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

TIM LUCAS. AUGUST 26, 2015

CONTENTS

1. Abstract 2
1.0.1. One or two sentences providing a basic introduction to the field 2
1.0.2. Two to three sentences of more detailed background 2
1.0.3. One sentence clearly stating the general problem (the gap) 2
1.0.4. One sentence summarising the main result 2
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 2
1.0.6. One or two sentences to put the results into a more general context. 2
1.0.7. Two or three sentences to provide a broader perspective, 2
2. Introduction 3
3. Methods 4
4. Results 20
5. Discussion 21
References 21

1. Abstract

1.0.1. *One or two sentences providing a basic introduction to the field.*

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,*

2. Introduction

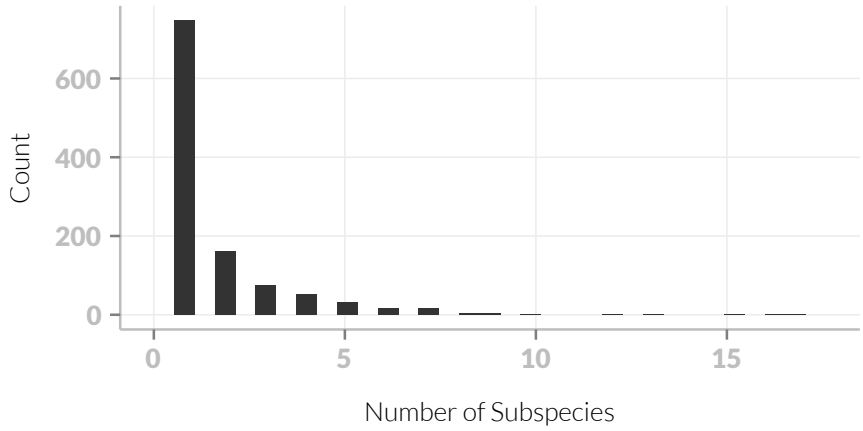

FIGURE 1. Histogram of number of subspecies

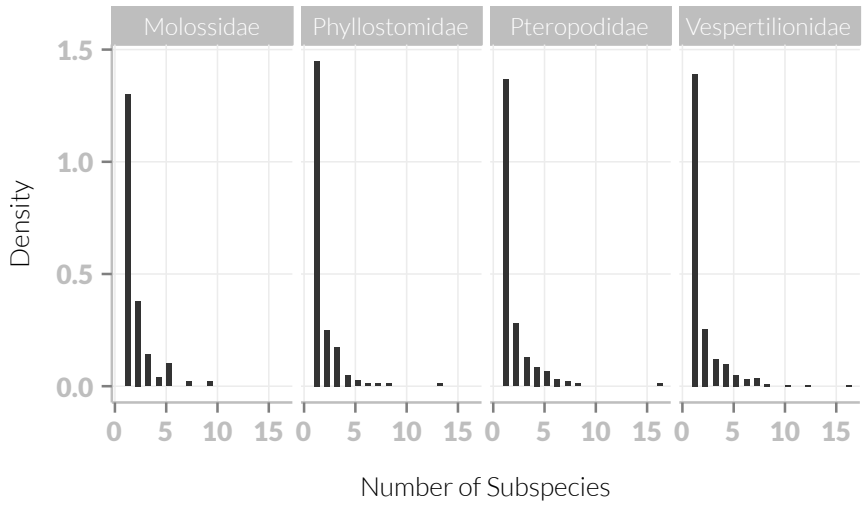

FIGURE 2. Histograms of number of subspecies for the families with many species.

3. METHODS

To measure pathogen richness I used data from [1]. 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. To control for study bias I collected the number of pubmed and scholar citations for each bat species including synonyms from ITIS [2] via the taxize package [3]. The counts were scraped using the rvest package [4].

I used two measures of population structure. F_{ST} and the number of subspecies. The number of of subspecies was counted using the Wilson and Reeder taxonomy [5].

Measures of body mass are taken from Pantheria [6]. They are log transformed due to the strong right skew.

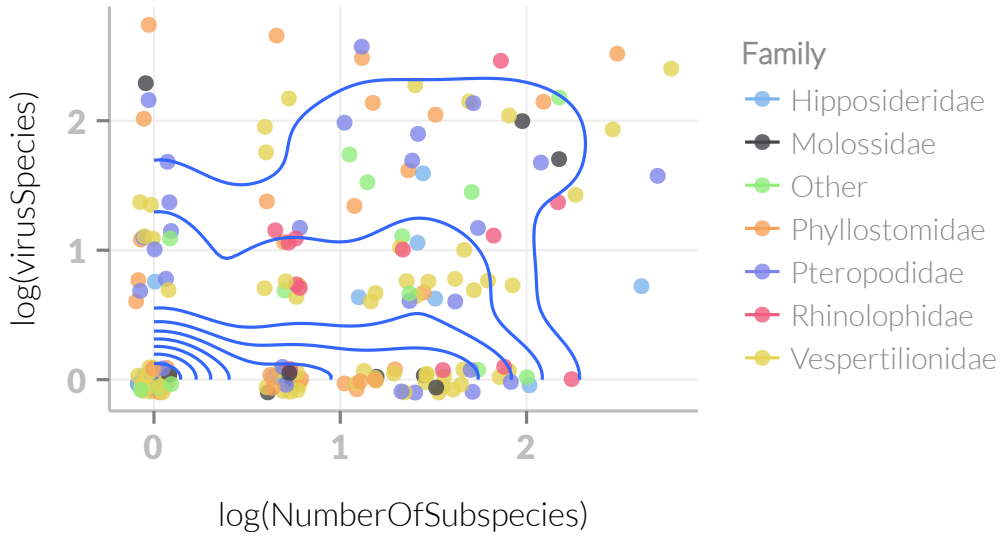

FIGURE 3. Number of viruses against number of subspecies. Points are coloured by family, with families with less than 10 species being grouped into "other". Contours show the 2D density of points and suggest a positive correlation.

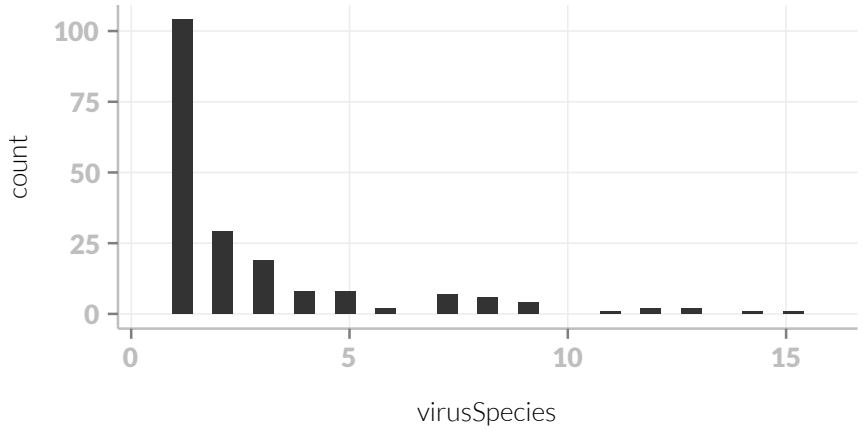

FIGURE 4. Histogram of known viruses per species

To control for phylogenetic nonindependance I used the best-supported phylogeny from [7] which is the supertree from [8] with names updated to match the Wilson & Reeder taxonomy [5]. Phylogenetic manipulation was performed using the ape package [9].

I wanted to run three models using the phylolm package testing the relationship between pathogen richness and log number of subspecies. I tried phylogenetically controlled, multivariate GLMs with poisson errors and identity links. This model was fitted both with and without an interaction term between number of subspecies and study effort. I also fitted a phylogenetically controlled, GLM with poisson errors and identity link to pathogen richness and study effort. The residuals from this

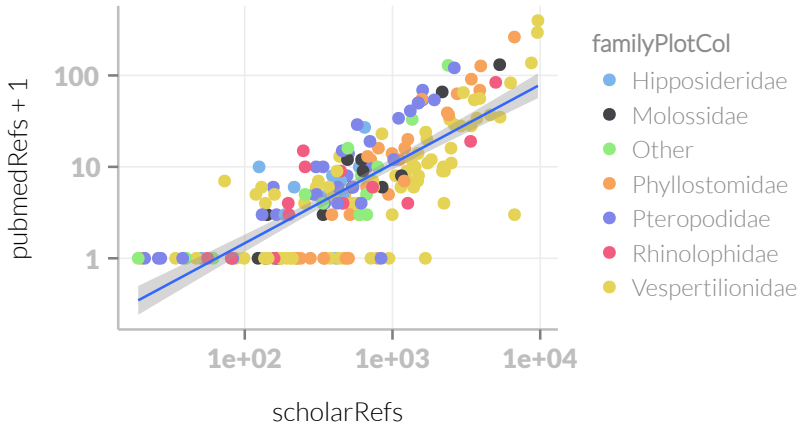

FIGURE 5. Logged number of references on scholar and pubmed, with a fitted (unphylogenetic) linear model. Colours indicate family.

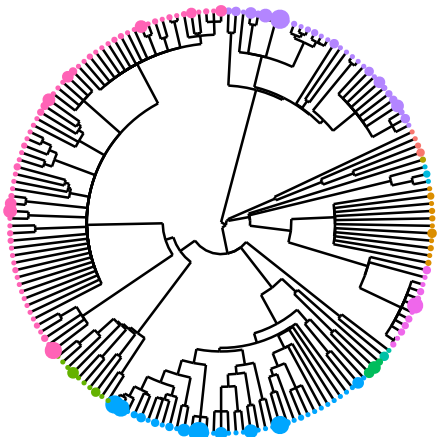

FIGURE 6. Pruned phylogeny with dot size showing number of pathogens and colour showing family.

model was then used as the response variable in a multivariate GLM. However, with these models the numerical optimisation failed to converge.

I ran three models using the caper package [10] testing the relationship between pathogen richness and log number of subspecies. All independant variables were log transformed — study effort was log(citations + 1). I ran phylogenetically controlled, multivariate linear models. This model was fitted both with and without an interaction term between number of subspecies and study effort. We also fitted a phylogenetically controlled, GLM with poisson errors and identity link to pathogen richness and study effort. The residuals from this model was then used as the response variable in a multivariate GLM.

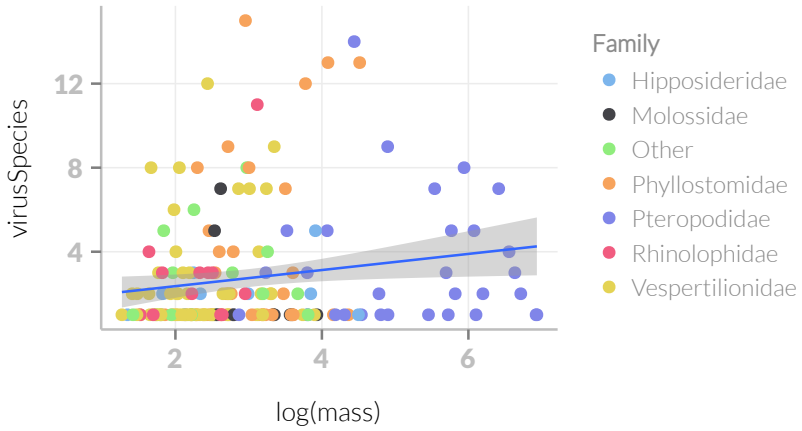

FIGURE 7. Unlogged number of virus species against log mass with a non-phylogenetic linear model added. Points are significantly jittered to try and reveal the sever overplotting in the bottom left corner in particular.

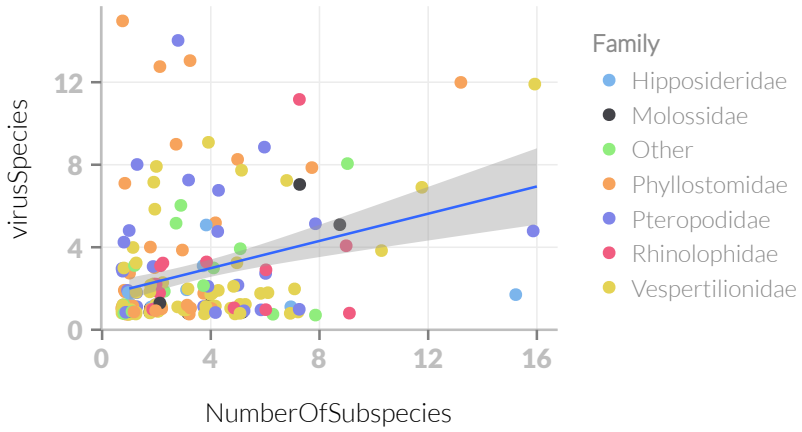

FIGURE 8. Number of virus species against logged number of subspecies (not marginal) with a non-phylogenetic linear model added. Points are significantly jittered to try and reveal the sever overplotting in the bottom left corner in particular.

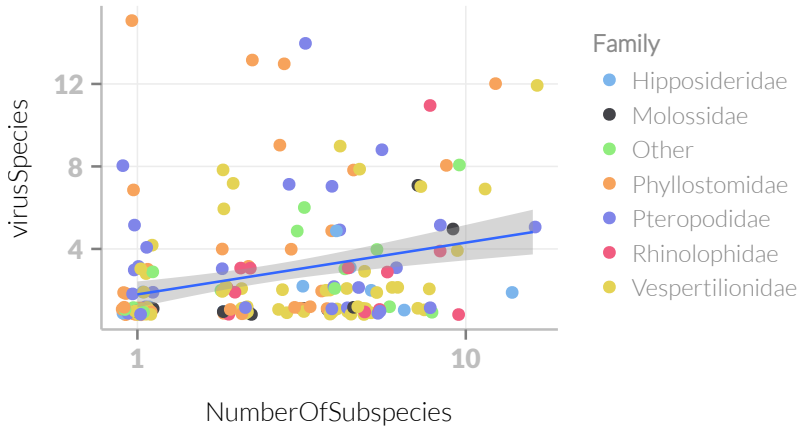

FIGURE 9. Number of virus species against logged number of subspecies (not marginal) with a non-phylogenetic linear model added.

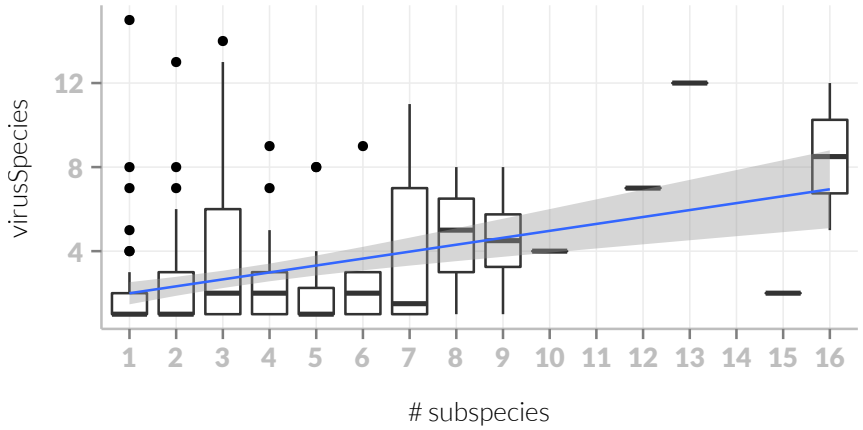

FIGURE 10. Number of virus species against number of subspecies. Data within a number of subspecies are plotted as boxplots with the dark bar showing the median, the box showing the interquartile range, vertical lines showing the range and outliers shown as seperate points. A non-phylogenetic linear model is shown in blue

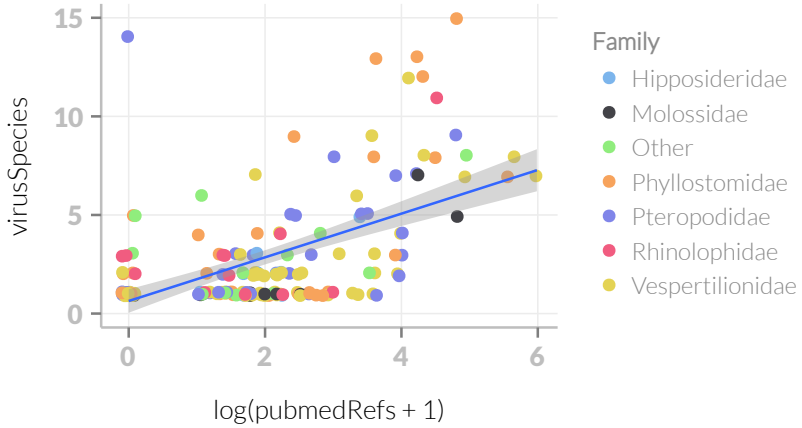

FIGURE 11. Virus species against study effort (log pubmed references +1)

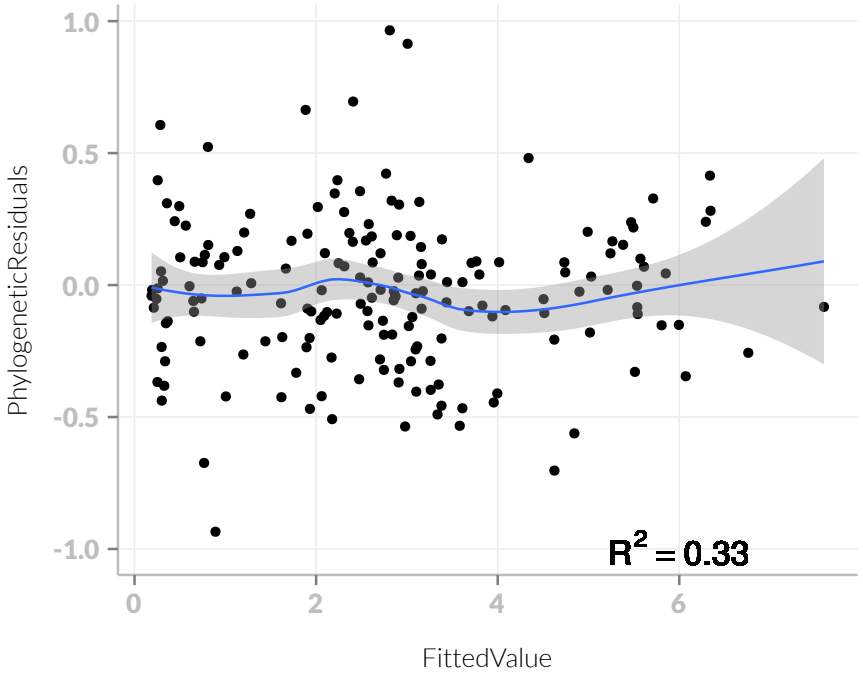

FIGURE 12. Fitted values against residuals from the full phylogenetic model (virusSpecies ~ log(pubmedRefs + 1) + log(NumberOfSubspecies) + log(mass)). A loess curve is shown in blue. The R^2 value give is for the full model (not the loess curve).

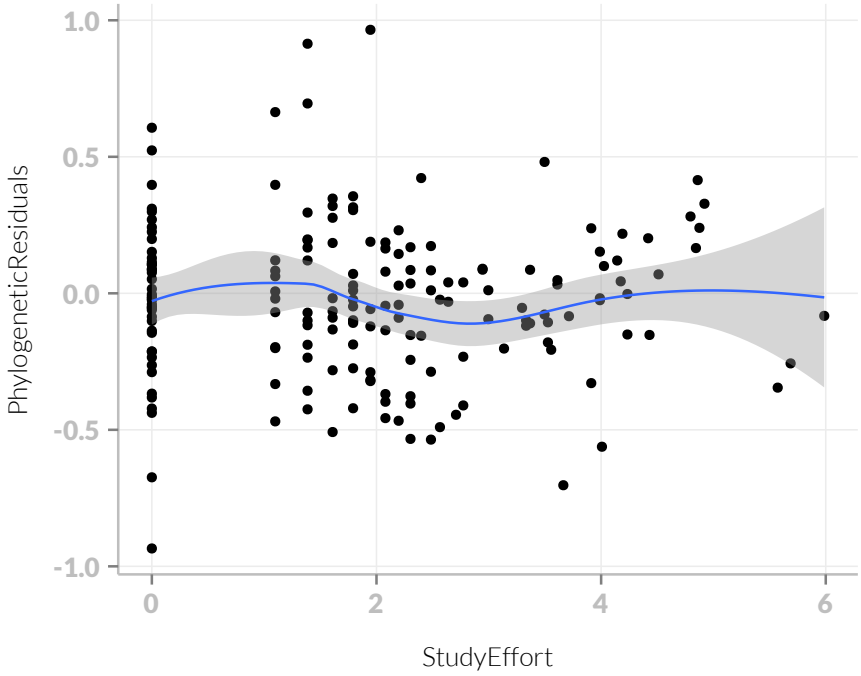

FIGURE 13. Study effort against residuals with a loess trend.

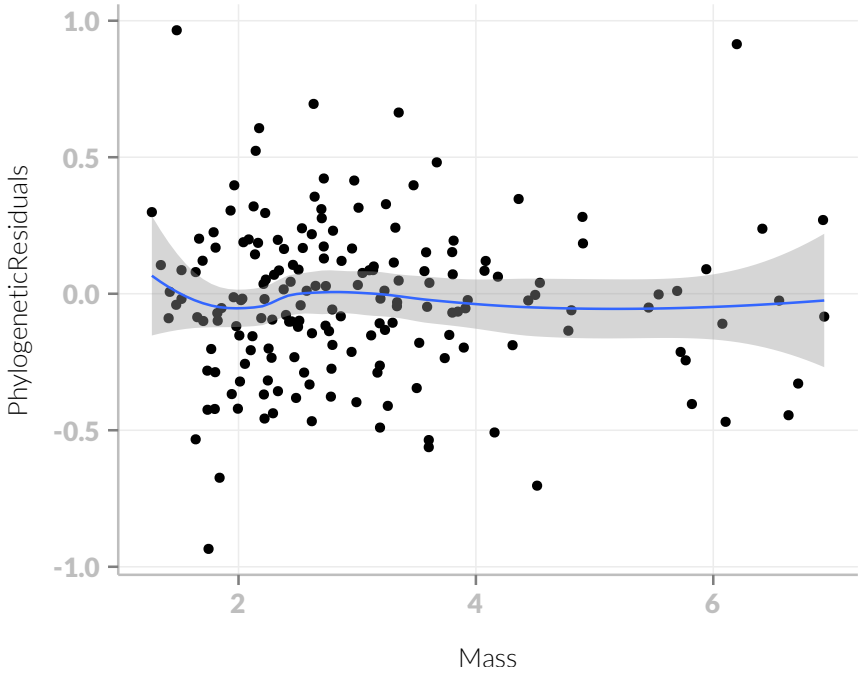

FIGURE 14. Mass against residuals with a loess trend shown in blue.

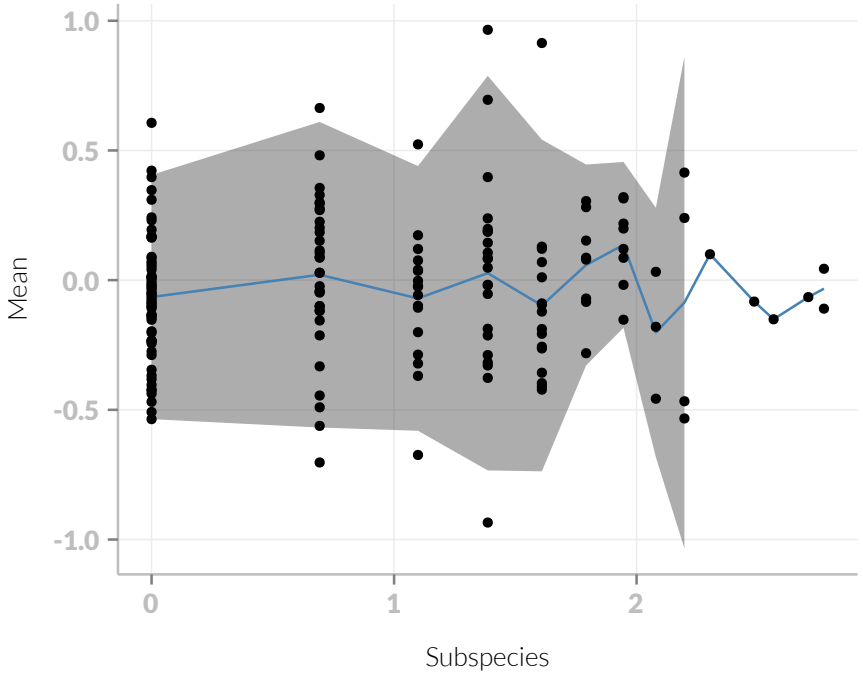

FIGURE 15. Logged number of subspecies against phylogenetic residuals. The mean for each value of logged subspecies is shown in blue. A ribbon showing the mean $\pm 1.96SD$ is shown in grey. The ribbon does not cover the full range of the x axis as there are not enough data points to calculate the SD towards the right.

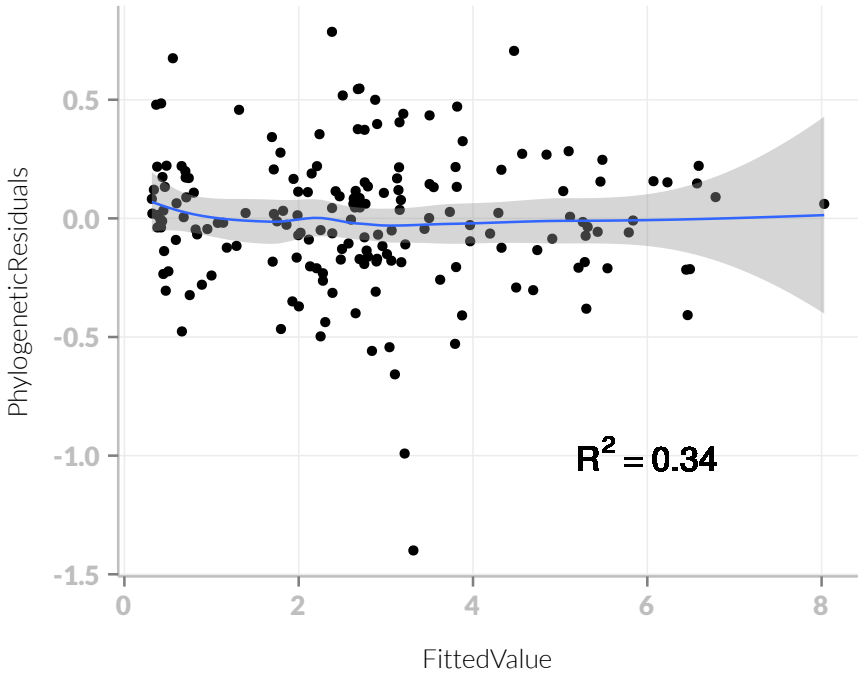

FIGURE 16. Fitted values against residuals from the full phylogenetic model (virusSpecies $\sim$ NumberOfSubspecies + log(pubmedRefs + 1) + log(mass)). A loess curve is shown in blue. The R^2 value give is for the full model (not the loess curve).

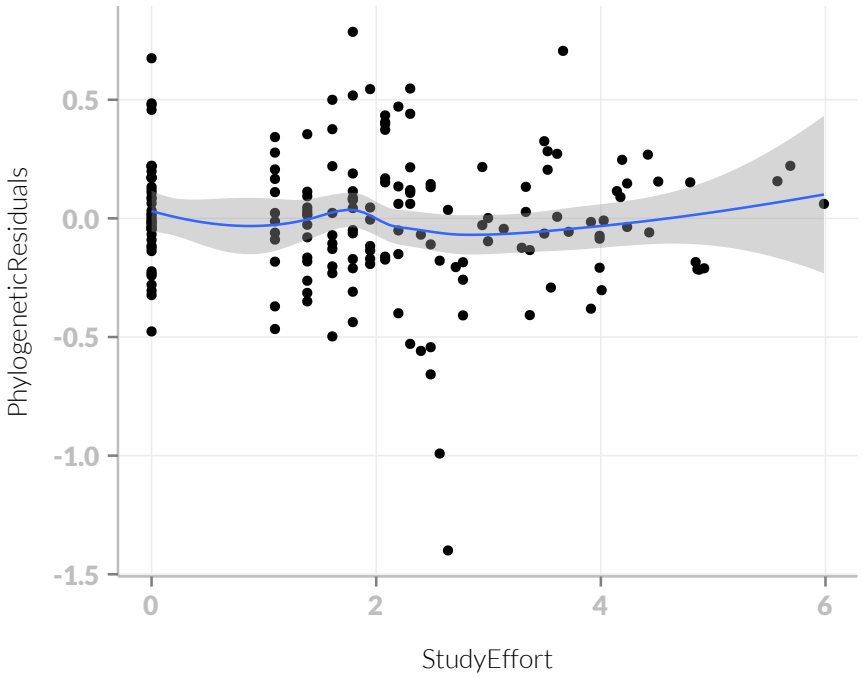

FIGURE 17. Study effort against residuals (unlogged subspecies) with a loess trend.

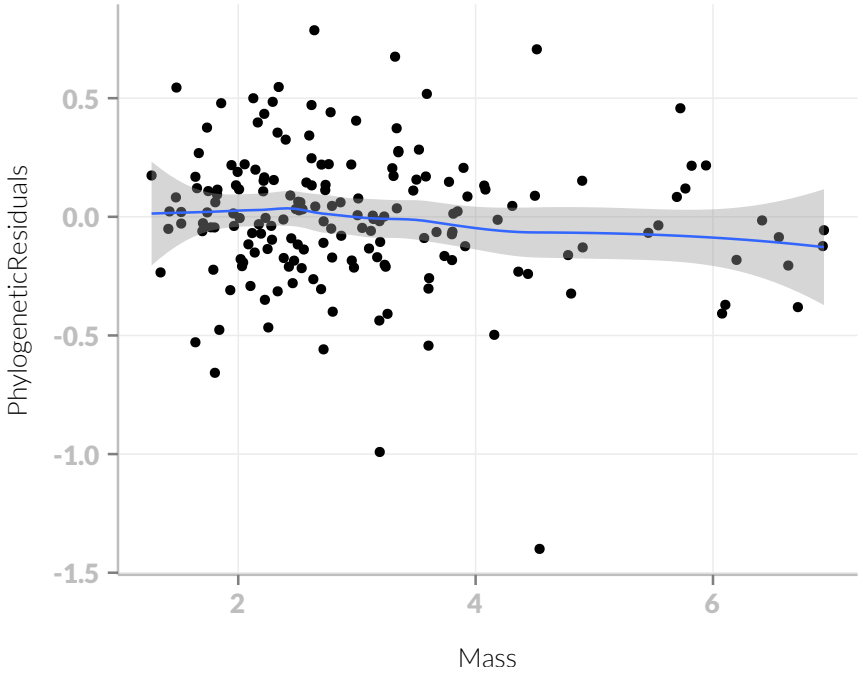

FIGURE 18. Mass against residuals (unlogged subspecies) with a loess trend shown in blue.

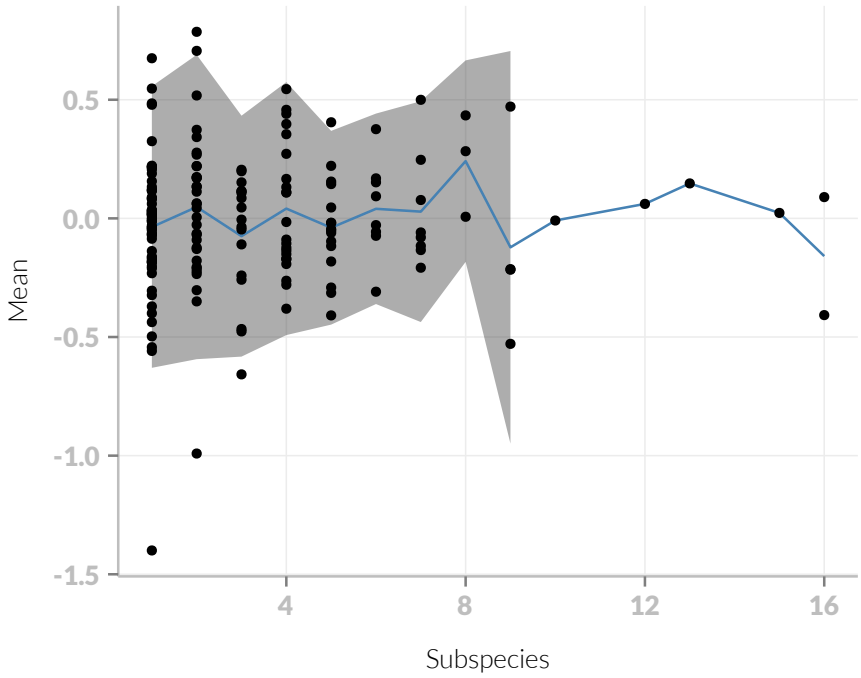

FIGURE 19. Number of subspecies against phylogenetic residuals. The mean for each value of logged subspecies is shown in blue. A ribbon showing the mean $\pm 1.96SD$ is shown in grey. The ribbon does not cover the full range of the x axis as there are not enough data points to calculate the SD towards the right.

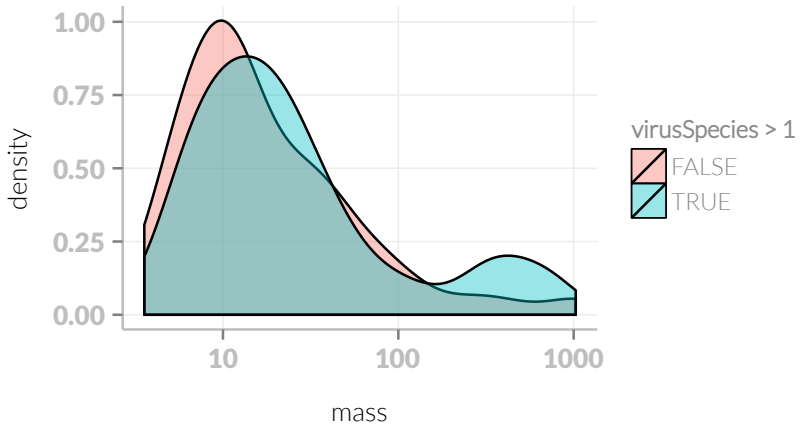

FIGURE 20. Density curves for mass of bat species with 1 or > 1 pathogen species. The hump of the *Pteropodidae* (large fruit bats) can be seen. It seems likely that this family are overstudied as they carry a number of important zoonotics. (Wilcox test: p = 0.086)

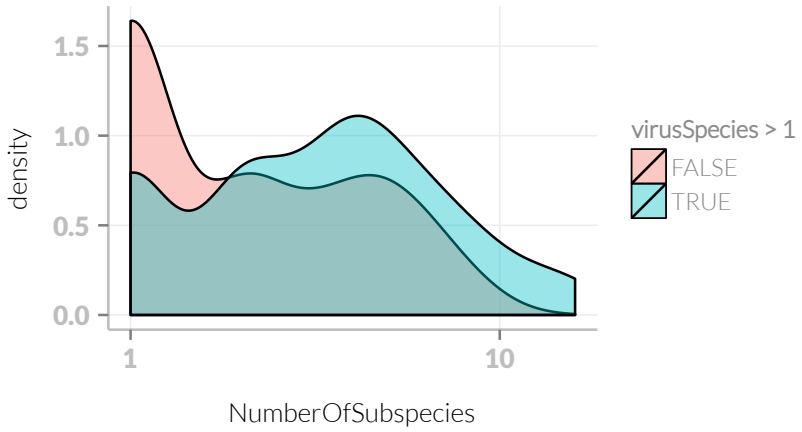

FIGURE 21. Density curves for number of subspecies of bat species with 1 or > 1 pathogen species. (Wilcox test: p = 2.3e-04)

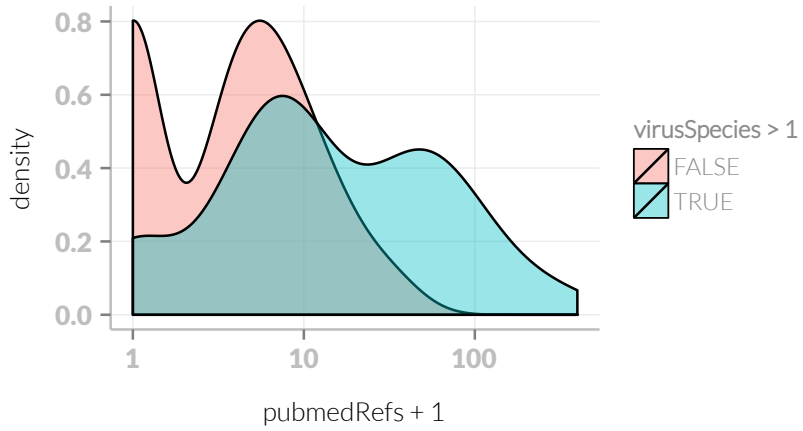

FIGURE 22. Density curves for number of pubmed references of bat species with 1 or > 1 pathogen species. There is a clear trend that many species with only 1 virus species, have 0 pubmed references.

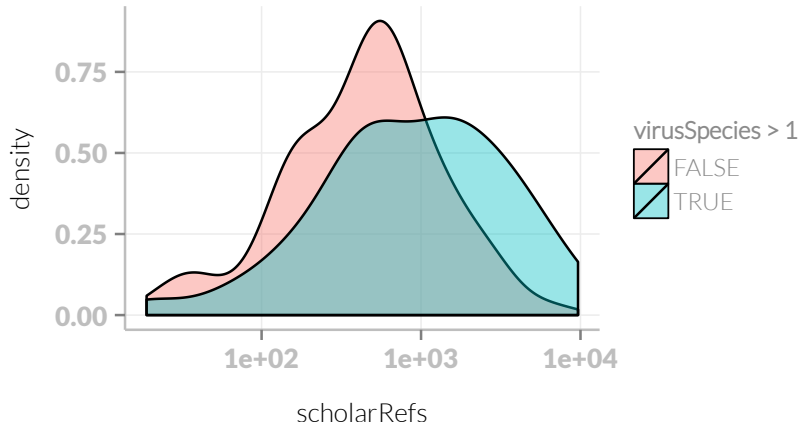

FIGURE 23. Density curves for number of scholar references of bat species with 1 or > 1 pathogen species. The strong trend in the pubmed data is not noticeable here. (t.test: p = 8e-05)

These are the fits from models fitted with the species with only 1 virus species removed. We can see that if number of subspecies is unlogged, it remains marginally significant. If subspecies is logged it is no longer significant.

```
# Models with species with only 1 virus removed

# Unlogged subspecies variable
unlogRm1.summary
##
## Call:
## pgls(formula = virusSpecies ~ log(pubmedRefs + 1) + NumberOfSubspecies +
##     log(mass), data = compSubspecies, lambda = "ML")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3996 -0.1690 -0.0071  0.1507  0.7863
##
## Branch length transformations:
##
## kappa  [Fix]  : 1.000
## lambda [ ML]  : 0.065
##    lower bound : 0.000, p = 0.1
##    upper bound : 1.000, p = <2e-16
##    95.0% CI   : (NA, 0.260)
## delta  [Fix]  : 1.000
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.0258     0.6815   -0.04     0.97
## log(pubmedRefs + 1)   0.9855     0.1355    7.27  1.2e-11 ***
## NumberOfSubspecies    0.1486     0.0679    2.19     0.03 *
## log(mass)             0.1305     0.1849    0.71     0.48
## ---
## Signif. codes:  0 '***' 1e-03 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.29 on 174 degrees of freedom
## Multiple R-squared: 0.339,Adjusted R-squared: 0.327
## F-statistic: 29.7 on 3 and 174 DF,  p-value: 1.45e-15
# logged subspecies variable
logRm1.summary
##
## Call:
## pgls(formula = virusSpecies ~ log(pubmedRefs + 1) + log(NumberOfSubspecies) +
##     log(mass), data = compSubspecies, lambda = "ML")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9345 -0.1998 -0.0249  0.1271  0.9651
##
## Branch length transformations:
##
## kappa  [Fix]  : 1.000
## lambda [ ML]  : 0.067
##    lower bound : 0.000, p = 0.1
##    upper bound : 1.000, p = <2e-16
##    95.0% CI   : (NA, 0.264)
## delta  [Fix]  : 1.000
##
```

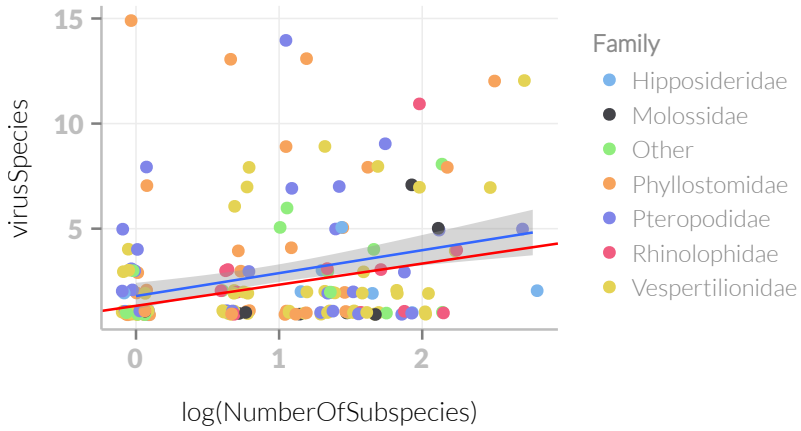

FIGURE 24. Number of virus species against log number of subspecies. Nonphylogenetic trend line in blue. Phylogenetic model (evaluated at mean body mass and mean study effort values) is shown in red.

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.0161     0.6921   -0.02    0.981
## log(pubmedRefs + 1)        1.0056     0.1348    7.46  3.9e-12 ***
## log(NumberOfSubspecies)    0.4807     0.2504    1.92    0.057 .
## log(mass)                  0.1381     0.1868    0.74    0.461
## ---
## Signif. codes:  0 '***' 1e-03 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.29 on 174 degrees of freedom
## Multiple R-squared: 0.335,Adjusted R-squared: 0.323
## F-statistic: 29.2 on 3 and 174 DF,  p-value: 2.42e-15
```

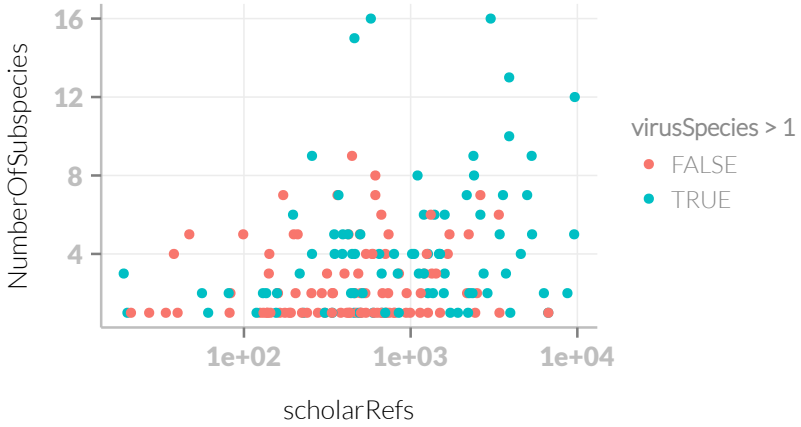

FIGURE 25. Number of subspecies by log study effort with colour indicating whether a species has 1 or more than 1 known virus species. There does not seem to be a huge difference. Species with many references often have many subspecies as expected. Species with many references, have more subspecies per effort if they have multiple viruses.

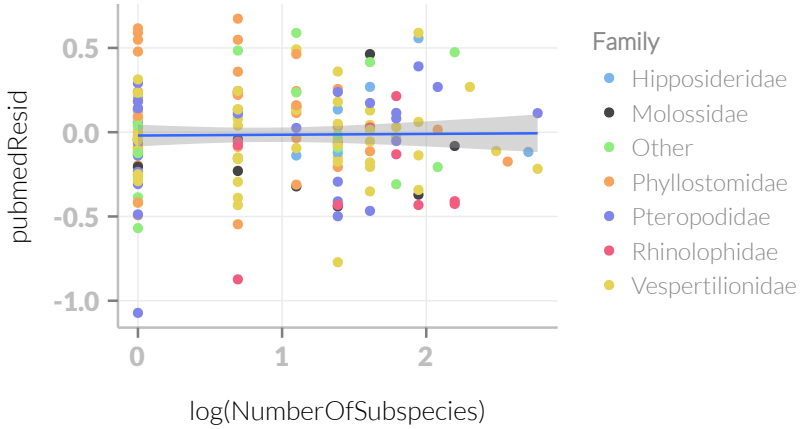

FIGURE 26. Plot using residuals from number of viruses against number of citations (study effort). Nonphylogenetic trend line added.

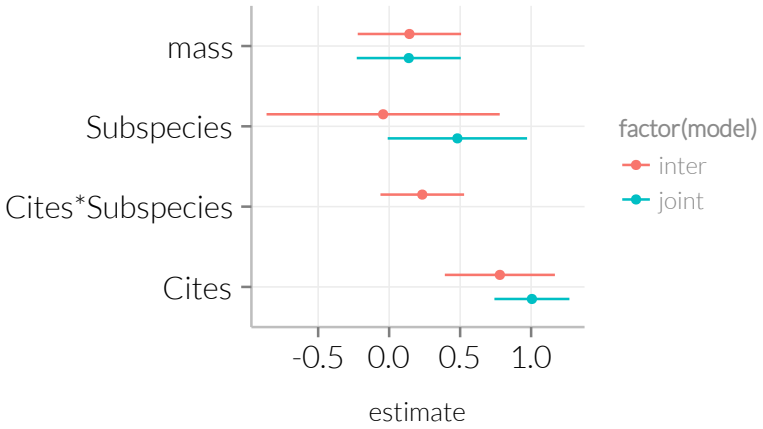

FIGURE 27. Plot of coefficient estimates and 95% confidence intervals for phylogenetic model with (inter) and without (joint) interactions between study effort and number of subspecies. Without interactions, number of subspecies is marginally significant.

4. RESULTS

See Figure 27 for a display of estimated coefficients for the two models using number of viruses as the response variable. The main model with mass, study effort and number of subspecies as predictors found study effort to be highly significant (β =1.01, p =3.9×10^{-12}). The number of subspecies was marginally significant (β = 0.48, p =0.06). The effect of nonindependance due to phylogeny was very small (λ =0.07, p =0.1).

The interaction term between study effort and number of subspecies, when included, was not significant (β = 0.23, p =0.12).

The model using the residuals from a regression between number of viruses and study effort as the response variable found no significant affect of number of subspecies. Mass was marginally significant.

5. Discussion

References

[1] Angela D Luis, David TS Hayman, Thomas J O'Shea, Paul M Cryan, Amy T Gilbert, Juliet RC Pulliam, James N Mills, Mary E Timonin, Craig KR Willis, Andrew A Cunningham, et al. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proceedings of the Royal Society B: Biological Sciences*, 280(1756):20122753, 2013.

[2] Integrated taxonomic information system (ITIS). `http://www.itis.gov`.

[3] Scott A Chamberlain and Eduard Szöcs. taxize: taxonomic search and retrieval in R. *F1000Research*, 2, 2013.

[4] Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2015. R package version 0.2.0.

[5] Don E Wilson and DeeAnn M Reeder. *Mammal species of the world: a taxonomic and geographic reference*, volume 12. JHU Press, 2005.

[6] Kate E Jones, Jon Bielby, Marcel Cardillo, Susanne A Fritz, Justin O'Dell, C David L Orme, Kamran Safi, Wes Sechrest, Elizabeth H Boakes, Chris Carbone, et al. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology*, 90(9):2648–2648, 2009.

[7] Susanne A Fritz, Olaf RP Bininda-Emonds, and Andy Purvis. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology letters*, 12(6):538–549, 2009.

[8] Olaf RP Bininda-Emonds, Marcel Cardillo, Kate E Jones, Ross DE MacPhee, Robin MD Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis. The delayed rise of present-day mammals. *Nature*, 446(7135):507–512, 2007.

[9] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[10] David Orme, Rob Freckleton, Gavin Thomas, Thomas Petzoldt, Susanne Fritz, Nick Isaac, and Will Pearse. *caper: Comparative Analyses of Phylogenetics and Evolution in R*, 2012. R package version 0.5.