

Testing for scale-dependent phylogenetic signal in intraspecific variation in intertegular distance across bee communities in Montana, USA

2024-10-17

Introduction

Species interactions vary across space and time due to both species turnover and to a phenomenon known as interaction rewiring. Interaction rewiring may be driven by temporal or spatial changes in phenology, abundance, or species traits (both intra- and interspecific). The impacts of interspecific trait matching, phenology, and abundance on ecological networks have been repeatedly demonstrated: changes in flowering time or tongue morphology have clear impacts on plant-pollinator interactions, for example. However, the role of intraspecific trait differences in shaping interactions—and their relative importance—remain poorly known.

Bees are a species-rich ($n > 20,000$) clade of Hymenopterans with a cosmopolitan distribution. Ecologically important pollinators, intraspecific variation in body size is common in some taxa. Because body size is highly correlated with tongue length and flying ability, it is plausible it mediates foraging choices and thus may explain a portion of plant-pollinator interaction rewiring across space.

Methods

- Bees and flowering plants were sampled when observed interacting across 4 years and 142 sites in Montana, USA. Sites were arranged into spatially nested design: 9 sites comprised a block and 8 blocks comprised a locality; there were two localities in the study.
- Bees were associated with their foraging substrate, collected and pinned for identification. A proxy for body size, intertegular distance (ITD)—the distance between tegulae (wing bases) on the—was measured in each female individual. (As 75% of collected individuals were female, male bees were excluded.)

Questions

Our study aims to answer the following questions:

- 1) How much phylogenetic signal is there in the magnitude of intraspecific variation in female bee body size?
- 2) At what spatial scale(s) does phylogenetic signal emerge?
- 3) Do bee species present at a greater number of sites have more intraspecific variation in female bee body size?
- 4) Is intraspecific variation in female bee body size predicted by interaction partners after controlling for phylogeny?
- 5) Does the strength of the relationship between intraspecific variation in female bee body size vary across scales?

Analysis

We'll begin by loading libraries for data manipulation, visualization, and Bayesian phylogenetic mixed models:

```
library(tidyverse)
library(ggplot2)
library(reshape2)
library(brms)
library(ape)
library(cowplot)
library(ggtree)
library(phytools)
```

Let's also write a function to calculate Bao's estimator of variation (CV_4):

```
bao <- function(data){
  N=length(data)
  y_bar=mean(data)
  s2_hat=var(data)
  cv_2=s2_hat/y_bar^2
  cv_1=sqrt(cv_2)
  gamma_1=sum(((data-y_bar)/s2_hat^0.5)^3)/N
  gamma_2=sum(((data-y_bar)/s2_hat^0.5)^4)/N
  bias=cv_2^(3/2)/N*(3*cv_2^0.5-2*gamma_1)
  bias2=cv_1^3/N-cv_1/4/N-cv_1^2*gamma_1/2/N-cv_1*gamma_2/8/N
  cv4=cv_1-bias2
  return(cv4)
}
```

Now we load our data:

```
bee_data <- read_csv("~/Dropbox/bee_phylo_itv/bee_females_for_ethan.csv") # bee trait data
bee_phylo <- read.tree("~/Dropbox/bee_phylo_itv/BEE_prunedtree.nwk") # pruned newick-format phlogeny of
```

Let's look at the trait data first:

```
bee_data
```

```
## # A tibble: 4,367 x 27
##   TRANSECT_COMBO LOCATION block_2 IT_microns IT_mm ORDER      FAMILY      GENUS
##   <chr>          <chr>    <chr>      <dbl>  <dbl> <chr>      <chr>      <chr>
## 1 HEMX1_01      HE       E          1776.   1.78 Hymenoptera Halictidae Agap~
## 2 HEMX1_02      HE       E          1857.   1.86 Hymenoptera Halictidae Agap~
## 3 HEMX1_06      HE       E          1857.   1.86 Hymenoptera Halictidae Agap~
## 4 HENEWHI1_04   HE       A          1959.   1.96 Hymenoptera Halictidae Agap~
## 5 HENEWHI2_02   HE       B          1959.   1.96 Hymenoptera Halictidae Agap~
## 6 HENEWHI2_04   HE       B          1939.   1.94 Hymenoptera Halictidae Agap~
## 7 HEOLDHI1_07   HE       M          1959.   1.96 Hymenoptera Halictidae Agap~
## 8 HEOLDHI2_06   HE       N          1816.   1.82 Hymenoptera Halictidae Agap~
## 9 HEOLDHI2_08   HE       N          1898.   1.90 Hymenoptera Halictidae Agap~
## 10 HEMX1_03      HE       E          2020.   2.02 Hymenoptera Halictidae Agap~
```

```
## # i 4,357 more rows
## # i 19 more variables: GENUS_SPECIES <chr>, SEX <lgl>, NESTING <chr>,
## #   NestingAlt <chr>, SOCIALITY <chr>, SocialityAlt <chr>, FLOWER_CODE2 <chr>,
## #   USDA_CODE <chr>, FLOWER_SPECIES <chr>, COUNTRY <chr>, STATE <chr>,
## #   COUNTY <chr>, LATITUDE <dbl>, LONGITUDE <dbl>, ELEVATION_m <dbl>,
## #   JULIAN_DATE <dbl>, DATE <chr>, TIME <time>, YEAR <dbl>
```

Important columns include TRANSECT_COMBO, LOCATION, block_2, IT_microns, GENUS_SPECIES, and FLOWER_SPECIES. We can summarize these data to get an idea of their size and diversity. First, the number of bee species in the dataset:

```
bee_data$GENUS_SPECIES %>% unique() %>% length()
```

```
## [1] 149
```

... the number of plant species:

```
bee_data$FLOWER_SPECIES %>% unique() %>% length()
```

```
## [1] 95
```

... and the total number of observations:

```
bee_data %>% nrow()
```

```
## [1] 4367
```

We may need to filter down these data to only include female bees. Let's check to see what levels are present in the "SEX" column:

```
bee_data$SEX %>% unique()
```

```
## [1] FALSE
```

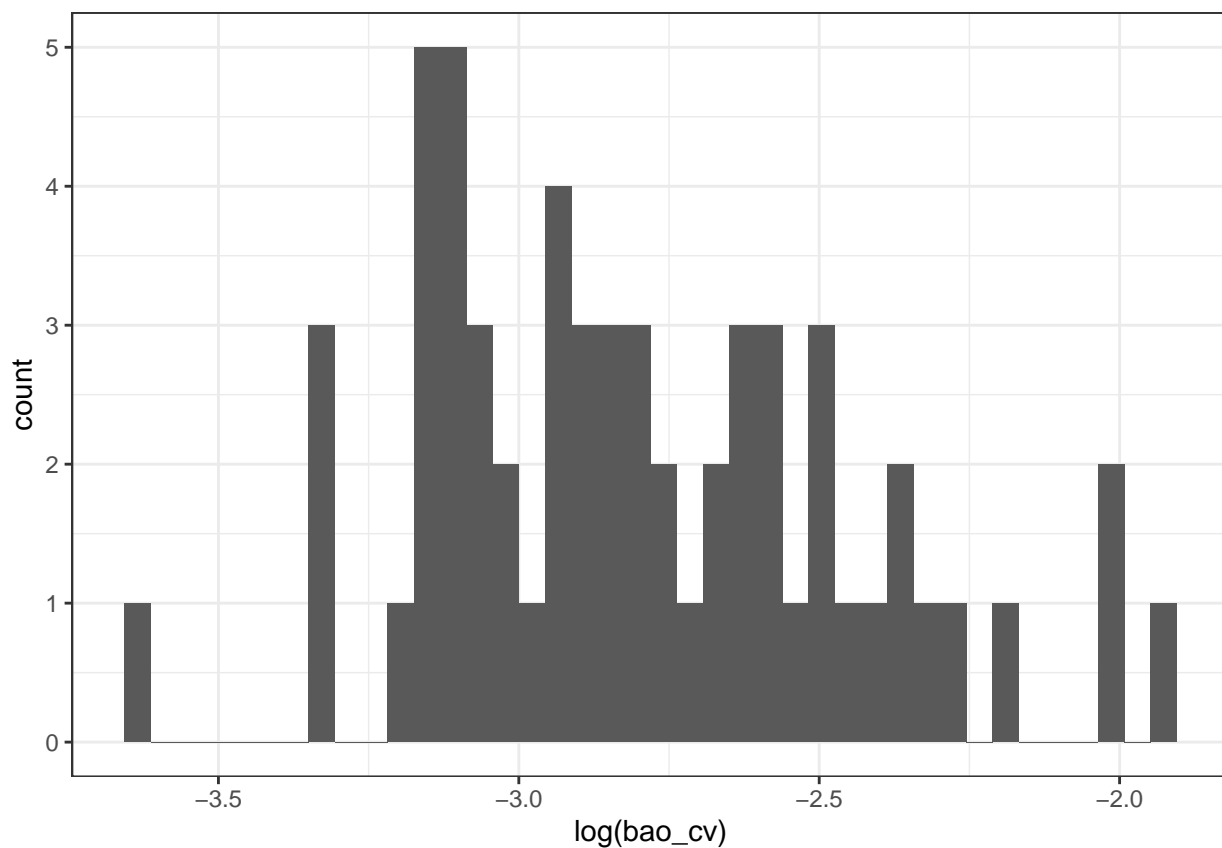
None, apparently—we must be working with a pre-filtered CSV. In that case, let's go ahead and calculate Bao's CV for all species with a sample size of $n \geq 20$:

```
set.seed(1) # set seed for reproducibility
bao_values <- bee_data %>%
  group_by(GENUS_SPECIES) %>%
  mutate(total_observations = n()) %>%
  filter(n() >= 20) %>% # Filter groups with 20 or more observations
  sample_n(20) %>% # Randomly sample 20 rows from the filtered data
  mutate(bao_cv = bao(IT_microns),
         num_sites = n_distinct(TRANSECT_COMBO)) %>%
  select(GENUS_SPECIES, bao_cv, num_sites, total_observations) %>%
  distinct()
bao_values
```

```
## # A tibble: 59 x 4
## # Groups:   GENUS_SPECIES [59]
##   GENUS_SPECIES      bao_cv num_sites total_observations
##   <chr>          <dbl>    <int>         <int>
## 1 Agapostemon_virescens 0.0364      16           26
## 2 Andrena_lawrencei    0.0446      10           82
## 3 Andrena_miranda      0.0444      16           30
## 4 Andrena_topazana     0.0552      14           20
## 5 Anthidium_utahense   0.0680      15           38
## 6 Anthophora_terminalis 0.0260      14           25
## 7 Apis_mellifera       0.0440      15           62
## 8 Ashmeadiella_buconis  0.0724      16           50
## 9 Ashmeadiella_cactorum 0.0474      15           36
## 10 Ashmeadiella_californica 0.0797      12           20
## # i 49 more rows
```

How are these data distributed?

```
p1 <- ggplot(bao_values, aes(x=log(bao_cv))) +
  theme_bw() +
  geom_histogram(bins=40)
p1
```



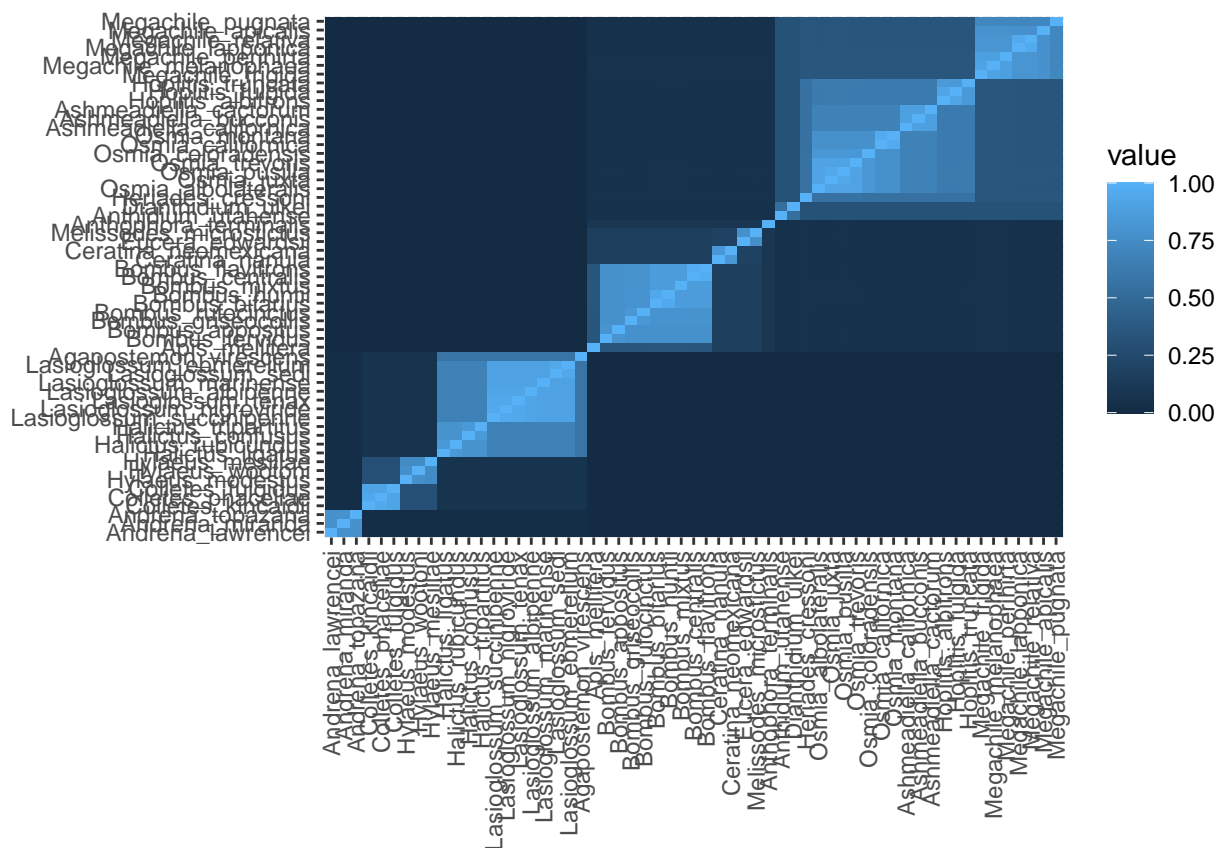
Next, let's prepare our phylogenetic data. We need to create a consensus phylogeny that retains branch lengths (by averaging), drop tips that aren't present in the 59 species in our summary dataset, and calculate and visualize a correlation matrix given a Brownian motion model of trait evolution (NOT a covariance

matrix; see dicussion here: <https://discourse.mc-stan.org/t/covariance-matrix-phylogenetic-models/20477/4>).

```
bee_phylo_consensus <- consensus.edges(bee_phylo)
bee_phylo_consensus <- keep.tip(bee_phylo_consensus, unique(bao_values$GENUS_SPECIES))
A <- ape::vcv.phylo(bee_phylo_consensus, corr = TRUE)
melted_A <- melt(A)
head(melted_A)
```

##		Var1	Var2	value
## 1	Andrena_lawrencei	Andrena_lawrencei	1.00000000	
## 2	Andrena_miranda	Andrena_lawrencei	0.77446907	
## 3	Andrena_topazana	Andrena_lawrencei	0.77443624	
## 4	Colletes_kincaidii	Andrena_lawrencei	0.02584236	
## 5	Colletes_phaceliae	Andrena_lawrencei	0.02584186	
## 6	Colletes_fulgidus	Andrena_lawrencei	0.02584443	

```
cov <- ggplot(data = melted_A, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1), axis.title = element_blank())
cov
```

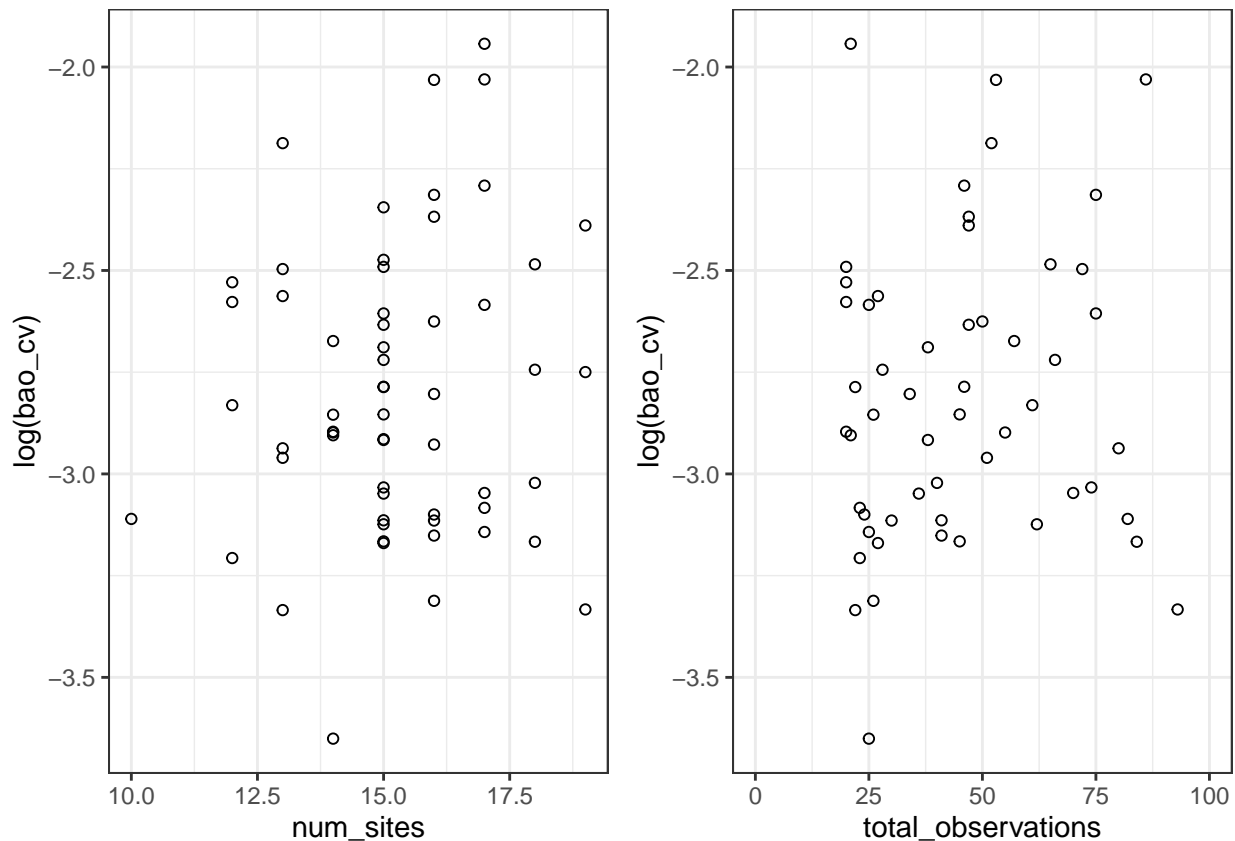


Here, areas of light blue show high correlation values (e.g., are pairwise comparisons of members of the same clade).

1) How much phylogenetic signal is there in the magnitude of intraspecific variation in female bee body size?

To test for phylogenetic signal in our ITV, we'll build a simple Bayesian linear model to predict Bao's CV from the number of sites where a given species is found. In out subsampled data, there *may* be a weak association between these variables, as well as the total number of observations of a species (before subsetting) and its Bao CV value (calculated after subsetting):

```
p2 <- ggplot(bao_values, aes(x=num_sites, y=log(bao_cv), )) +  
  theme_bw() +  
  geom_point(pch=21)  
p3 <- ggplot(bao_values, aes(x=total_observations, y=log(bao_cv), )) +  
  theme_bw() +  
  geom_point(pch=21) +  
  xlim(0,100) # note trimmed axis to ignore outliers  
plot_grid(p2, p3)
```



In a Bayesian statistics, a linear model will tell us the posterior probability of its parameters (i.e., the effect size of a predictor) given the data. Ignoring the marginal likelihood (or “probability of data”), this is what that looks like:

$$P(\text{model parameters}|\text{data}) \propto P(\text{data}|\text{model parameters}) * P(\text{model parameters})$$

It's useful to formally define our model. We start with the likelihood, or the probability of a particular parameter value given the observed data (i.e., the first half of the right-hand side of the equation). It

is a little tricky to think about the most appropriate probability distribution for Bao's CV. The data-generating process—local adaptation to different niches—might lead to gamma-distributed body sizes (in this case, right-skewed data). Gamma distributed data seem tricky to deal with in linear models. By log-transforming Bao's CV for intertegular distance we should be able to use a Gaussian distribution, which will make calculating phylogenetic signal easier. The likelihood thus tells us Bao's CV is stochastically related to a normal distribution with parameters μ_i (the mean) and σ (variance). In this case, μ_i is itself *deterministically* related to the parameter β ("number of sites"), with an overall intercept (α) plus a set of phylogenetically correlated intercepts (α_j):

$$\begin{aligned}\log(\text{Bao's CV}) &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \alpha_j + \beta_{\text{abundance},i}\end{aligned}$$

We then need to define our priors, or the distribution of sensible values for model parameters. We know that α_j should be drawn from a normal distribution with mean α and correlation structure σ_A :

$$\alpha_j \sim \text{Normal}(\alpha, \sigma_A)$$

At this point it's probably easiest to get the rest of the priors we need from the `get_prior` function from `brms`:

```
priors <- get_prior(bao_cv ~ 1 + num_sites + (1|gr(GENUS_SPECIES, cov = A)),
                   data = bao_values, data2 = list(A = A), family=gaussian())
prior

## function (prior, ...)
## {
##   call <- as.list(match.call())[-1])
##   seval <- rmNULL(call[prior_seval_args()])
##   call[prior_seval_args()] <- NULL
##   call <- lapply(call, deparse_no_string)
##   do_call(set_prior, c(call, seval))
## }
## <bytecode: 0x1438e0940>
## <environment: namespace:brms>
```

This gives us a flat prior for the effect of `num_sites` on the outcome variable, e.g. a uniform distribution over all reals:

$$\beta_{\text{abundance},i} \sim \text{Uniform}(-\infty, \infty)$$

It also suggests that we use a standard Student's T distribution for σ , the standard deviation of the residuals:

$$\sigma \sim \text{Student}(3, 0, 2.5)$$

Now, we'll run our model, using an MCMC chain with the default number of steps:

```
model_1 <- brm(
  log(bao_cv) ~ num_sites + (1|gr(GENUS_SPECIES, cov = A)),
  data = bao_values,
  family = gaussian(),
  data2 = list(A = A),
  prior = priors
)
```

(Output curtailed to keep this document a manageable size.)

Let's summarize the results:

```
summary(model_1)
```

```
## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: log(bao_cv) ~ num_sites + (1 | gr(GENUS_SPECIES, cov = A))
## Data: bao_values (Number of observations: 59)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##         total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~GENUS_SPECIES (Number of levels: 59)
##           Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)    0.35      0.09    0.17    0.53 1.00      517      994
##
## Regression Coefficients:
##           Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -3.24      0.35   -3.93   -2.56 1.00     2694     2769
## num_sites      0.02      0.02   -0.02    0.07 1.00     3020     3170
##
## Further Distributional Parameters:
##           Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma        0.24      0.04    0.16    0.34 1.00      509      576
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

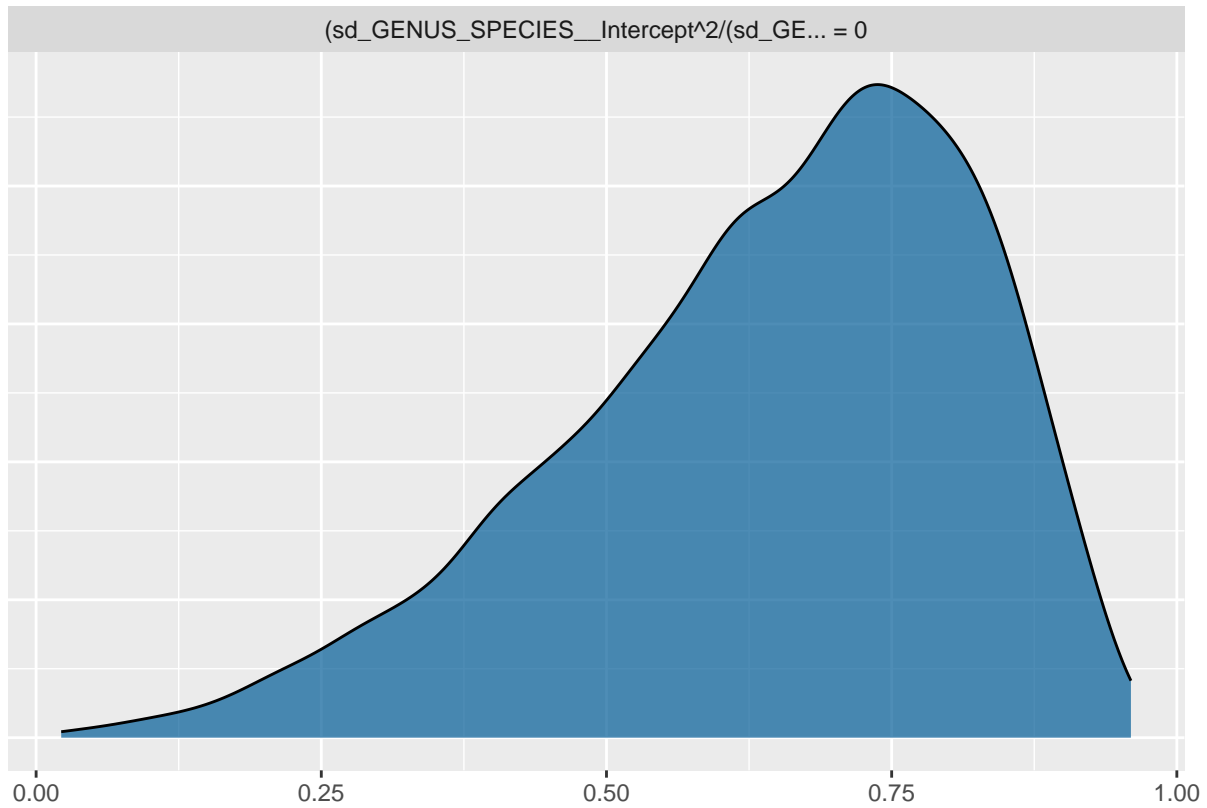
This suggests there is a positive effect of abundance on Bao's CV of intertegular distance—but that it is not credible at the 95% level. It also suggests that phylogenetically correlated intercepts explain a lot of the variation in Bao's CV we see here. We can formally calculate Pagel's λ —the proportion of variance in the model attributable to phylogeny—using its residuals. We can also perform a one-sided hypothesis test to evaluate the evidence that $\lambda > 0$. To do so, we calculate the ratio of the posterior probability that $\lambda > 0$ compared with the posterior probability that $\lambda = 0$:

```
hyp <- "sd_GENUS_SPECIES__Intercept^2 / (sd_GENUS_SPECIES__Intercept^2 + sigma^2) = 0"
(hyp <- hypothesis(model_1, hyp, class = NULL))
```

```
## Hypothesis Tests for class :
##           Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio
## 1 (sd_GENUS_SPECIES... = 0    0.64      0.18    0.24    0.91      NA
##   Post.Prob Star
## 1      NA      *
## ---
## 'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
## '*' : For one-sided hypotheses, the posterior probability exceeds 95%;
##       for two-sided hypotheses, the value tested against lies outside the 95%-CI.
## Posterior probabilities of point hypotheses assume equal prior probabilities.
```



```
plot(hyp)
```



This appears to be greater than Laura's estimates from **phytools**. The obvious next steps would be to see whether signal declines using small subsets of the data (e.g., ask whether λ only emerges at broader spatial scales). But we should chat first, I think!