THE ROLE OF POPULATION STRUCTURE IN PATHOGEN DIVERSITY IN WILD BAT POPULATIONS

TIM LUCAS. NOVEMBER 9, 2015

1. Abstract

1.0.1. *One or two sentences providing a basic introduction to the field.* Bats are an important reservoire of zoonotic diseases. It is still unclear what factors determine the number of pathogens a wild bat species carries. But once understood, these factors could provide a way to priorities surveillance of bat populations.

1.0.2. *Two to three sentences of more detailed background.*

1.0.3. *One sentence clearly stating the general problem (the gap).*

1.0.4. *One sentence summarising the main result.*

1.0.5. *Two or three sentences explaining what the main result reveals in direct comparison to what was thought to be the case previously.*

1.0.6. *One or two sentences to put the results into a more general context.*

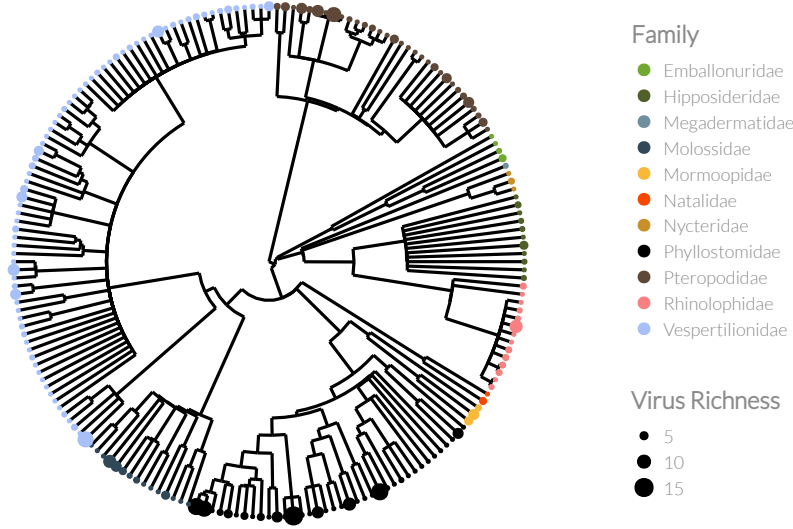1.0.7. *Two or three sentences to provide a broader perspective,*

Figure 1. Pruned phylogeny with dot size showing number of pathogens and colour showing family.

2. Introduction

3. Methods

To measure pathogen richness I used data from (Luis et al. 2013). These simply include known infections of a bat species with a pathogen species. Only species with at least one pathogen were included in the analysis. Rows with host species that were not identified to species level were removed. Many viruses were not identified to species level or their identified species was not in the ICTV virus taxonomy (ICTV 2014). I counted a virus if it was the only virus, for that host species, in the lowest taxonomic level identified in the ICTV taxonomy. That is, if a host carries an unknown Paramyxoviridae virus, then it must carry at least one Paramyxoviridae virus. If a host carries an unknown Paramyxoviridae virus and a known Paramyxoviridae virus, then it is hard to confirm that the unknown virus is not another record of the known virus. In this case, this would be counted as one virus species.

I used two measures of population structure. F_{ST} and the number of subspecies. The number of subspecies was counted using the Wilson and Reeder taxonomy (Wilson and Reeder 2005). F_{ST} and other measures were collated from the literature. Studies are from a wide range of spatial scales, from local (~ 10 km) to continental. As F_{ST} inevitably increases with spatial scale I controlled for this by only using data from studies where a large proportion of the species range was studied. I used the ratio of the furthest distance between F_{ST} samples to the width of the species range and only used studies if this ratio was greater than 0.3. To allow comparison between different measures (F_{ST}, ϕ_{ST}) and data from different molecular regions I converted all data to diploid gene flow. WILL ADD EXTRA METHODS LATER. These two measures of population structure were analysed separately as the number of subspecies has 196 data points while there is only F_{ST} data for ~ 30 bat species.

To control for study bias I collected the number of Pubmed and Google Scholar citations for each bat species including synonyms from ITIS (*Integrated Taxonomic Information System (ITIS)* n.d.) via the taxize package (Chamberlain and Szöcs 2013). The counts were scraped using
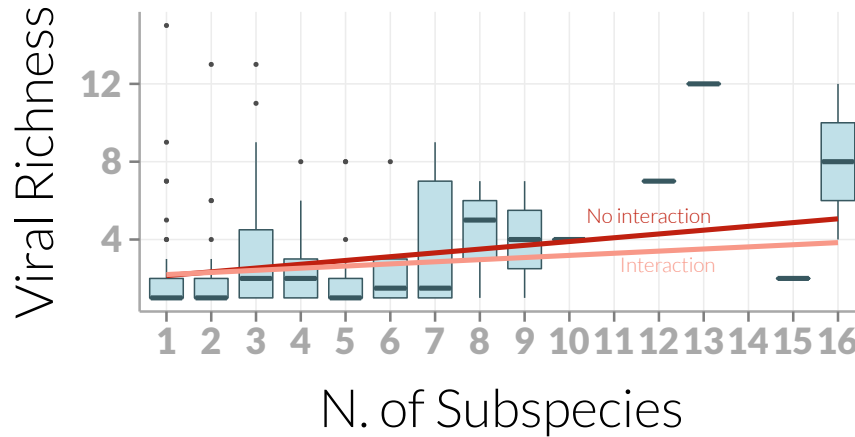
Figure 2. Number of virus species against number of subspecies. Data within a number of subspecies are plotted as boxplots with the dark bar showing the median, the box showing the interquartile range, vertical lines showing the range and outliers shown as seperate points. Regression lines are from multivariate phylogenetic models with other independant variables set at their median value. The models shown are thos with (pink) and without (red) an interaction between study effort and number of subspecies.

the rvest package (Wickham 2015). I log transformed these variables as they were strongly right skewed. The log number of citations on Pubmed and Google scholar were highly correlated (pgls: $t = 19.32$, df $= 194$, $p = 0$). The results here are for analyses using only Google Scholar citations. See the appendix for analyses run using Pubmed citations.

Measures of body mass are taken from Pantheria (Jones, Bielby, et al. 2009) and primary literature (Canals et al. 2005; Arita 1993; López-Baucells et al. 2014; Orr and Zuk 2013; Lim and Engstrom 2001; Aldridge 1987; Ma et al. 2003; Owen et al. 2003; Henderson and Broders 2008; Heaney et al. 2012; Oleksy et al. 2015; Zhang et al. 2009). *Pipistrellus pygmaeus* was assigned the same mass as *P. pipistrellus* as they indistinguishable by mass. Body mass measurements were log transformed due to the strong right skew. Distribution size was estimated by downloading range maps for all species from IUCN () and were also logged due to right skew.

To control for phylogenetic nonindependance I used the best-supported phylogeny from (Fritz et al. 2009) which is the supertree from (Bininda-Emonds et al. 2007) with names updated to match the Wilson & Reeder taxonomy (Wilson and Reeder 2005). Phylogenetic manipulation was performed using the ape package (Paradis et al. 2004). The importance of the phylogeny on each variable separately was estimated using (Revell 2012).
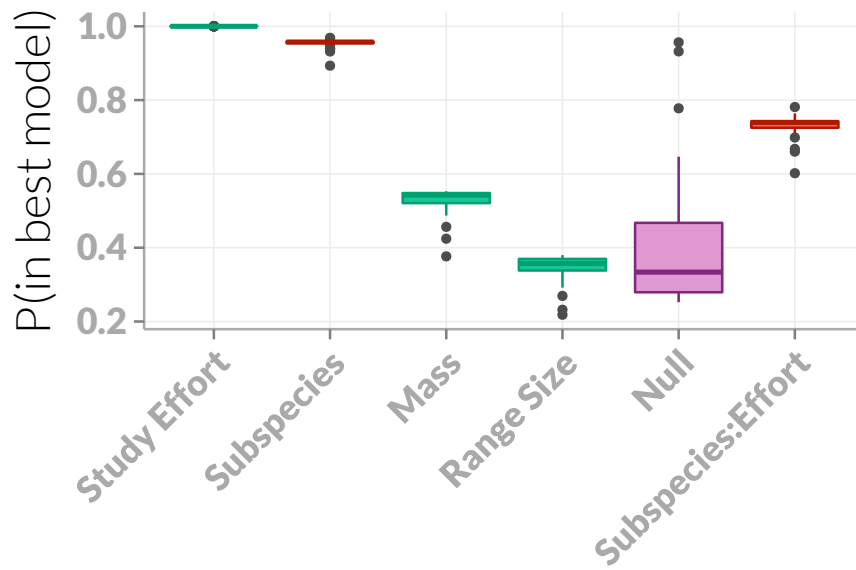
Figure 3. Akaika variable weights for number of subspecies analysis. The probability that each variable will be in the best model if the data were recollected is shown for each of the bootstrap analyses. The purple "Null" box is a uniform random variable used as a null. Population structure (Number of subspecies) and the interaction between subspecies and study effort, shown in red, are more likely to be in the best model than this random variable.
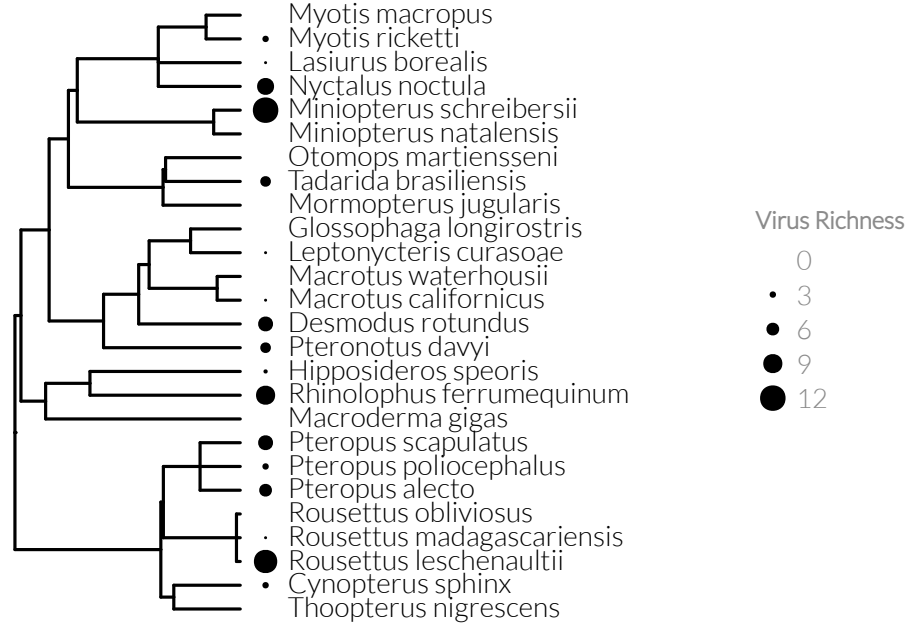
Figure 4. Pruned phylogeny with dot size showing number of pathogens and colour showing family.

F_{ST} studies are conducted at a range of spatial scales, but F_{ST} often increases with distance studied (). To minimise the effects of this I only used data from studies that cover 20% of the diameter of the species range. This is a largely arbitrary value that could be considered to reflect a 'global' estimate of F_{ST} while keeping a reasonable number of datapoints available. I calculated the diameter of the species range by finding the furthest apart points in the IUCN species range () even if the range is split into multiple polygons. The width covered by each study was the distance between the most distant sampling sites. When this was not explicit in the paper, the centre of the lowest level of geographic area was used.

3.1. **Statistical analysis.** Statistical analysis for both dependant variables were conducted using a information theory/model averaging approach (Burnham and Anderson 2002) specifically following (Whittingham, Swetnam, et al. 2005; Whittingham, Stephens, et al. 2006). I chose a credible set of models including all combinations of independent variables. In the analysis using the number of subspecies dependant variable I also included an interaction term between study effort and number of subspecies as I believe *a priori* that this interaction may be present. The interaction was only included in models with both study effort and number of subspecies as an individual term.

I fitted phylogenetic regressions using caper (Pinheiro et al. 2015) to all models. In each case I simultaneously fitted the λ parameter as this avoids mispecifying the model (Revell 2010). κ and δ were constrained to one.
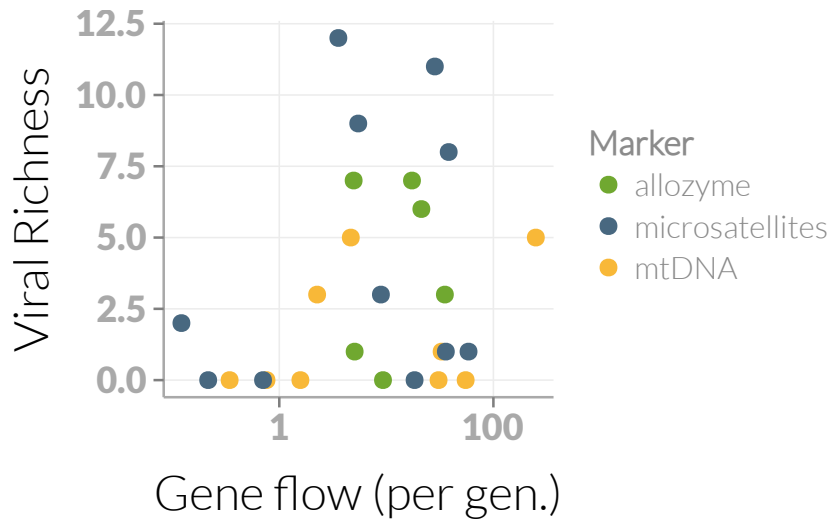
Figure 5. Gene flow per generation (on a log scale) against viral richness with the genetic marker used shown with colour.

As the number of data points is close to or much less than the number of variables times 40 I calculated small sample corredted AIC (AICc) for each model and then calculated Akaiki weights. This value can be interpreted as the probability that a model would be the best model if the data were recollected. For each variable, the sum of the Akaiki weights for models containing that variable are summed. This value can be interpreted as the probability that the given variable is in the best model. Following (Whittingham, Swetnam, et al. 2005) I included a uniformally random variable as a null variable as even unimportant variables can have Akaiki weights significantly greater than zero. The whole analysis was run 50 times, resampling the random variable each time. We calculated $A\bar{I}Cc$ by averaging AICc scores within models. ΔAICc was calculated as $\min(A\bar{I}Cc) - A\bar{I}Cc$, not the mean of the individual ΔAICc scores, to guarantee that the best model has ΔAICc $= 0$.

Plots were created with a combination of (Wickham 2009; Lucas 2015; Solt and Hu 2015; Yu 2015)
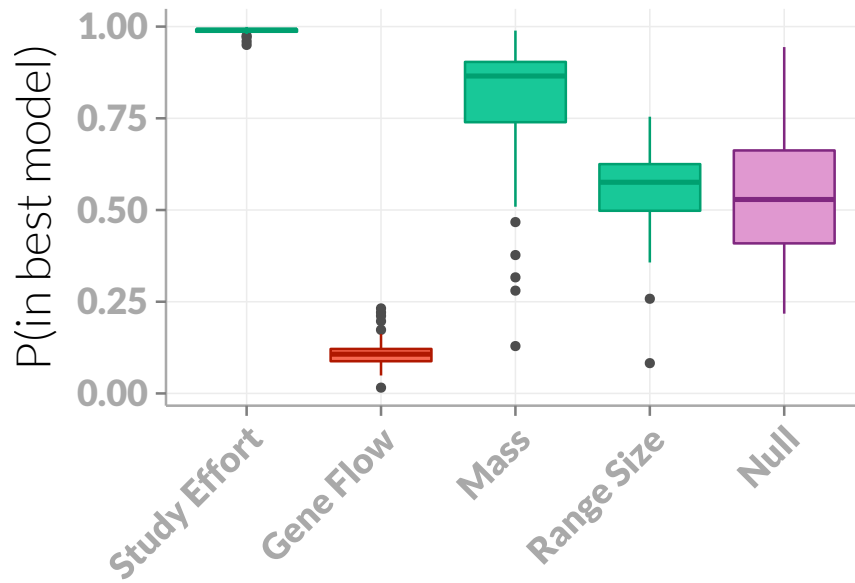
Figure 6. Akaika variable weights for F_{ST} analysis. The probability that each variable will be in the best model if the data were recollected is shown for each of the bootstrap analyses. The purple "Null" box is a uniform random variable used as a null. Population structure (F_{ST}), shown in red, is much less likely to be in the best model than this random variable.

Model	AĪCc	ΔAICc	w_i	$\sum w_i$
log(Scholar)*NSubspecies + rand	896	0.00	0.14	0.14
log(Scholar)*NSubspecies	896	0.15	0.13	0.27
log(Scholar)*NSubspecies + rand + log(Mass)	896	0.15	0.13	0.40
log(Scholar)*NSubspecies + log(Mass)	896	0.35	0.12	0.52
log(Scholar)*NSubspecies + log(Mass) + log(RangeSize)	897	0.95	0.09	0.61
log(Scholar)*NSubspecies + rand + log(RangeSize)	898	2.03	0.05	0.66
log(Scholar)*NSubspecies + log(RangeSize)	898	2.15	0.05	0.71
log(Scholar) + NSubspecies	898	2.35	0.04	0.75
log(Scholar) + NSubspecies + log(Mass)	899	2.51	0.04	0.79

Table 1. Model selection results. AĪCc is the mean AICc score across 50 resamplings of the null random variable. ΔAICc is the AĪCc score minus the lowest score. w_i is the Akaike weight and can be interpreted as the probability that the model is the best model (of those in the plausible set). $\sum w_i$ is the cumulative sum of the Akaike weights.

4. Results

4.1. **Number of Subspecies.**

4.1.1. *More descriptive.* After data cleaning there was data for 196 bat species in 11 families. The number of described virus species for a bat host ranged up to 15 viruses in *Carollia perspicillata*. Figure 1 shows the phylogeny used and the number of viruses for each species. The mean number of viruses across families is fairly constant with a lower range of 1.67 for Nycteridae. The highest mean is Mormoopidae with 5 virus species per bat species, but this is based on a sample size of 3. The Phyllostomidae have the second highest mean (n = 37) of 3.49.

The small change in mean pathogen richness across families and the lack of clear pattern in Figure 1 implies that viral richness is not strongly phylogenetic. This is corroborated by the small estimated size of λ ($\lambda = 0.04$, $p = 0.21$). This fact implies that other factors must control pathogen richness. It also implies that pathogens are not directly inherited down the phylogeny, although this is to be expected by the fast evolution of viruses.

Of the explanatory variables, the number of subspecies has no phylogenetic autocorrelation ($\lambda = 10^{-6}$, $p = 1$), study effort and distribution size have weak but significant autocorrelation (Study Effort: $\lambda = 0.1$, $p = 9.12 \times 10^{-3}$, Distribution size: $\lambda = 0.32$, $p = 8.82 \times 10^{-8}$) and mass is strongly phylogenetic ($\lambda = 0.93$, $p = 0$).

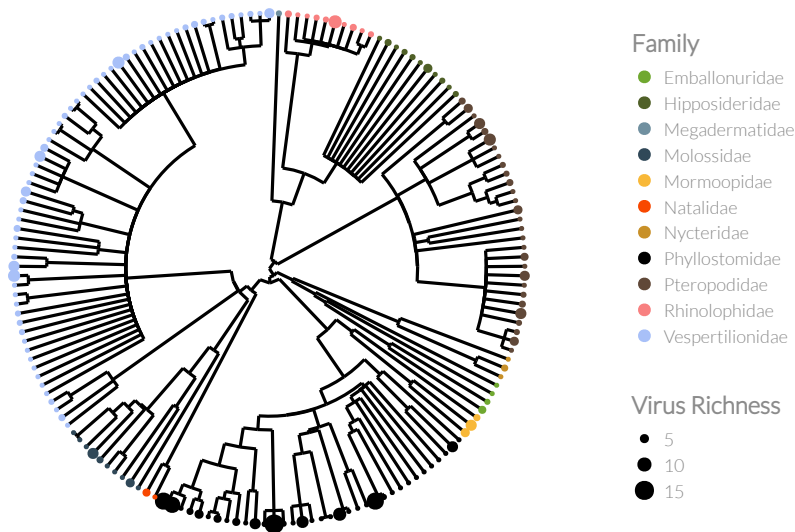4.1.2. *Model results.*

4.2. F_{ST}.

5. Discussion

Figure 7. Pruned phylogeny (Jones, Bininda-Emonds, et al. 2005) with dot size showing number of pathogens and colour showing family.

6. Appendix

References

Luis, A. D. et al. (2013). "A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special?" In: *Proceedings of the Royal Society B: Biological Sciences* 280.1756, p. 20122753.

ICTV (2014). *International Committee on Taxonomy of Viruses Master Species List*. url: `http://talk.ictvonline.org/files/ictv_documents/m/msl/5208/download.aspx`.

Wilson, D. E. and D. M. Reeder (2005). *Mammal species of the world: a taxonomic and geographic reference*. Vol. 12. JHU Press.

Integrated Taxonomic Information System (ITIS). `http://www.itis.gov`.

Chamberlain, S. A. and E. Szöcs (2013). "taxize: taxonomic search and retrieval in R". In: *F1000Research* 2.

Wickham, H. (2015). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.2.0. url: `http://CRAN.R-project.org/package=rvest`.

Jones, K. E., J. Bielby, et al. (2009). "PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184". In: *Ecology* 90.9, pp. 2648–2648.

Canals, M. et al. (2005). "Relative size of hearts and lungs of small bats". In: *Acta Chiropterologica* 7.1, pp. 65–72.

Arita, H. T. (1993). "Rarity in Neotropical bats: correlations with phylogeny, diet, and body mass". In: *Ecological Applications*, pp. 506–517.

López-Baucells, A. et al. (2014). "Echolocation of the big red bat *Lasiurus egregius* (Chiroptera: Vespertilionidae) and first record from the Central Brazilian Amazon". In: *Studies on Neotropical Fauna and Environment* 49.1, pp. 18–25.

Orr, T. J. and M. Zuk (2013). "Does delayed fertilization facilitate sperm competition in bats?" In: *Behavioral Ecology and Sociobiology* 67.12, pp. 1903–1913.

Lim, B. K. and M. D. Engstrom (2001). "Bat community structure at Iwokrama forest, Guyana".
    In: *Journal of Tropical Ecology* 17.05, pp. 647–665.

Aldridge, H. (1987). "Turning flight of bats". In: *Journal of Experimental Biology* 128.1, pp. 419–
    425.

Ma, J. et al. (2003). "Dietary analysis confirms that Rickett's big-footed bat (*Myotis ricketti*) is a
    piscivore". In: *Journal of Zoology* 261.03, pp. 245–248.

Owen, S. F. et al. (2003). "Home-range size and habitat used by the northern myotis (*Myotis
    septentrionalis*)". In: *The American midland naturalist* 150.2, pp. 352–359.

Henderson, L. E. and H. G. Broders (2008). "Movements and resource selection of the Northern
    Long-Eared Myotis (*Myotis septentrionalis*) in a forest–agriculture landscape". In: *Journal of
    Mammalogy* 89.4, pp. 952–963.

Heaney, L. R. et al. (2012). "*Nyctalus plancyi* and *Falsistrellus petersi* (Chiroptera: Vespertilion-
    idae) from northern Luzon, Philippines: ecology, phylogeny, and biogeographic implications".
    In: *Acta Chiropterologica* 14.2, pp. 265–278.

Oleksy, R., P. A. Racey, and G. Jones (2015). "High-resolution GPS tracking reveals habitat
    selection and the potential for long-distance seed dispersal by Madagascan flying foxes *Pteropus
    rufus*". In: *Global Ecology and Conservation* 3, pp. 678–692.

Zhang, L. et al. (2009). "Recent surveys of bats (Mammalia: Chiroptera) from China. I. Rhi-
    nolophidae and Hipposideridae". In: *Acta Chiropterologica* 11.1, pp. 71–88.

Fritz, S. A., O. R. Bininda-Emonds, and A. Purvis (2009). "Geographical variation in predictors
    of mammalian extinction risk: big is bad, but only in the tropics". In: *Ecology letters* 12.6,
    pp. 538–549.

Bininda-Emonds, O. R. et al. (2007). "The delayed rise of present-day mammals". In: *Nature*
    446.7135, pp. 507–512.

Paradis, E., J. Claude, and K. Strimmer (2004). "APE: analyses of phylogenetics and evolution in
    R language". In: *Bioinformatics* 20, pp. 289–290.

Revell, L. J. (2012). "phytools: An R package for phylogenetic comparative biology (and other
    things)." In: *Methods in Ecology and Evolution* 3, pp. 217–223.

Burnham, K. P. and D. R. Anderson (2002). *Model selection and multimodel inference: a practical
    information-theoretic approach.* Springer Science & Business Media.

Whittingham, M. J., R. D. Swetnam, et al. (2005). "Habitat selection by yellowhammers Emberiza
    citrinella on lowland farmland at two spatial scales: implications for conservation management".
    In: *Journal of applied ecology* 42.2, pp. 270–280.

Whittingham, M. J., P. A. Stephens, et al. (2006). "Why do we still use stepwise modelling in
    ecology and behaviour?" In: *Journal of animal ecology* 75.5, pp. 1182–1189.

Pinheiro, J. et al. (2015). *nlme: Linear and Nonlinear Mixed Effects Models.* R package version
    3.1-122. url: `http://CRAN.R-project.org/package=nlme`.

Revell, L. J. (2010). "Phylogenetic signal and linear regression on species data". In: *Methods in
    Ecology and Evolution* 1.4, pp. 319–329.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer New York. isbn: 978-0-
    387-98140-6. url: `http://had.co.nz/ggplot2/book`.

Lucas, T. (2015). *palettetown: Use Pokemon Inspired Colour Palettes.* R package version 0.1.0.
    url: `http://CRAN.R-project.org/package=palettetown`.

Solt, F. and Y. Hu (2015). *dotwhisker: Dot-and-Whisker Plots of Regression Coefficients from
    Tidy Data Frames.* Available at The Comprehensive R Archive Network (CRAN). url: `http:
    //CRAN.R-project.org/package=dotwhisker`.

Yu, G. (2015). "ggtree: an R package for versatile annotation and visualization of phylogenetic
    tree". In:

Jones, K. E., O. R. Bininda-Emonds, and J. L. Gittleman (2005). "Bats, clocks, and rocks: diversification patterns in Chiroptera". In: *Evolution* 59.10, pp. 2243–2255.