User's Manual — Random Tanglegram Partitions (Random TaPas): An Alexandrian approach to the cophylogenetic Gordian knot

Juan Antonio Balbuena[1], Óscar Alejandro Pérez-Escobar[2], Cristina Llopis-Belenguer[1], Isabel Blasco-Costa[3]

[1]Ecology and Evolution of Symbionts Lab, Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, Official P.O. Box 22085, 46071 Valencia, Spain; [2]Identification and Naming Department, Royal Botanic Gardens, Kew, Richmond, TW9 3AB, U.K., [3]Department of Invertebrates, Natural History Museum of Geneva, P.O. Box 6134, CH-1211 Geneva, Switzerland.

This manual illustrates how to implement Random TaPas with a script written in R (R Core Team 2018). The prior installation of the following packages: distory (Chakerian and Holmes 2010), paco (Hutchinson et al. 2017) and phytools (Revell 2012) is required. We demonstrate the use of the script with data of mitochondrial haplotypes of the cytochrome oxidase subunit 1 of the trematode, *Coitocaecum parvum* (Crowcroft, 1945), and those of its amphipod host, *Paracalliope fluviatilis* (Thomson, 1879), from seven locations in South Island, New Zealand (Lagrue et al. 2016). A clean version of the script can be downloaded at https://github.com/Ligophorus/RandomTaPas.

INPUT FILES

The input is represented by a triple (*H*, *S*, *A*), where *H* and *S* represent the phylogenies of hosts and symbionts, and **A** is a binary matrix with rows and columns corresponding to terminals in *H* and *S*, respectively, in which extant associations between each terminal are coded as 1 and no associations as 0. The input required is two Nexus or Newick files containing the phylogenetic relationships of hosts (*H*) and their symbionts (*S*) and one ASCII file that codes the associations between hosts (rows) and symbionts (columns) (**A**). Should the user wish to incorporate phylogenetic uncertainty in the analysis, two additional Nexus or Newick files with the set of posterior-probability trees used to build the consensus trees of hosts (*mH*) and symbionts (*mS*) need to be read. Matrix **A** should include row and column labels to identify the taxa. These labels have to match exactly those of the terminals in *H* (rows) and *S* (columns), but their order does not need to be the same than in the corresponding phylogenetic trees.

LIBRARIES AND CUSTOM-MADE FUNCTIONS

First, load the packages required in the R session:

```
library(paco)
library(phytools)
library(distory)
library(parallel)
```

Note that package `parallel` is convenient for multi-core computing for estimation of confidence intervals of the host-symbiont frequencies from the sets of posterior-trees.

Next, a series of custom-made functions need to be read:

`trimHS.maxC` For *i* in 1 to *N*, `trimHS.maxC` randomly chooses *n* unique associations in **A**, so that each terminal in *H* is associated to one, and only one, terminal in *S*, and vice versa. It trims **A** to produce $\mathbf{A}_i$, which includes only the *n* associations. The function returns a list of the *N* trimmed matrices. The user can optionally remove duplicated $\mathbf{A}_i$s produced over the *N* runs by setting `check.unique= TRUE`. We recommend this alternative if *n* is small, because the likelihood of obtaining the same $\mathbf{A}_i$ in different runs increases as *n* decreases.

```
trimHS.maxC <- function (N, HS, n, check.unique= FALSE) {
  trim.int <- function (x, HS, n) {
    HS.LUT <- which(HS == 1, arr.in= TRUE)
    HS.LUT <- cbind(HS.LUT, 1:nrow(HS.LUT))
    df <- as.data.frame(HS.LUT)
    hs.lut <- subset(df[sample(nrow(df)), ],
                     !duplicated(row) & !duplicated(col))
    if (nrow(hs.lut) < n) hs <- NULL else {
      hs.lut <- hs.lut[sample(nrow(hs.lut), n), ]
      hs <- diag(nrow(hs.lut))
      rownames(hs) <- rownames(HS[hs.lut[ ,1], ])
      colnames(hs) <- colnames(HS[ ,hs.lut[ ,2]])
      return(hs)
    }
  }
  trim.HS <- lapply(1:N, trim.int, HS= HS, n= n )
  if (check.unique == TRUE) trim.HS <- unique(trim.HS)
  if (length(trim.HS) < N) warning("No. of trimmed H-S assoc. matrices < No. of runs")
  return(trim.HS)
}
```

`geo.D` For any $\mathbf{A}_i$ produced with `trimHS.maxC`, `geo.D` prunes *H* and *S* to leave only terminals represented in $\mathbf{A}_i$, computes and returns the geodesic distance (Chakerian and Holmes 2010) between the pruned trees. A requirement is that *H* and *S* are strictly bifurcating trees.

```
geo.D <- function (hs, treeH, treeS) {
  treeh <- ape::drop.tip(treeH, setdiff(treeH$tip.label, rownames(hs)))
  trees <- ape::drop.tip(treeS, setdiff(treeS$tip.label, colnames(hs)))
    # foo distory requires same labels in both trees. Dummy labels are produced.
    # 1st reorder hs as per tree labels:
  hs <- hs[treeh$tip.label, trees$tip.label]
    # 2nd swap trees labels with corresponding ones in treeh:
  hs.lut <- which(hs[treeh$tip.label, trees$tip.label]==1, arr.ind = TRUE)
  dummy.labels <- rownames(hs.lut)
  trees$tip.label <- dummy.labels
  combo.tree <- list(treeh, trees)
  gd <- distory::dist.multiPhylo(combo.tree)
  return(gd)
}
```

`paco.ss` For any $\mathbf{A}_i$ produced with `trimHS.maxC`, `paco.ss` prunes *H* and *S* to leave only terminals represented in $\mathbf{A}_i$. Then it applies Procrustes Approach to Cophylogeny (PACo) (Balbuena et al. 2013) to $\mathbf{A}_i$ and the pruned trees, and computes and returns the sum of squared residuals of the Procrustes superposition of the host and symbiont configurations in Euclidean space. `symmetric= FALSE` (the default) indicates that the superposition is applied asymmetrically (*S* depends on *H*). `symmetric= TRUE` will apply PACo symmetrically (dependency between *S* and *H* is reciprocal). `proc.warns= FALSE` (the default) suppresses trivial warnings returned when the phylogenies differ in the numbers of tips (see Hutchinson et al. (2017) for details). The user can modify `ei.correct` to set how to correct potential

negative eigenvalues resulting from the conversion of phylogenetic distances into Principal Coordinates (Balbuena et al. 2013; Hutchinson et al. 2017): `sqrt.D` (the default) takes the element-wise square-root of the phylogenetic distances (de Vienne et al. 2011). Other options are `lingoes`, `cailliez` (Lingoes and Cailliez corrections, respectively) and `none` (when no correction is required, particularly if H and S are ultrametric, de Vienne et al. 2011).

```
paco.ss <- function (hs, treeH, treeS, symmetric= FALSE,
                         proc.warns= FALSE, ei.correct= "sqrt.D") {
  eigen.choice <- c("none", "lingoes", "cailliez", "sqrt.D")
  if (ei.correct %in% eigen.choice == FALSE)
                   stop(writeLines("Invalid eigenvalue correction parameter.\r
              Correct choices are 'none', 'lingoes', 'cailliez' or 'sqrt.D'"))
  treeh <- ape::drop.tip(treeH, setdiff(treeH$tip.label, rownames(hs)))
  trees <- ape::drop.tip(treeS, setdiff(treeS$tip.label, colnames(hs)))
  # Reorder hs as per tree labels:
  hs <- hs[treeh$tip.label, trees$tip.label]
  DH <- cophenetic(treeh)
  DP <- cophenetic(trees)
  if (ei.correct == "sqrt.D"){DH <- sqrt(DH) ; DP <- sqrt(DP); ei.correct="none"}
  D <- paco::prepare_paco_data(DH, DP, hs)
  D <- paco::add_pcoord(D, correction= ei.correct)
  if (proc.warns == FALSE) D <- vegan::procrustes(D$H_PCo, D$P_PCo,
                                                  symmetric = symmetric) else
    D <- suppressWarnings(vegan::procrustes(D$H_PCo, D$P_PCo,
                                                  symmetric = symmetric))
  return(D$ss)
}
```

`link.freq` After applying `geo.D` or `paco.ss` to each $\mathbf{A}_i$ produced by `trimHS.maxC`, this function determines the frequency (as percentage) of each host-symbiont association occurring in a given percentile of cases that maximize phylogenetic congruence, i.e. either the fraction of lowest geodesic distances (`geo.D`) or lowest sum of squared residuals (`paco.ss`). The output is a data frame with host-symbiont associations in rows. The first and second columns display the names of the host and symbiont terminals, respectively. The third column designates the host-symbiont association by pasting the names of the terminals and the fourth one represents the frequency of each association. The percentile applied (0.05 by default) can be specified with `percentile`. The user can also establish the character separating the host and symbiont names (`sep`). For future usage, frequencies of host-symbiont associations above a given percentile value can also be computed setting `below.p= FALSE`.

```
link.freq <- function (x, fx, HS, percentile= 0.05, sep= "-", below.p= TRUE) {
  if (below.p == TRUE) percent <- which(fx <= quantile(fx, percentile)) else
    percent <- which(fx >= quantile(fx, percentile))
  trim.HS <- x[percent]
  paste.link.names <- function(X, sep) {
    X.bin <- which(X>0, arr.in=TRUE)
    Y <- diag(nrow(X.bin))
    Y <- diag(nrow(X.bin))
    rownames(Y) <- rownames(X)[X.bin[,1]]
    colnames(Y) <- colnames(X)[X.bin[,2]]
    pln <- paste(rownames(Y), colnames(Y), sep=sep)
    return(pln)
  }
  link.names <- t(sapply(trim.HS, paste.link.names, sep=sep))
  lf <- table(link.names)
  lf <- as.data.frame(lf*100 / length(trim.HS))
  HS.LUT <- which(HS ==1, arr.in=TRUE)
```

```
  linkf <-  as.data.frame(cbind(rownames(HS)[HS.LUT[,1]],
                                colnames(HS)[HS.LUT[,2]]))
  colnames(linkf) <- c('H', 'S')
  linkf$HS <- paste(linkf[,1], linkf[,2], sep=sep)
  linkf$Freq <- rep(0, nrow(linkf))
  linkf[match(lf[,1], linkf[,3]), 4] <- lf[,2]
  return(linkf)
}
```

`One2one.f` For a particular **A**, the function returns the maximum *n* for which all $\mathbf{A}_i$s can be built over a number of replicates (`reps`), 10^4 by default. It can be used to decide the best *n* (number of unique associations in **A**) prior to application of `trimHS.maxC`. (See example below.)

```
One2one.f <- function (hs, reps=1e+4) {
  HS.LUT <- which(hs ==1, arr.in=TRUE)
  HS.LUT <- cbind(HS.LUT,1:nrow(HS.LUT))
  df <- as.data.frame(HS.LUT)
  V <- rep(NA,reps)
  for(i in 1:reps){
    hs.lut <- subset(df[sample(nrow(df)),],
                     !duplicated(row) & !duplicated(col))
    n <- sum(HS)
    while (n >0) {
      n <- n-1;
      if (nrow(hs.lut) == n) break
    }
    V[i]<- n
  }
  V <- min(V)
  return(V)
}
```

`tangle.gram` This wrapper of `plot.cophylo` (package `phytools`) is used for mapping as heatmap the host-symbiont frequencies estimated by Random TaPas on the tanglegram. It also plots the average frequency of occurrence of each terminal and optionally, the fast maximum likelihood estimators of ancestral states of each node if `node.tag= TRUE` (the default). For a given color scale, the user can choose whether it is relative (`rcolgrad= TRUE`, the default), i.e. spanning from the minimum to the maximum frequencies observed, or absolute (`rcolgrad= FALSE`), i.e. spanning from 0% to 100%. The user can also customize the number of breaks on the color gradient (`nbreaks`), the size of the color points at terminals and nodes (`cexpt`) and additional visualization options included in `plot.cophylo` (see https://www.rdocumentation.org/packages/phytools/versions/0.6-60/topics/cophylo).

```
tangle.gram <- function(treeH, treeS, hs, colgrad, rcolgrad= TRUE, nbreaks=50,
                        fqtab, node.tag=TRUE, cexpt=1, ...) {
  rescale.range <- function(x) {
    if (rcolgrad==TRUE) {
      x <- round(x)
      y <- range(x)
      col_lim <- (y[1]: y[2])-y[1]+1
      x <- x-y[1]+1
      new.range <- list(col_lim, x)
    } else {
      new.range <- list(1:101, round(x)+1)
    }
    return(new.range)
  }
  NR <- rescale.range(fqtab[,4])
  rbPal <- colorRampPalette(colgrad)
  linkcolor <- rbPal(nbreaks)[as.numeric(cut(NR[[1]],breaks = nbreaks))]
  HS.lut <- which(hs ==1, arr.ind=TRUE)
```

```
  linkhs <- cbind(rownames(hs)[HS.lut[,1]], colnames(hs)[HS.lut[,2]])
  obj <- phytools::cophylo(treeH,treeS, linkhs)
  phytools::plot.cophylo(obj, link.col=linkcolor[NR[[2]]], ...)

  Hfreq <- aggregate(fqtab[,4], by=list(freq = fqtab[,1]), FUN=mean)
  Sfreq <- aggregate(fqtab[,4], by=list(freq = fqtab[,2]), FUN=mean)

  Hfreq <- Hfreq[match(obj$trees[[1]]$tip.label, Hfreq$freq),]
  Sfreq <- Sfreq[match(obj$trees[[2]]$tip.label, Sfreq$freq),]

  if(node.tag==TRUE){
  fit.H<-fastAnc(obj$trees[[1]],Hfreq[,2])
  fit.S<-fastAnc(obj$trees[[2]],Sfreq[,2])
  NLH <- rescale.range (fit.H)
  NLS <- rescale.range (fit.S)
  nodelabels.cophylo(pch=16, col=linkcolor[NLH[[2]]], cex=cexpt)
  nodelabels.cophylo(pch=16, col=linkcolor[NLS[[2]]], cex=cexpt, which="right")
  }
  TLH <- rescale.range (Hfreq[,2])
  TLS <- rescale.range (Sfreq[,2])
  tiplabels.cophylo(pch=16, col=linkcolor[TLH[[2]]], cex=cexpt)
  tiplabels.cophylo(pch=16, col=linkcolor[TLS[[2]]], cex=cexpt, which="right")
}
```

INPUT AND PRELIMINARY EXPLORATION OF DATA

The input files need to be read. The syntax below implies that the input files are in the working directory of the R session. Otherwise, the user should also supply the path to the file or alternatively use `file.choose()` to open them interactively.

```
TreeH <- read.nexus("Pfluviatilis.tre")        # Amphipod (host) consensus tree
TreeS <- read.nexus("Cparvum.tre")             # Trematode (symbiont) consensus tree
mTreeH <- read.nexus("m_Pfluviatilis.tre")     # 1,000 prob. amphipod trees
mTreeS <- read.nexus("m_Cparvum.tre")          # 1,000 post. prob. trematode trees
HS <- as.matrix(read.table("PfxCp_assoc.txt", header = TRUE)) # Association matrix
```

Optionally, the user can plot the tanglegram to visualize the host-symbiont relationships. This is implemented with `cophylo` and `plot.cophylo` of package `phytools` (Fig. 1).

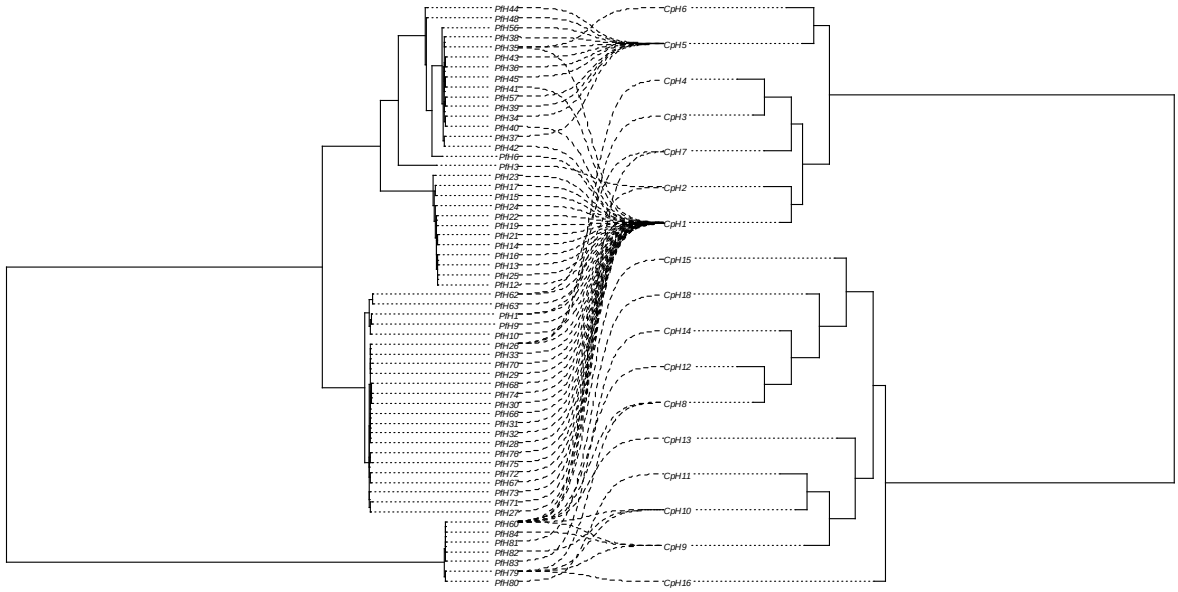

Figure 1. Tanglegram showing the relationships between haplotypes of of the trematode, *Coitocaecum parvum* (Crowcroft, 1945), and its amphipod host, *Paracalliope fluviatilis.*

```
HS.lut <- which(HS ==1, arr.ind=TRUE)
linkhs <- cbind(rownames(HS)[HS.lut[,1]], colnames(HS)[HS.lut[,2]])
obj <- cophylo(TreeH,TreeS, linkhs)
plot.cophylo(obj, link.lwd=1, link.lty=2, fsize=0.5, pts=FALSE, link.type="curved")
```

We advise to reset the graphics device, as `plot.cophylo` can alter the default margin settings of subsequent plots:

```
dev.off()
```

In this example, we set *N*, number of Random TaPas runs, at 10^4.

```
N= 1e+4
```

Setting the number of unique host-associations, *n*, can be tricky in some host-symbiont systems. Below we demonstrate a procedure to achieve this.

First, we run `One2one.f` 10^4 times with the given association matrix:

```
One2one.f(HS,N)
[1] 5
```

Thus, 5 is the maximum *n* that would allow producing 10^4 $\mathbf{A}_i$s. (Because Random TaPas is based on randomization, occasionally the output with the present data can be 6). In the accompanying publication, we recommend using an *n* close to 20%, or at least 10%, of the total number of associations (75 in this example). Therefore, *n*=5 is not optimal. So we explore the number of $\mathbf{A}_i$s that can be produced with *n* ranging from 5 to 10:

```
X <- 5:10
Y <- rep(NA, length(X))
for(i in 1:length(X)) {
  THS <- trimHS.maxC(N, HS, n=X[i], check.unique=TRUE)
  Y[i] <- length(THS)
}
Warning in trimHS.maxC(N, HS, n = X[i], check.unique = TRUE): No. of
trimmed H-S assoc. matrices < No. of runs

Warning in trimHS.maxC(N, HS, n = X[i], check.unique = TRUE): No. of
trimmed H-S assoc. matrices < No. of runs

Warning in trimHS.maxC(N, HS, n = X[i], check.unique = TRUE): No. of
trimmed H-S assoc. matrices < No. of runs

Warning in trimHS.maxC(N, HS, n = X[i], check.unique = TRUE): No. of
trimmed H-S assoc. matrices < No. of runs

Warning in trimHS.maxC(N, HS, n = X[i], check.unique = TRUE): No. of
trimmed H-S assoc. matrices < No. of runs

Warning in trimHS.maxC(N, HS, n = X[i], check.unique = TRUE): No. of
trimmed H-S assoc. matrices < No. of runs
```

Note that since `check.unique= TRUE`, the function evaluates the number of unique $\mathbf{A}_i$s produced with each *n*. The warnings tell that this number is, in all cases, less than *N*. The relationship between *n* and the number of unique $\mathbf{A}_i$s can be plotted (Fig. 2):

```
plot(X,Y, type="b", xlab="Number of unique H-S associations",
                    ylab="Number of runs accomplished")
```

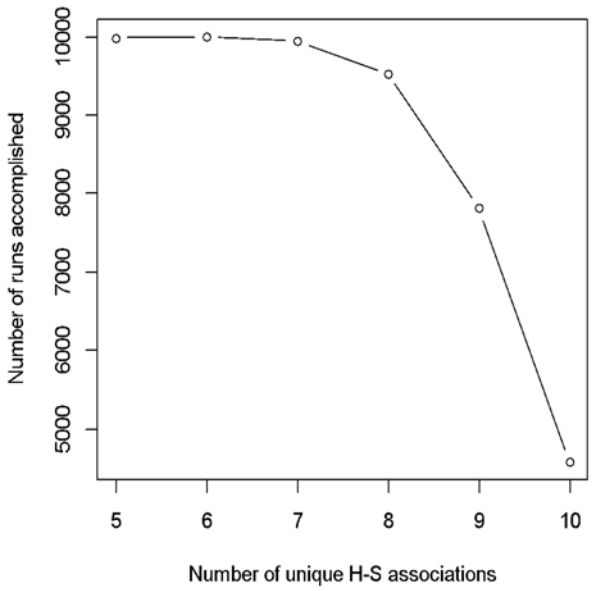

Figure 2. Number of Random TaPas runs that can be accomplished with a given number of unique host-symbiont associations (*n*).

We choose $n = 8$ since it offers a good compromise between the number of unique $\mathbf{A}_i$s that can be evaluated (> 9,500) and the fraction (>10%) of total number of host-symbiont associations.

Applying Random TaPas to the example

Once N and n are set, the $\mathbf{A}_i$s are generated with `trimHS.maxC`.

```
n=8
THS <- trimHS.maxC(N, HS, n=n, check.unique=TRUE)
```

```
Warning in trimHS.maxC(N, HS, n = 8, check.unique = TRUE): No. of trimmed
H-S assoc. matrices < No. of runs
```

```
THS[sapply(THS, is.null)] <- NULL
```

The warning is the same as above. Since the number of valid $\mathbf{A}_i$s < N, the last line of code removes all null cases in THS.

Now, `geo.D` and `paco.ss` are applied over the set of valid $\mathbf{A}_i$s to obtain geodesic distances and sum of squared residuals, respectively. Note that because codivergence between amphipod and trematode lineages was assumed to be reciprocal, we set PACo to run in symmetric mode (`symmetric=TRUE`).

```
GD <- sapply(THS, geo.D, treeH=TreeH, treeS= TreeS)
PACO <- sapply(THS, paco.ss, treeH=TreeH, treeS= TreeS, symmetric=TRUE)
```

Then, we extract the frequencies of host-symbiont associations represented in the 5% percentile:

```
LFGD05 <- link.freq(THS, GD, HS, percentile=0.05)
LFPACO05 <- link.freq(THS, PACO, HS, percentile=0.05)
```

Optionally the user can compute 95% confidence intervals for these frequencies using sets of posterior probability trees. We use here 1,000 pairs of such trees. This task is compute-intensive and we recommend using parallel computing. First, we define two empty matrices to save the frequencies of host-symbiont associations for each tree pair produced with both, geodesic distances and PACo:

```
GD05 <- matrix(NA, length(mTreeH), nrow(LFGD05))
PACO05 <- matrix(NA, length(mTreeH), nrow(LFPACO05))
```

Then, determine the number of CPU cores in the system and we choose here to use all except two. (A text progress bar is also set to monitor the progress of the operation.)

```
cores <- detectCores()
pb <- txtProgressBar(min = 0, max = length(mTreeH), style = 3)
cl <- makeCluster(cores-2)
for(i in 1:length(mTreeH))       # CIs geodesic distances
{
  GD.CI<-parallel::parSapply(cl, THS, geo.D, treeH=mTreeH[[i]],
                             treeS= mTreeS[[i]])
  LFGD05.CI <- link.freq(THS, GD.CI, HS, percentile=0.05)
  GD05[i,] <- LFGD05.CI[,4]
  setTxtProgressBar(pb, i)
}
close(pb)
#
pb <- txtProgressBar(min = 0, max = length(mTreeH), style = 3)
for(i in 1:length(mTreeH))       # CIs PACo
{
  PA.CI<-parallel::parSapply(cl, THS, paco.ss, treeH=mTreeH[[i]],
                             treeS= mTreeS[[i]], symmetric=TRUE)
  LFPA05.CI <- link.freq(THS, PA.CI, HS, percentile=0.05)
  PACO05[i,] <- LFPA05.CI[,4]
  setTxtProgressBar(pb, i)
}
close(pb)
stopCluster(cl)
#
colnames(GD05) <- LFGD05[,3]
colnames(PACO05) <- LFPACO05[,3]
```

Compute 95% confidence intervals and mean values of frequencies of host-symbiont associations across posterior probability trees estimated with Random TaPas using geodesic distances and PACo:

```
GD.LO <- apply(GD05, 2, quantile, 0.025)
GD.HI <- apply(GD05, 2, quantile, 0.975)
GD.AV <- apply(GD05, 2, mean)

PACO.LO <- apply(PACO05, 2, quantile, 0.025)
PACO.HI <- apply(PACO05, 2, quantile, 0.975)
PACO.AV <- apply(PACO05, 2, mean)
```

VISUALIZATION OF RESULTS

The mean host-symbiont frequencies of the 1,000 pairs of posterior probability trees produced by Random TaPas, with their confidence intervals are displayed as barplots with error bars (Fig. 3):

```
op <- par(mfrow=c(2,1),mgp = c(1, 0.2, 0), mar=c(2.8,2.5,0.2,1.5), tck = 0.02)
link.fq <-barplot(GD.AV, names.arg = LFGD05$HS,
        horiz=FALSE, cex.names = 0.5, las=2, cex.axis=0.8, ylab="Frequency (%)",
        ylim=c(0, max(GD.HI)), col="lightblue")
suppressWarnings(arrows(link.fq, GD.HI, link.fq, GD.LO, length= 0,
                        angle=90, code=3, col="darkblue", lwd=0.5))
abline(h=n/sum(HS)*100, col="red")

link.fq <-barplot(PACO.AV, names.arg = LFPACO05$HS,
        horiz=FALSE, cex.names = 0.5, las=2, cex.axis=0.6, ylab="Frequency (%)",
        ylim=c(0, max(PACO.HI)),col="lightblue")
suppressWarnings(arrows(link.fq, PACO.HI, link.fq, PACO.LO, length= 0,
                        angle=90, code=3, col="darkblue", lwd=0.5))
abline(h=n/sum(HS)*100, col="red")
par(op)
```

`suppressWarnings` is used to omit warnings related to zero-length error bars not being drawn.

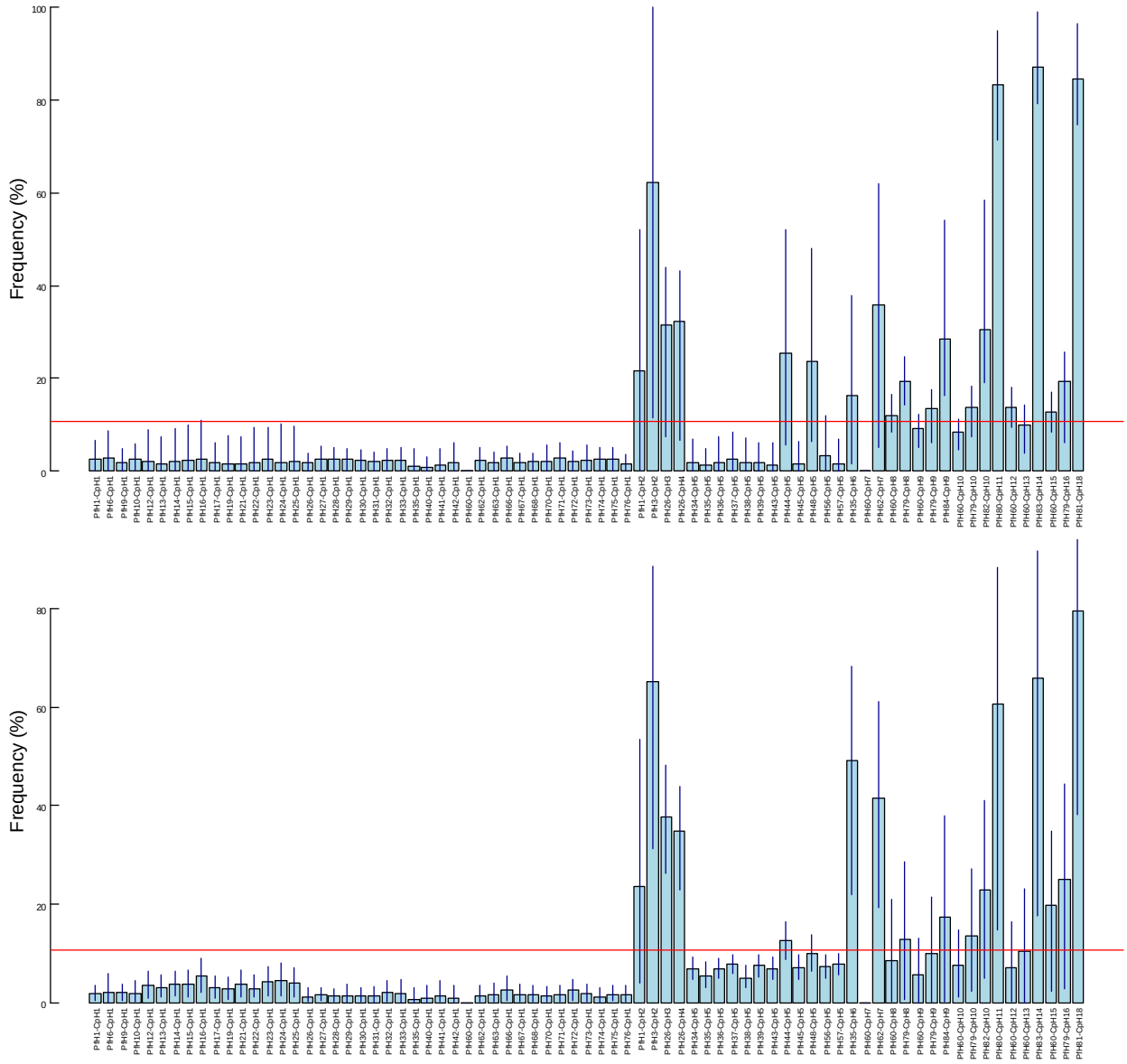

Figure 3. Average frequency distributions and their 95% confidence intervals computed with 1,000 randomly chosen pairs posterior probability trees used to build the consensus trees of trematode and amphipod haplotypes applying Random TaPas with geodesic distances (above) and PACo (below). Horizontal lines represent the expected mean if all associations were contributing equally to cophylogenetic signal (i.e., a distribution with mean $n/N_{hs} \times 100$, where N_{hs} is the number of host-symbiont associations).

Comparison with perfect cospeciation, can also be done by box plotting the variance to mean ratio (VMR) of the frequency distributions of the consensus and the set of posterior probability trees (Fig. 4):

```
V2M <- function(x) var(x)/mean(x)
vmrGD <- V2M(LFGD05[,4])
vmrPA <- V2M(LFPACO05[,4])
vmrMGD <- apply(GD05, 1, V2M)
vmrMPA <- apply(PACO05, 1, V2M)
boxplot(vmrMGD, vmrMPA, names = c("GD", "PACo"), ylab="Variance to mean ratio",
        col="lightblue", las=3)
text(1,vmrGD,"*",cex=2, col="darkblue")
text(2,vmrPA,"*",cex=2, col="darkblue")
```

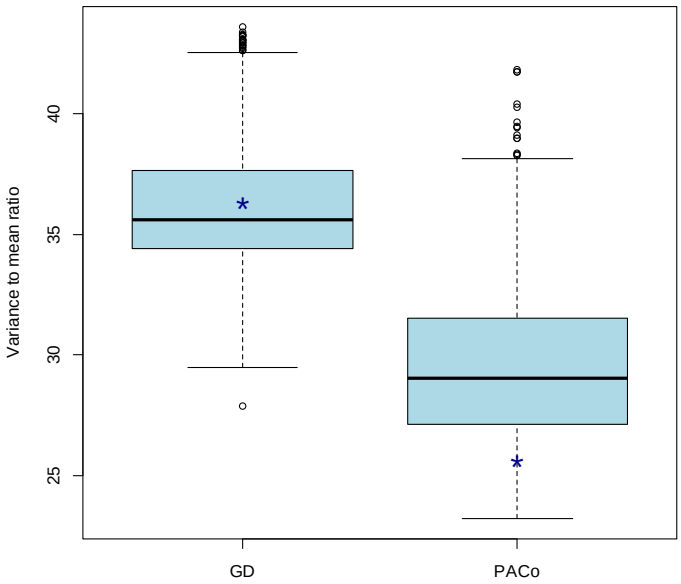

Figure 4. Variance to mean ratio (VMR) of the frequency distribution frequency distributions applying Random TaPas with geodesic distances and PACo to the consensus trees (asterisk) and to 1,000 randomly chosen pairs posterior probability trees (boxplots).

The expected VMR in a perfect codiversification system is 1, which is much lower than the present values (Fig. 4).

```
dev.off() #resets graphics device for next tanglegram plot
```

Finally, the estimated host-symbiont frequencies are mapped as heatmap on the tanglegram. We also plot the average frequency of occurrence of each terminal and the fast maximum likelihood estimators of ancestral states of each node. This is demonstrated here with geodesic distances only.

First, since some branches are very short, both trees are rendered ultrametric to facilitate visualization of data at the nodes, and a color scale ranging from dark read (lowest frequency) to dark blue (highest frequency) is set:

```
trh <- compute.brtime(TreeH, TreeH$Nnode:1)
trs <- compute.brtime(TreeS, TreeS$Nnode:1)

col.scale <- c("darkred", "lightblue","darkblue")
```

Plot the tanglegram (Fig. 5):

```
tangle.gram(trh, trs, HS, colgrad=col.scale, rcolgrad= TRUE,
            nbreaks=50, LFGD05, link.lwd=1, link.lty=1, fsize=0.5,
            pts=FALSE, link.type="curved", node.tag=TRUE,
            cexpt=1.2)
```

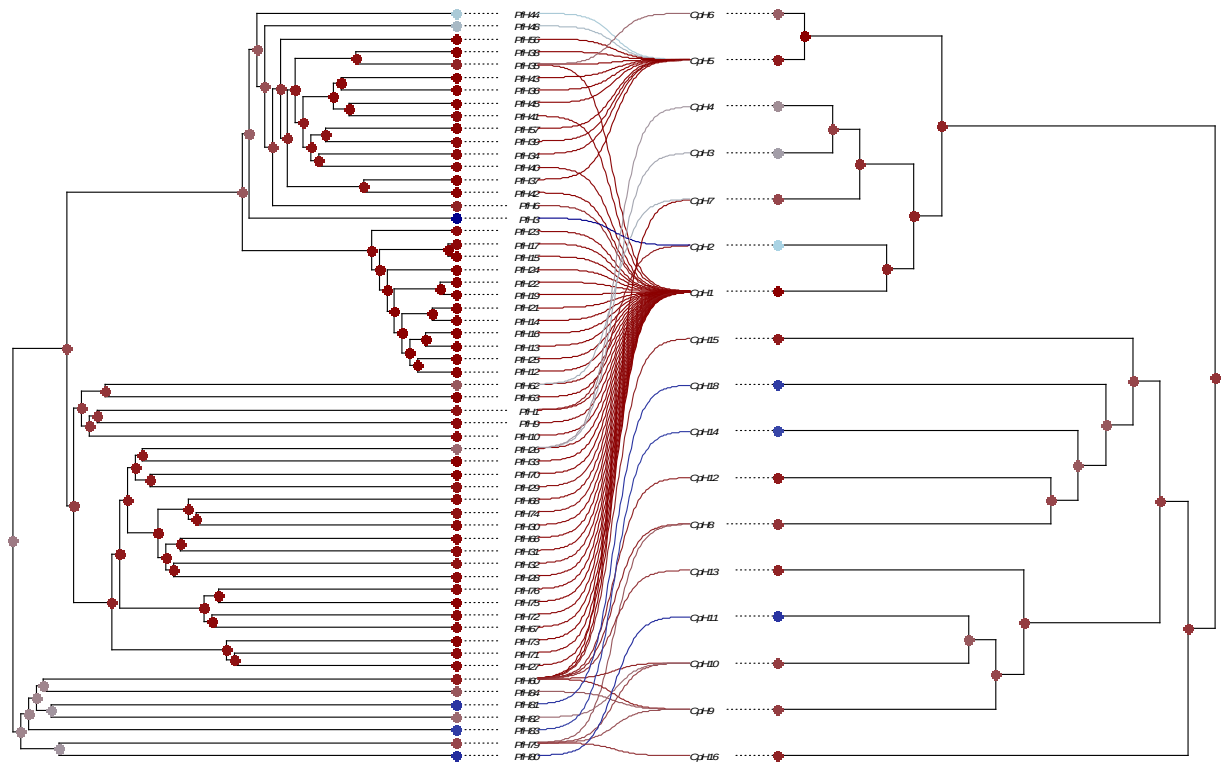

Figure 5. Tanglegram representing the association between haplotypes of the trematode Coitocaecum parvum with those of its amphipod host, Paracalliope fluviatilis. The frequency corresponding to each trematode-amphipod association shown in Fig. 7 is mapped using a color scale ranging from dark blue (highest) to dark red (lowest). The average frequency of occurrence of each terminal and fast maximum likelihood estimators of ancestral states of each node are also mapped according to the same scale.

REFERENCES

Balbuena J.A., Míguez-Lozano R., Blasco-Costa I. 2013. PACo: A Novel procrustes application to cophylogenetic analysis. PLoS ONE 8:e61048.

Chakerian J., Holmes S. 2010. Computational tools for evaluating phylogenetic and hierarchical clustering trees. arXiv:1006.1015

de Vienne D.M., Aguileta G., Ollier S. 2011. Euclidean nature of phylogenetic distance matrices. Syst. Biol. 60:826–832.

Hutchinson M.C., Cagua E.F., Balbuena J.A., Stouffer D.B., Poisot T. 2017. paco: implementing Procrustean Approach to Cophylogeny in R. Meth. Ecol. Evol. 8:932-940.

Lagrue C., Joannes A., Poulin R., Blasco-Costa I. 2016. Genetic structure and host – parasite co-divergence: evidence for trait-specific local adaptation. Biol. J. Linn. Soc. 118:344–358.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Revell L.J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Meth. Ecol. Evol. 3:217–223.