Supplementary Discussion

Olival et al. 'Host and viral traits predict zoonotic spillover from mammals'

Total and Zoonotic Viral Richness in Domestic Species

Historical analysis suggests that early agricultural expansion underlies a large number of once-zoonotic, now-endemic human pathogens, due to increasingly intimate contact with animals during domestication, and an increasingly dense and stationary human population[1]. It is also likely that surveillance effort among domestic animals is skewed by their economic importance, and the number of zoonoses is affected by intensity of production, length of time since domestication, or other species traits. To test these hypotheses, we ran separate model selection for 32 domestic mammal species in our dataset based on a separate set of variables (Methods; Extended Data Figure 1). All code, data, and model outputs needed to replicate and evaluate these domestic animal models are provided at http://doi.org/10.5281/zenodo.569079

The best model for total viral richness in domestic mammal species explained 94.2% of the total deviance yet was primarily driven by research effort with a small effect of intensity of production and mammal order – highlighting large biases in viral surveillance relative to wild species. The number of disease-related publications per domestic animal species scales positively with total viral richness and explained 72.4% of the relative deviance in the best-fit total viral richness model using all data.

The best-fit model for zoonotic viruses in domestic species included total viral richness (offset), intensity of production and host order – with a significant negative effect for Cetartiodactyla relative to the other orders and variables included. Using the stringent dataset for domestic animals, the proportion of zoonotic viruses per host was also predicted by the number of years since domestication (log) and phylogenetic distance to humans yet these variables were not included consistently in the top models (within $\Delta AIC < 2$). Cross-validation tests for the total and zoonotic virus models for domestic animal species had a poor model fit for some folds of the data, indicating that the small sample size limits predictive power (Supplementary Table 2).

Comparative Performance of Host-specific Models

We assess the variation explained by our models with the metric of deviance explained (D_{null} – D_{model})/D_{null}, where D_{null} is the deviance of an intercept (or intercept and offset) only model. We present both deviance explained including all variables and excluding our measures of research effort (number of disease-related citations). While many measures are used for assessing model performance, deviance explained is an appropriate measure for a broad class of likelihood-based models, and may be considered a generalization of R^2 in ordinary least-squares regression.

In the discussion, we state that our model has performance greater than or comparable to previously-published studies that examined patterns of viral richness within a narrower taxonomic set of hosts (i.e. within a mammalian order)[2-4]. As model performance is measured and reported differently in different studies, and many *pseudo-R^2* methods that are used are often not reported, our comparison with these studies relies on selecting performance metrics that are published and also in deriving measures that approximate the performance scale we use.

Luis et al. 2013 use generalized least squares to identify host traits correlates of zoonotic viral richness in bats and rodents[3]. They provide AICc values, from which model deviance may be derived, for best and null models, but we are unable to separate deviance explained by citations based on the information provided. Based on the provided values, we found that the best model with rodents and bats for predicting number of zoonoses at species-level explained ~12% of total deviance, including research effort. The best model with rodents and bats predicting number of total viruses at species-level explained ~11% of total deviance, including research effort. Luis et al. also calculate an *R* (correlation between observed and predicted values) value to describe model performance (i.e. the correlation between viral richness predictions and observations for each species). This value is *R*=0.66 for their best model to explain zoonotic viral richness. While we believe this is not best approach for measuring our GAM performance[5], but following the same approach for our models, we would calculate R of 0.94 for our all-zoonoses GAM, 0.72 for our all-viruses GAM, 0.81 for our strict-zoonoses GAM, and 0.66 for our strict-viruses GAM (see code for these calculations at http://doi.org/10.5281/zenodo.569079).

Han et al. 2015 used a binary response variable and boosted regression trees to determine whether or not a given rodent species is host to at least one zoonotic virus[2]. This approach was not directly comparable to our study design, and BRTs are not likelihood-based and thus cannot be measured in the same way by deviance explained. The pseudo-r^2 of their best model was 0.80 on training data, and 0.48 on test data, with 78% of relative importance coming from literature-bias variable. Our model fit is more comparable with the training data fit, so we approximate the variation explained by biological traits as 17.6%.

Davies and Pedersen 2008 examined virus sharing only among primates and did not account for research effort in their model[4]. Their approach to model viral sharing was structured differently from ours, however their model that included a phylogenetic and ecological trait had a pseudo $r^2 = 0.16$.

We separately fit all our models using two datasets, i.e. the entire dataset and 'stringent' set of data (see Methods). As previously noted, the stringent dataset included only viruses identified in mammal hosts using viral isolation, PCR, or other methods of nucleic acid sequence confirmation. Using the stringent data resulted in a reduction in the number of data points (host-species associations) from 2805 to 1460 observations, and as a result we observed several small differences in our model outcomes between the stringent and all data models. For example, we found that several mammalian orders were included in the all data zoonoses model, including Cetartiodactyla, Chiroptera, Perissodactyla, Scandentia, Peramelemorphia, and Diprotodontia. While all these orders provided some predictive power, only the first three orders were significant. However, using only the stringent data in the zoonoses model, the effect of Chiroptera does not provide predictive power, and the effect of Perissodactyla is reduced so as to be non-significant. In the stringent zoonoses model we also observed an increase in the strength of the negative effect for Cetartiodactyla, and Lagomorph and Primates were additionally included as they provide some predictive power, but the effects were not significant (Extended Data Table 1). While the strength of the effect associated with these taxonomic order inclusions and exclusions between models were relatively minor, or not statistically significant in several cases, these findings highlight the value of examining large, taxonomically-curated datasets for all mammals. These differences also indicate the need for systematic sampling across taxa with consistent methodology to better understand these sources of variation in viral richness.

The effect of research effort on the proportion of zoonotic viruses, for both the all data and stringent data models, scales positively at first, but decreases and becomes negative in more heavily researched species (e.g. Fig. 2f). While we cannot rule out that observing zoonotic viruses in a species may drive additional research, our finding suggests that initial viral discovery efforts are biased towards detection of human pathogens, and is supported by empirical studies that show extensive sampling effort is required to discover the majority of a wild mammal species' unique viral diversity[6].

Supplementary Discussion References:

1 Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279-283 (2007).

2 Han, B. A., Schmidt, J. P., Bowden, S. E. & Drake, J. M. Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences* **112**, 7039-7044, doi:10.1073/pnas.1501598112 (2015).

3 Luis, A. D. *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**, doi:10.1098/rspb.2012.2753 (2013).

4 Davies, T. J. & Pedersen, A. B. Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proceedings of the Royal Society B-Biological Sciences* **275**, 1695-1701, doi:10.1098/rspb.2008.0284 (2008).

5 Willmott, C. J. Some Comments on the Evaluation of Model Performance. *Bulletin of the American Meteorological Society* **63**, 1309-1313, doi:http://dx.doi.org10.1175/1520-0477(1982)063%3C1309:SCOTEO%3E2.0.CO;2 (1982).

6 Anthony, S. J. *et al.* A strategy to estimate unknown viral diversity in mammals. *mBio* **4**, doi:10.1128/mBio.00598-13 (2013).