# Supplementary Information:

# Terrestrial habitats decouple the relationship between species and subspecies diversification in mammals (van Holstein & Foley, 2020)

**A. Supplementary Figures**

|  |  |
| --- | --- |
| ***Figure 1:*** *Phylogenetic signal in average subspecies richness* | |
| *a) Distribution of values of lambda and K for average subspecies richness calculated on 50 genus-level trees* |  |
| *b) Distribution of values of lambda and k for experimental trait value distributions for 50 randomly generated trees* |  |

|  |  |  |
| --- | --- | --- |
| ***Figure 2:*** *Distribution of p-values for a) interaction terms in phylogenetic regressions, b) models with interaction terms in phylogenetic ANOVAs* | | |
|  | a) | b) |
| Subspecies richness and species richness |  |  |
| Subspecies richness and range size | a) | b) |

|  |
| --- |
| ***Figure 3:*** *Results of randomly subsetting our dataset 10,000 times. A: p-values for Kendall’s tau correlation between species and average subspecies richness in a random subset of 270 genera. B: Histogram of Kendall’s tau values for the same random subsets of 270 genera. Blue dashed line marked “NT”: Kendall’s tau for non-terrestrial mammals. Orange dashed line marked “T”: Kendall’s tau for terrestrial mammals. Grey dashed line: Kendall’s tau for all mammals.* |

|  |
| --- |
| Figure 4: *The effect of taxonomic inflation on the correlation between species and subspecies richness*. *Under “light” inflation (in which only one species per genus undergoes inflation), the correlation between species and subspecies richness increases with the proportion of genera that undergo inflation; this pattern is reversed under the regime of “heavy” inflation (in which all species in a genus undergo inflation). The start condition, of Tau=0.15, is indicated in purple. All correlations were significant (p<0.05) except for the scenario in which 10% of the dataset underwent “heavy” inflation.* |

**B. Supplementary Materials and Methods**

**PGLS models**

Revell (2010) recommends applying suitable branch length transformations, if necessary simultaneously optimized, in regression models. The phylogenetic regressions we report in the manuscript were run with lambda and kappa optimized (i.e. lambda=”ML” and kappa=”ML”), because AIC scores of these models were the second lowest (872.8) of all possible branch length transformation combinations (see Table 1). Models in which all three branch length transformations were optimized performed best (AIC=872.2), but the difference in AIC score between these models and those with only lambda and kappa optimized is less than 2, and a preference for the more complex model with all three transformations optimized is therefore not justified.

***Table 1:*** Branch length transformation in PGLS regressions and average AIC scores of models

|  |  |
| --- | --- |
| **BL transformation** | **Median AIC score across 50 trees** |
| None | 1253.1 |
| Lambda | 874.9 |
| Kappa | 1150.2 |
| Delta | 1196.4 |
| **Lambda, Kappa** | **872.8** |
| Lambda, Delta | 873.9 |
| Kappa, Delta | 1072.5 |
| **Lambda, Kappa, Delta** | **872.2** |

**Behaviour of Blomberg’s K and Pagel’s lambda depending on data distribution**

We simulated 50 random pure-birth trees with 1000 tips each using the *pbtree* function in phytools, then generated a right-skewed dataset by randomly sampling from a lognormal distribution with a mean of 1.9 (since the mean of our dataset is 1.9 subspecies). Under this experimental condition, all values of lambda and K were both low (<0.1), so a simple positive skew in our data does not explain the observed asymmetry.

The distribution of actual observed subspecies richness, however, is heavy tailed, so we explored the potential impact of this by simulating datasets comprising 90, 70, and 50% values sampled from the lognormal distribution, and 10, 30, and 50% values (respectively) sampled from a uniform distribution with a lower bound of 5 and an upper bound of 15. The results of these regimes are shown in Figure 1b. Even when only 10% of the dataset contains random values from 5-15 (i.e., it is weakly heavy-tailed), the distribution of lambda shifts considerably rightward from its position under a pure lognormal (i.e. no heavy tail) regime, while K remains stationary.

As a control, we ran the experiment again with data drawn randomly from a normal distribution (i.e. the distribution of data was not right-skewed and had no heavy tail). This normal distribution had a mean of 1.9 and a standard deviation of 1. Since our data cannot be negative, we removed all values below 0 and replaced them by re-sampling from the normal distribution. The marked asymmetry in the distributions of values for lambda and K disappears under this regime (see Figure 1b). In this way we can explain the asymmetry in our estimates of lambda and K purely as a result of the distribution of our trait without invoking explanations based on evolutionary process, which has been shown to be problematic given the complex relationship between process and phylogenetic signal (Revell, Harmon, & Collar, 2008).

**Testing the effect of taxonomic inflation**

1. Created a random dataset of 5000 observations of species richness (for which the median was 4) and subspecies richness (for which the median was 2), which had a Kendall correlation of 0.14 (similar to “all mammals”)
2. **Light inflation regime (one subspecies in a genus gets inflated to species status):** 
   1. Subsetted 2.5, 5, 7.5, 10% of simulated dataset, then “inflated” them:
      1. Species richness (*n*) increased by 1 *(n+1)*
      2. Subspecies richness decreased accordingly *(n2/n+1)*
3. **Heavy inflation regime (all subspecies in a genus get inflated to species status):** 
   1. Subsetted 2.5, 5, 7.5, 10% of simulated dataset, then “inflated” them:
      1. All species were split, so new species richness = n2
      2. Because all species were completely split, all subspecies are now 1
4. Calculated new Kendall’s tau

**C. Code**

**Kendall correlations and p-values for 10,000 random subsets of 270 genera**

set.seed(5)

result = replicate(n = 10000,

expr = {

r1<-sample\_n(data, 270)

output <- c(cor.test(r1$Species\_per\_genus, r1$Mean\_subsp, method="kendall")$estimate, cor.test(r1$Species\_per\_genus, r1$Mean\_subsp, method="kendall")$p.value)

})

**Behaviour of Blomberg’s K and Pagel’s lambda depending on data distribution**

*# the regime for data comprising 70% samples from a skewed distribution, 30% from a uniform distribution is given. This was repeated with different percentages of the two distributions (see methods description).*

set.seed(3)

seventy\_thirty\_output = replicate(n = 50, expr = {

tree.sim<-pbtree(n=1000)

a<-rlnorm(700,meanlog=0.6, sdlog=0.3)

b<-runif(300, min = 5, max = 15)

data<-c(a,b)

names(data)<-names\_1000$taxa

lambda<-phylosig(tree.sim, data, method="lambda", test=T)

k<-phylosig(tree.sim, data, method="K", test=T)

output<-list(lambda$lambda,lambda$P, k$K, k$P)

})

*# the regime for data comprising 100% samples from a normal distribution:*

normal\_output\_1000 = replicate(n = 50, expr = {

tree.sim<-pbtree(n=1000)

data<-abs(rnorm(1000, mean=1.9, 1))

names(data)<-names\_1000$taxa

lambda<-phylosig(tree.sim, data, method="lambda", test=T)

k<-phylosig(tree.sim, data, method="K", test=T)

output<-list(lambda$lambda,lambda$P, k$K, k$P)

})

**Testing the effect of taxonomic inflation**

*# Generating a random dataset of 5000 observations of species and subspecies richness, with a Kendall correlation of 0.14 between these variables*

set.seed(2)

samples = 5000

r = 0.22

library('MASS')

data = mvrnorm(n=samples, mu=c(4, 2), Sigma=matrix(c(1, r, r, 1), nrow=2), empirical=TRUE)

X = data[, 1] # standard normal (mu=0, sd=1)

Y = data[, 2] # standard normal (mu=0, sd=1)

*# Light inflation regime (applied to 7.5% of genera)*

Results\_7.5\_Light = replicate(n = 100, expr = {

grp <- sample( LETTERS[1:2], 5000, replace=TRUE, prob=c(0.075, 0.925) ) *#make random string of length 5000 with 7.5% “a” (to be inflated) and 92.5% “b”. This was changed to 2.5% “a”, 5% “a”, and 10% “a” to see the effect of this regime when applied to different proportions of the dataset (see Supplementary Methods).*

grp1<-data.frame(grp)

df1 <- data.frame(grp1,randominflation10) *#attach random sample of a and b to data*

d1<-df1 %>% *#take out 7.5% and "inflate"*

filter(grp=='A') %>%

mutate(Y=((X)^2)/(X+1))%>%

mutate(X= X+1)

d2<-df1 %>%

filter(grp=='B')

d3<- bind\_rows(d1, d2)

tau<-cor.test(d3$X, d3$Y, method="kendall")$estimate

p<-cor.test(d3$X, d3$Y, method="kendall")$p.value

output <- data.frame(tau,p)

})

*# Heavy inflation regime (applied to 7.5% of genera)*

Results\_7.5\_Heavy = replicate(n = 100, expr = {

grp <- sample( LETTERS[1:2], 5000, replace=TRUE, prob=c(0.075, 0.925) ) *#make random string of length 5000 with 7.5% “a” (to be inflated) and 92.5% “b”. This was changed to 2.5% “a”, 5% “a”, and 10% “a” to see the effect of this regime when applied to different proportions of the dataset (see Supplementary Methods).*

grp1<-data.frame(grp)

df1 <- data.frame(grp1,randominflation10) *#attach random sample of a and b to data*

d1<-df1 %>% *#take out 7.5% and "inflate"*

filter(grp=='A') %>%

mutate(Y=1)%>%

mutate(X= X\*X)

d2<-df1 %>%

filter(grp=='B')

d3<- bind\_rows(d1, d2)

tau<-cor.test(d3$X, d3$Y, method="kendall")$estimate

p<-cor.test(d3$X, d3$Y, method="kendall")$p.value

output <- data.frame(tau,p)

})