



**HAL**  
open science

# Analysis of haplotype networks: the randomized minimum spanning tree method

Emmanuel Paradis

► **To cite this version:**

Emmanuel Paradis. Analysis of haplotype networks: the randomized minimum spanning tree method. *Methods in Ecology and Evolution*, 2018, 9 (5), pp.1308 - 1317. 10.1111/2041-210X.12969 . ird-01822368

**HAL Id: ird-01822368**

**<https://ird.hal.science/ird-01822368>**

Submitted on 8 Mar 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

January 5, 2018

# Analysis of haplotype networks: the randomized minimum spanning tree method

Emmanuel Paradis

ISEM, IRD, Univ. Montpellier, CNRS, EPHE, Montpellier, France

E-mail: [Emmanuel.Paradis@ird.fr](mailto:Emmanuel.Paradis@ird.fr)

## Summary

1. Haplotype network construction is a widely used approach for analysing and  
3 visualising the relationships among DNA sequences within a population or  
species. This approach has some problems such as how to quantify alternative  
links among sequences, or how to plot efficiently networks to compare them easily.
- 6 2. In this paper, a new method is presented: the randomized minimum spanning tree  
method, based on randomizing the input order of the data in order to produce  
alternative branchings in the haplotype network. It is shown that this new method  
9 can produce, at least in some situations, networks with less alternative links than  
the minimum spanning network method.
- 12 3. A new graphical display of haplotype networks is introduced here. This is based  
on calculating the coordinates of the haplotypes from a multidimensional scaling  
of the haplotype distance matrix. The display can be done in two or three  
dimensions. The eigenvalues extracted from the multidimensional scaling analysis  
15 give an indication of the relevant number of dimensions.
- 18 4. These tools are illustrated with the analyses of published data on the leopard and  
on the jaguar. These analyses show interesting and contrasting patterns between  
these two species of big cats.
5. All tools are implemented in R and available in the package *pegas*.

**Keywords:** Hamming distance, haplotype network, microevolution, minimum spanning  
21 tree, *Panthera*

## Introduction

The analysis of DNA sequences from individuals sampled in one or several populations makes possible to address different questions on microevolutionary processes such as population structure, gene flow, past demographic bottlenecks and expansions, or geographical events of colonizations and extinctions (Bossart & Prowell, 1998; Emerson *et al.*, 2001; Buerkle & Lexer, 2008). An important aspect of these analyses is the inference of ancestry–descendance relationships among haplotypes. Typically, if the sequences are assumed to be contemporaneous, two alternative approaches can be adopted: inferring either a phylogenetic tree, or a network. The phylogenetic approach assumes that the ancestral sequences are unobserved and associated with the internal nodes of the tree while the observed sequences are associated with its terminal nodes. On the other hand, if some observed sequences may be ancestral to others, a network approach is more appropriate, in which case these sequences will be associated with the internal nodes of the network. It is clear that choosing an approach or the other will depend on the time frame and the mutation rate of the sequences. If the sequences are sampled sequentially through time (e.g., pathogens sampled through an epidemic, or ancient DNA), several approaches have been developed to take this temporal dimension into account (e.g., Jombart *et al.*, 2011).

Network methods can be classified according to different criteria (Table 1). For instance, it is possible to define two categories depending on their main objective: the ‘explicit’ networks depict reticulated processes of evolution such as hybridization, horizontal gene transfer, or admixture, whereas the ‘implicit’ networks represent some form of uncertainty over a tree (i.e., without reticulation) representation (Huson & Bryant, 2006; Klopper & Huson, 2008). Another way to classify network methods considers the reconstruction method: these include parsimony, distances, maximum likelihood, Bayesian inference, split decomposition, or consensus methods (Holland

48 *et al.*, 2004). Another important criterion is whether unobserved haplotypes can be  
included in the network: distance-based methods cannot generally do this because they  
do not consider explicitly the evolving characters (Table 1). It is crucial to assess the  
51 reliability of a phylogeny or network of haplotypes since its construction may be affected  
by sampling biases (i.e., missing haplotypes). A badly estimated network is likely to lead  
to wrong inference of population or history.

54 The present paper focuses on the distance-based, implicit network approach. Two  
methods are commonly used to build a haplotype network: the minimum spanning tree  
(MST) method (Kruskal, 1956) and the method from Templeton *et al.* (1992, TCS). The  
57 MST method has been applied in many fields (Nešetřil *et al.*, 2001). Its principle is to  
first build a matrix of pairwise distances among sequences (or haplotypes), and then find  
the shortest set of paths that link all observations where the length of each link is taken  
60 from the pairwise distance. The TCS method, often referred to as statistical parsimony,  
is based on a model of evolution of the genetic characters measured on each  
individual—originally restriction fragment lengths but the method can be applied to  
63 DNA sequences. The main difference between both methods is that an MST is a network  
with no reticulation, thus for  $n$  sequences, the resulting MST will have  $n - 1$  links. On  
the other hand, a TCS network may have reticulations defining alternative branchings,  
66 and include unobserved haplotypes in the network. Thus, this method may be used to  
infer micro-evolutionary events such as putative recombinations (Posada & Crandall,  
2001). Bandelt *et al.* (1999) developed another method called the median-joining  
69 network (MJN) where alternative branchings are found by examining potential ancestral  
sequences for each triplet of sequences. MJN belongs to a class of methods which  
includes several variants such as reduced median network or quasi-median network  
72 (Bandelt *et al.*, 1995, 1999). Thus, the MST usually cannot represent possible  
ambiguous or alternative branchings. Another limitation of the MST construction is that,

for a given data set, the MST may not be unique. Figure 1 shows a simple illustration of  
75 this problem: the three represented data sets are identical but the inferred MSTs are  
different because the observations are ordered differently. This is a consequence of the  
presence of ties in the distance matrix (see the algorithm descriptions below). With real  
78 data, this problem and its consequences may be very hard to detect with a large number  
of observations. Bandelt *et al.* (1999) pointed out that instead of constructing a single  
tree, it is possible to construct a minimum spanning network (MSN) by modifying  
81 slightly Kruskal's (1956) algorithm as explained below.

In this paper, I propose a new algorithm to construct a network which can be seen as  
intermediate between the MST and MSN methods. I also present tools to compare  
84 networks programmed in R (R Core Team, 2017).

## Methods

### MINIMUM SPANNING TREE AND NETWORK

The MST algorithm can be sketched as follows (Kruskal, 1956):

- 87 1. Compute the matrix of pairwise distances among the  $n$  observations and sort them  
in increasing order.
2. Assign each observation to its own group; there are thus  $n$  initial groups.
- 90 3. Set  $i \leftarrow 1$ .
4. Take the  $i$ th distance from step 1: if the two corresponding observations are not in  
the same group, then create a link between them and pool the two groups.
- 93 5. Set  $i \leftarrow i + 1$ .
6. Repeat steps 4 and 5 until there is a one group.

The number of groups is decreased by one at each iteration of step 4. The issue with this  
96 algorithm is how ties in the distance matrix are treated in step 1, and this is, obviously,  
dependent on the implementation. The MSN tries to solve this problem (Bandelt *et al.*,  
1999); its algorithm is:

- 99 1. Compute the matrix of pairwise distances among the  $n$  observations, extract the  
unique values, and sort them in increasing order (denoted as  $\delta_1, \delta_2, \dots$ ).
2. Set  $i \leftarrow 1$ .
- 102 3. Create the links for all pairs of observations with distance equal to  $\delta_i$ .
4. If all observations are linked in a single group, then stop.
5. Set  $i \leftarrow i + 1$ , and go to step 3.

#### THE RANDOMIZED MINIMUM SPANNING TREE

105 The input data are a set of aligned sequences from which a distance matrix is computed.  
The sequences could be DNA or other kinds as long as there is a method to compute  
pairwise distances. In most applications, a simple Hamming distance (or Manhattan in  
108 the case of binary characters) will be relevant. In order to remove the influence of the  
order of the input data, the procedure is based on a randomization of this order. This is  
repeated many times and for each replication an MST is constructed. The MSTs are then  
111 post-processed in order to return a single network including all the links observed among  
the replications. Because of the nature of the proposed algorithm, it is called here the  
*randomized minimum spanning tree* (RMST) method.

114 The RMST usually has less links than the MSN. Figure 2 shows a simple example  
with four binary sequences. The first step of the network construction is to consider links  
of length one A–C and A–B; the second step considers the links of length two A–D and

117 B–C. During the MST construction, the link B–C is never included because B and C  
were already grouped together during the first step, so the RMST does not include this  
link. On the other hand, the MSN includes this link thus resulting in two alternative  
120 paths, B–A–C and B–C, both of the same length. The RMST avoids this ambiguity. Note  
the difference between the present example and the one in Figure 1: in the latter, the  
additional link output by the MSN creates a path shorter than the one created by the  
123 MST.

## GRAPHICAL TOOLS

Plotting networks is a notoriously difficult problem (e.g., Klopper & Huson, 2008).  
Several computer programs perform graphical display of various types of evolutionary  
126 networks, such as igraph (Csardi & Nepusz, 2006), network (Butts, 2008), phangorn  
(Schliep, 2011), pegas (Paradis, 2010), or SplitsTree (Huson, 1998), among many others.  
In practice, it would be very useful to graphically compare networks constructed under  
129 different methods or assumptions. However, this is usually not possible (or very difficult)  
because there is no standard procedure for plotting haplotype networks. To propose a  
solution to this problem, an implementation based on multidimensional scaling (MDS;  
132 Torgerson, 1952) is developed here. The procedure is to first perform an MDS on the  
distance matrix in order to extract two or three sets of coordinates. These coordinates are  
then used to plot the observed sequences or haplotypes in 2-D or in 3-D, and the links  
135 inferred from the network are then drawn. Thus, this procedure contrasts with most  
existing ones which compute the layout of haplotypes trying to minimize line crossings  
(e.g., Klopper & Huson, 2008). The proposed procedure has several advantages. First,  
138 an MDS on the original distance matrix will arrange the sequences depending on their  
similarity. Second, the eigenvalues extracted from the MDS make possible to assess  
whether it is relevant to use two or three dimensions in this projection. Third, the



141 coordinates of the observations will be the same for all networks since they depend only  
on the distance matrix, making graphical comparisons easier. Fourth, the procedure is  
computationally straightforward since an MDS is usually fast to perform even with  
144 several hundreds observations.

The tools presented in this article have been implemented in the R package *pegas*  
(Paradis, 2010). This package has already implemented a standard 2-D plot using an  
147 “energy-minimisation” algorithm to optimise the layout. Plots in three dimensions have  
been implemented using the *rgl* package (Adler *et al.*, 2016).

## SIMULATION STUDY

To assess how the RMST is able to find alternative links in a network of haplotypes,  
150 some simulations were run under different situations of mutation rate ( $\mu$ ), sequence  
length ( $l$ ), and number of sequences ( $n$ ). A set of  $n$  sequences was simulated under the  
JC69 model of sequence evolution (Jukes & Cantor, 1969) along a random network with  
153 no reticulation and link lengths taken from a standard uniform distribution. The network  
was simulated by generating a random binary tree where the internal nodes were  
considered contemporaneous to the leaves. The sequences were then analysed with the  
156 RMST using different numbers of randomizations (5, 10, 20, 50, 100): the number of  
additional links found by the RMST as well as the number of unique distances were  
recorded. The other parameters were:  $n = 50, 100, 500, \text{ or } 1000$ ,  $l = 500 \text{ or } 1000$ , and  
159  $\mu = 0.01 \text{ or } 0.1$ . These parameter values were chosen to result in substantial numbers of  
ties in the distance matrices. The simulations were replicated 100 times for each  
combination of  $n$ ,  $l$ , and  $\mu$ . The code used is provided in the Supplementary Information.

## DATA

162 Two data sets were considered to apply the methods introduced in this paper: a set of  
mtDNA sequences from 33 leopards (*Panthera pardus*) published by Uphyrkina *et al.*  
(2001), and a set of mtDNA sequences from 37 jaguars (*P. onca*) published by Eizirik  
165 *et al.* (2001). Both data sets were downloaded from GenBank (accession numbers:  
AY035227–AY035292 and AF244814–AF244887, for each species, respectively). The  
sequences were aligned separately for the different genes (using information from  
168 GenBank and from the original publications) with MUSCLE (Edgar, 2004), and then  
combined into two global alignments with 726 and 707 sites, respectively. All sequences  
were unique for the leopard data, but 22 unique sequences were identified for the jaguar  
171 data. For each alignment, a matrix of Hamming distances was calculated with *ape*  
(Paradis *et al.*, 2004). These matrices were used as input for the construction of the  
networks. The individual labels from the original studies were kept for the present  
174 analyses. The R scripts used for these analyses is provided in the Supplementary  
Information.

## Results

### SIMULATION STUDY

177 To simplify the presentation of the results, the number of additional links was calculated  
as successive differences with increasing number of randomizations (i.e., the numbers of  
links found with five randomizations compared to zero, with ten randomizations  
180 compared to five, and so on). The results were clearly related to the number of unique  
distances among the simulated distances. Considering that the total number of distances  
is given by  $n(n-1)/2$ , the percentage of unique distances was always less than 1%  
183 (Table 2). Increasing  $l$  and/or  $\mu$  resulted in more variation among the sequences and,

consequently, less additional links for the same  $n$ . On the other hand, increasing  $n$  for the same values of  $l$  and  $\mu$  resulted in more closely related sequences, and thus more  
186 additional links found by the RMST. In three cases, some additional links were still found with 100 randomizations. In the scenario simulating the most diversity ( $l = 1000$ ,  $\mu = 0.1$ ), five randomizations were enough to find the additional links of the RMST for  
189 all values of  $n$ .

## APPLICATION

The MDS analysis of the distance matrices resulted in slightly different patterns of eigenvalues. For the leopard data, the first eigenvalue was much larger than the others  
192 though the second and third ones were substantially larger than the remaining ones (Fig. 3a). For the jaguar data, the first and the second eigenvalues were much larger than all the other ones (Fig. 3b). Thus, we may anticipate that three dimensions may represent  
195 the distribution of sequences for the leopard data whereas two dimensions may be enough for the jaguar data.

The MSN and RMST analyses revealed large numbers of additional links compared  
198 to the MST ones (Table 3). In the case of the leopards, the number of links was multiplied by 6.7 from the MST to the MSN, and by 1.9 to the RMST. This increase in number of links was slightly smaller in the case of the jaguar: 5 to the MSN and 1.3 to  
201 the RMST. The RMST analyses were repeated with different numbers of randomizations. For the leopard data, 59 links were found with 10 randomizations while 60 links were found with 50 randomizations or more. For the jaguar data, 28 links were  
204 found with 10 randomizations or more.

The MDS-based plots of the leopard sequences showed a clear continental separation with the African individuals (SHO\*) on the right-hand side of the plot, and the Asian  
207 ones on the left-hand side (Fig. 4). Interestingly, two individuals laid outside of these

two groups: the one from the Arabian Peninsula (NIM1) and the one from Java, Indonesia (MEL1). Remarkably, the RMST did not add any further link from NIM1 to the others, whereas two additional links were observed between MEL1 and the Asian ones. The MSN kept NIM1 with a single link to SAX2, but added one further link between MEL1 and the others. The 3-D displays revealed that the African individuals, which are apparently aligned in the 2-D plots, are actually arranged along an arc in the third dimension of the MDS (videos provided in the Supplementary Information).

For the jaguar data, the arrangement of the individuals on the first axis followed a North–South axis with individuals from the South on the right-hand side of the plot (Fig. 5). Two individuals remained single-linked with the MSN and RMST analyses: Pon23 from Nicaragua which was linked with two individuals from Nicaragua and from Costa Rica, both with the same haplotype, and Pon63 from Venezuela which was linked with Pon73, an individual of unknown origin but presumably from Brazil (Eizirik *et al.*, 2001). Except for Pon63, very little dispersion was observed in the third dimension as expected from the eigenvalues of the MDS (videos provided in the Supplementary Information).

## Discussion

The analysis of the relationships among DNA sequences and haplotypes within and among populations is crucial for testing hypotheses on microevolutionary processes. However, such analyses often suffer from shortcomings. Typically, two issues are often observed. First, practitioners usually construct a single haplotype network which is then interpreted depending on the context of the study. This can be a problem when the assumptions of the method are not met, which can usually be assessed by comparing different constructions, for instance, by using different distances. The second problem is that there seems to be a confusion in the literature between MST and MSN. It is common

to read “minimum spanning network” when an MST is obviously shown since no loop is  
234 present. This is problematic since the MST may not be unique as already mentioned by  
Bandelt *et al.* (1999).

The RMST method is an alternative to the MSN with the advantage of creating less  
237 links among haplotypes while fully taking into account the ambiguities induced by data  
ordering. In the applications with real data presented in this paper, 100 randomizations  
were done and the results were identical than with smaller numbers. The computing  
240 times of the method is thus proportional to the product of  $n$  (since  $n - 1$  links are built at  
each replication of the MST algorithm) with the number of randomizations. The  
implementation in `pegas` resamples the distance matrix by reordering its rows and  
243 columns simultaneously, instead of reordering the original data matrix, and thus avoids  
to recalculate the distance matrix at each iteration of the MST (which requires a  
computing time proportional to  $n^2$ ).

246 The simulation study showed that the number of randomizations required to reach  
convergence of the RMST procedure is affected by the number of sequences, the  
sequence length, and the mutation rate. Because it is not easy to define a priori how  
249 many randomizations are required for a given data set, it is recommended to repeat the  
analyses with increasing numbers of randomizations and check that the constructed  
networks are identical (see code in Supplementary Information).

252 The respective merits of the methods used to construct haplotype networks are still  
debated (e.g., Mardulyn, 2012). However, comparing different methods is not without  
difficulties because some of them construct networks that have inherently different  
255 structures. Parsimony-based methods seek to combine all most parsimonious  
phylogenetic trees into a network which, as a consequence of this estimation procedure,  
have the observed sequences only at its terminal nodes (Branders & Mardulyn, 2016).  
258 This contrasts with, on one hand, the MST-based methods where the observed sequences

are at both the internal and the terminal nodes of the network, and, on the other hand, the TCS or MJN method which includes unobserved sequences in the network. For instance, 261 the MJN constructed with the data in Figure 1 would have four nodes and three links all of length one, but this network is not strictly different from the MSN and RMST ones (Fig. 1c) because they all define paths of length two between each pair of observed 264 sequences. Clearly, the presence of loops in RMST or MSN networks must be interpreted cautiously with respect to potentially unsampled haplotypes (e.g., Joly *et al.*, 2007). On the other hand, the MJN inferred from the data in Figure 2 would be identical 267 to the RMST (Fig. 2d) but different from the MSN (Fig. 2c).

Plotting networks (i.e., graphs with reticulations) is notoriously difficult for graphical software developers. This is another difficulty in the analysis of haplotype relationships. 270 The graphical approach proposed here is a solution to this problem. By using the coordinates inferred from the MDS applied on the distance matrix, the haplotypes are always positioned in the same way, for a given distance matrix, whatever the links 273 among them. Furthermore, the analysis of the eigenvalues extracted from the MDS gives information on the general structure of the data as illustrated by the examples above.

The analyses of the leopard and the jaguar data were mainly illustrative, although 276 they show some interesting results. For the leopards, the contrast between African and Asian individuals was very clear. Two individuals were outside the bulk of the other individuals: one from the Arabian Peninsula, and the other from Java. Both represent 279 two subspecies (*P. pardus nimr* and *P. pardus melas*) that are morphologically markedly different from the others (Stein & Hayssen, 2013). Variation within the African group was also substantial and appeared in the third dimension of the plot. For the jaguars, 282 variation was much less than for the leopards, and the MSN and the RMST showed much more additional links than the MST. Interestingly, Eizirik *et al.* (2001) reported an MSN with only one additional link. However, the present analysis showed that the

285 RMST had seven additional links and the MSN even more.

Another interesting result is the presence of a “horseshoe effect” with the leopard data but not with the jaguar data. This effect, which is sometimes observed in  
288 multivariate analyses like MDS or principal component analysis (PCA), is a consequence of the dominance of local structures in the data. Such dominance can be the result of the inherent structure of the original data matrix (Ahmed *et al.*, 1974), a  
291 transformation of the distances that gives more emphasis on the most similar observations (e.g., an exponential decay function; Diaconis *et al.*, 2008), or, typically for population genetic data, local processes such as isolation by distance (Novembre &  
294 Stephens, 2008). Whatever the origin of such structures, the decomposition of the data matrix (in the case of PCA) or of the distance matrix (in the case of MDS) results in the second axis to be related to the first one in a polynomial-like manner (actually a  
297 sinusoidal function; see Ahmed *et al.*, 1974; Diaconis *et al.*, 2008), and the subsequent axes with increasing degrees of the polynomials. In practice, the proximities of the sequences on the second and third axes must therefore be interpreted with caution.  
300 However, this does not affect the interpretation of the network layouts which are the same for all networks as long as the distance matrix is the same.

As rightly pointed out by Leigh & Bryant (2015), the haplotype network  
303 methodology does not generally rely on an evolutionary model. However, a distance-based approach is very valuable because distances can be computed for different kinds of data, and they are straightforward to interpret in terms of number of changes.  
306 An interesting perspective will be to develop an approach to incorporate models of DNA sequence evolution into haplotype network analyses. A challenge will be to find how to compute a likelihood in the presence of loops in the network (Maynard Smith, 1989).

309 **Acknowledgements**

I am grateful to the Associate Editor and three anonymous reviewers for their constructive comments on a previous version of this paper. The simulations benefited from the ISEM computing cluster platform. This is publication ISEM 2017-277.

**Data accessibility**

The data are archived in GenBank (accession numbers: AY035227–AY035292 and AF244814–AF244887).

**References**

- Adler, D., Murdoch, D., Nenadic, O., Urbanek, S., Chen, M., Gebhardt, A., Bolker, B., Csardi, G., Strzelecki, A., Senger, A. & R Core Team (2016) *rgl: 3D visualization using OpenGL*. R package version 0.95.1441.
- Ahmed, N., Natarajan, T. & Rao, K.R. (1974) Discrete cosine transform. *IEEE Transactions on Computers*, **C-23**, 90–93.
- Bandelt, H.J., Forster, P. & Röhl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.
- Bandelt, H.J., Forster, P., Sykes, B.C. & Richards, M.B. (1995) Mitochondrial portraits of human populations using median networks. *Genetics*, **141**, 743–753.
- Bossart, J.L. & Prowell, D.P. (1998) Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology & Evolution*, **13**, 202–206.
- Branders, V. & Mardulyn, P. (2016) Improving intraspecific allele networks inferred by maximum parsimony. *Methods in Ecology & Evolution*, **7**, 90–95.



- Buerkle, C.A. & Lexer, C. (2008) Admixture as the basis for genetic mapping. *Trends in Ecology & Evolution*, **23**, 686–694.
- 333 Butts, C. (2008) network: a package for managing relational data in R. *Journal of Statistical Software*, **24**, 2.
- Csardi, G. & Nepusz, T. (2006) The igraph software package for complex network  
336 research. *InterJournal*, **Complex Systems**, 1695.
- Diaconis, P., Goel, S. & Holmes, S. (2008) Horseshoes in multidimensional scaling and local kernel methods. *Annals of Applied Statistics*, **2**, 777–807.
- 339 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.
- Eizirik, E., Kim, J.H., Menotti-Raymond, M., Crawshaw, Jr, P.G., O'Brien, S.J. &  
342 Johnson, W.E. (2001) Phylogeography, population history and conservation genetics of jaguars (*Panthera onca*, Mammalia, Felidae). *Molecular Ecology*, **10**, 65–79.
- Emerson, B., Paradis, E. & Thébaud, C. (2001) Revealing the demographic histories of  
345 species using DNA sequences. *Trends in Ecology & Evolution*, **16**, 707–716.
- Holland, B.R., Huber, K.T., Moulton, V. & Lockhart, P.J. (2004) Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular  
348 Biology and Evolution*, **21**, 1459–1461.
- Huson, D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
- 351 Huson, D.H. & Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.
- Joly, S., Stevens, M.I. & van Vuuren, B.J. (2007) Haplotype networks can be misleading  
354 in the presence of missing data. *Systematic Biology*, **56**, 857–862.
- Jombart, T., Eggo, R.M., Dodd, P.J. & Balloux, F. (2011) Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, **106**, 383–390.

- 357 Jukes, T.H. & Cantor, C.R. (1969) Evolution of protein molecules. H.N. Munro, ed.,  
*Mammalian Protein Metabolism*, pp. 21–132. Academic Press, New York.
- Kloepper, T.H. & Huson, D.H. (2008) Drawing explicit phylogenetic networks and their  
360 integration into SplitsTree. *BMC Evolutionary Biology*, **8**, 22.
- Kruskal, Jr, J.B. (1956) On the shortest spanning subtree of a graph and the traveling  
salesman problem. *Proceedings of the American Mathematical Society*, **7**, 48–50.
- 363 Leigh, J.W. & Bryant, D. (2015) POPART: full-feature software for haplotype network  
construction. *Methods in Ecology & Evolution*, **6**, 1110–1116.
- Mardulyn, P. (2012) Trees and/or networks to display intraspecific DNA sequence  
366 variation? *Molecular Ecology*, **21**, 3385–3390.
- Maynard Smith, J. (1989) Trees, bundles or nets? *Trends in Ecology & Evolution*, **4**,  
302–304.
- 369 Nešetřil, J., Milková, E. & Nevsetřilová, H. (2001) Otakar Borůvka on minimum  
spanning tree problem Translation of both the 1926 papers, comments, history.  
*Discrete Mathematics*, **233**, 1–36.
- 372 Novembre, J. & Stephens, M. (2008) Interpreting principal component analyses of  
spatial population genetic variation. *Nature Genetics*, **40**, 646–649.
- Paradis, E. (2010) pegas: an R package for population genetics with an  
375 integrated–modular approach. *Bioinformatics*, **26**, 419–420.
- Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and  
evolution in R language. *Bioinformatics*, **20**, 289–290.
- 378 Posada, D. & Crandall, K.A. (2001) Intraspecific gene genealogies: trees grafting into  
networks. *Trends in Ecology & Evolution*, **16**, 37–45.
- R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R  
381 Foundation for Statistical Computing, Vienna, Austria.
- Schliep, K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

- Stein, A.B. & Hayssen, V. (2013) *Panthera pardus* (Carnivora: Felidae). *Mammalian Species*, **45**, 30–48.
- 384
- Templeton, A.R., Crandall, K.A. & Sing, C.F. (1992) A cladistic analysis of phenotypic association with haplotypes inferred from restriction endonuclease mapping and DNA
- 387 sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–635.
- Torgerson, W.S. (1952) Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–419.
- 390 Uphyrkina, O., Johnson, W.E., Quigley, H., Miquelle, D., Marker, L., Bush, M. & O'Brien, S.J. (2001) Phylogenetics, genome diversity and origin of modern leopard, *Panthera pardus*. *Molecular Ecology*, **10**, 2617–2633.

393 **Supporting Information**

**sim.R.** Code to run the simulations.

**analysis\_cats.R.** Code to reproduce the analyses of the leopard and jaguar data.

396 **pardus\_MST.mp4.** Animation of the MST from the leopard data.

**pardus\_MSN.mp4.** Animation of the MSN from the leopard data.

**pardus\_RMST.mp4.** Animation of the RMST from the leopard data.

399 **onca\_MST.mp4.** Animation of the MST from the jaguar data.

**onca\_MSN.mp4.** Animation of the MSN from the jaguar data.

**onca\_RMST.mp4.** Animation of the RMST from the jaguar data.

Table 1: Comparison of some features of different methods to construct haplotype networks. MST: minimum spanning tree; MSN: minimum spanning network; RMST: randomized minimum spanning tree; TCS: statistical parsimony; MP: maximum parsimony; MJN: median-joining network;  $n$ : number of haplotypes;  $L$ : number of links in the network.

Method	Input data	$L$	Unobserved haplotypes	Reference
MST	Distances	$n - 1$	No	Kruskal (1956)
MSN	"	$\geq n - 1$	No	Bandelt <i>et al.</i> (1999)
RMST	"	"	No	This paper
TCS	Sequences	"	Possibly	Templeton <i>et al.</i> (1992)
MP	"	"	Yes, at internal nodes	Branders & Mardulyn (2016)
MJN	"	"	Possibly, as median-vectors	Bandelt <i>et al.</i> (1999)

Table 2: Simulation results: mean number of additional links found by increasing the number of randomizations in the RMST algorithm ( $l$ : sequence length;  $\mu$ : mutation rate;  $n$ : number of sequences; NUD: mean number of unique distances).

$l$	$\mu$	$n$	Number of randomizations					NUD
			5	10	20	50	100	
500	0.01	50	29.96	7.79	3.24	1.16	0.04	39.11
		100	60.25	15.62	6.78	2.75	0.31	50.40
		500	302.88	74.31	33.65	12.29	1.61	80.24
		1000	613.10	156.06	75.05	28.22	3.44	94.85
	0.1	50	2.48	0.18	0.02	0.00	0.00	216.09
		100	5.71	0.52	0.01	0.00	0.00	270.74
		500	33.26	2.89	0.25	0.04	0.00	340.92
		1000	63.49	4.96	0.38	0.04	0.00	363.05
1000	0.01	50	14.58	1.78	0.68	0.07	0.01	71.25
		100	26.42	3.61	0.79	0.14	0.00	96.87
		500	147.65	23.23	7.15	1.58	0.00	149.60
		1000	292.88	45.67	11.77	2.46	0.16	177.73
	0.1	50	1.33	0.21	0.00	0.00	0.00	390.53
		100	2.74	0.23	0.00	0.00	0.00	520.35
		500	15.37	1.58	0.05	0.00	0.00	671.83
		1000	31.22	2.18	0.11	0.00	0.00	707.06

Table 3: Number of links in the networks constructed with the two data sets analysed.  $n$ : number of haplotypes.

Species	$n$	Number of links		
		MST	MSN	RMST
Leopard	33	32	214	60
Jaguar	22	21	105	28

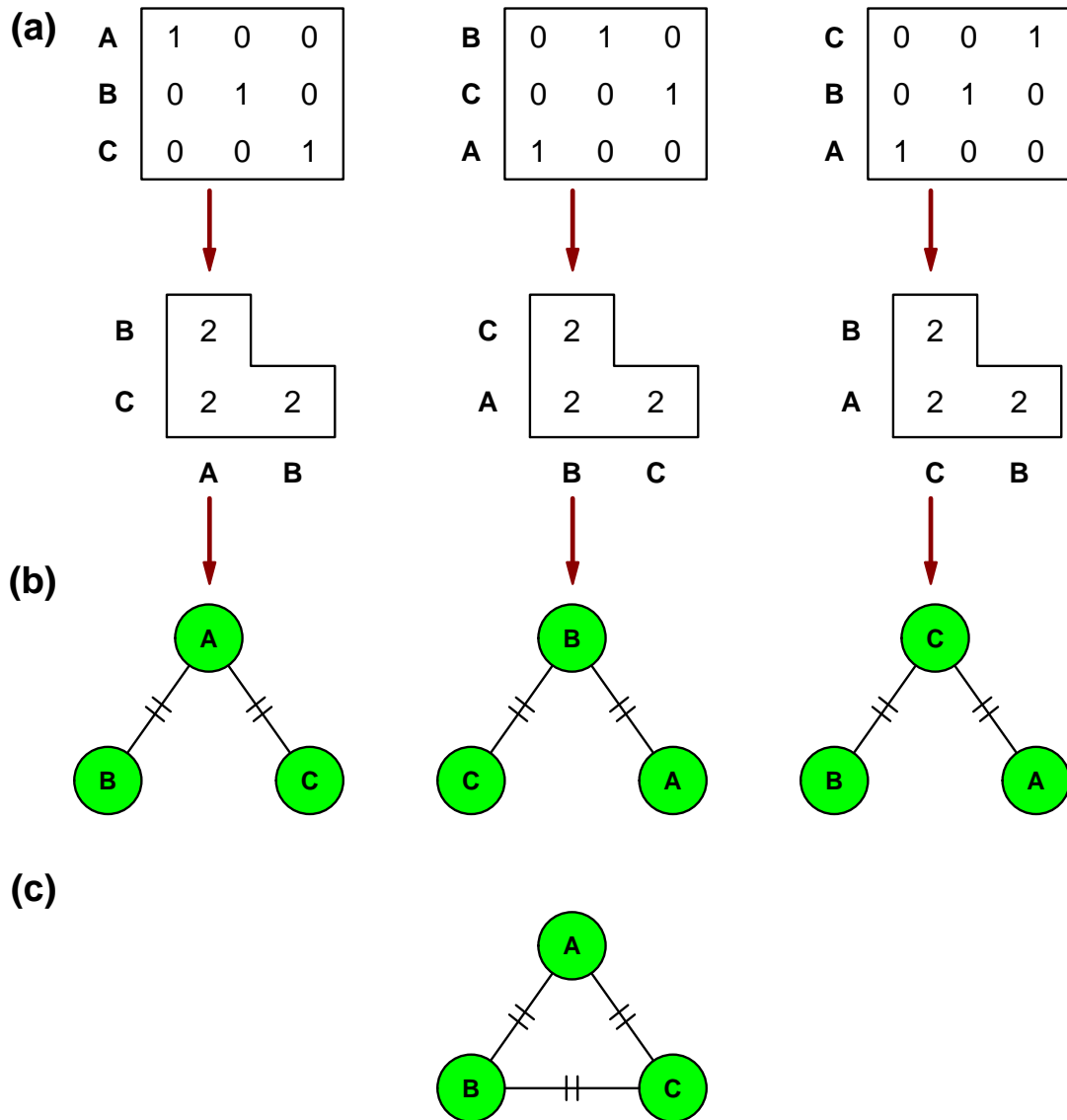


Figure 1: (a) Three identical data sets but with rows in different order and the corresponding distance matrices. (b) The three minimum spanning trees (MST) are different. (c) The minimum spanning network (MSN) and the randomized minimum spanning tree (RMST) are identical for the three data sets.

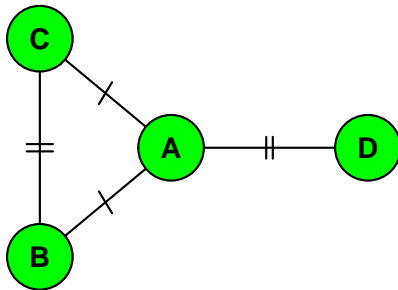
(a)

<b>A</b>	1	1	0	0
<b>B</b>	0	1	0	0
<b>C</b>	1	0	0	0
<b>D</b>	1	1	1	1

(b)

<b>B</b>	1		
<b>C</b>	1	2	
<b>D</b>	2	3	3
	<b>A</b>	<b>B</b>	<b>C</b>

(c)



(d)

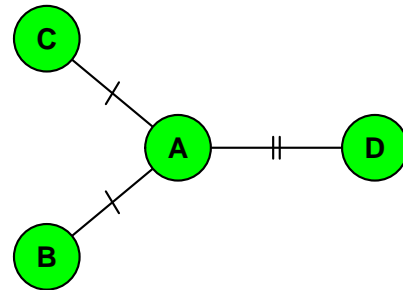


Figure 2: (a) A matrix of four sequences with four sites. (b) The inferred distances. (c) The minimum spanning network (MSN) creates a link between B and C when checking the distances of length two which has the same length than the path B–A–C. (d) The randomized minimum spanning tree (RMST) does not have additional link and is identical to the minimum spanning tree (MST).

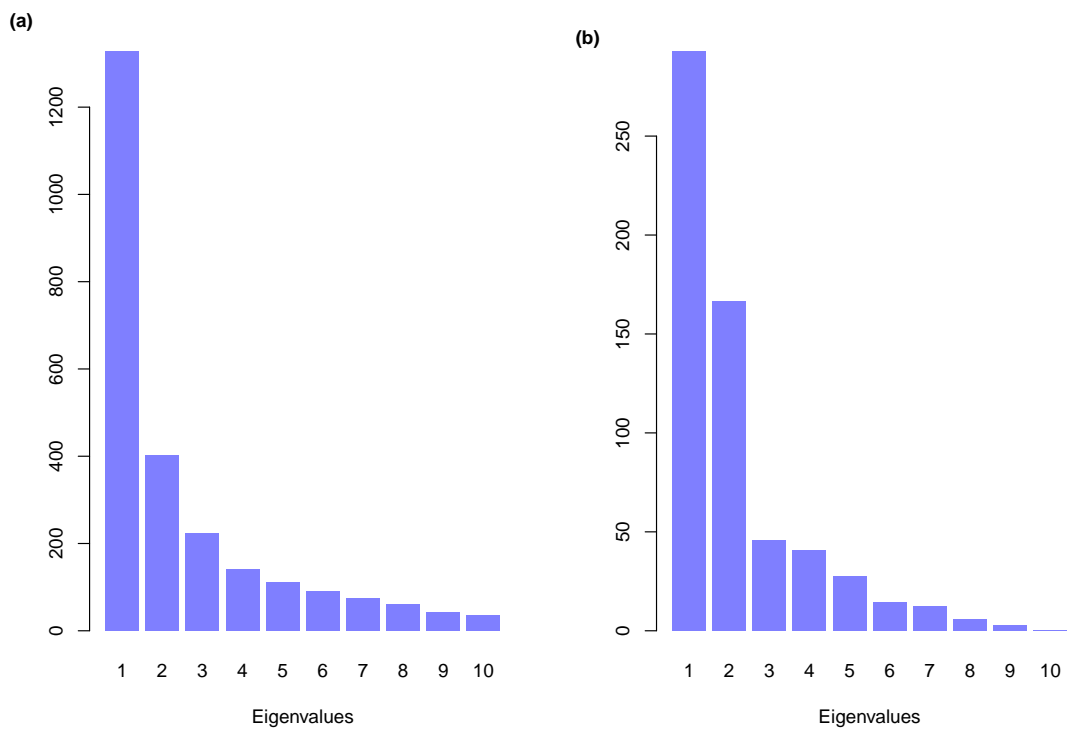


Figure 3: Eigenvalues extracted from the distance matrix for (a) leopards and (b) jaguars.



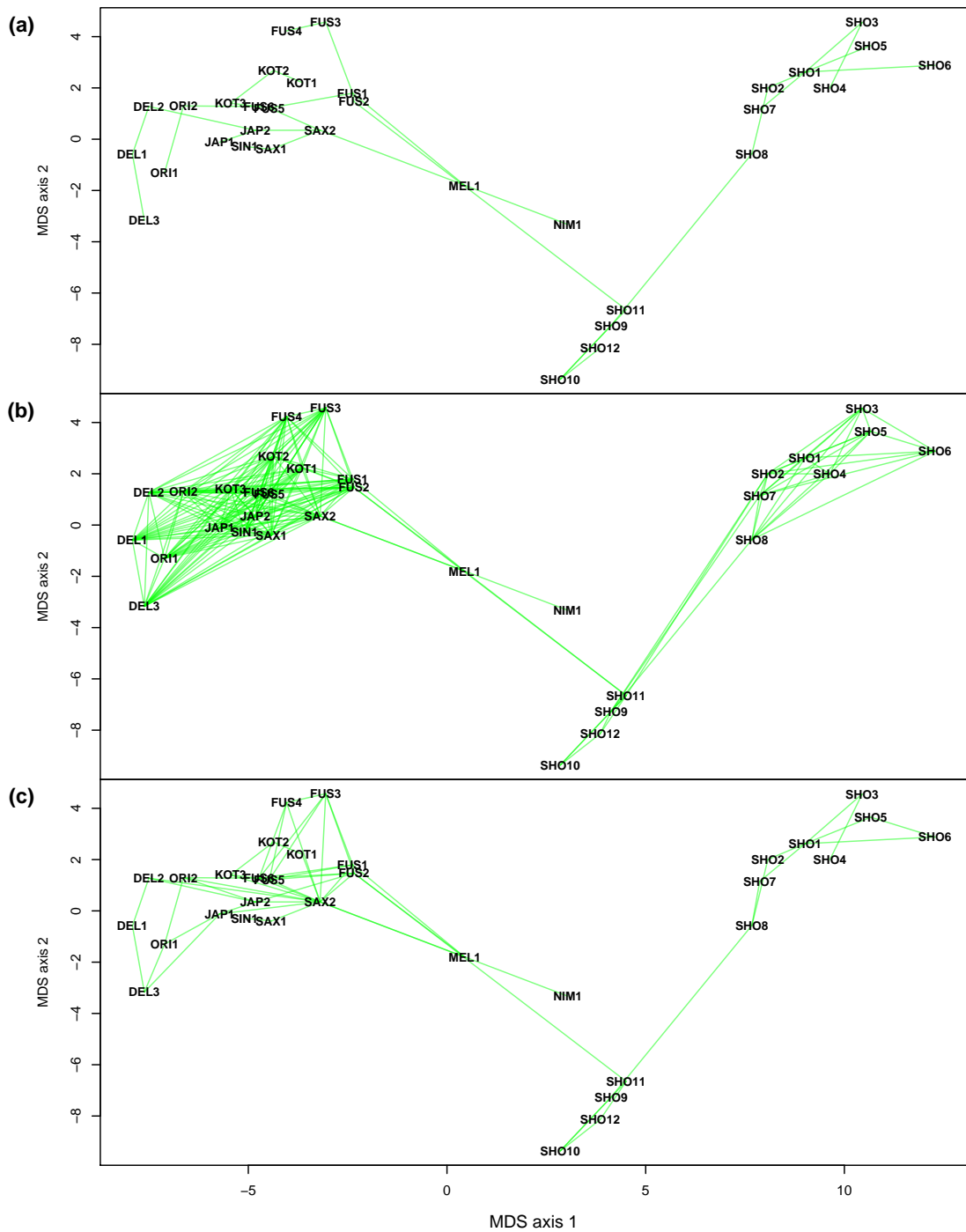


Figure 4: (a) Minimum spanning tree, (b) minimum spanning network, and (c) randomized minimum spanning tree for the leopard data.

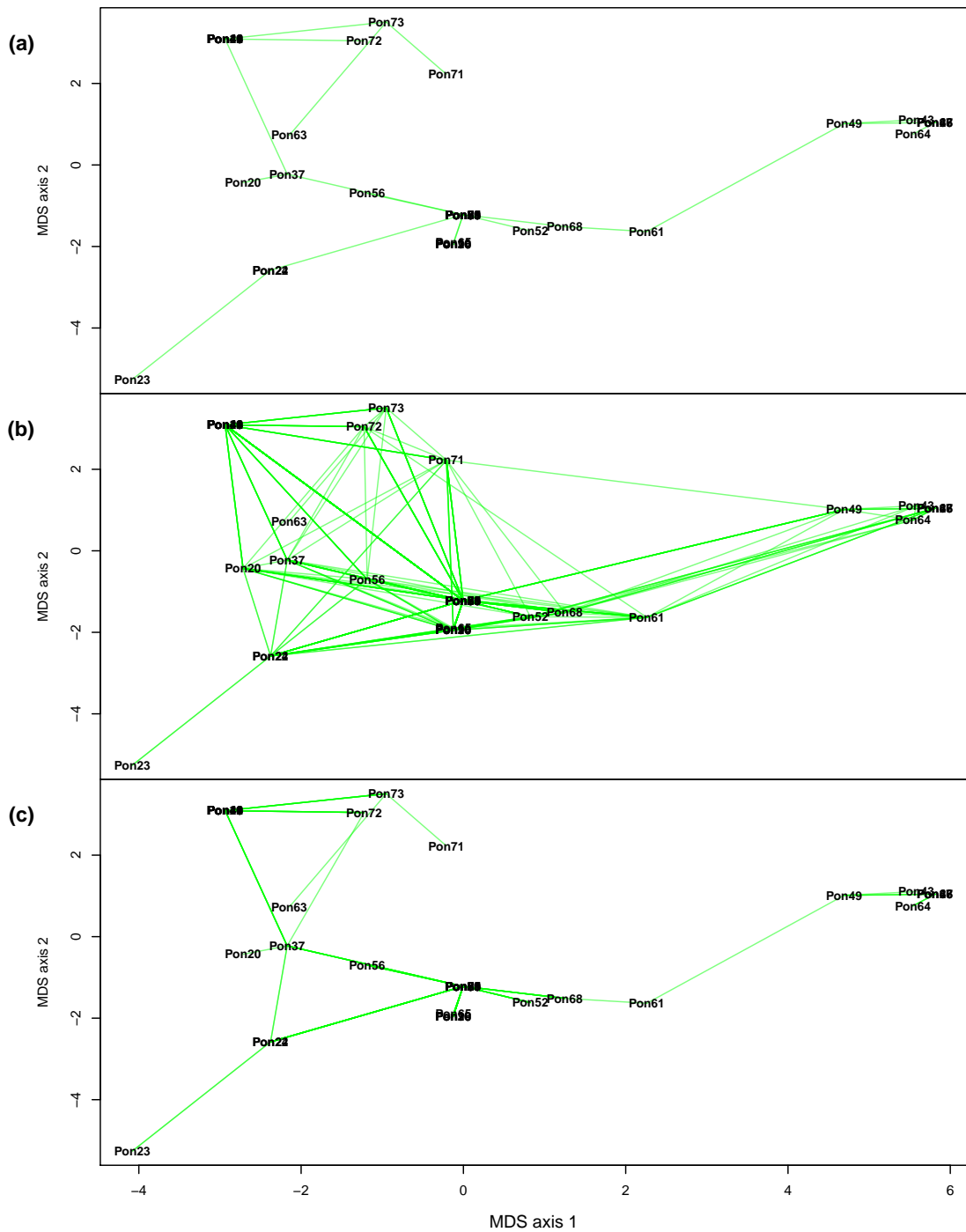


Figure 5: (a) Minimum spanning tree, (b) minimum spanning network, and (c) randomized minimum spanning tree for the jaguar data.