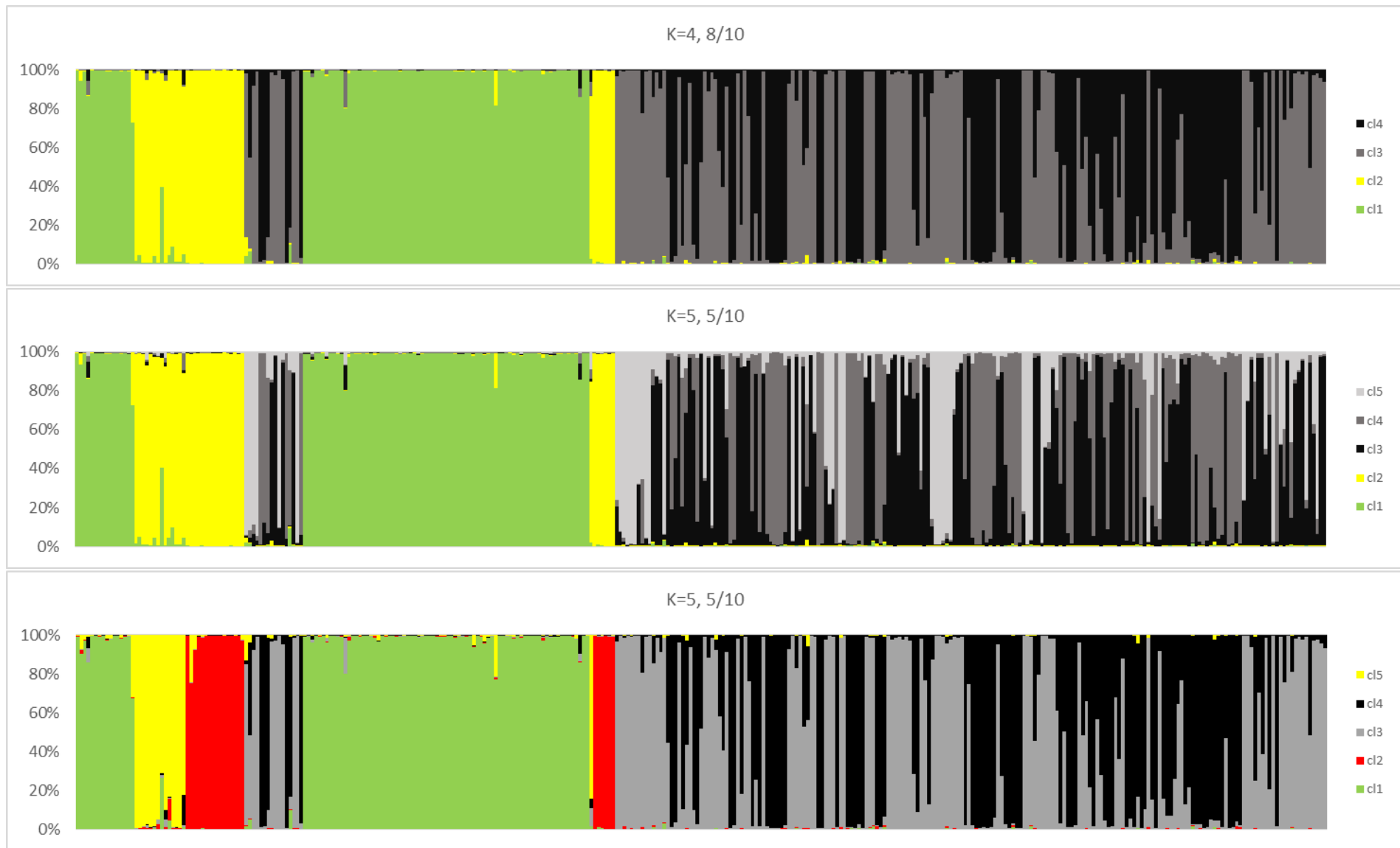


Fig. 14: Structure runs without priors for *T. nigerrimum* group. Parameters for Structure Runs: 10 runs, 200k Burn-in Iterations, 200k Iterations, no population prior, admixture. Upper panel shows results for K=4, major clustering model from Clumpak (8 runs/10). Middle and lower panels show results for K=5, major clustering models (5/10 runs each). Order of the individuals is from left to right: Reference dataset (*darioi*, *ibericum*, *magnum*, *nigerrimum*), Test dataset in order of species obtained from Consensus between PCA, DAPC and GENECLASS 2 (*darioi*, *ibericum*, *magnum*, *nigerrimum*), then in alphabetical order of sample name. In the upper panel Cluster 1 =

*darioi*, Cluster 2 = *ibericum* + *magnum*, Clusters 3 and 4 = *nigerrimum*. In the middle panel, Cluster 5 = *nigerrimum*. In the lower panel, Cluster 1 = *darioi*, Cluster 2 = *ibericum*, Cluster 3 = *magnum*, Cluster 4 + 5 = *nigerrimum*.

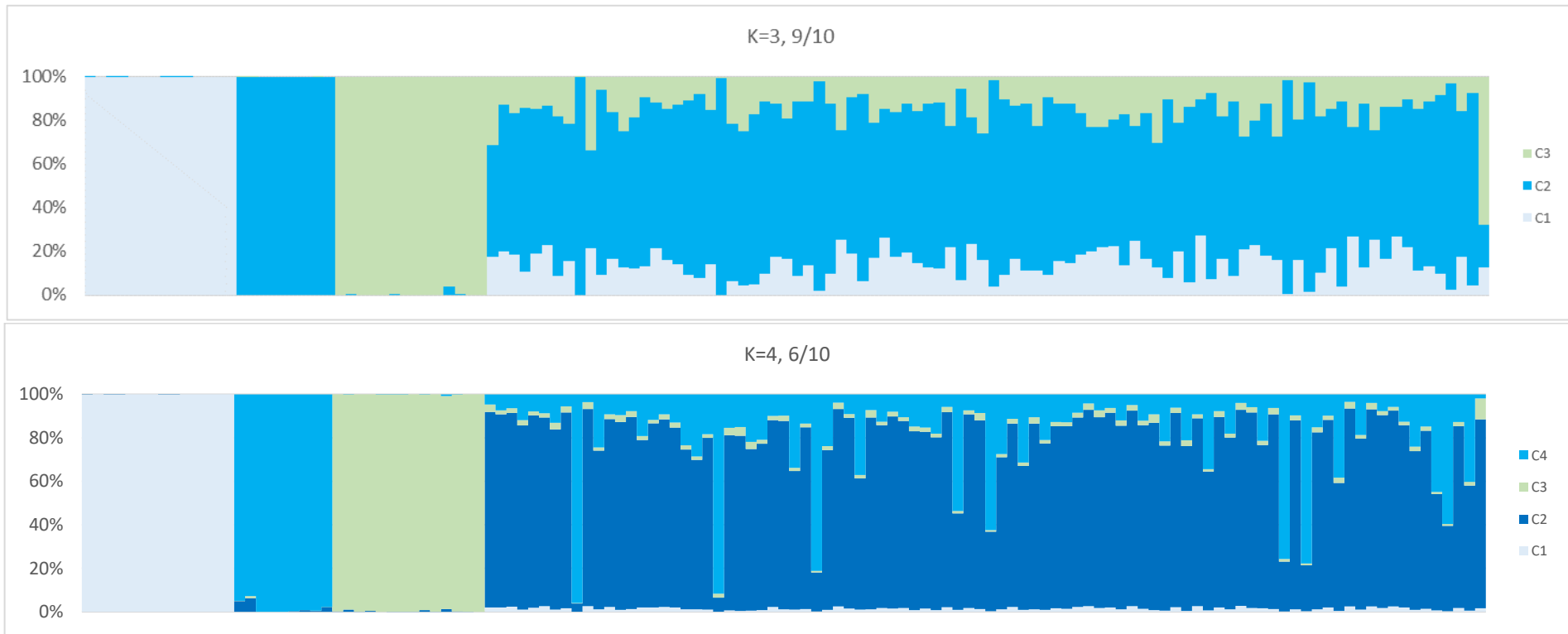


Fig. 15: Structure runs with priors for *T. erraticum* group. Parameters for Structure Runs: 10 runs, 200k Burn-in Iterations, 200k Iterations, model updated only with Reference dataset (POPFLAG on), no admixture. Upper panel shows results for K=3, major clustering model from Clumpak (9 runs/10). Lower panel shows results for K=4, major clustering model (6/10). Order of the individuals is from left to right: Reference dataset (*erraticum*, *madeirense*, *subboreale*) Test dataset in order of species obtained from Consensus between PCA, DAPC and GENECLASS 2 (*erraticum*, *madeirense*, *subboreale*), then in alphabetical order of sample name. Here in the upper panel Cluster 1 = *erraticum*, Cluster 2 = *madeirense*, Cluster 3 = *subboreale*. In the lower panel, the additional Cluster 4 = *madeirense*.





Fig. 16: Structure runs without priors for *T. erraticum* group. Parameters for Structure Runs: 10 runs, 200k Burn-in Iterations, 200k Iterations, no population prior, admixture. Upper panel shows results for K=3, single clustering model from Clumpak (10 runs/10). Lower panel shows results for K=4, single clustering model (10/10). Order of the individuals is from left to right: Reference dataset (*erraticum*, *madeirense*, *subboreale*) Test dataset in order of species obtained from Consensus between PCA, DAPC and GENECLASS 2 (*erraticum*, *madeirense*, *subboreale*), then in alphabetical order of sample name. Here in the upper panel Cluster 1 = *erraticum*, Cluster 2 = *madeirense/subboreale*, Cluster 3 = *subboreale*. In the lower panel, Cluster 1 = *erraticum*, Cluster 2 = *madeirense*, Cluster 3 = *madeirense*, Cluster 4 = *subboreale*.