Supplementary notes

Michelle Kendall, Caroline Colijn

In these supplementary notes we give further details of the data and methods used in the main text (Section 1). We then provide supplementary results (Section 2): we explain how to select summary trees and how to visualize tree distances with multidimensional scaling. We provide further analysis of anole lizard and ebolavirus phylogenies, and additional analysis of phylogenies of chorus frogs. Finally we provide a brief supplementary Discussion (Section 3).

1 Supplementary methods

1.1 Anole lizards

We used species trees from a recent *BEAST analysis of the *distichus* species group in the lizard genus Anolis. Geneva et al. [1] made species trees available [2]. They sampled 54 individuals from the *brevirostris* (8) and *distichus* (46) complexes, both within the *distichus* species group. For each individual they sequenced DNA from seven exonic nuclear loci and from one mitochondrial locus. They used gene trees to identify putative species and generated species trees in *BEAST [3] using four independent analyses, each with two billion generations.

We sampled 1000 trees uniformly at random from the latter half of the available *BEAST posterior (file `Anoles_StarBEAST_posterior.species.trees`, with maximum clade credibility (MCC) tree `Anoles_StarBEAST_MCC.species.tre` [2]). We computed all pairwise tree distances according to our metric ($\lambda = 0$) in this sample of 1000 posterior trees. To detect clusters we used k-means clustering (`kmeans` from the `stats` package in R [4]), and compared clustering solutions with the Bayesian Information Criterion (BIC), as described in the `adegenet` package in R [5]. We found that a choice of $k = 8$ clusters minimized the BIC. We visualized the distances using MDS (`dudi.pco` in the `ade4` package in R [6]). Each point represents a tree, and the distances between the points approximate the distances in our metric. We colored points according to their k-means cluster. An MCC tree was found for the whole posterior and for each cluster using TreeAnnotator [7] and plotted with FigTree [8].

1.2 Ebolavirus

We analyzed sequence data from Ebolavirus samples, both historical and from the 2014 outbreak, published recently by Gire et al. [9]. A full description of the data collection, library construction, sequencing, SNP calling and alignments is available in that work (with sequence data and Beast inference settings in file `2014_GN.SL_SRD.HKY_strict_ctmc.exp.xml`). We selected the following 20 taxa:

`BDBV_2007_FJ217161, BDBV_2012_KC545393, EBOV_1976_KC242801,`
`EBOV_1994_KC242792, EBOV_1995_AY354458, EBOV_1996_KC242794,`
`EBOV_2002_KC242800, EBOV_2007_HQ613403, EBOV_2014_EM095,`
`RESTV_1990_AF522874, RESTV_1996_AB050936, RESTV_2008_FJ621583,`
`RESTV_2008_FJ621584, RESTV_2008_FJ621585, SUDV_1976_FJ968794,`
`SUDV_2000_AY729654, SUDV_2011_JN638998, SUDV_2012_KC545389,`
`SUDV_2012_KC589025, TAFV_1994_FJ217162.`

These had no duplicated sequences for any gene among the other selected taxa, allowing inference of trees from each gene separately. Following Gire et al. [9], we used a coalescent prior with exponential growth, a random starting tree, a strict molecular clock with uniform rate across branches and prior mean of 0.0001. We used an HKY substitution model with equal rates. We ran 10 million MCMC iterations and confirmed the results with multiple runs in BEAST v1.8. We used a partitioned Beast analysis to infer trees from all genes together (fixing the trees to be shared across partitions).

We sampled 150 trees from the posterior for each gene (1200 trees in total; Figure 2 in the main text). We computed pairwise distances ($\lambda = 0$) between all these trees and visualized them with MDS, coloring the points according to the gene giving rise to the corresponding tree. The results (the three clusters of trees from VP30, and congruence of the other trees) do not depend on the random sampling of 150 trees from the posterior.

2 Supplementary Results

2.1 Navigating islands and selecting summary trees

[1]Tree inference methods use data to constrain the set of possible trees to a relatively small region of tree space. The fact that data may support trees in separated regions or 'islands' of tree space has deep implications for tree inference and analysis [11, 12]. A further complicating factor is that when taxa have incomplete data at some loci there can be 'terraces' of many equally likely trees, with trees in a terrace all supporting the same subtrees for the taxa with data at a given locus [13]. However, the difficulty of detecting and interpreting tree islands has meant that the majority of analyses, particularly on large datasets, remain based on a single summary tree. This may be a maximum clade credibility (MCC) tree with posterior support values illustrating uncertainty, or a maximum likelihood or parsimony tree with bootstrap supports [14].

Our approach includes a natural way to group trees into clusters. Since distance is defined by the metric that is used, these are different from previously described tree islands [11, 12]. We note that islands are of particular concern for tree inference and for outcomes that require the topology of tree, which will affect ancestral character reconstruction and consequently the interpretation of many phylogenetic datasets [15]. However, other analyses, and tree estimation methods themselves, take trees' branch lengths as well as topology into account. We find that the clusters typically merge together in the metric as λ approaches 1; the posterior becomes unimodal (Figure S2).

There are many challenges to summarising complex tree spaces [14]. Maximum clade credibility (MCC) trees are used to summarize posterior distributions by collecting the clades with the strongest posterior support. However, where these are not concordant the MCC tree can have negative branch lengths. Furthermore, the MCC tree itself may never have been sampled by the MCMC chain, casting doubt on its ability to reflect the relationships in the data.

In the main text we used the familiar method of MCC trees with posterior support values to demonstrate the way that each cluster corresponds to a possible, likely resolution of uncertain clades. In fact, we can also use the metric directly to find 'central' trees within any collection of trees using barycentric methods such as the geometric median [16]. That is, we can exploit the fact that our metric is simply the Euclidean distance between the two vectors $v_\lambda(T_a)$ and $v_\lambda(T_b)$. Among N trees T_i $(i = 1, \ldots, N)$ in a posterior sample, we can find the tree closest to the average vector $\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v_\lambda(T_i)$. The average vector $\bar{v}$ may not in itself represent a tree, but we can then find the tree vector(s) from our sample closest to this average, minimizing the distance between $\bar{v}$ and $v_\lambda(T_i)$. Each of these corresponds to an actual tree T_c from the original sample, with non-negative branch lengths. The minimal distance between the central vector and closest tree vectors is a measure of the quality of the summary: if it is small, each T_c is close to 'average' in the posterior. Such a tree T_c is known as the geometric median tree. Geometric

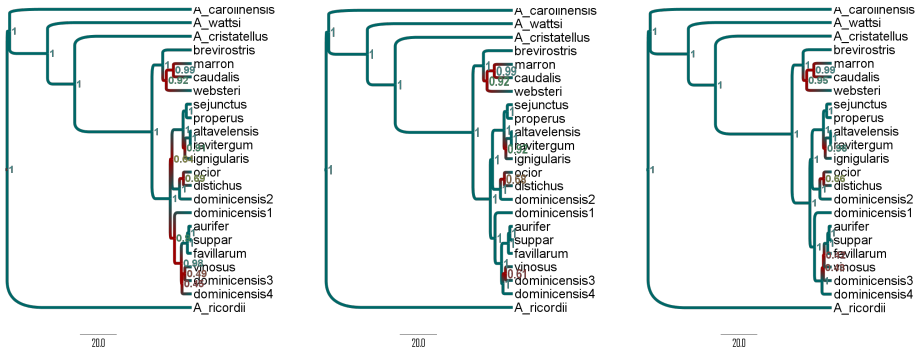
[1]Parts of Sections 2.1 and 2.2 also appear in [10] but are included here for clarity of explanation.

median trees will always have been sampled by the MCMC, and will not have negative branch lengths. It is also straightforward to weight trees by likelihood or other characteristics when finding the geometric median. We found that within clusters, geometric median trees are very close in topology to the MCC tree for the cluster.

2.2 Visualization with MDS

Visualization techniques like MDS have been used to explore tree space using other metrics, but have been challenged by poor-quality projections [17, 18]. When a multidimensional set of distances is projected into a low-dimensional picture there is typically some loss of information which may result in a poor-quality visualization. For example, if 10 points are all 3 units away from each other this will not project well into two dimensions; some will appear more closely grouped than others. However, if there are only 3 such points they can be arranged on a triangle, capturing the distances in two dimensions. One approach to checking the quality of a visualization is a Shepard plot [19], which is a scatter plot of the projected (MDS) distance vs the true distance (from the metric). We include Shepard plots in our supplementary figures to demonstrate their quality.

2.3 Anole lizards



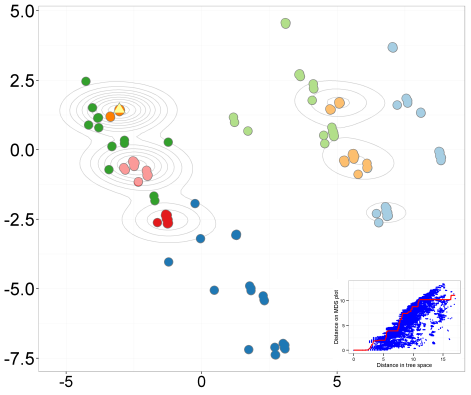
(a) Overall MCC (b) Dark orange group (c) Red group

(d) Light orange group (e) Light green group

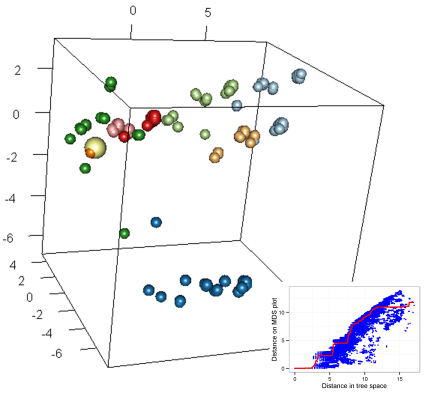
Figure S1: Anoles consensus trees from our $\lambda = 0$ analysis, showing lengths in millions of years.

Recall that the trees in Figure 1 of the main text are displayed as cladograms (branch lengths = 1) because we compared tree topologies ($\lambda = 0$) and this makes differences more clear. In Figure S1 we show the same trees with the correct branch lengths. Figure S2 shows the MDS plots of posterior species trees for several values of λ, increasing from $\lambda = 0$ (as in the main text) to $\lambda = 0.1$ (where branch lengths are weighted quite highly because the lengths are often much larger than 1, with a mean of 13 and median of 3.7 in units of millions of years before the present). As λ increases, clusters spread and merge together, though at $\lambda = 0.1$ they remain

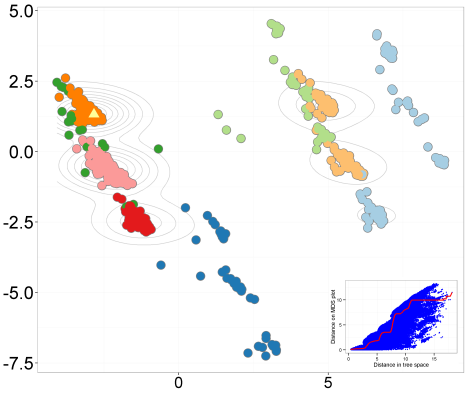
visible as distinct 'strips', particularly in the 3D plot. The division between the left and right sides of the plot persists.



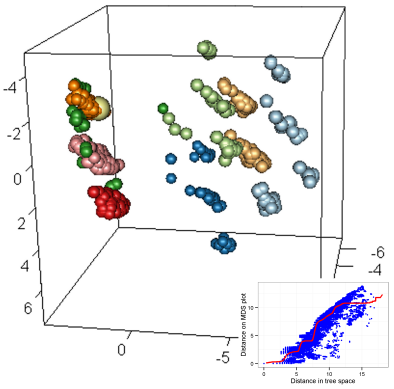
(a) $\lambda = 0$



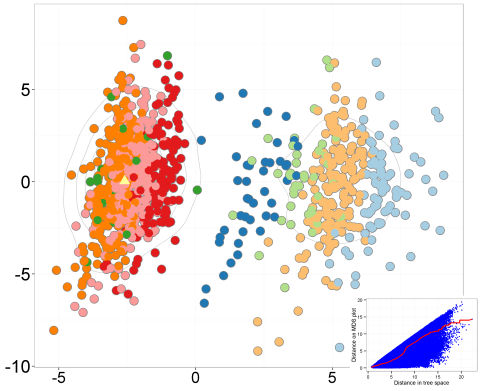
(b) $\lambda = 0$



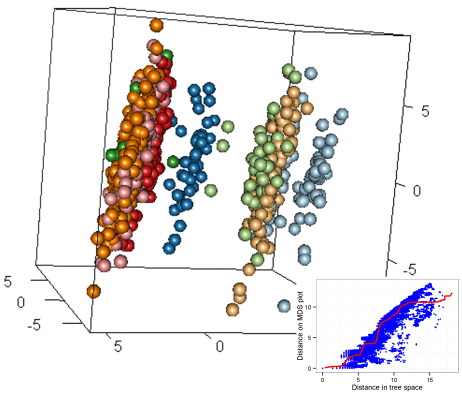
(c) $\lambda = 0.05$



(d) $\lambda = 0.05$



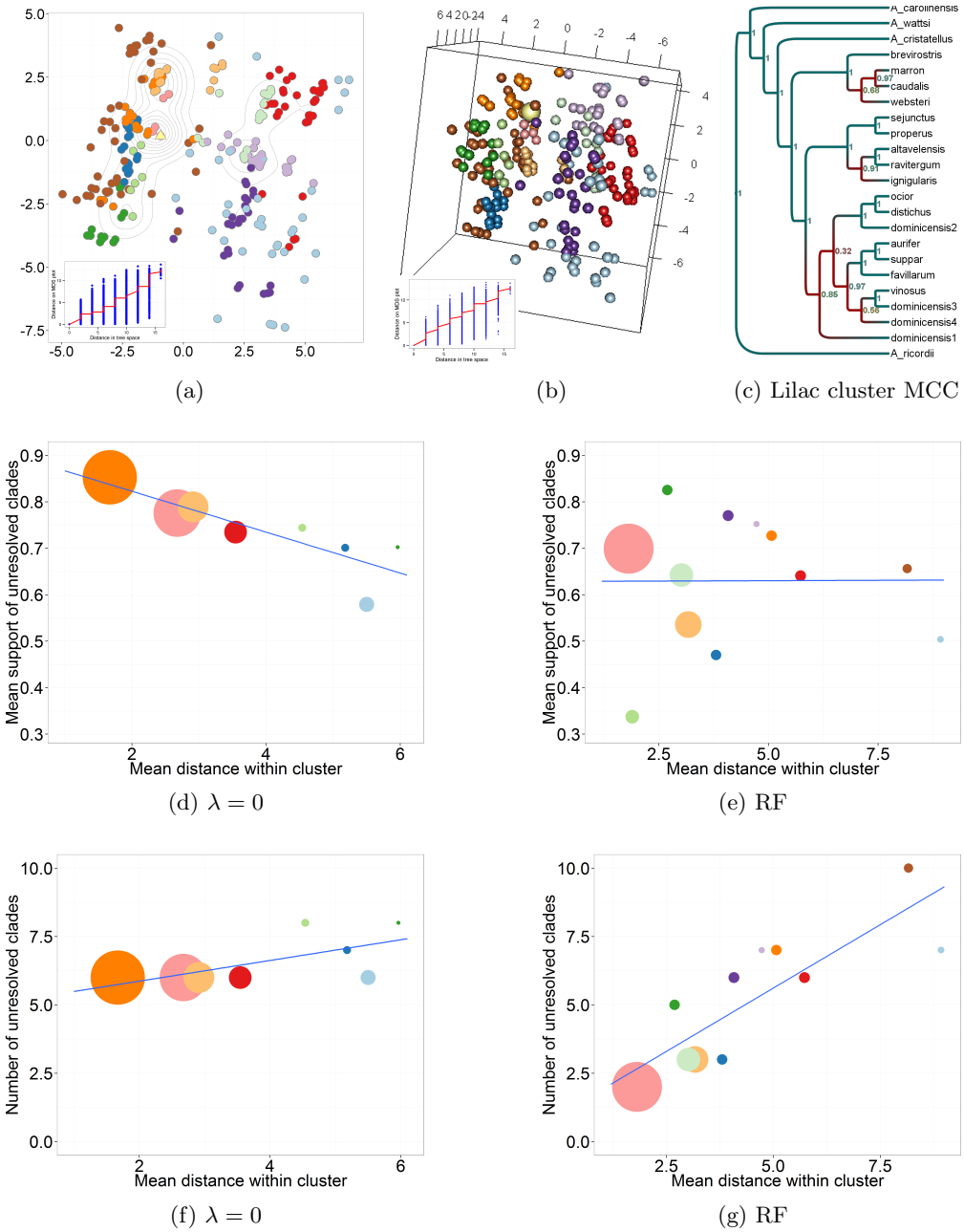
(e) $\lambda = 0.1$



(f) $\lambda = 0.1$

Figure S2: MDS plots of the posterior anoles species trees for several choices of λ. (a) is the same as Figure 3C in the main text; (b) is a 3D MDS plot of the same clusters with the same colors, which better shows the separation, particularly between the dark blue cluster and the others. (c) and (d) are 2D and 3D MDS plots of the tree distances when $\lambda = 0.05$. The inclusion of lengths spreads the trees out; whereas in (a), all trees with the same topology are plotted on top of each other, here, variation in the branch lengths contributes to the distances, spreading the clusters out. (e) and (f) have $\lambda = 0.1$. Clusters are merging together somewhat, but are still distinctive, which the 3D visualization illustrates. 3D plots have an additional degree of freedom, allowing MDS projected distances to be more closely correlated with the input distances than in 2D (inset Shepard plots).

Individual gene trees in the anole data had large polytomies, and did not mirror these alternative resolutions of the uncertainty in the posterior [1], though it is possible that these alternatives would be mirrored by gene trees with a sufficiently high number of samples. It is also possible that there are islands of tree topologies that the *BEAST algorithm did not reach.



(a) (b) (c) Lilac cluster MCC

(d) $\lambda = 0$ (e) RF

(f) $\lambda = 0$ (g) RF

Figure S3: An analysis of the anoles posterior species trees using the RF metric. MDS plots in 2D (a) and 3D (b) do not show well-separated clusters. (c) MCC tree from the lilac RF cluster, showing multiple areas of uncertainty. (d) and (e) relationship between the spread (mean pairwise distance) within a cluster and the level of certainty in the cluster's MCC tree. The level of certainty is measured by the mean of the cluster's MCC support values that are less than 1. In our metric, the mean support of unresolved clades is highest in tightly-defined clusters (d); this is not the case in the RF metric (e). (f) and (g) compare the spread of a cluster with the number of unresolved clades in its MCC. Our metric shows no relationship whereas in the RF metric, clusters with higher spread have more unresolved clades. In (d)-(f), sizes of points correspond to the number of trees in the cluster, and the colors correspond to the MDS clusters for each metric.

The same approach (pairwise distances followed by k-means clustering) that we have used with our distances can, in principle, be used with any metric or quantitative tree comparison tool. We used k-means clustering and comparison of BIC values to cluster the anoles posterior species trees using the Robinson Foulds metric (RF) [20], which is the most widely used. Figure S3 shows the results. k-means clustering will always allow some clusters to be obtained; we found that $k = 12$ groups minimized the BIC.

Two of the RF clusters broadly correspond to the groups we identified (in the sense that the MCCs have the same topology), namely our orange cluster containing the posterior MCC, and our pale orange cluster. The RF clusters are not visibly tightly grouped or well-separated in either 2D or 3D MDS plots (Figures S3a and S3b). Shepard plots show that the correlation between the projected distances and the RF distances is not strong; visible grouping is not a good test of how meaningful the clusters are. We compared the 'tightness' of clusters, measured by the mean tree-tree distances within clusters, to the level of certainty (posterior support values) in MCC trees for the clusters. Our metric has higher MCC support in tighter clusters. In other words, groups of trees a small distance from each other have more highly resolved clades than those far from each other (Figure S3d). The pink RF cluster is the only large, well-supported cluster that the RF metric detects and its topology is the same as the MCC. The other RF clusters have posterior support values that are more uncertain than the posterior itself. In other words, the RF metric does not resolve uncertainty into distinct, well-supported alternative trees (e.g. Figure S3c). In contrast, our metric identifies large, tight clusters with high posterior supports.

2.4 Ebolavirus

Figure S4 supplements Figure 2 from the main text. Shepard plots are provided to indicate the quality of the projections. The distinct VP30 clusters detected by our metric are discernible directly from the distribution of pairwise tree distances (see the Shepard plot: our metric found no tree pairs whose distance was between 7.7 and 10.8). To illustrate the groups of trees (instead of the groups' MCC trees), we have provided their DensiTree [21] plots (Figure S4). These show the same distinct placements of the Sudan clade as are described in the main text.

One difference among the clusters can be interpreted as three different choices of the root for the tree of the three major clades. The fact that this uncertainty in the timing of diversification in the ancient history occurs in VP30 and not in the other 6 genes is a substantial difference. It is also not the only difference between clusters, as the Sudan 2011 and Reston 1990 placements also vary. The VP30 trees also remain distinctive when lengths are taken into account.

Beast estimates of clock rates and the root height differed among the Ebolavirus genes (horizontal scales in the DensiTree plots in Figure S4). When $\lambda = 1$, it is this difference that is primarily detected by the metric and it can also be detected directly from the log files and tree heights. We also compared the Billera, Holmes, Vogtmann (BHV) metric [22] to our $\lambda = 1$ (the most comparable alternative as BHV compares rooted trees and captures branch lengths). In both, differences in root heights overwhelmed structural differences in the trees.

2.5 Chorus Frogs

We present an additional demonstration of our method. Estimating species trees from multiple genes sequenced in multiple taxa is a formidable challenge [23, 24, 3, 25]. Recently, Barrow et al. [26] used anonymous nuclear loci to estimate a phylogeny for the North America genus of chorus frogs, *Pseudacris*. As in the case of ecomorph trunk anoles (main text), this genus is a model system with evidence of reproductive character displacement, allopatric divergence, and hybridization and reinforcement [27, 28, 29, 30, 31, 32]. Data included sequences for 44 individuals, from 3 mitochondrial loci and 27 nuclear loci. Full methods and data are available in [26, 33] respectively. Barrow et al. found that four major clades of frogs were supported consistently but that there was discordance between trees derived from nuclear and mitochondrial data. They interpreted this as a signal of a possible selective sweep, or mitochondrial introgression [26].

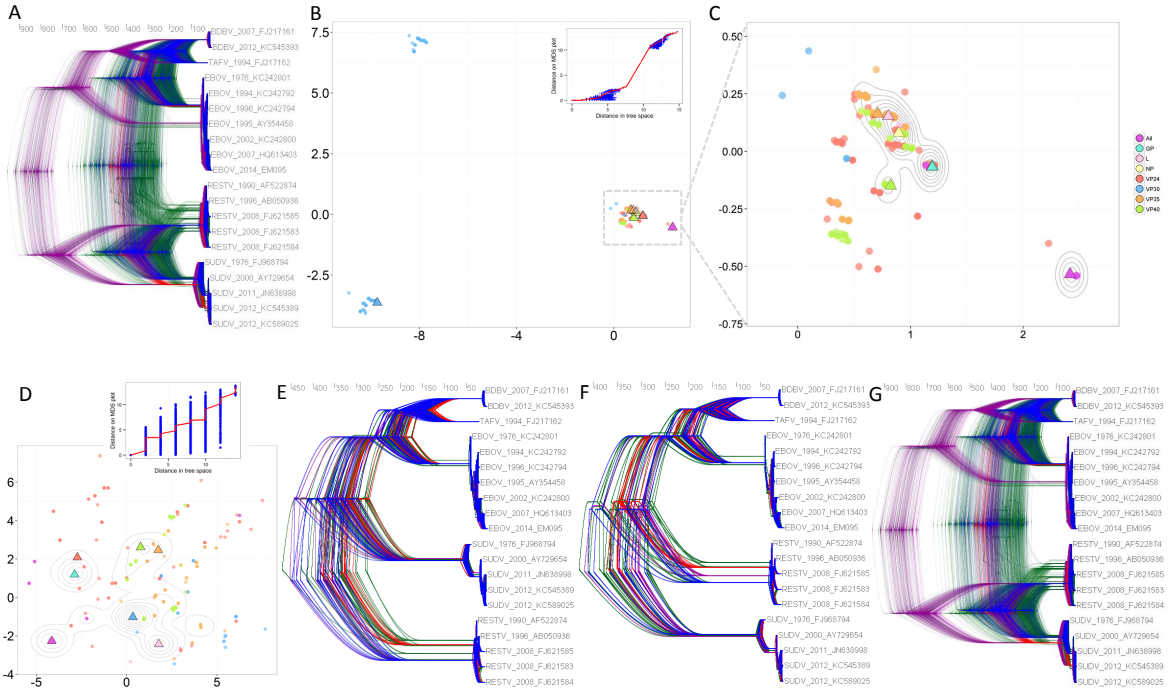


Figure S4: Ebola analysis, parallel to Figure 2 in the main text, showing Shepard plots for our metric's MDS and RF. DensiTree [21] images show the resolution of the Sudan clade within each cluster. In a reflection of Figure 2 from the main text, (A) displays all 1200 trees from individual genes and concatenated alignments. Trees from each cluster are shown in (E) bottom left VP30, (F) top right VP30 and (G) main.

The trees are broadly concordant but there are a number of points of uncertainty in the posterior MCC (Figure S5). We used posterior tree file `species_1367351410414.trees` with MCC `2allele-44taxa-2Kburn.tree`, available in [33]). We sampled 1000 trees from the posterior, computed tree-tree distances, and visualized the posterior with MDS and k-means clustering.

There are several tightly-defined clusters of distinct trees. Clusters have much higher MCC support values than the MCC for the whole posterior. In particular, clusters differ in whether *triseriata* and *kalmi* are sister clades (as in the posterior MCC, the orange cluster and the purple cluster) or, alternatively, *kalmi*, *ferariumC* and *ferariumI* form a sister clade to *triseriata* (e.g. red cluster, light purple cluster). They also differ in the timing of *clarki*'s divergence from the *brimlei*/*brachyphona* clade, and at several other points. The clusters represent alternative, well-supported patterns in the frogs' evolution.

3 Discussion

Our metric compares trees with the same set of taxa (i.e. the same tips). As a consequence, it is suited for the kinds of questions we have described, in which there is one set of taxa and trees can be compared from different genes, inference methods, and sources of data. Our metric does not capture distances between trees with different taxa; where the taxa overlap between two trees, our approach can compare the subtrees restricted to the taxa present in both trees. In contrast, comparisons between *unlabeled* trees take a different form (e.g. kernel methods [34]), suitable to comparing trees on different sets of taxa.

Many phylogenetic analyses are, implicitly or explicitly, conducted in the context of a rooted tree. In the context of macroevolution, examples include estimates of times to divergence, ancestral relationships and ancestral character reconstruction. In more recent literature, most methods to link pathogen phylogenies to epidemic dynamics (phylodynamics) [35, 36, 37] are based on

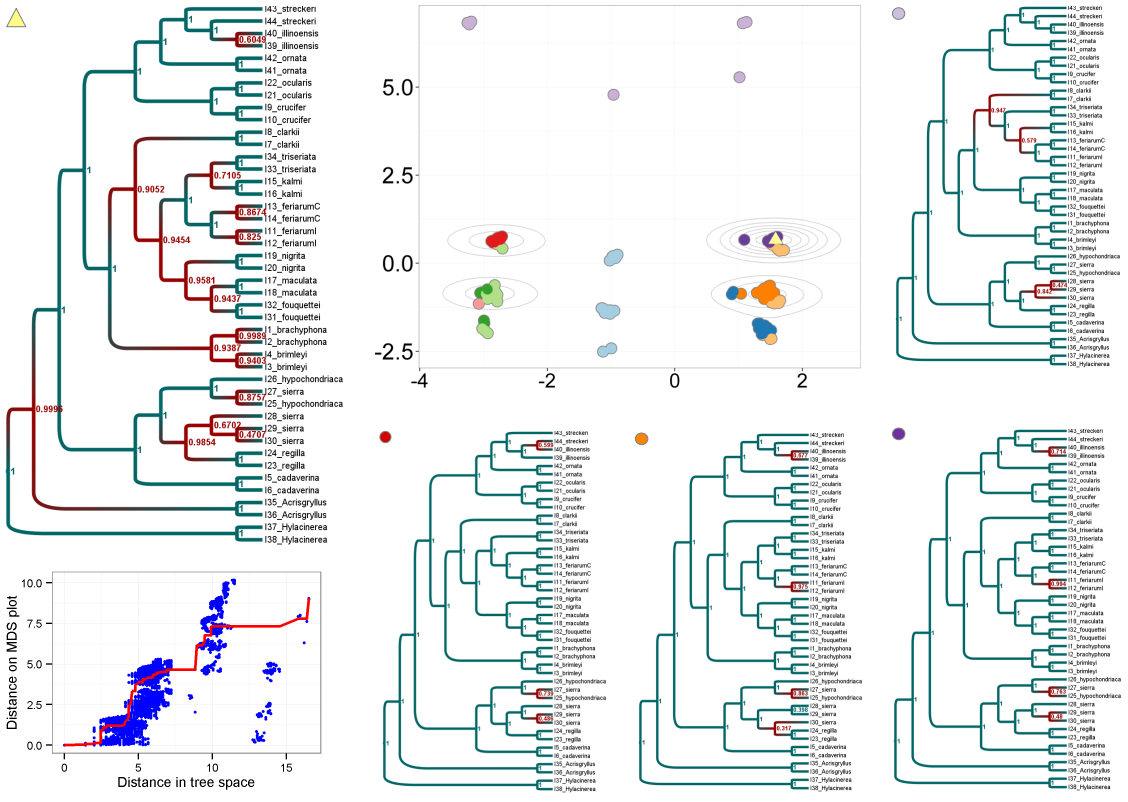


Figure S5: Chorus tree frog analysis showing the correspondence between clusters and distinct topologies. The overall MCC tree is marked by a yellow triangle, and examples of cluster MCC trees are shown, as indicated. Again, the existence of clusters is visible in the Shepard plot in the distribution of distances along the horizontal axis.

rooted phylogenetic trees. For these reasons, the fact that the relationships to the root of the tree play a central role in our metric allows it to capture intuitive similarities in groups of trees in a way that other metrics do not.

Sequencing technologies continue to decrease in cost, with the result that it is now feasible to sequence up to tens of thousands of taxa in multiple genes, at least in viruses. As the computational challenges associated with haplotype phasing are resolved [38, 39, 40], phylogenetic methods will be used for large datasets on higher organisms. As an example, the Epidemiology Network Ag1000G has 765 Anopheles mosquito genomes visible to the public [41]. Currently, Bayesian tree inference is not feasible beyond hundreds of taxa, and maximum-likelihood, maximum-parsimony and neighbor-joining methods are used with add-on estimation of the root and timing information by software such as Path-O-Gen [42] and LSD [43]. While this presents challenges for tree comparison, it also provides compelling motivation to develop appropriate tree comparison metrics. Our computation time per tree pair is polynomial (k^2) in the number of tips. For very large numbers of tips, the vectors become infeasibly long and may need to be represented in a more efficient format (as many pairs of tips have the same m and M values). Exploiting different ways to navigate the tree may also yield more efficient computation of $v(T)$.

The fact that our metric is a Euclidean distance between two vectors whose components have an intuitive description means that simple extensions are straightforward to imagine and to compute. For example, it may be the case that the placement of a particular tip is a key question. This could occur, for example, in a real-time analysis of an outbreak, where new cases need to be placed on an existing phylogeny to determine the likely source of infection. We could form a metric that emphasizes differences in the placement of a particular tip (say, A), by weighting A's entries of m and M highly compared to all other entries. In this new metric, trees would appear

similar if their placement of A was similar; patterns of ancestry among the other tips would contribute less to the distance. Indeed, it is possible to design numerous metrics, extending this one and others, and using linear combinations of existing metrics [44]. Ours has the advantages we have presented here.

References

[1] Geneva, A. J., Hilton, J., Noll, S. & Glor, R. E. Multilocus phylogenetic analyses of Hispaniolan and Bahamian trunk anoles (distichus species group). *Molecular Phylogenetics and Evolution* **87**, 105–117 (2015).

[2] Geneva, A. J., Hilton, J., Noll, S. & Glor, R. E. Data from Dryad Digital Repository (2015). doi: 10.5061/dryad.622h6.

[3] Heled, J. & Drummond, A. J. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27**, 570–580 (2010).

[4] R Core Team. R: a language and environment for statistical computing (2015). http://www.r-project.org/.

[5] Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).

[6] Chessel, D., Dufour, A. B. & Thioulouse, J. The ade4 package-I-One-table methods. *R news* **4**, 5–10 (2004).

[7] Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214 (2007).

[8] Rambaut, A. FigTree (2006). http://tree.bio.ed.ac.uk/software/figtree.

[9] Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).

[10] Kendall, M. & Colijn, C. A tree metric using structure and length to capture distinct phylogenetic signals. Preprint (2015). http://arxiv.org/abs/1507.05211.

[11] Maddison, D. R. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology* **40**, 315–328 (1991).

[12] Salter, L. A. & Pearl, D. K. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Systematic Biology* **50**, 7–17 (2001).

[13] Sanderson, M. J., McMahon, M. M. & Steel, M. Terraces in phylogenetic tree space. *Science* **333**, 448–450 (2011).

[14] Heled, J. & Bouckaert, R. R. Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology* **13**, 221 (2013).

[15] Sullivan, J., Holsinger, K. E. & Simon, C. The effect of topology on estimates of among-site rate variation. *Journal of Molecular Evolution* **42**, 308–312 (1996).

[16] Haldane, J. B. S. Note on the median of a multivariate distribution. *Biometrika* **35**, 414–415 (1948).

[17] Hillis, D. M., Heath, T. A. & St John, K. Analysis and visualization of tree space. *Systematic Biology* **54**, 471–482 (2005).

[18] Berglund, D. *Visualization of Phylogenetic Tree Space.* Ph.D. thesis, Stockholm University (2011).

[19] Shepard, R. N., Romney, A. K. & Nerlove, S. B. *Multidimensional Scaling: Theory and applications in the behavioural sciences: I. Theory.* (Seminar press, 1972).

[20] Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131–147 (1981).

[21] Bouckaert, R. R. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**, 1372–1373 (2010).

[22] Billera, L. J., Holmes, S. P. & Vogtmann, K. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics* **27**, 733–767 (2001).

[23] Nichols, R. Gene trees and species trees are not the same. *Trends Ecol. Evol.* **16**, 358–364 (2001).

[24] Degnan, J. H. & Rosenberg, N. A. Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2**, e68 (2006).

[25] Anderson, C. N. K., Liu, L., Pearl, D. & Edwards, S. V. Tangled trees: the challenge of inferring species trees from coalescent and noncoalescent genes. *Methods Mol. Biol.* **856**, 3–28 (2012).

[26] Barrow, L. N., Ralicki, H. F., Emme, S. a. & Lemmon, E. M. Species tree estimation of North American chorus frogs (Hylidae: Pseudacris) with parallel tagged amplicon sequencing. *Molecular Phylogenetics and Evolution* **75**, 78–90 (2014).

[27] Fouquette Jr, M. J. Speciation in Chorus Frogs. I. Reproductive Character Displacement in the Pseudacris Nigrita Complex. *Systematic Biology* **24**, 16–23 (1975).

[28] Gartside, D. F. Analysis of a Hybrid Zone between Chorus Frogs of the Pseudacris nigrita Complex in the Southern United States. *Copeia* **1980**, 56–66 (1980).

[29] Lemmon, E. M., Lemmon, A. R. & Cannatella, D. C. Geological and climatic forces driving speciation in the continentally distributed trilling chorus frogs (Pseudacris). *Evolution* **61**, 2086–2103 (2007).

[30] Lemmon, E. M., Lemmon, A. R., Collins, J. T., Lee-Yaw, J. a. & Cannatella, D. C. Phylogeny-based delimitation of species boundaries and contact zones in the trilling chorus frogs (Pseudacris). *Molecular Phylogenetics and Evolution* **44**, 1068–1082 (2007).

[31] Lemmon, E. M. Diversification of conspecific signals in sympatry: Geographic overlap drives multidimensional reproductive character displacement in frogs. *Evolution* **63**, 1155–1170 (2009).

[32] Lemmon, E. M. & Lemmon, A. R. Reinforcement in chorus frogs: Lifetime fitness estimates including intrinsic natural selection and sexual selection against hybrids. *Evolution* **64**, 1748–1761 (2010).

[33] Barrow, L. N., Ralicki, H. F., Emme, S. A. & Moriarty Lemmon, E. Data from Dryad Digital Repository (2014). doi: 10.5061/dryad.23rc0.

[34] Poon, A. F. Y. *et al.* Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PloS One* **8**, e78122 (2013).

[35] Stadler, T. *et al.* Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution* (2011). `http://mbe.oxfordjournals.org/content/early/2011/09/02/molbev.msr217.full.pdf+html`.

[36] Rasmussen, D. A., Volz, E. M. & Koelle, K. Phylodynamic inference for structured epidemiological models. *PLoS Computational Biology* **10**, e1003570 (2014).

[37] Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution* **31**, 1869–1879 (2014).

[38] Pan, W., Zhao, Y., Xu, Y. & Zhou, F. {WinHAP2}: an extremely fast haplotype phasing program for long genotype sequences. *BMC Bioinformatics* **15**, 164 (2014).

[39] Zhi, D. & Zhang, K. Genotype Calling and Haplotype Phasing from Next Generation Sequencing Data. In *Statistical Analysis of Next Generation Sequencing Data*, Frontiers in Probability and the Statistical Sciences, 315–333 (Springer International Publishing, 2014).

[40] Regan, J. F. *et al.* A rapid molecular approach for chromosomal phasing. *PLoS One* **10**, e0118270 (2015).

[41] MalariaGEN. Oxford, UK. http://www.malariagen.net/projects/vector/ag1000g, accessed July 9th 2015.

[42] Rambaut, A. Path-O-Gen (2009). http://tree.bio.ed.ac.uk/software/pathogen.

[43] Gascuel, O., To, T. H., Jung, M. & Lycett, S. http://www.atgc-montpellier.fr/LSD.

[44] Liebscher, V. Gromov meets phylogenetics — new animals for the zoo of biocomputable metrics on tree space. Preprint (2015). arXiv:1504.05795v1.