Genetic hubs: a phylogeographer's widget for pinpointing of ancestral populations

Mikula Ondřej[1,2,3,*]

1. *Institute of Vertebrate Biology of the Czech Academy of Sciences, Brno, Czech Republic*
2. *Institute of Animal Physiology and Genetics of the Czech Academy of Sciences, Brno, Czech Republic*
3. *Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic*

* correspondence: onmikula@gmail.com

Abstract

Statistical phylogeography benefits from the development of increasingly realistic models of spatially structured genetic variation. Their fitting, however, is computationally demanding and requires population and/or genomic sampling that is not available for many species of interest. 'Genetic hubs' is a method that can be used for exploratory analyses of various kinds of genetic data, including those as typical in mitochondrial phylogeography, i.e. many small samples of single locus genotypes scattered throughout the species distributional range. 'Genetic hubs' allows to quantify and visualize gradients of genetic variation with the aim to pinpoint possible origin of expansion. It estimates local genetic variability as an accessibility of all genetic variation from the site in question and it allows to take dissimilarity of genotypes into account. The method represents fast and versatile tool that can be used whenever history of range expansion is assumed to shape the observed distribution of genetic variation and it is useful especially for preliminary analyses whose purpose is to provide sound basis for formulation of testable hypotheses and design of follow-up studies.

Background

Phylogeography attempts to understand processes and historical events that shaped spatial structure of genetic variation and, compared to other branches of population genetics, it focuses on their timing and geographical setting (Avise, 2000; Knowles & Maddison, 2002). One important goal in such endeavor is to estimate the origin of expansion, i.e. the location of ancestral population, from which the rest of distribution range was colonized.

The range expansion results in a serial founder effect that can be thought as a spatial analog of genetic drift (Slatkin & Excoffier, 2012), which creates gradients of decreasing genetic diversity, increasing linkage disequilibrium and flattening ancestral allele frequency spectrum. All these phenomena can be used when searching for the origin of expansion of humans and the observed trends match expectations based on paleontologically well-evidenced out-of-Africa dispersal scenario (DeGiorgio, Jakobsson & Rosenberg, 2009; Jakobsson et al., 2008; Ramachandran et al., 2005). These analyses could be performed because of at least hundreds of loci genotyped in hundreds of individuals from tens of populations. The source population can be also identified by analysis of fewer loci genotyped in many individuals from several populations, an approach applied in island phylogeography (Kuo et al., 2015; Rodríguez et al., 2013). There are, however, numerous phylogeographic data sets whose sampling is comprehensive, yet sparse. They cover more or less evenly most of the distribution range, but consist of one to a few individuals per population and, at best, a handful of loci. In such cases, no estimates of local genetic diversity are available and conclusions about range expansions are largely based on researcher's intuition and common wisdom.

Genetic hubs

Here, I present a method that approximates trends in genetic diversity by integrating information over such globally comprehensive, yet locally sparse sampling. The local diversity is approximated by a distance that *must not* to be travelled from a particular place to access all the observed genetic variation. If all alleles are present at the site, a virtual agent must not to take any travel to access the whole variation. The more depleted is gene pool at the site, the longer are travels of the agent, especially if the site is at the periphery of species distribution. The algorithm is called Genetic hubs as the place of highest diversity (="genetic hub") has the same property of "being close to everything" as a hub in public transport network.

In practice, the species distribution is represented by a spatial graph whose vertices correspond to sites and edges to travels between them. The diversity score is estimated separately for each locality. First, it is figured out, which alleles are missing there and what is the shortest path to each of them. Then the graph is effectively reduced just to edges that appear in any of the shortest paths and each edge in the reduced graph is assigned its own weight, which is equal to proportion of variation accessed through the edge. For instance, if the edge appears on the shortest paths to one out of four alleles, its weight is 0.25. The road to be taken for all the remaining genetic variation is then calculated as $\sum_{i=1}^{r} d_i w_i$, where d_i is the length of ith edge, w_i is its weight and the summation goes across *r* edges of the reduced graph. This sum is contrasted to its theoretical maximum, which is calculated in the same way, but assuming all variation at the very most distant locality and no variation at the locality in question. This is equivalent to $\sum_{i=1}^{m} d_i$, sum of lengths of *m* edges forming the shortest path to the most distant site. Finally, the diversity score is calculated as $1 - \frac{\sum_{i=1}^{r} d_i w_i}{\sum_{i=1}^{m} d_i}$. In other words, the length of travel to be taken by the agent is calculated, expressed as a proportion of its theoretical maximum and subtracted from unity to make the value proportional (not inversely proportional) to local diversity. The calculation of diversity score is demonstrated in Figure 1.

An obvious, and often essential, extension of the algorithm is to take dissimilarities between alleles into account. This is achieved by modification of the weight calculation, where the proportion of variation accessed through the edge is calculated not only from presences and absences of alleles but also from their dissimilarity matrix. The matrix is subjected to multidimensional scaling and every allele is assigned a vector of values that specify its position in the resulting multidimensional Euclidean space. The space dimensionality is equal to the number of positive eigenvalues in spectral decomposition of the matrix. If the dissimilarity measure is a true metric, all variation can be represented in this space, otherwise some information is lost. In fact, the decision whether or not to weight edges by allele dissimilarities is a delicate one. If we assume recent and fast spread from a single panmictic population, allele dissimilarities do not tell us anything about the expansion. All alleles were well mixed at the beginning and those retained in the same or a nearby population may be very similar as well as largely different from each other. On the other hand, if the average rate of expansion is slow enough to be comparable with effective mutation rate, new alleles are arising along the way and their similarity bears imprint of the expansion process.

Another issue is the choice of spatial graph. The first obvious possibility is the fully wired graph, where all sites are direct neighbors separated just by their physical distance on the Earth surface. It should be appreciated, however, this is not a 'neutral', 'uninformative' or even 'universal' choice. The reason is that any edge passing through inhospitable environment introduces bias as it spuriously indicates contact where none exists. An ideal choice would be probably a fully wired graph with edge lengths modified to reflect landscape resistance to migration (McRae, 2006). Information about the long-term landscape resistance is not easily available, however. A viable alternative is

therefore to modify graph structure, either *ad hoc* or according to some predefined criterion, to avoid long-ranging shortcuts that are more risky to introduce substantial bias. The only requirement is the graph has to remain fully connected so that every vertex can be reached, directly or indirectly, from any other vertex. In any case it is advisable to decide about the graph structure prior to any calculation.

The knowledge of genetic hub location is useful especially if it is not contingent on presence of a single sample but supported by the pattern of diversity as a whole. Thus, it is advisable to rerun the algorithm on reduced data sets with sample present at the genetic hub site or at the neighboring sites omitted. This is in fact a jackknifing procedure, although limited to the neighborhood of the genetic hub. Every jackknifed sample is omitted, once at time, but its site is retained as empty one so even after omission of the hub sample itself, the hub's position may be unchanged. When jackknifing is completed, it is useful to compare genetic hub locations and quantify congruence of diversity trends, e.g. by Spearman correlation coefficient calculated on diversity scores. The omission of samples from sites far apart from the hub is unlikely to change its position and their inclusion into jackknifing could in fact create spurious support for the genetic hub location.

Although the possibility of using various dissimilarities and spatial graphs makes the method remarkably versatile, it has also its inherent limitations. Most importantly, the local genetic diversities are not estimated independently of each other, but instead, they are approximated under the assumption of a single range expansion (with no range expansion as a special case). If the spatial pattern was created, for instance, by population growth at some places and population decline at others, it would not be appropriate to estimate local diversities by integrating information across the whole area. In such case the local diversity is determined by local factors and has to be estimated from local data. Another common case when the algorithm is misled by its crucial assumption is the secondary contact of two expanding populations. Here, the hotspot of diversity is at the frontlines rather than at the origins of the expansions. Due to this limitation Genetic hubs do not allow formal comparison of different historical scenarios. The algorithm also pinpoints just a single site as the hub, although it is likely that ancestral population occupied some larger area. The hub can be thought, therefore, as a centroid of the ancestral range. Finally, and most obviously, more or less even sampling intensity across the whole area is assumed, otherwise the peak of local diversity may be an artefact. The weighting by allele dissimilarity makes the method more robust in this respect.

Genetic hubs are available in the form of open-source package GenHubs for R (R Core Team 2018), which is accessible via CRAN (...). It allows to estimate genetic hub location under a range of settings, namely using different spatial graphs with or without weighting by allele dissimilarity. Apart from the core GenHubs function it offers also jackknifing procedure and associated plotting methods.

Examples

Overall, Genetic hubs are intended mostly for data exploration and visualization. They are expected to be useful for at least three purposes: (1) a visualization making researcher's impressions explicit; (2) a preliminary hypothesis formulation where the goal is to pinpoint areas and populations worth of more detailed study; (3) comparative analyses of multiple species or loci where repeatedly spotting the same place as a center of expansion gives higher credit to the implicit assumption of a single expansion scenario. Note also that the virtual agent can travel from places with no data. The algorithm can be therefore used in a predictive manner to estimate local diversities at spots where the species occurrence is known or assumed but from which samples are not available. This feature enhances usefulness of Genetic hubs for the preliminary hypothesis formulation as it allows to assess, which non-sampled locality might be of the greatest interest.

First, we demonstrate usefulness of Genetic hubs on a well-studied example of post-glacial colonization of Europe. Comparative phylogeography of various vertebrate species suggested putative refugia to be located in the Mediterranean (Hewitt, 1999; Schmitt, 2007; Taberlet, Fumagalli, Wust-Saucy & Cosson, 1998), but also in more northerly regions (McDevitt et al., 2012; Kotlík et al., 2006). This view is supported by the fossil record (Knitlová & Horáček, 2017; Sommer & Nadachowski, 2006; Tougard, Renvoise, Petitjean & Quere, 2008).

Two species of hedgehogs (*Erinaceus*) live in Europe: the Northern White-breasted hedgehog (*E. roumanicus*) lives in the east of Europe, Balkans and in the central Europe, where it meets with its western counterpart, the Western European Hedgehog (*E. europaeus*). In addition there is a pronounced phylogeographic pattern within *E. europaeus* with three distinct mitochondrial lineages (Seddon, Santucci, Reeve & Hewitt, 2001). The whole pattern was interpreted as a result of colonization from refugia located in the Balkans (*E. roumanicus*), Italy (E1 lineage of *E. europaeus*) and Iberia (E2 lineage of *E. europaeus*). The third lineage of *E. europaeus* is confined to Sicilia and won't be further considered here. The data set reanalyzed here consists of 423 georeferenced records of 100 haplotypes of 426 bp long sequences of mitochondrial control region, originally published by Seddon et al. (2001; 2002), Bolfíková and Hulva (2012) and Černá Bolfíková et al. (2017). The calculation was done in both unweighted and weighted fashion, the latter being based on Kimura two-parameter distances (Kimura, 1980) between haplotypes. In place of spatial graph I used Gabriel graph (Gabriel & Sokal, 1969) with some links crossing the sea manually deleted. Results of the genetic hub analysis are presented in Figure 2. The genetic hub of E2 lineage of *E. europaeus* is located either in Iberia (weighted variant) of in southern France (unweighted variant) as expected from the existing biogeographic scenarios. Location of the genetic hub was surprising in the E1 lineage (in both variants) as it was found in southern Germany instead of Italy. This does not indicate, however, Italy was not a refugium of *E. europaeus* during the last glacial. It only suggests the population from which the rest of the distribution range of E1 lineage was colonized could live more northerly. It is also worth to consider an effect of unbalanced sampling as there are much fewer sites in Italy. The genetic hub of *E. roumanicus* (in both variants) was at the Adriatic coast of Croatia, in accord with the assumed Balkan location of the glacial refugium. In the weighted variant the gradient of diversity is apparent especially in E2 lineage of *E. europaeus* (in the expected north-eastern direction), but in the other two units it also shows interpretable features. In the E1 lineage its minimum is in Scandinavia which is sure to be colonized late and in *E. roumanicus* it has its minimum at the very east suggesting eastward colonization of regions that were inhospitable for long time due to its continental climate. In the unweighted variant (not shown) Iberia also appeared to be colonized late by E2 lineage, which is arguably an artefact caused by not taking allele dissimilarities into account.

The phylogeographic structure of the Wood Mouse (*Apodemus sylvaticus*) also likely results from postglacial colonization. Using 981 cytochrome *b* sequences, Herman et al. (2017) identified six phylogeographic lineages, three of which are analyzed here. The south-eastern lineage is distributed in Italy and Balkans, suggesting glacial refugium somewhere in that regions (Michaux, Magnanou, Paradis, Nieberding & Libois, 2003), the central lineage dominating the western part of continental Europe might spread from a refugium in Iberia or southern France (Michaux et al., 2003) and the newly discovered peripheral lineage is distributed in the British Isles and the eastern Europe which was interpreted to be due to replacement by the central lineage (Herman et al., 2017). Overall, Herman and co-workers were sceptic about possibility of locating glacial refugia from their data set. Indeed, in spite of being impressive in size it comprised only a single mitochondrial locus, which inevitably bears only limited information about population-level processes. However, it is exactly that kind of data set for which Genetic hubs are suited best and where they can bring answer that is provisional, but obtained in a transparent and reproducible manner. The re-analyzed data set

includes 445 georeferenced records of 383 cytochrome *b* haplotypes (1140 bp long) from three out of six phylogeographic lineages. The other three were either narrowly localized (Sicilia, Channel Islands) or extraterritorial (northern Africa) and were not considered here. The procedure was the same as in the hedgehog data set. As may be seen in Figure 3, central and south-eastern lineages have their genetic hubs in the presumed refugial regions, southern France and Italy, respectively. In the weighted variant gradient of diversity is well apparent (Figure 3), but the same was the case in unweighted variant (not shown here). Interpretation of the genetic hub location is more complicated in the case of the peripheral lineage. The algorithm unequivocally pinpoints sites in Wales (weighted variant) or even Scotland (unweighted variant), that is in regions which are expected inhospitable for wood mice until the beginning of the Holocene. If the replacement hypothesis of Herman et al. is true, however, the history of these populations is at odds with assumptions of the method, because in such case the geographic pattern of variation was not shaped only by expansion, but also the subsequent wave of local extinctions. Therefore, the genetic hub cannot be interpreted as co-incident with the origin of expansion.

In these two examples Genetic hubs served to examine whether geographical distribution of genetic variation fits *a priori* expectations. More explorative (rather than confirmatory) use of the algorithm is illustrated by a small comparative study, which involves three rodents associated with forests and woodlands in the Tanzanian Eastern Arc Mountains – namely *Grammomys surdaster*, *Mus triton* and *Praomys delectorum*. The Eastern Arc is a chain of more or less isolated mountain massifs surrounded by savanna, today often turned into agricultural landscape, while in higher elevations they are covered by forests (Platts et al., 2011). More specifically, the focus was on the southern part of the chain as the species considered are absent or represented by genetically distinct lineages in more northerly massifs.

The analyzed data sets were relatively small (18–36 georeferenced records of 16–30 cytochrome *b* haplotypes), but they still allow explicit phylogeographic hypotheses to be formulated (Figure 4). I used the same options as before, except for the Gabriel graph was not manually modified and only weighted genetic hubs were calculated. The data were taken from several published studies (Bryja et al., 2014; 2017; Krásová et al., 2018; Sabuni, Aghová, Bryjová, Šumbera & Bryja, 2018) and supplemented by new records (sequences available in GenBank, accession numbers: XXXXXXXX–XXXXXXXX). Genetic hubs of *G. surdaster* and *P. delectorum* are in the north, while genetic hub of *M. triton* is in the south of the mountain chain. Note, however, that distribution of variability in *M. triton* is not entirely monotonic which calls into question the assumption of a single range expansion as the only force operating behind. In contrast, patterns observed in the other two species seem to be pronounced and monotonic, which is consistent with relatively recent expansion in the southwards direction.

Jackknifing largely supported the location of genetic hubs in the cases of *Erinaceus* and *A. sylvaticus*. It was conducted with the second order neighborhood, i.e. with sites that were in the graph either directly linked to the genetic hub or separated by at most one other site. In *Erinaceus*, only a few of the jackknifed genetic hubs were displaced from their original location and they always stayed in the same region (not shown). Also the rank correlation between original and jackknifed scores was always very high ($\rho \geq 0.99$). In *A. sylvaticus* the correlation was also high: $\rho \geq 0.98$ in the central and peripheral lineages and $\rho \geq 0.84$ in the south-eastern lineage. Again, only a minor proportion of jackknifed genetic hubs were shifted in location and only one of them deserves special attention. Namely, in both weighted and unweighted variant of the south-eastern lineage analysis, one out of six jackknife replicates resulted in the shift of genetic hub across the sea, from central Italy to the coast of Montenegro.

Results calculated for the three Eastern Arc Mountain species were less robust (Figure 5). The jackknifing here was conducted with just the first order neighborhood (=direct neighbors in the graph), but it still resulted in substantial shifts of genetic hub location. In both *G. surdaster* and *P. delectorum* two out of three jackknife replicates were substantially shifted southward and in M. triton two out of four were shifted northward. Score correlations might be as low as 0.50 in *P. delectorum* and 0.53 in *M. triton*, although they were reasonably high ($\rho \geq 0.90$) in *G. surdaster*. The results are therefore more dependent on the particular samples in hand and more individuals as well as sampling sites should be employed to obtain more reliable estimates of genetic diversity gradients.

Discussion

Genetic hubs method is introduced here as a tool for exploratory data analysis for phylogeography, but also landscape genetics. The two disciplines have similar goals, but they work on different time scales. Landscape genetics is focused on the present and very recent past and thus it assumes spatial arrangement of populations to be more or less static, ancestral alleles to be possibly present in data and spatial variation to be determined mostly by segregation and ongoing migration. If the signal of expansion is present in such data, its origin can be approximated by the unweighted variant of the genetic hub analysis, i.e. without taking allele dissimilarities into account. Phylogeography, instead, is focused on historical processes. It assumes populations shifted largely in location on the time scale of interest, ancestral alleles to be already replaced by their mutational variants and spatial variation to be strongly affected by colonization-extinction process. In such case, the weighted variant of Genetic hubs is more appropriate.

As already mentioned, the most powerful techniques for inference of colonization routes require either multiple individuals to be sampled from every population or a large number of genomic markers to be sampled from every individual. If there was a large set of populations with precise estimates of genetic diversity, one could use any spatial interpolation method (e.g. Miller & Wood, 2014) to identify gradient of variation and its peak. This is seldom the case, however. Alternatively, the origin of range expansion can be also identified from asymmetries in spatial distribution of binary alleles (Peter & Slatkin, 2013), which requires large number of markers but not so many individuals per site. If there are few populations with multiple individuals structured coalescent (aka isolation-with-migration) models (Beerli & Felsenstein, 2001; De Maio, Wu, O'Reilly & Wilson, 2015; Kühnert, Stadler, Vaughan & Drummond, 2016) can be used to identify the source population. If there are few individuals per population, but with a large number of loci genotyped, one can use admixture modelling to get an idea which population was the ancestral one. This can be achieved by proper interpretation of clusters delimited on the basis of Hardy-Weinberg expectations (Pritchard, Stephens & Donnelly, 2000; Guillot, Estoup, Mortier & Cosson, 2005) or by exploiting information about physical location of loci on chromosomes and analysis of introgression blocks (Hellenthal et al., 2014).

Genetic hubs may be used for exploration of any data, which can be converted to lists of alleles present at particular sites (unweighted variant) or to distances between unique genotypes and ultimately components of variation attributable to them (weighted variant). Nevertheless, the method is arguably most useful when dealing with data sets that cannot be analyzed by the abovementioned methods. This is the case for mitochondrial phylogeography that flourished for two decades since 1990 and still represents an important initial step in biogeographical and systematic studies. Typical sampling here is globally comprehensive, but locally sparse: it often includes tens of sampling sites covering substantial part of the species distribution, but only one to a few individuals per site. Whereas the high number of sites precludes the use of parameter-rich structured

coalescent, the low number of loci precludes the use of methods that rely on genomic sampling and sequencing of a few more loci does not change the matters substantially. In contrast, genetic hub analysis may provide meaningful results even in such unfavorable situation. Although any single locus carries only partial information about population level processes, some loci can be informative on their own and their conflicting signals may be interpretable as it is the case for maternally, paternally and bi-parentally inherited markers (Toews & Brelsford, 2012).

Nevertheless, there is yet another methodology for ancestral location inference, able to process the very same data sets as just discussed. Lemey, Rambaut, Drummond and Suchard (2009) treated location as a discrete trait, whose evolution unfolds along with the phylogeny itself. In other words, the migration between geographic sites is modelled in the very same way as mutation process. In continuous space an analogous approach treats location as a bivariate trait evolving by diffusion over a plane (Lemey, Rambaut, Welch & Suchard, 2010) or on a sphere (Bouckaert, 2016). Both approaches can be further bridged by modelling migration as a diffusion over a graph of sites covering densely the area suitable for migration (Bouckaert, Bowern & Atkinson, 2018). This methodology, implemented in the Bayesian framework, is open to numerous extensions and modifications including informative priors on particular rates, fixing some of them to zero and allowing them to be asymmetric (Bouckaert et al., 2018; Edwards et al., 2011; Lemey et al., 2009). As such it holds a promise to provide probabilistic estimates of ancestral locations, not only algorithmic ones as provided by Genetic hubs. It remains to be integrated, however, with models taking into account population effective sizes (Kühnert et al., 2016). When population size is omitted, presence of highly divergent haplotypes at the same site is implicitly interpreted as the result of intensive migration rather than as the remnant of ancestral polymorphism. This simplification can greatly bias results, although it may be appropriate for viruses whose mutation and migration rates are comparable (Lemey et al., 2009; Pybus et al., 2012) or for organisms, whose population sizes at remote discrete sites are very small compared to those between the sites (e.g. polar bears analyzed by Edwards et al., 2011).

In summary, the probabilistic inference of ancestral populations and colonization routes is possible when population and/or genomic sampling is intensive and it might become possible for data sets with extensive sampling as well. Genetic hubs method focuses on a part of this problem, namely, on the location of the expansion origin. In this respect, it wants to serve the same purpose in phylogeography as the neighbor-joining method (Saitou & Nei, 1987) does in phylogenetics. It is intended as an approximate, yet fast and versatile, alternative to model-based methods, useful especially for exploratory and preliminary analyses, which can provide good starting points for future studies.

Acknowledgements

Financial support for this study was provided by the Czech Science Foundation, project no. 18-17398S. I thank to J. Bryja for commenting the first version of the manuscript.

References

Avise, J.C. (2000). *Phylogeography: The history and formation of species*. Cambridge, MA: Harvard Univ. Press.

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the*

National Academy of Sciences of the United States of America, 98(8), 4563–4568. doi: 10.1073/pnas.081068098

Bolfíková, B., & Hulva, P. (2012). Microevolution of sympatry: landscape genetics of hedgehogs *Erinaceus europaeus* and *E. roumanicus* in Central Europe. *Heredity*, 108(3), 248-255. doi: 10.1038/hdy.2011.67

Bouckaert, R. (2016). Phylogeography by diffusion on a sphere: whole world phylogeography. *PeerJ*, 4, e2406. doi: 10.7717/peerj.2406

Bouckaert, R. R., Bowern, C., & Atkinson, Q. D. (2018). The origin and expansion of Pama-Nyungan languages across Australia. *Nature Ecology & Evolution*, 2(4), 741-749. doi: 10.1038/s41559-018-0489-3

Bryja, J., Mikula, O., Patzenhauerová, H., Oguge, N., Šumbera, R., & Verheyen, E. (2014). The role of dispersal and vicariance in the Pleistocene history of an East African mountain rodent, *Praomys delectorum*. *Journal of Biogeography*, 41(1), 196-208. doi: 10.1111/jbi.12195

Bryja, J., Šumbera, R., Kerbis Peterhans, J.C., Aghová, T., Bryjová, A., Mikula, O., ..., Verheyen, E. (2017). Evolutionary history of the thicket rats (genus *Grammomys*) mirrors the evolution of African forests since late Miocene. *Journal of Biogeography*, 44(1), 182–194. doi: 10.1111/jbi.12890

Černá Bolfíková, B., Eliášová, K., Loudová, M., Kryštufek, B., Lymberakis, P., Attila, D., & Hulva, P. (2017). Glacial allopatry vs. postglacial parapatry and peripatry: the case of hedgehogs. *PeerJ*, 5, e3163. doi: 10.7717/peerj.3163

DeGiorgio, M., Jakobsson, M., & Rosenberg, N.A. (2009). Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38), 16057-16062. doi: 10.1073/pnas.0903341106

De Maio, N., Wu, C.-H., O'Reilly, K.M., & Wilson, D. (2015). New Routes to Phylogeography: A Bayesian structured coalescent approximation. *PLoS Genetics*, 11(8), e1005421. doi:10.1371/journal.pgen.1005421

Edwards, C.J., Suchard, M.A., Lemey, P., Welch, J.J., Barnes, I., Fulton, T.L., ..., Shapiro, B. (2011). Ancient hybridization and an Irish origin for the modern Polar Bear matriline. *Current Biology*, 21(15), 1251-1258. doi: 10.1016/j.cub.2011.05.058

Gabriel, K.R., & Sokal, R.R. 1969. A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18(3): 259–278. doi: 10.2307/2412323

Guillot, G., Estoup, A., Mortier, F., & Cosson, J.F. (2005). A spatial statistical model for landscape genetics. *Genetics*, 170(3), 1261-1280.

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, 343(6172), 747-751. doi: 10.1126/science.1243518

Herman, J.S., Johannesdottir, F., Jones, E.P., McDevitt, A.D., Michaux, J.R., White, T.A., Wojcik, J.M., Searle, J.B. (2017). Post-glacial colonization of Europe by the wood mouse, *Apodemus sylvaticus*: evidence of a northern refugium and dispersal with humans. *Biological Journal of the Linnean Society*, 120(2), 313-332. doi: 10.1111/bij.12882

Hewitt, G.M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, 68(1-2), 87-112. doi: 10.1111/j.1095-8312.1999.tb01160.x

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.-C., ..., Singleton, A.B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181), 998–1003. doi: 10.1038/nature06742

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2), 111-120. doi: 10.1007/BF01731581

Knitlová, M., & Horáček, I. (2017). Late Pleistocene-Holocene paleobiogeography of the genus *Apodemus* in Central Europe. *PLoS ONE*, 12(3), e0173668. doi:10.1371/journal.pone.0173668

Knowles, L.L., & Maddison, W.P. (2002). Statistical phylogeography. *Molecular Ecology*, 11(12), 2623-2635. doi: 10.1046/j.1365-294X.2002.01637.x

Kotlík, P., Deffontaine, V., Mascheretti, S., Zima, J., Michaux, J.R., & Searle, J.B. (2006). A northern glacial refugium for bank voles (*Clethrionomys glareolus*). *Proceedings of the National Academy of Sciences of the United States of America*, 103(40), 14860-14864. doi: 10.1073/pnas.0603237103

Krásová, J., Mikula, O., Mazoch, V., Bryja, J., Říčan, O., & Šumbera R. 2018. Evolution of the Grey-bellied Pygmy Mouse group: highly structured molecular diversity with predictable geographic ranges but morphological crypsis. *Molecular Phylogenetics and Evolution*,

Kühnert, D., Stadler, T., Vaughan, T.G., & Drummond, A.J. (2016). Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Molecular Biology and Evolution*, 33(8), 2102-2116. doi: 10.1093/molbev/msw064

Kuo, H.C., Chen, S.F., Fang, Y.P., Cotton, J.A., Parker, J.D., Csorba, G., ..., Rossiter, S.J. (2015). Speciation processes in putative island endemic sister bat species: false impressions from mitochondrial DNA and microsatellite data. *Molecular Ecology*, 24(23), 5910-5926, doi: 10.1111/mec.13425

Lemey, P., Rambaut, A., Drummond, A.J., & Suchard, M.A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9), e1000520. doi: 10.1371/journal.pcbi.1000520

Lemey, P., Rambaut, A., Welch, J.J., & Suchard, M.A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8), 1877-1885. doi: 10.1093/molbev/msq067

McDevitt, A.D., Zub, K., Kawałko, A., Oliver, M.K., Herman, J.S., & Wojcik, J.M. (2012). Climate and refugial origin influence the mitochondrial lineage distribution of weasels *Mustela nivalis* in a phylogeographic suture zone. *Biological Journal of the Linnean Society*, 106(1): 57–69. doi: 10.1111/j.1095-8312.2012.01840.x

McRae, B.H. (2006). Isolation by resistance. *Evolution*, 60(8), 1551–1561. doi: 10.1111/j.0014-3820.2006.tb00500.x

Michaux, J.R., Magnanou, E., Paradis, E., Nieberding, C., & Libois R. (2003). Mitochondrial phylogeography of the Woodmouse (*Apodemus sylvaticus*) in the Western Palearctic region. *Molecular Ecology*, 12(3), 685-697. doi: 10.1046/j.1365-294X.2003.01752.x

Miller, D.L., & Wood, S.N. (2014). Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics*, 21(4), 715-731. doi: 10.1007/s10651-014-0277-4

Peter, B.M., & Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution*, 67(11), 3274-3289. doi: 10.1111/evo.12202

Platts, P.J., Burgess, N.D., Gereau, R.E., Lovett, J.C., Marshall, A.R., McClean, C.J., ..., Marchant, R. (2011). Delimiting tropical mountain ecoregions for conservation. *Environmental Conservation*, 38(3), 312-324. doi: 10.1017/S0376892911000191

Platts, P.J., Burgess, N.D., Gereau, R.E., Lovett, J.C., Marshall, A.R., McClean, C.J., ..., Marchant, R. (2011). Data from: Delimiting tropical mountain ecoregions for conservation. *Dryad Digital Repository*. doi: 10.5061/dryad.c5310

Pritchard, J.K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.

Pybus, O.G., Suchard, M.A., Lemey, P., Bernardin, F.J., Rambaut, A., Crawford, F.W., ..., Delwart, E.L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37), 15066–15071. doi: 10.1073/pnas.1206598109.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44), 15942–15947. doi: 10.1073/pnas.0507611102

Rodríguez, V., Brown, R.P., Terrasa, B., Pérez-Mellado, V., Castro, J.A., Picornell, A., Ramon, M.M. (2013). Multilocus genetic diversity and historical biogeography of the endemic wall lizard from Ibiza and Formentera, *Podarcis pityusensis* (Squamata: Lacertidae). *Molecular Ecology*, 22(19), 4829-4841. doi: 10.1111/mec.12443

Sabuni, C., Aghová, T. Bryjová, A., Šumbera, R., & Bryja, J. (2018). Biogeographic implications of small mammals from Northern Highlands in Tanzania with first data from the volcanic Mount Kitumbeine. *Mammalia*, 82(4), 360-372. doi: 10.1515/mammalia-2017-0069

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406-425. doi: 10.1093/oxfordjournals.molbev.a040454

Seddon, J.M., Santucci, F., Reeve, N.J., & Hewitt G.M. (2001). DNA footprints of European hedgehogs, *Erinaceus europaeus* and *E. concolor*: Pleistocene refugia, postglacial expansion and colonization routes. *Molecular Ecology*, 10(9), 2187-2198. doi: 10.1046/j.0962-1083.2001.01357.x

Seddon, J.M., Santucci, F., Reeve, N.J., & Hewitt, G.M. (2002). Caucasus Mountains divide postulated postglacial colonization routes in the white-breasted hedgehog, *Erinaceus concolor*. *Journal of Evolutionary Biology*, 15(3), 463-467. doi: 10.1046/j.1420-9101.2002.00408.x

Schmitt, T. (2007). Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, 4, 11. doi: 10.1186/1742-9994-4-11

Slatkin, M., & Excoffier, L. (2012). Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics*, 191(1), 171–181. doi: 10.1534/genetics.112.139022

Sommer, R.S., & Nadachowski, A. (2006). Glacial refugia of mammals in Europe: evidence from fossil records. *Mammal Review*, 36(4), 251-265. doi: 10.1111/j.1365-2907.2006.00093.x

Taberlet, P., Fumagalli, L., Wust-Saucy, A., & Cosson, J. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, 7(4), 453–464. doi: 10.1046/j.1365-294x.1998.00289.x

Toews, D.P.L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, 21(16), 3907-3930. doi: 10.1111/j.1365-294X.2012.05664.x

Tougard, C., Renvoise, E., Petitjean, A., & Quere, J.-P. (2008). New Insight into the colonization processes of Common Voles: Inferences from molecular and fossil evidence. *PLoS ONE*, 3(10), e3532. doi: 10.1371/journal.pone.0003532

Figure legends

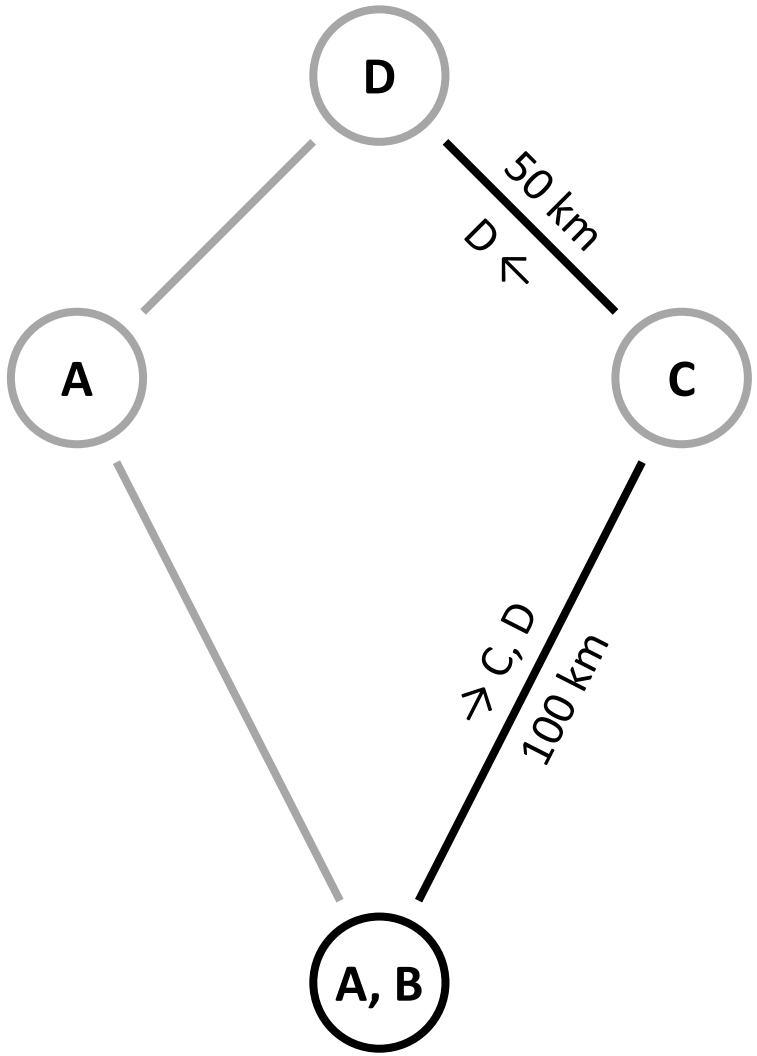
Figure 1. Example of diversity score calculation. Four sites (southern, western, northern and eastern) are shown, each with a different sample of A, B, C, D alleles. In the left panel, the arrows show the shortest paths from the southern site (with A, B alleles) to sites with other alleles. Every allele represent one fourth of total variation and the longer link is on the shortest path to two alleles (C, D), whereas the shorter link on the path to just a single allele (D). The diversity score is therefore calculated as 1 – (100 * 0.50 + 50 * 0.25) / 150 = 0.58. In the right panel, the same is shown for the northern site, whose diversity score turns out to be 0.50.

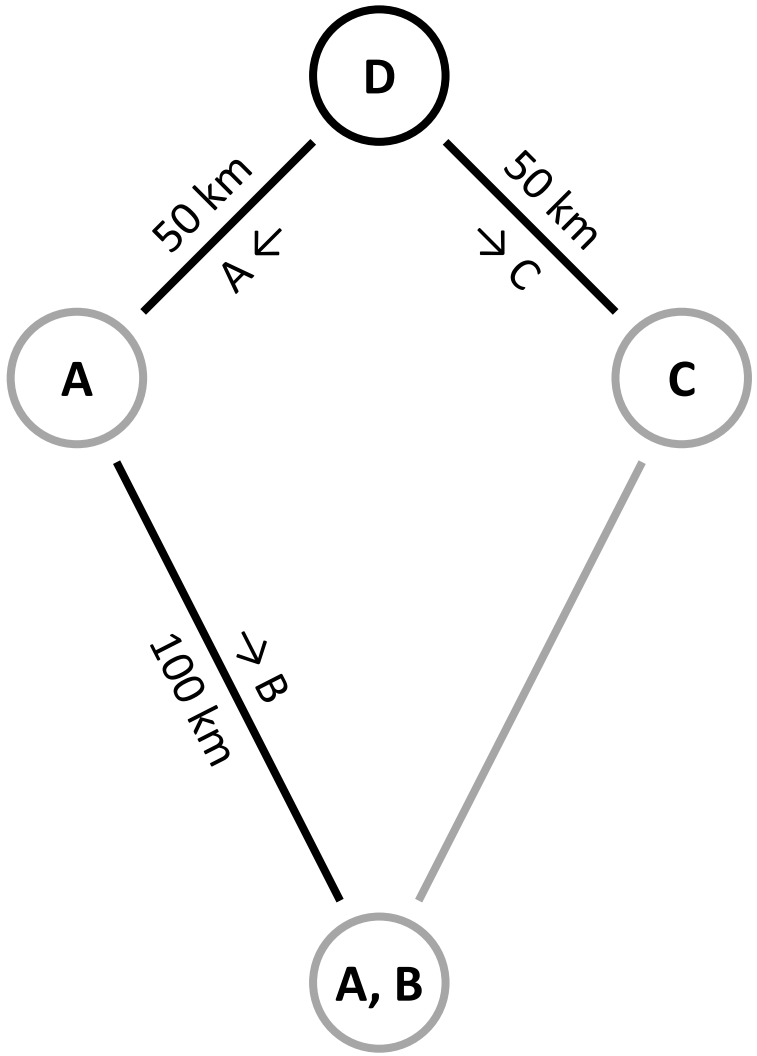
Figure 2. Genetic hubs of three species and lineages of hedgehogs (*Erinaceus*). The shades of colors indicate gradients of diversity for the weighted analysis, whose genetic hub is marked by the black star. Genetic hub from the unweighted analysis is marked by the purple star. Links belong to the three (partially overlapping) graphs upon which the calculation was based.

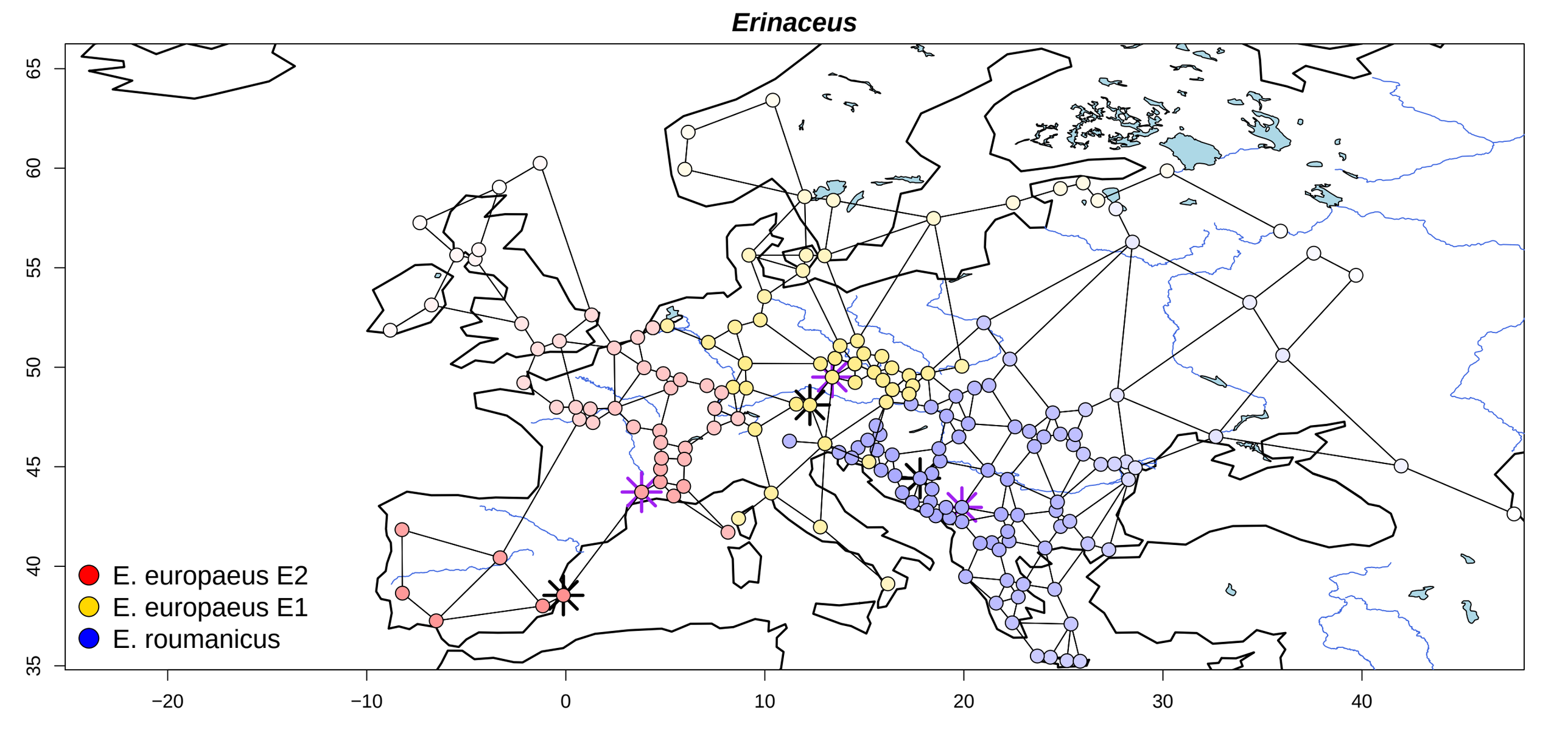
Figure 3. Genetic hubs of three lineages of wood mouse (*A. sylvaticus*). The sampling sites of the peripheral lineage are not shown completely due to overlap with the other two lineages. Symbols are the same as in Figure 2.

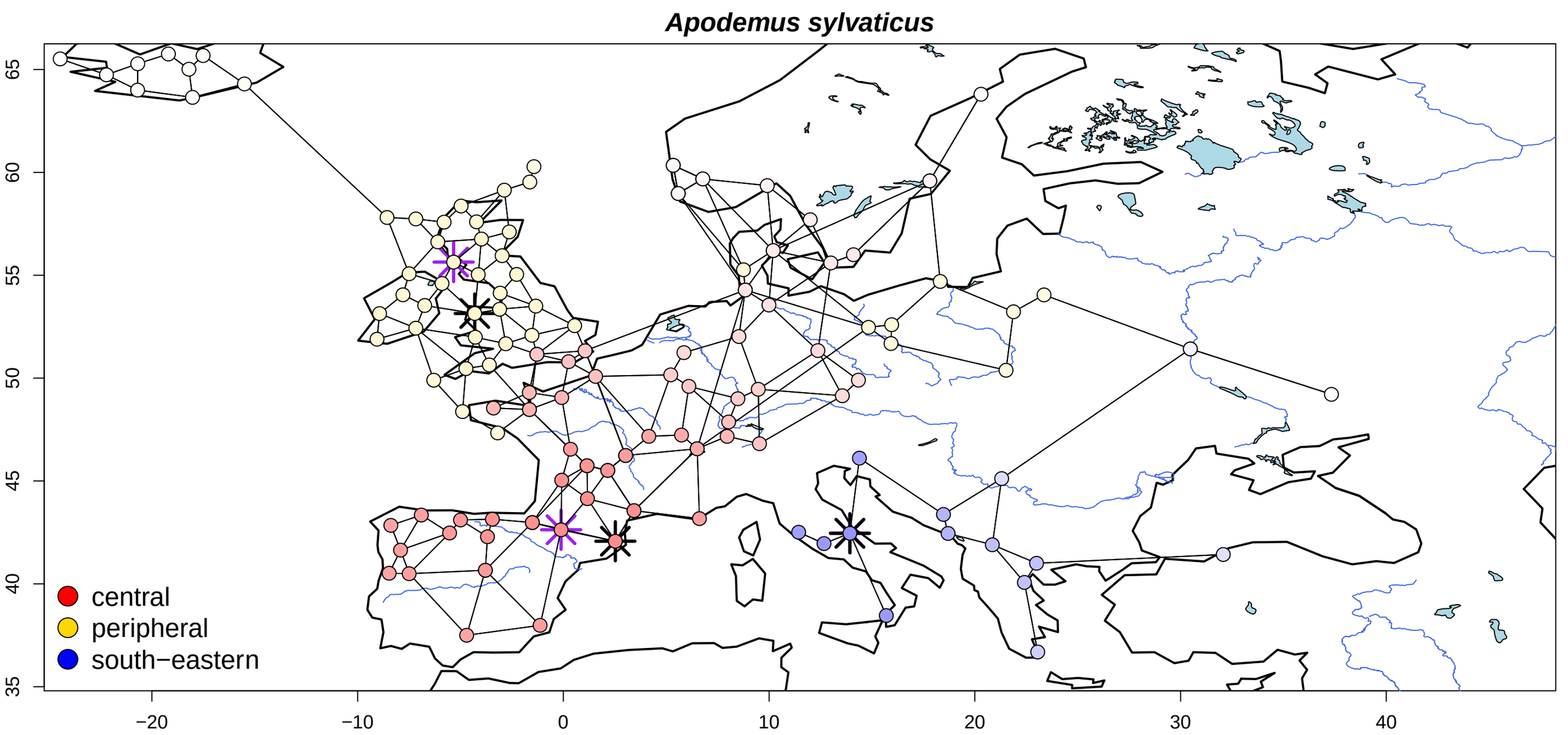
Figure 4. Genetic hubs of three rodent species living the forests and woodlands of the Eastern Arc Mountains.

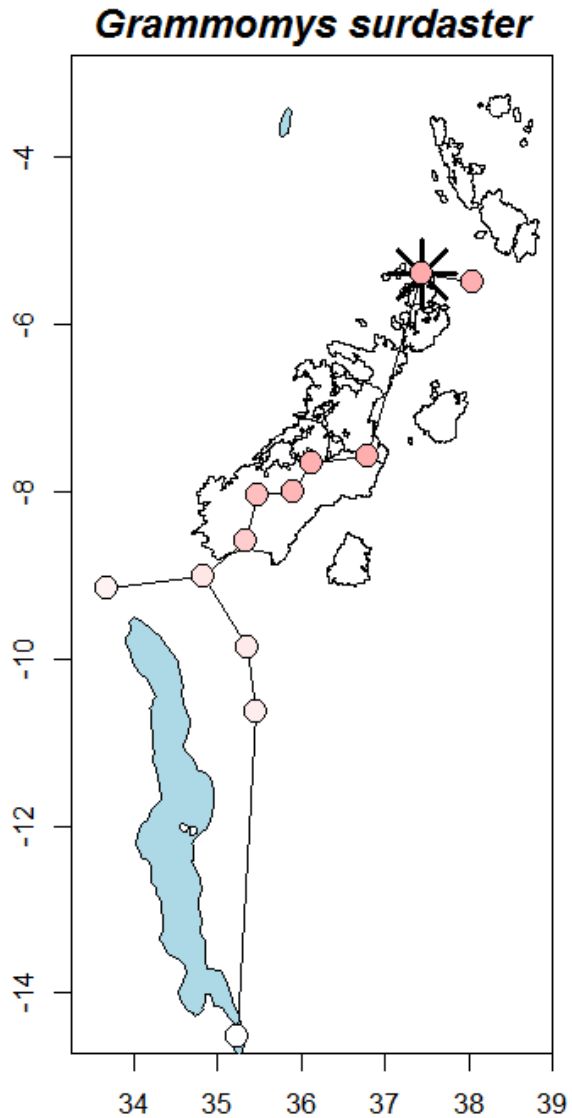
Figure 5. Jackknifing of genetic hubs of the Eastern Arc species. The size of the colored circle indicate the proportion of jackknifed genetic hubs estimated to lie at the particular site.

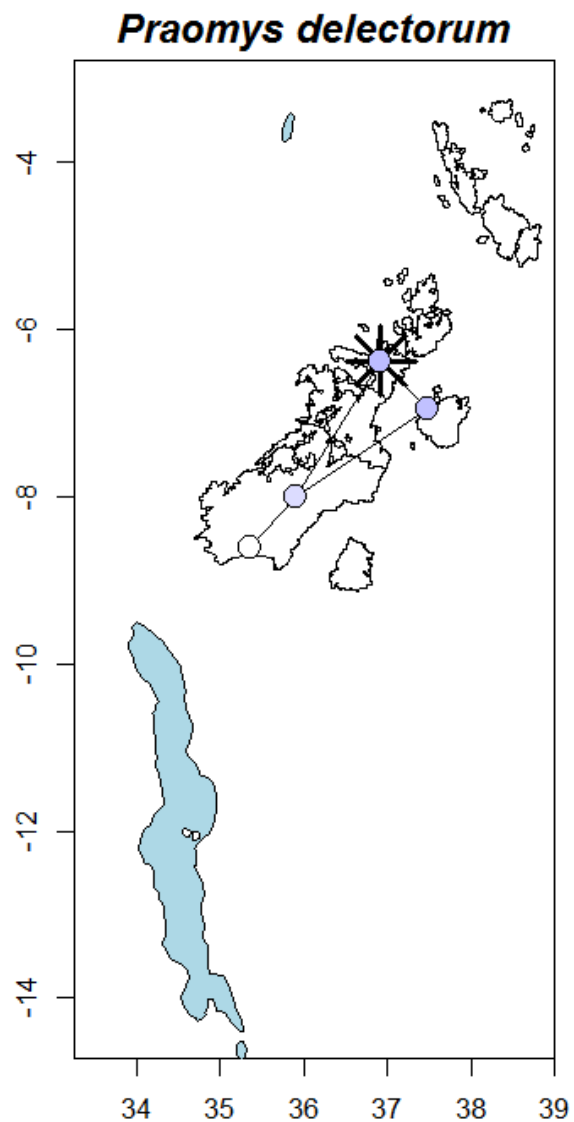

D
50 km
D ↖
A
C
↗ C, D
100 km
A, B

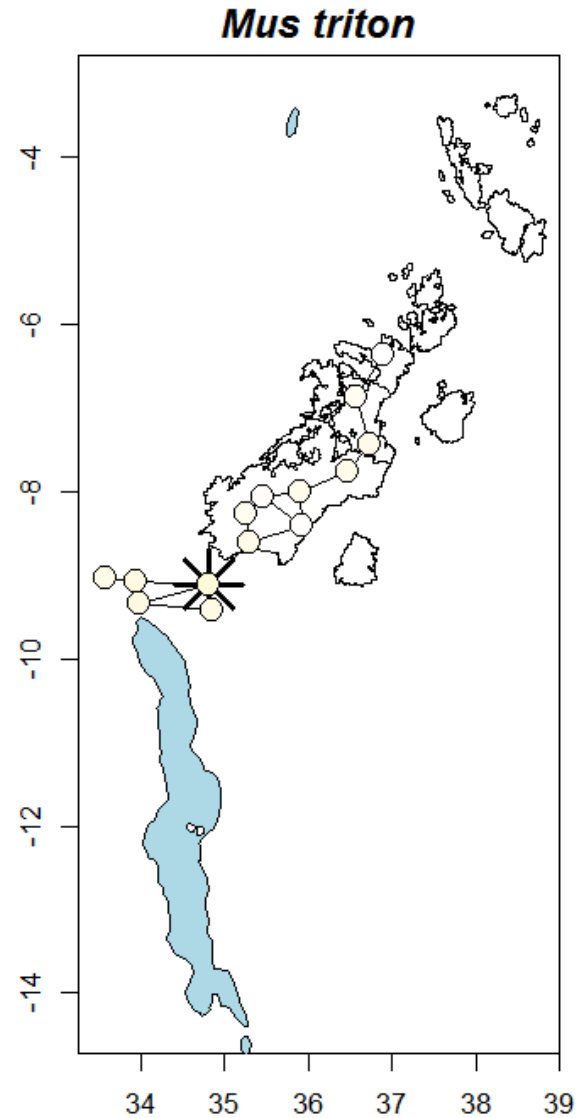

D
50 km
A ↖
50 km
↘ C
A
C
100 km
↘ B
A, B

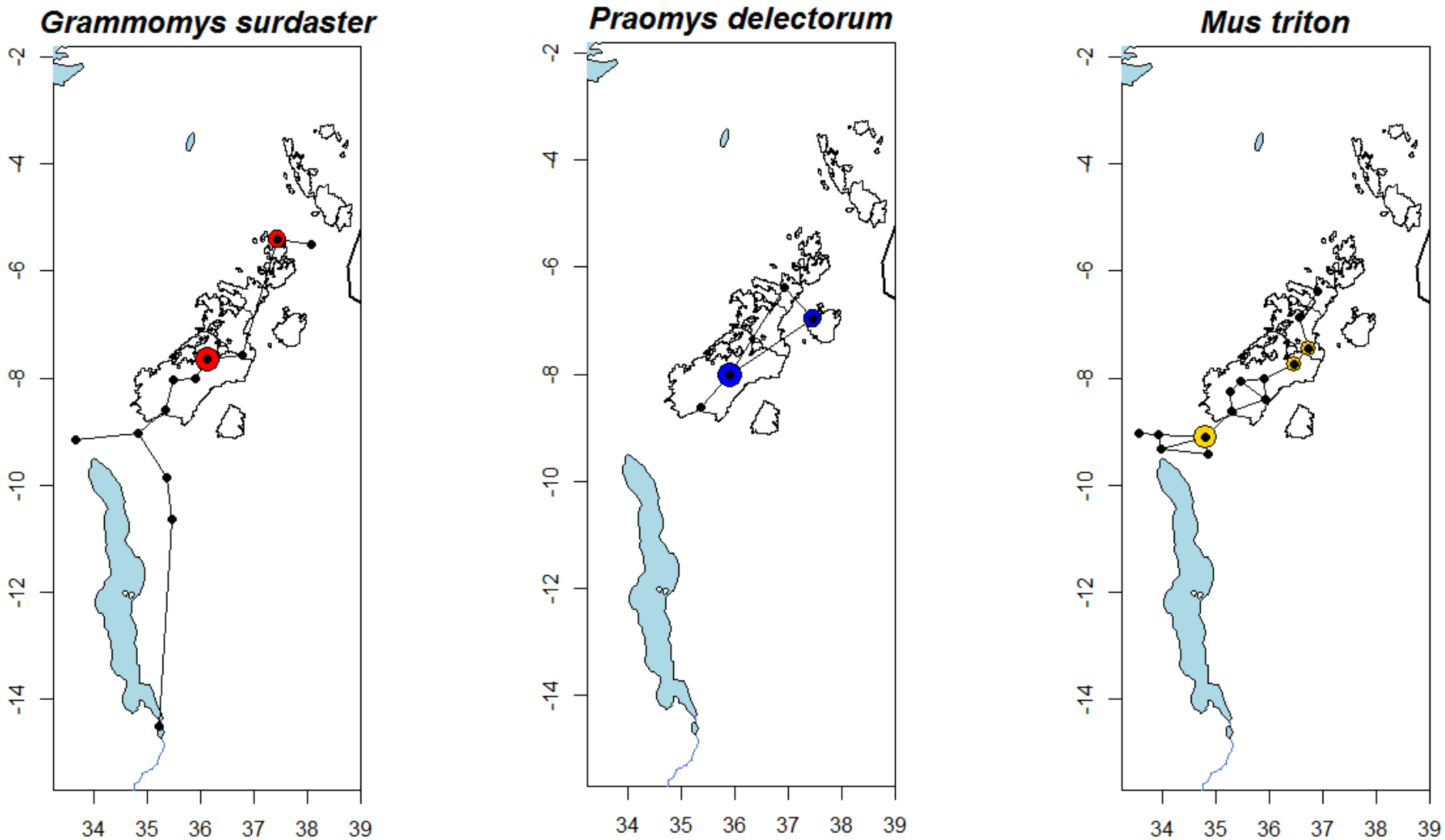

Erinaceus
E. europaeus E2
E. europaeus E1
E. roumanicus
65
60
55
50
45
40
35
−20
−10
0
10
20
30
40


Apodemus sylvaticus
central
peripheral
south−eastern
65
60
55
50
45
40
35
−20
−10
0
10
20
30
40


Grammomys surdaster
-4
-6
-8
-10
-12
-14
34
35
36
37
38
39


Praomys delectorum
-4
-6
-8
-10
-12
-14
34
35
36
37
38
39


Mus triton
-4
-6
-8
-10
-12
-14
34
35
36
37
38
39


Grammomys surdaster
Praomys delectorum
Mus triton
-2
-4
-6
-8
-10
-12
-14
34
35
36
37
38
39