ALL THREE REPLIES TO REVIEWERS ARE INCLUDED BELOW:

RE: Rodent reservoirs of future zoonotic diseases

Dear Dr. Levin and Dr. McMichael,

Thank you for editing our manuscript and providing us with the opportunity to submit a revised version. We have addressed all of the comments and suggestions by the two reviewers, which you will find below in blue text. We feel that these constructive comments have improved the manuscript and we hope that you will now find the paper suitable for publication in PNAS.

Thank you very much for your efforts in the communication of our work.

All the best,

Barbara Han

REVIEWER 1:

This paper is a methodical and statistically robust appraisal of one important group of zoonotic reservoirs, rodents. Machine based learning appears to be a valuable tool. Key results indicate zoonotic rodent hosts and super hosts are characterised by a "fast-paced" life history strategy, larger geographic ranges, and overlap with at leat 50 people/km2 human density. These are characteristics previously reported (although derived differently) in other studies. I think it would be preferably to acknowledge that these results confirm earlier findings or hypotheses (particularly as they used other data sources), e.g. a selection:

Generalist hosts and generalist pathogens, broad overlapping host ranges are discussed in work by Cleaveland et al. 2001; Taylor et al. 2001; ME Woolhouse, including:

Woolhouse MEJ & Gowtage-Sequeria S (2005) Host range and emerging and reemerging pathogens. Emerging Infect. Dis. 11:1842-1847.

Davies TJ & Pedersen AB (2008) Phylogeny and geography predict pathogen community similarity in wild primates and humans Proc. R. Soc. B 275(1643):1695-1701.

the association with human modified environments generally:

McFarlane, Ro, Adrian Sleigh, and Tony McMichael. "Synanthropy of Wild Mammals as a Determinant of Emerging Infectious Diseases in the Asian-Australasian Region." EcoHealth 9.1 (2012): 24-35.

and more specifically for rodents and including other life history traits:

Bordes, Frederic, Vincent Herbreteau, Stephane Dupuy, Yannick Chaval, Annelise Tran, and Serge Morand. "The diversity of microparasites of rodents: a comparative analysis that helps in identifying rodent-borne rich habitats in Southeast Asia." Infection ecology & epidemiology 3

(2013).

> ***We're grateful to Reviewer 1 for taking the time to compile this list. We have now added several of the suggested references (see reference numbers 1, 5, 6, 25, 26) and we draw attention to the ways in which the results of our clade-level analysis corroborate the findings of other studies in specific hosts and disease systems which draw upon independent data sources (see in particular line 107-111 with reference to a new study by Ostfeld and colleagues, ref #28). Here we also added text to mention patterns of synanthropy and zoonotic reservoirs (including McFarlane et al. and Bordes et al. references suggested above; refs #22-23),***
>
> ***We didn't explicitly investigate host range (the diversity of host species a parasite may infect) in our analyses or patterns of parasite sharing and overlap among wild host species, so we have refrained from discussing our results in this particular context.***

Current literature providing a rationale for the association with fast living strategies is convincing.

I have some concern regarding source data despite the high accuracy of the models to predict zoonotic reservoirs. The study is based on the PANtheria database (the principle author of which is also the first author of an influential paper in this field, and cited by this study i.e. Jones KE et al., 2008). This might raise issues of independance of findings, particularly for the variables generated for each species describing the anthropogenic and environmental conditions within each geographic range based on the extent of digital species range maps - each step with its own assumptions.

> ***We had difficulty interpreting this comment, and we are not sure what the reviewer means by "independence of findings". We found the vast majority of the environmental and anthropogenic variables were not important predictors of zoonotic reservoir status (see trait profiles in Table S2). The environmental and anthropogenic variables are calculated in PanTHERIA as a central tendency across the geographic range of the host species, so the species ranges would have to suffer from large inaccuracies in order to throw off statistical results. In addition, inaccuracies in PanTHERIA would tend to obscure statistical relationships rather than generate them, unless all of the data happens to be wrong in the same direction. While it is possible, it is hard to imagine what inaccuracies in the range shape files could be propagated across both the environmental and the anthropogenic variables to cause them to be wrong in a correlated way.***

Some further critical thinking or clearer writing could improve this work. I have particular issue with the following:

"Current hot spots of rodent reservoir diversity occur in North America, the Atlantic coast of South America, Europe, Russia, and parts of Central and EastAsia (Figure 2). These hot spots generally coincide with regions of high mammal biodiversity (Figure S2) where the risk of zoonotic disease emergence is greatest (2), but there are fewer rodent reservoirs in South America and Africa than expected given the high levels of mammal biodiversity in these regions".

Do the authors mean to use high mammal diversity as a proxy for rodent diversity? Obviously some areas favour rodents, reservoir or otherwise. For example, there are globally 2261 rodent species, 609 are found in south America- largest number per continent or land region (in fact accounting for about a half of S.American mammal diversity: 609/1295 total mammal spp. Rodents are a very small proportion of total mammal diversity in parts of Oceania.

> ***In this section we meant to place rodent reservoir diversity in context with biodiversity writ large to acknowledge that patterns of reservoir diversity could be driven by overall patterns of mammal biodiversity, and to place our results in context with conclusions from Jones et al. which suggest that zoonotic disease emergence should coincide with regions of high biodiversity. We had not previously considered placing our results in the context of rodent biodiversity specifically, but following reviewer suggestions we have discussed our results in the context of recently published rodent biodiversity maps (Jenkins, Pimm & Joppa 2013), line 93.***

Do the authors accept these findings from Jones et al 2008 uncritically ("regions of high mammal biodiversity ....where the risk of zoonotic disease emergence is greatest (2)")? That was a broad brush analysis that could be refined with the help of this study, i.e. rodents are the most diverse mammal group, they are important in many reemerging diseases as well as first emergences of new pathogens - the event of interest in Jones et al 2008. This issue is far more interestingly and critically dealt with elsewhere where the authors state that:

"predicted rodent reservoirs occur broadly, spanning all biomes, in regions that exhibit a wide range of mammal species richness and in middle to high income economies. Nearly all of the predicted hyper-reservoirs occur in upper latitudes in developed nations with relatively low biodiversity..... in ecoregions that experience an appreciable degree of seasonality"

It would be interesting to discuss the implications of these findings further.

> ***In lines 91-94 and 103-107 we have included additional discussion of our results with these comments in mind. In particular, we note that while geographic ranges of the predicted reservoirs diverge from those reported in Jones et al. 2008, they do coincide with regions where human EIDs are most concentrated (as reported by Jones et al. 2008, Suppl. Fig. 1, 2a). As noted by the reviewer, this is particularly interesting given that rodents are often associated with first emergence of new human pathogens, which is the focal point of the Jones et al. paper.***

Also, how the findings of this study match with zoonoses incidence and/or burden of disease in these hot spots would be important to clarify, and the difference between the two (occurrence of host and incidence of disease) should be made explicit.

> ***We thought this was a great suggestion and an important addition to the paper. To examine the human burden of disease we compiled data on the number of rodent-borne disease outbreaks in humans since 1990 at the country scale, as well as the total number of rodent-borne zoonoses that are found within each country. From these data we created a new map illustrating regions exhibiting the highest***

> ***diversity of rodent-borne zoonoses, and where the most outbreaks of these diseases have occurred globally. This map also shows the geographic ranges of rodent species predicted to be undiscovered carriers of zoonoses.***
>
> ***This new map corroborates supplementary findings of Jones et al. (Supplementary Figure 1, 2a) and shows that predicted reservoir species are distributed in countries with high burdens of rodent-borne zoonoses (countries that either show a large number of distinct zoonoses, or countries which exhibit numerous outbreaks).***

I would feel more confident in the methodology and results of this study if these were tempered by consideration of points raised above. However I congratulate the authors for their innovative approach to advancing this area of research.

REVIEWER #2:

Comments:
Han et al. using machine learning methods to identify the probability of a rodent species being a novel zoonotic reservoir host, or if the species is already a host, becoming a super-reservoir, based on species' life history, ecology and geographic traits. Han et al. create and describe the model and then use the model to predict 20 species as either being novel potential reservoir species or new super-reservoir hosts and plotted on species richness maps as hotspots.

The manuscript is certainly topical and addresses an important topic of general interest - that of creating methods to predict reservoir hosts of new, or hosts with more, zoonotic pathogens and thereby better understand the process of zoonotic emergence. The ms is clearly written and the figures are of good quality. I liked the machine learning approach and although not completely novel in life history or conservation analyses (see Bland et al. 2014 Conservation Biol) it is definitely interesting and valuable. However, there are some questions about the analyses that are not clear and may be potentially confounding, there is also a lack of clarity in the presentation of the results and in general I am not convinced that the ms presents the conceptual leap needed for publication in PNAS. My specific points are as follows:

1. Analyses - (a) The analyses appears to have taken 216 species' traits that are reported as zoonotic (having a pathogen that is shared with humans) and split the data into a training and testing set to build the machine learnt algorithm and test its performance (zoonotic/non-zoonotic and zoonotic/super-reservoir zoonotic). It is not clear how the zoonotic/non-zoonotic analyses were performed as conceptually to build a model you would need to have some 'zeros' in the analyses. Zeros in this context would be species which have been investigated for pathogens but don't have ones shared with humans. There are lots of species in the analyses without zoonotic pathogens (2277 minus 216 species) but it is not clear if these are a) true zeros, or b) have been included to train the model to recognise zoonotic species. This is in some ways analogous to niche modelling's presence/absence data, and where real absence data is not available, pseudo-absences are often created. It is not clear to the reader if these types of absence data are included or needed to be included.

> ***The reviewer identifies several methodological questions that that we have taken care to address in our revision.***

The 216 reservoir species and 2061 other rodent species were combined for a total of 2277 species (lines 57-58). These species were divided into a training set (80% of all 2277 species) and a test set (the remaining 20% of all species), with both training and test sets containing reservoir species (lines 151-152). Thus, we adopted a conservative approach in designating a single contrast class (reservoirs) vs. everything else. In other words, we considered species that have been sampled and found not to harbor zoonotic pathogens the same as those which have no prior surveillance history. We did this to acknowledge the uncertainty inherent in designating species as 'non-reservoirs' (e.g., in the worst case, species for which a single individual tests negative for a given zoonotic pathogen probably should not be treated as a 'non-reservoir' species). A more conservative designation which applies the same label to non-reservoirs and unknown species would lead also to more conservative models whose classification performance can only increase with the addition of newly discovered reservoir species in the future (discussed in lines 154-159).

As an additional check, we conducted another boosted regression analysis with the reservoir species duplicated in the data set as "unknowns" (i.e., labeled the same as the other 2061 species; referred to here as Method B). This gives the effect of looking at presence vs. background (all rodent species), as suggested by the reviewer, and contrasts with our approach (Method A). Results show that both analyses give similar trait profiles, but the rate of detecting true or known positives is much higher with Method A. At the same time, Method A yields a higher false positive rate, and may therefore be flagging more species currently labeled as unknown but likely to turn out positive if targeted for surveillance. In our view, a higher false positive rate (classifying a species as a reservoir when it is not) is preferable given the potential costs of missing a novel reservoir. We plot the comparison between Methods A and B below using ROC curves, a common diagnostic of how well a method is able to distinguish between two classes.

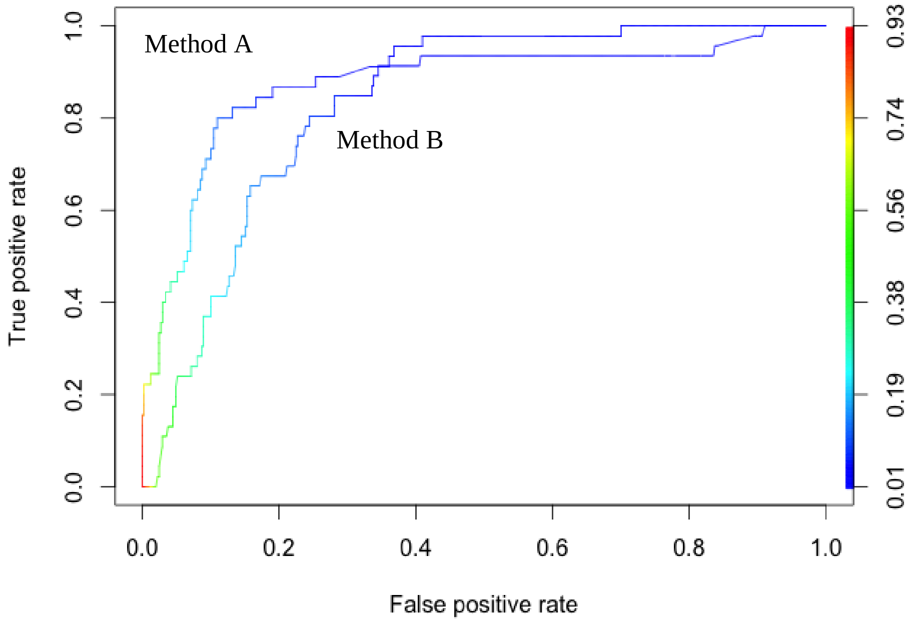


Figure. **ROC curves showing the relationship of sensitivity (true positive rate) to specificity (false positive rate) for Method A vs. Method B.**

(b) Traits seem to have been included completely indiscriminately and I think more thought is needed here to include ones that directly help to address the hypotheses and are non-repetitive.

For example, LittersperYr and LittersperYr_EXT are actually the same variable but with the latter variable has a bit more data in it by making some assumptions about methods of counting infants. There are many of these duplicate variables.

> ***We included all variables published in the PanTHERIA dataset, which represent both intrinsic host traits as well as biogeographical characteristics for each species. As noted by the reviewer, some of the variables in PanTHERIA are similar to each other. While there is appreciable correlation across some of the predictors, the variables themselves are not redundant. For example, we included X5.1_AdultBodyMass_g and X5.5_AdultBodyMass_g_EXT; and X16.1_LittersPerYear and X16.2_LittersPerYear_EXT. The _EXT suffix designates a derived variable in PanTHERIA which was calculated by fitting a GLM that accounts for known intercorrelated variables (e.g., a large degree of variation in Adult Body Mass is explained by Adult Body Length; similarly, a large degree of variation in Litters Per Year is explained by Interbirth Interval; explained in Table 3, Metadata portion of*** *(Jones et al. 2008)****). Thus, for X5.5_AdultBodyMass_g_EXT, this approach yields a single fitted value representing Adult Body Mass after accounting for correlation with Adult Body Length. Generalized boosted regression algorithms include methods for regularization ("shrinkage") to prevent overfitting so that simply including more predictor variables (Elith, Leathwick & Hastie 2008) will not lead to spurious improvements in prediction. By including all variables we allow the boosting algorithm to iteratively improve classification accuracy by learning from the additional information that may be represented in minor variables. However, as can be seen from the trait profiles, the majority of the derived variables are not important for prediction and we found that removing these variables from the analysis altogether did not improve nor degrade predictive performance of the models.***
>
> ***More importantly, our purpose in taking a machine learning approach was to learn as directly as possible from the data that exists rather than limit the analysis to what we expected, a priori, would be important predictors. In other words, a major goal was to use the data to generate rather than test hypotheses linking life history to zoonotic host status.***

(c) It wasn't really that clear why the authors had just focused on rodents. Yes they have a lot of zoonotics but so do other mammalian orders such as bats, ungulates, carnivores etc. It didn't seem a huge amount of more work to do it across mammals.

> ***We focused on rodents because, historically, they have been identified as carrying a disproportionate number of zoonoses. As the reviewer suggests, we originally considered combining all clades into a single analysis but found fundamental data differences among the mammal clades (primates, carnivores, ungulates, rodents, and bats) that would impact our ability to interpret trait profiles. Using bats as an example clade - bats carry a large number of zoonoses just like the rodents, but the vast majority of these are viral whereas the distribution among protozoa, bacteria, helminths, and viruses are less skewed for the other clades. Bats also have quite complete data on adult forearm length (which is often recorded to estimate age and size), whereas this variable has near zero coverage in rodents and is less meaningful. These and similar clade-specific***

differences in data coverage, and the biological information therein, led to different analyses, figures, and discussions.

(d) It really wasn't clear how the 60-70% values were arrived at for the 13 species highlighted as novel zoonotic species or why that probability was chosen, e.g. why not >80%.

The highest probability was 70%. We tried to strike a balance of selecting species with high probabilities (those in the 99th percentile) but not so many species that the map depicting their geographic ranges would appear too busy. We therefore chose a natural break in the probability results to arrive at 13 species (i.e., a nice round probability of 0.60), which includes those species which were ≥ 99.4th percentile. Of course, there are many possible ways of selecting species. Exact probabilities can be replicated using our published code and data, if desired.

(e) It wasn't clear why that particular machine learning classification method was chosen, there are a LOT out there! What makes this particular one good for this analysis? How sensitive are the results to the method? Would you get different results if you chose another one?

These are excellent questions – there are indeed several methods available. Among the classification algorithms, neural nets and boosted trees are consistently in the top performers in terms of accuracy but boosted trees in our experience tend to be much more robust (less sensitive to tuning). Additionally, the local nature of classification trees means they should be less biased by data that are not evenly distributed across the input space. As a bonus, boosted regression trees are among the most accessible of the updated classification and regression algorithms, particularly to ecologists, due to the availability of tutorial-like resources including a working guide *(Elith et al. 2008)* ***and several actively supported R packages and websites that support gbm analysis (gbm, dismo, caret).***

(f) What is the effect of missing data? There are a lot of species which only have body mass and this needed to be discussed.

Boosted regression trees and other tree-based machine learning methods can readily accommodate missing values by treating "missingness" as an attribute of a predictor variable which the algorithm learns during the training process (described in detail in Section 9.6, p.332-333 of (Hastie, Tibshirani & Friedman 2009); also see *(Elith et al. 2008; Kunz, M. & Johnson, K. 2013)*. ***Importantly, missing values are not discarded or imputed, as commonly done for traditional parametric methods and some other machine learning methods.***

We double-checked our data for species where body mass was the sole reported variable and were unable to find any species for which this was the case. In any case, we agree that readers may be inquisitive about data coverage so we have included a new supplementary table reporting the percent coverage for each variable across all rodent species and for all rodent reservoir species (Table S3).

2. Results - (a) Some of the figures don't appear to be that novel and it isn't clear what value these have in a PNAS paper. For example, Figure 2a and 3a is just plotting out the richness map of zoonotic rodent species found in GIDEON, Figure S2 is a map of mammalian species richness which has been shown elsewhere.

> ***We were also surprised to be unable to find a published global map of the geographic distributions of zoonotic reservoir species, either generally (for mammals) or specifically for rodents. Thus we thought that, in addition to making a novel contribution, Figure 2a might be useful for sparking hypotheses for future work. For example, from this map we were curious as to why there weren't more rodent reservoirs in regions with disproportionately high biodiversity (South America and Africa), especially in light of the conclusions of Jones et al. 2008 in which regions with high wildlife host species richness (developing countries at lower latitudes) are at greatest risk of future zoonotic disease emergence (discussed beginning in line 91). We included a map of mammal biodiversity in the supplementary materials for ease of reference, especially since regions of high mammal diversity coincide with rodent reservoir hot spots. We have dropped this map and replaced it with a citation for online maps of biodiversity available at mappingbiodiversity.org*** *(Jenkins et al. 2013)****.***

(b) Table S1 seems to be missing the pathogen type next to the name of the pathogen and Table S2 needs more work to make it understandable, there are abbreviations not mentioned in the legend and the different results for the error distributions are never mentioned in the text but seem important?

> ***We have added back the pathogen type to Table S1.***
>
> ***The error distributions in Table S2 correspond to the type of response variable for each analysis. For example, reservoir and super-reservoir status are binary response variables with Bernoulli distributed errors. The number of zoonotic diseases and the number of WOS citations per rodent species are both count variables and have Poisson distributed errors. We clarify this in lines 139-140 of the Methods, and also in the Figure legend of Table S2. We also added text mentioning the pseudo-R^2 results from a model treating reservoir status as a count variable and mention explicitly that both count and binary treatments of reservoir status led to similar trait profiles (lines 63-65 of Results).***

Also I don't understand why there are only families in the table - where are the per species results? Also shouldn't there be misclassification rate results?

> ***The families in Table S2 reflect the use of families as additional binary variables in analyses. This allowed us to identify whether species in particular families were more likely to be zoonotic reservoirs (explained in lines 131-132 under Data). In contrast, the 'per species results' are the species predicted by the boosted regression models (reservoir and hyper-reservoir status, both as binary variables) to be novel reservoirs of zoonotic disease based on their trait profiles.***

> ***The misclassification rate is subsumed in AUC values, which sums the relationship of sensitivity (true positive rate) to specificity (false positive rate), and is a common diagnostic of how well a method is able to distinguish between two classes. These values are reported in Table S2. However, which misclassification indices are most salient when evaluating a model depend on how costly one considers false positives relative to false negatives. For our purposes, a false positive (classifying a species as a reservoir when it is not) is far less costly than a false negative (classifying a species as a non-reservoir when it has been demonstrated to harbor zoonoses) especially given the large number of species which have not been investigated.***

(c) It is not clear which taxonomy was used and how species from GIDEON were matched to those in Pantheria.

> ***Taxonomy for GIDEON and PanTHERIA were joined using (Wilson & Reeder 2005). We now note this explicitly at the end of the Methods section (lines 139-140).***

3. Importance - The machine learning approach is an interesting one, however in the end it is correlative not mechanistic method. I am not against correlative analyses at all but zoonotic emergence is a complicated process! What the authors are trying to model is the risk of a species becoming zoonotic by looking at host characteristics when in reality the probability of a species hosting a zoonotic pathogen also involves human ecology and host-human interactions (contact rates, human demography, socioeconomic status, phylogenetic relatedness of the host to humans). So I am not convinced that this (although interesting) study is a big conceptual step forward in understanding zoonotic emergence.

> ***Absolutely – the disease emergence process is hugely complicated, involving many additional factors including those listed here by the reviewer. Numerous other studies, including recent work exploring how poverty and biodiversity interact with infectious disease*** *(Bonds, Dobson & Keenan 2012; Ngonghala et al. 2014)****, have collectively made enormous contributions to our understanding of the diverse factors influencing zoonotic emergence and ecology (lines 112-113; we have also cited 11 representative studies in line 113). We hope that our unique examination of highly conserved intrinsic host traits will add value to this quickly evolving body of literature.***

Literature Cited

Bonds, M.H., Dobson, A.P. & Keenan, D.C. (2012) Disease Ecology, Biodiversity, and the Latitudinal Gradient in Income. *PLoS Biol*, **10**, e1001456.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York.

Jenkins, C.N., Pimm, S.L. & Joppa, L.N. (2013) Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences*, **110**, E2602–E2610.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L., Safi, K., Sechrest, W., Boakes, E.H., Carbone, C., Connolly, C., Cutts, M.J., Foster, J.K., Grenyer, R., Habib, M., Plaster, C.A., Price, S.A., Rigby, E.A., Rist, J., Teacher, A., Bininda-Emonds, O.R.P., Gittleman, J.L., Mace, G.M. & Purvis, A. (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals (ed WK Michener). *Ecology*, **90**, 2648–2648.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. & Daszak, P. (2008) Global trends in emerging infectious diseases. *Nature*, **451**, 990–993.

Kunz, M. & Johnson, K. (2013) *Applied Predictive Modeling*. Springer, New York.

Ngonghala, C.N., Pluciński, M.M., Murray, M.B., Farmer, P.E., Barrett, C.B., Keenan, D.C. & Bonds, M.H. (2014) Poverty, Disease, and the Ecology of Complex Systems. *PLoS Biol*, **12**, e1001827.

Wilson, D.E. & Reeder, D.M. (2005) *Mammal Species of the World: A Taxonomic and Geographic Reference*. JHU Press.

RE: Rodent reservoirs of future zoonotic diseases

Dear Editors,

Thank you for providing us the opportunity for a second revision. The reviewers have, again, provided thoughtful and constructive comments and questions, all of which we address below in blue text.

This process has really helped to improve the communication of this work, and we hope that you will now find the paper suitable for publication in PNAS.

All the best,

Barbara Han

REVIEWER 1:

This is a very interesting manuscript on the forward-thinking topic of identifying potential sources of zoonotic disease before diseases emerge. I applaud the authors on their innovative approach.

I am not fully convinced that the issue of sampling has been adequately addressed. The fact that traits didn't predict 'studiedness' out of sample is compelling. But it does do really well in-sample. How did you choose the samples for the in-sample versus out-of-sample tests? I wonder how well the models would have done if you had chosen a different sample.

The training (in-sample) and test (out of sample) data were partitioned randomly (by setting the seed) and stratifying by reservoir status, and the distribution of WOS hits (for studiedness). Training accuracy was determined through 10-fold cross-validation, providing results representing the averaging the results of 10 models trained on 10 random partitions of the data.

The main reasons that make me cautious are, as you said, the number of citations for each species increases monotonically with the number of zoonoses they harbor. And secondly, in support of your conclusions, you point out that the geographic ranges of the rodents you predict to be novel reservoirs are located in countries where human EID events are the most concentrated (according to Jones et al 2008). However, Jones et al. point out that the geographic distribution of EID events has been highly biased by sampling. And the relative risk maps also presented in Jones et al (fig 3) are quite different from the current EID distribution (fig 2) because sampling bias was taken into account.

I would be more convinced if there was some kind of secondary analysis that further demonstrates that sampling bias is not important. For example, if you put citations in your analysis where does it fall out? Is it toward the top of the tree? I know this doesn't get at the issue you are interested in- prediction of novel reservoirs, but it would give an indication of how important sampling is. You said that the number of citations for each species increases monotonically with the number of zoonoses they harbor. What about using the residuals of a regression of # of zoonoses ~ citations, for your response variable in the regression trees- How different are your results then?

We included citation counts in earlier analyses to confirm sampling effects, and they were the top predictors of reservoir status/#zoonoses. This is not surprising, especially given that citation counts increase exponentially rather than linearly with the number of zoonotic pathogens found per species. There are additional, possibly interacting, factors driving sampling bias. As one example, **Mus musculus** ***carries 11 unique zoonotic pathogens, is an important model system for biomedical research, and also exhibits interesting behavioral changes due to zoonotic infection (being attracted to cat urine) by*** **Toxoplasma gondii** ***(the etiologic agent for Toxoplasmosis).***

These issues motivated us to consider more directly the possibility that we are describing the trait profile of well-studied rodent species rather than highly permissive zoonotic reservoirs. If this is so, we would expect that the traits that best predict citation counts are very similar (in both the order and magnitude of relative importance scores) to the traits predicting zoonotic status/#zoonoses. The last 2 columns of Table S2 show that the best-tuned models for predicting citation count returned very low pseudo-R^2 values (0.07). Even when we modeled a subset of the best-studied species (those with citation counts > 10), prediction accuracy only reached pseudo-R2=0.17.

More generally, we realize that these and similar uncertainties raised during the review process could benefit from more lucid explanation concerning the relationship between sampling bias and intrinsic traits, which we have now added in lines 56-59. There is little doubt that sampling biases are inherent in these kinds of data (metrics of infection in humans and animals), but it is difficult to see how biases in our ability to sample infection in wild species will bias the estimates of species intrinsic biology or life history characteristics.

Other comments/questions:

I would have liked to see more of the main results that are in table S2 in the main text, perhaps a figure of variable importance for the top 7 or so variables for the different analyses? Is fig 4 in order of importance? Why some here and some in the supplemental?

We put more of the main results from Table S2 into a new Figure (Fig. 4), which now shows, in order of importance, partial plots for all features with relative importance scores > 1. Originally, we had produced more partial plots and split them up by intrinsic biological features (for the main figure) and biogeographical features (which we tucked into supplementary section). We have deleted the biogeographical plots because the patterns they suggest were more effectively visualized as maps (Figs 2-3).

We now present new maps visualizing geographic ranges for all species comprising the 90th percentile of species predicted to be new reservoirs (Fig. 2B) or hyper-reservoirs (Fig. 3B). This provides a more intuitive depiction of hotspots of the top 10% of predicted new reservoirs, and a list of these species and their probabilities are now given in Table S4.

The trait profiles given by both analyses (reservoir status vs. the number of zoonoses as response variables) after including species density are similar – the features and the directionality of the predictions tell a consistent story of fast life history traits being

robust predictors of reservoir status. We display the partial plots from the Poisson model in Fig 4, and include the partial plots from the Bernoulli model in a new Figure S1. More plots, if desired, can also be drawn using the code and data that will be deposited in Dryad.

How did you classify a species as a reservoir? "... standard diagnostic procedures ..." Does that include detection of antibodies (that could cross-react)?

We checked each putative reservoir species against published literature (those explicitly cited in GIDEON as well as more recently published studies) to ascertain that the scientific community had reached consensus that the species was indeed a sylvatic carrier of the infectious pathogen causing the zoonotic disease. In most cases this was easy given the large number of citations. For those species with fewer citations (e.g., recently identified reservoirs, or species carrying a rare or neglected zoonotic disease) we found that the papers of first report (those declaring a new zoonotic reservoir species) generally reflected a more rigorous diagnostic standard and were always peer-reviewed.

The authors mention biodiversity a lot, but why not include it in the model? For example, Luis et al 2013 use as a covariate the number of other (rodent/mammal) species within a species' distribution.

Thanks for the suggestion. We counted the number of mammal species found within the geographic ranges for 2086 of 2277 species. 191 species were excluded, either because they had not been assessed by IUCN, or their species binomials do not match the most recent Wilson and Reeder (2005) mammal taxonomy.

For these 2086 species we derived species density (1), the number of unique mammal species divided by the area of the species' geographic range (n/km^2), which corrects somewhat for the strong positive correlation between geographic range size and species richness. Including this new covariate ("SpeciesDensity" n/km2 in Table S2; partial plots labeled as "log Species (n/km^2)") did not improve prediction accuracy, and gave very similar model outputs (trait profiles in Table S2 and species predictions Table S4). The partial plot for this covariate shows that while the majority of rodent species reflect medium to high mammal richness within their geographic ranges, zoonotic reservoirs tend to have disproportionately lower mammal richness within their geographic ranges. However, one caveat is that these outputs must be considered in light of decelerating species area curves – ie, species richness will nearly always be lower in smaller ranges. We discuss the implications of this new result (and caveats) in lines 97 and lines 108+ of the revised manuscript, which we think could be an interesting topic for future work.

I took me a while to understand figure 1. Because the legend starts with "types and frequencies of parasites", I at first thought that the numbers inside the circles were the number of parasites, e.g., there were 27 bacterial zoonoses that had 1 reservoir species. But if I now understand correctly, it's 27 rodent species that had 1 bacterial zoonosis. I would clarify in the legend.

We have rewritten this legend for clarity.

Table S1: Why are the arenaviruses listed separately by species, but the hantaviruses are simply classified as Old World- HFRS and New World-HPS? This lumps several species of

viruses together. A couple of species of arenavirus are not listed: Guanarito (Venezuelan hemorrhagic fever), and Flexal viruses.

The hantaviruses are classified by GIDEON as "Hantavirus - Old World" vs. "hantavirus pulmonary syndrome", which is the New World strain. The Old World strains cause hemorrhagic fever with renal syndrome whereas the New World strains cause Hantavirus pulmonary syndrome. There is essentially a unique strain of hantavirus for each rodent host species, and from the human health perspective, strain nomenclature is less important than how clusters of related strains present clinically in human patients.

Thank you for catching the VHF error – it was inadvertently excluded from our list and has been added back (Table S1). According to recent correspondence with GIDEON scientific support, GIDEON currently does not report Flexal virus as a pathogen of zoonotic relevance. I was able to find reports of 2 human cases reported briefly in book chapters (where they were listed in tables without primary citations and no additional detail), but these reports do not identify the putative rodent reservoir down to the species level, which excludes Flexal virus from our analysis even if GIDEON had included it.

What software/functions were used for the analyses?

We used the R environment, and the **gbm** ***package for generalized boosted regressions, which we now specify in lines 184-185.***

Some kind of metadata for Tables S2 and 3 are needed. I saw no definition of the variables. E.g., What's PET? What does EXT mean?

We have created a new Table S3 that lists all of the PanTHERIA variables included in our analysis along with their definitions, units of measure, and the percent coverage of each variable across all rodent species and across rodent reservoir species.

REVIEWER 2:

This is an interesting and topical manuscript, that is well presented.

I have two major concerns:

1) The term 'reservoir' is used misleadingly. Reservoir implies a definitive host responsible for the maintenance of endemic infection (see Haydon et al. 2002 EID). Whereas, the analyses presented show that potential for a zoonotic pathogen to be shared with particular rodent species, but does not predict reservoir status. At a stretch, I can see that the methods may predict potential candidates for reservoirs. But determining reservoir status for multi-host pathogens is a considerable undertaking even when potential hosts have been identified. This is an issue, which can easily be addressed by using less loaded terms (shared pathogens?). A definition of reservoirs would also help if it really has to be used. But I would strongly recommend putting a caveat about the capabilities to predict reservoir status.

Yes, we wish to be clear about what we mean by 'reservoir', especially since it differs from the ecological definition. Since an ultimate goal of our study was to offer predictions about which species should be targeted to determine their status as possible zoonotic reservoirs (**sensu strictu*****), we have kept the term but made clear our definitions***

(line 178): "In contrast to other ecological definitions of reservoir (38), we apply the term more generally to encompass wild species capable of carrying infections transmissible to humans; Predicted reservoir species are undiscovered potential sources of infection known to be transmissible to humans."

2) I was glad to see that the authors demonstrate their methods do not just identify the most studied species. However, I am concerned that a similar problem may be apparent in terms of their methods identifying the areas of the world where the most extensive sampling and diagnosis of pathogens and parasites has been carried out (Fig 2A looks a bit like that - except for Australia which is a bit of an exception in terms of biogeography anyway). In the poorest parts of the world, most cases of disease are never diagnosed. Recent work (Crump et al. 2013) shows that the vast majority of febrile illness (of which there are very many) in parts of Africa are likely to be from bacterial zoonoses. But who knows what the causes of these illnesses are and what are the reservoir hosts for these pathogens/parasites. Further investigation of Lepto for instance shows that many potential species are involved and that rodents as previously presumed to be the reservoir may not actually be such important hosts. This is obviously not that easy to address, since this study is reliant on data of varying quality and extent from elsewhere. But I would really like to see this issue addressed somehow. It would be good to see whether there are any correlations between the density of laboratories/ PCR machines (or some other good proxy for pathogen detection)/ published studies using PCR/ sequencing and number of reservoir species shown in Fig 2A. I also think this point relates to that made by Reviewer 1 which the authors responded well to regarding the burden of zoonotic diseases. Again it needs to be made very clear that these are diagnosed outbreaks and that the likely burden of disease is much much higher in parts of the world with limited diagnostic infrastructure and capacity.

Absolutely, the disparities in diagnostic capacity among countries are enormous. This was another reason we chose to examine intrinsic features of host biology and natural history because they are much less susceptible to issues of sampling bias. In our estimation there is little reason to suspect that a country's capacity to diagnose infection (or, similarly, their GDP, or primary research productivity) will bias intrinsic properties of host species (like litter size, body size, or age to sexual maturity). We try to make this point more clearly in lines 58-62 of the revised manuscript.

Out of curiosity, we did search for suitable proxies for "diagnostic capacity" among countries, but found that the specificity required for diagnostic tests for each infection, and the number of different kinds of diagnostic tests required for different pathogen types meant choosing one representative disease or diagnostic technology across countries would have introduced bias of a different kind (e.g., counting PCR machines would bias the dataset to include only infections that are diagnosable through PCR). In addition, the majority of countries in poorer regions of the world are served through regional centers that process samples pooled across multiple countries, so the geographical boundaries are less discernible.

REVIEWER 3:

Comments (Required):: […] for responding to my suggestions. I now feel more confident that the study was undertaken amongst critical thinking on the subject, albeit trying to generate hypotheses from existing data without a priori assumptions.

However, I believe that your introduction and discussion lacks punch to convince me why and where else we should use MBL to advance our understanding of wild reservoir hosts and emerging zoonoses including hot spots - if MBL helps to generate important hypotheses you should present these more convincingly.

We have added text throughout the manuscript to further highlight the utility of MBL for hypothesis generation. See the Introduction (lines 47, 62) and the Discussion (lines 123, 127).

Better structure may address this. You discuss what the results may mean at a host level (L181-187) and then there is speculative discussion of many issues before you start to discuss what hypotheses the results may raise at a geographic (including social, economic factors) level. That is, I find the discussion of the possible reasons for high latitude distribution of novel rodent reservoirs useful. It would be good to clarify whether 'human emerging infectious disease events' L225 means non-zoonotic EIDs. I presume it does and if so this points towards an interesting hypothesis.

I remain skeptical of the worth of seeking to demonstrate concurrence of your interesting and specific results with those of Jones et al 2008. Some of this makes for confusing reading. There are many differences in the data and assumptions of these two analyses. For example, the following statement is also inaccurate 'to some degree' :L202-204 "To some degree these hot spots coincide with regions of high mammal diversity where the risk of zoonotic disease emergence is estimated to be greatest (2)".

We have reorganized the results and discussion of this paragraph (lines 93-134) to present prominent patterns succinctly and moved up discussion of hypotheses. We clarify that "human emerging infectious disease events" refer to all EIDs (both non-zoonotic and zoonotic EIDs).

In this section we have also removed some of the text where we tried to place our results in the context of Jones et al. 2008, and deleted the sentence referred to above.

Cited

1. ***Magurran AE, McGill BJ eds. (2010) Biological Diversity: Frontiers in Measurement and Assessment.***

2. ***Haydon DT (2002) Identifying Reservoirs of Infection: A Conceptual and Practical Challenge. Emerg Infect Dis 8(12):1468–1473.***

RE: Rodent reservoirs of future zoonotic diseases

Dear Editors,

Thank you for the opportunity to submit a third revision. We hope that we have now fully addressed all of the reviewers' concerns (below, in blue text).

Thanks very much for your efforts in the communication of our work.

All the best,

Barbara Han

Reviewer Comments:

Reviewer #1:

Comments:

I have enjoyed seeing the progression of this paper to its current high standard. Your results generated are very interesting. In a subsequent study it would be very interesting to identify what the trait overlap is between host, agricultural and urban pest rodent spp., as identified by MBL. This may provide useful insight into reasons for increased (rodent-borne) zoonotic disease emergence and /or the factors which may tip your predicted rodent hosts into actual hosts. There are many potential uses for such information.

> ***A good idea - thanks for the suggestion and for the thoughtful reviews.***

Reviewer #2:

Comments:

This paper is very exciting and innovative. Yet. I remain to be convinced that sampling bias was adequately addressed.

Just because you are no longer comparing your results to those of Jones et al 2008 does not mean that the issue has been resolved. The areas that you predict to have more reservoirs are where human EID events are the most concentrated, yet the risk maps by Jones et al 2008 that take into account sampling bias are quite different. This point was not addressed.

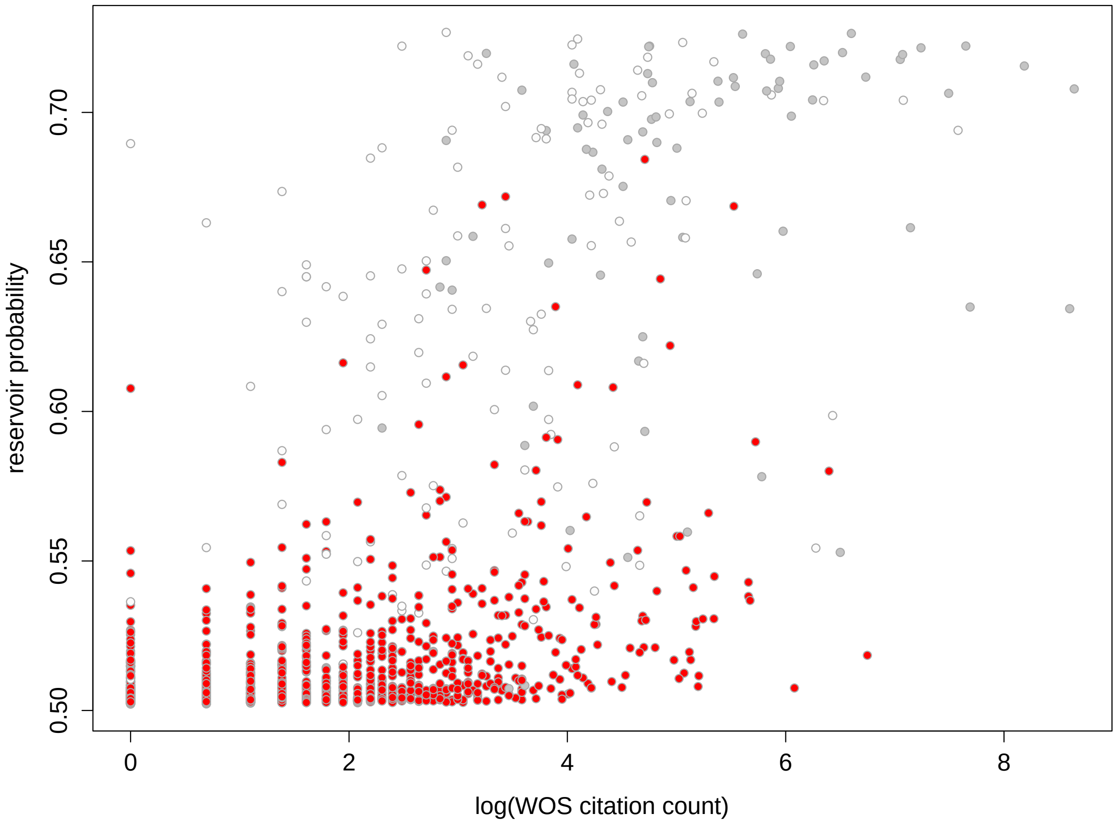
> ***The risk maps produced by Jones et al. 2008 (Figure 3) do not take into account reporting bias. The legend for Figure 3 states, "The relative risk is calculated from regression coefficients and variable values in Table 1 (omitting the variable measuring reporting effort), categorized by standard deviations from the mean and mapped on a linear scale from green (lower values) to red (higher values)." To be clear, their logistic regression models (presence/absence of human outbreaks per country) do incorporate reporting bias, but relative risk for the maps in Figure 3 is calculated from regressions that do not include reporting bias. Jones et al.***

> ***then compare global maps of reporting bias (reported in Supplementary Figure 3) to global maps of disease risk without reporting bias (reported in main Figure 3) to point out the geographical discrepancies between countries with high rates of disease reporting and those with predicted EID hotspots (see p.992, paragraph 2), and ultimately concluding that our surveillance resources are poorly allocated.***
>
> ***This aside, our Figure 2b and the Jones et al. Figure 3a are disparate given fundamental differences between studies: our study was specific to rodents recorded to carry zoonoses, while their calculations of relative risk were based on the number of human outbreaks caused by any zoonotic pathogen carried by any non-human animal (including reptiles, birds, and all mammal species). Thus, the maps depict very different things: Jones et al. Figure 3a is a global map of the relative risk of human zoonotic outbreaks, while our Figure 2b maps geographic ranges of predicted new rodent reservoirs of zoonotic disease.***

AUC scores tend to be higher than pseudo-R2. Predicting a yes/no is much easier than predicting a number. When you predict the number of zoonoses in Table S2 (Poisson reservoir status) the out-of-sample prediction is not much better than the out-of-sample study effort (10+citations), 0.21 vs 0.17.

With just a quick look at the data, I tallied up the number of citations on Web of Science for the top 12 species for both predicted new reservoirs and new hyper-reservoirs from table S4. (Also including appropriate synonyms for species, e.g. Arvicola amphibious = Arvicola terrestris.) Only a few of these species has <50 citations, showing that they are generally well studied species. The predicted hyper-reservoirs have about twice as many citations as the predicted reservoirs. There are a few predicted species with very few citations, but the overall picture doesn't convince me that citations aren't a big factor in these predictions. But this was just a quick look at the top 12. What would the bigger picture look like?

> ***To give a bigger picture of the relationship between citation count and reservoir status, we produced a scatterplot of the probability of being a reservoir (the output of the Bernoulli gbm) vs. the number of WOS citations, where each point represents a species and the color represents whether that species is known to carry 1 zoonosis (white), 2+ zoonoses (gray), or unknown reservoir status (red).***
>
> ***This plot (shown below) makes clear that while there is skew (there are more well-studied reservoirs and hyper-reservoirs), there are also several species that are well studied and not reservoirs, as well as unstudied species that are reservoirs.***



Another analysis that would be even more convincing: when you include citations (and it's one of the top predictors) do the trait profiles look the same as these results (still larger ranges, shorter age at sexual maturity & gestation length, larger litter size, etc)? This could go in the supplemental.

> ***Yes, when citation count is included, the trait profile of a zoonotic reservoir remains very similar except that citation count is the top predictor. The tuning parameters, pseudo-R2, and trait profile from this model have been added to Supplementary Table 2.***
>
> ***The trait profile and the corresponding partial dependence plots (now in a new Supplementary Figure 3) show that while citation count has the highest relative importance for correctly predicting the number of zoonoses carried by rodent species, the intrinsic traits which follow are consistent with our findings that rodent reservoirs have larger geographic ranges and are distinguished by a "fast-paced" life history strategy compared to non-reservoirs. Reservoirs tend to reach sexual maturity early and produce large litters more times per year, and the mean mass of offspring produced per year (normalized by adult body size; production (36)) is also greater. The results of this additional analysis have been added in lines 181-185.***

If this point was addressed, I would be fully supportive of publication in PNAS.