Mammalian biogeography and the Ebola virus in Africa

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Mammal Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>MAMMAL-15-57.R2</td>
</tr>
<tr>
<td>Manuscript Type</td>
<td>Review</td>
</tr>
<tr>
<td>Keywords:</td>
<td>zoonotic disease, tropical rainforests, Africa, chorotypes, favourability models</td>
</tr>
<tr>
<td>Subject Areas (select one):</td>
<td>Distribution</td>
</tr>
<tr>
<td>Mammalian Orders (select all that apply):</td>
<td>Chiroptera, Primates, Rodentia, Artiodactyla, Carnivora, Erinaceomorpha</td>
</tr>
</tbody>
</table>
Mammal Review Friday, 19 February 16

Mammalian biogeography and the Ebola virus in Africa
ABSTRACT

1. Ebola virus is responsible for the fatal Ebola virus disease (EVD).

2. Identifying the distribution area of the Ebola virus is crucial to understand risk factors conditioning the emergence of new EVD cases. Existing distribution models have underrepresented the potential contribution that hosts and vulnerable species make in sustaining the virus presence.

3. In this paper, we map favourable areas for Ebola virus in Africa according to environmental and zoogeographic descriptors, independent of human-to-human transmissions. We combine two different biogeographic approaches: analysis of mammalian distribution types (chorotypes), and distribution modelling of the Ebola virus.

4. We first obtain a model defining the distribution of environmentally favourable areas for the presence of Ebola virus. Based on a review of mammal taxa affected or suspected of exposure to the Ebola virus, we model favourable areas again, this time according to mammalian chorotypes. We then build a combined model in which both the environment and mammalian distributions explain the favourable areas for Ebola virus in the wild.

5. We demonstrate that mammalian biogeography contributes to explain the distribution of Ebola virus in Africa, although vegetation may also underscore clear limits to the presence of the virus. Our model suggests that the Ebola virus distribution may be even more widespread than previously suspected, given that additional favourable areas are found throughout the coastal areas of West and Central Africa, stretching from Cameroon to Guinea, and extend further East into the East African Lakes region.
Our findings show that the most favourable area for the Ebola virus is significantly associated with the presence of the virus in animals. Such core areas are surrounded by regions of intermediate favourability in which human infections of unknown source were found. This difference in association between human and animals and the virus may offer further insights on how EVD can spread.

INTRODUCTION

EVD is a zoonosis caused by filoviruses of the genus *Ebolavirus* (hereafter Ebola virus), of which 4 species (*Zaire ebolavirus*, *Sudan ebolavirus*, *Taï Forest ebolavirus* and *Budibugyo ebolavirus*) are known in Africa (Kuhn et al. 2010). These viruses cause often-fatal haemorrhagic fever upon infection in humans (Kuhn et al. 2011). Ebola virus transmission from wildlife has mostly been linked to people handling and butchering wild animals for bushmeat (Leroy et al. 2004a). Because bushmeat is important for human nutrition in Sub-Saharan Africa (Fa et al. 2015), it is fundamental to understand how host factors, together with ecological conditions and human behaviour contributes to Ebola virus outbreaks (Groseth et al. 2007). Recent biogeographical analyses have highlighted the importance of potential hosts in explaining the spatial assemblage of human infectious diseases worldwide (Murray et al. 2015). Biogeography has contributed broadly to questions of infectious disease ecology, management and surveillance (e.g. Cliff & Hagget 1995, Guernier et al. 2004, Smith & Guégan 2010). In this study, we build on this and propose a way to integrate virological, zoogeographical and environmental information through a combination of biogeographical approaches. Using these approaches we define the areas where Ebola virus may find suitable conditions to occur in the wild.
Despite methodological advances and increasing data availability, a limited understanding of the animals potentially implicated in the zoonosis has hampered mapping the extent of Ebola virus. Although likely hosts for the Ebola virus have been highlighted by some authors (Peterson et al. 2007), existing models describing the virus distribution have either not considered the contribution of hosts in sustaining the virus presence (Peterson et al. 2004), or have assumed that only a small number of species, suspected to be the reservoirs for the virus, are meaningful in the biogeography of the virus (Pigott et al. 2014). However, the ecology of the Ebola virus is complex and widely unresolved (Groseth et al. 2007). Thus, imposing restrictions to the selection of animal species considered in a distribution model might underrepresent the zoological substrate that could be determining the virus distribution. In fact, the role of particular bat species as true reservoirs of Ebola virus is still under discussion, and it is almost certain that there is a significant virus spillover among mammal species not suspected to be the natural hosts (Leroy et al. 2004a, 2004b, Groseth et al. 2007, Lahm et al. 2007, Olival et al. 2014).

Murray et al. (2015) suggested that mammalian biodiversity, as a whole, could be the strongest predictor explaining similarities between pathogeographic regions of the world. In this study, we analysed the spatial distribution of Ebola virus in Africa, independent of human-to-human transmissions, under the hypothesis that it is influenced by how mammal species are distributed throughout the region. Thus, instead of taking sides with an uncertain selection of probable natural hosts and victims, we tested the explanatory potential of the biogeographic patterns of mammals (involving species co-occurrence and thus potential interactions) in Africa. In order to test this hypothesis, our prediction was that a distribution model of Ebola virus, based on variables defining the existing types of mammalian distributions in Africa, should better describe the virus occurrences recorded in wildlife than...
a model based on environmental descriptors alone. In this paper, we examined all known literature regarding events of Ebola virus emergence, either EVD outbreaks or recorded presence of the virus in animals (Appendix S1 in Supporting Information). We then combined two different biogeographic approaches: the analysis of distribution types for mammals (chorotypes), and distribution modelling of Ebola virus, to develop a map of favourable areas for the virus according to both environmental conditions and mammal distributions in Africa.

METHODS

Identifying spatial links between Ebola virus and wildlife

Our model to define the favourable areas for Ebola virus was derived from occurrence data of any recorded presence of the virus in wildlife, disregarding whether it was detected in victims (i.e. organisms experiencing EVD symptoms, including human index cases) or in potential hosts (i.e. species without symptoms, in which the virus is harboured). From the literature, we found a total of 91 geo-referenced events (Appendix S1) in which the Ebola virus was transmitted to humans from wildlife or was present in other mammal species. These events were recognized either via the: 1) detection of viral antibodies or viral nucleic acid (n = 58 events, including 22 in animal carcasses), 2) observed abnormal increase of mortality in animal populations, associated with EVD outbreaks (n = 51 events), or 3) identification of index case of EVD in humans (n = 40 events) (note that the total number of occurrence records is 91 because recognition methods overlap). We found published evidence from cases of serological and/or PCR positivity of EVD in animals, or of animal EVD-linked mortality, in 28 mammal species: 10 primates, 3 rodents, 1 shrew, 8 bats, 1 carnivore and 5 ungulates (Fig 1; Appendix S1).
In 45 of the 58 laboratory-confirmed events of Ebola virus presence in wildlife, the virus was either isolated or was detected by polymerase chain reaction (PCR). In 11 out of the other 12 events, the virus was detected using enzyme-linked immunosorbent assay (ELISA) serology. The high sensitivity of this method, recommended for laboratory diagnosis of EVD by the WHO (2014), was demonstrated by Ksiazek et al. (1992, 1999). The remaining case was serologically identified by both immunofluorescence antibody test and western blot analysis (Hayman et al. 2010).

A large number of incidents of abnormally high animal mortality attributed to Ebola virus were reported since November 1994 (Formenty et al. 1999). These data are of high significance for the development of a model of Ebola virus distribution in wildlife; the detection of unusual animal mortality can uncover the propagation of Ebola virus (WHO 2003a, 2003b, Rouquet et al. 2005). Indeed, increased animal mortality preceded the first human cases in most EVD outbreaks that occurred in Gabon and the Republic of Congo between 2001 and 2003 (Leroy et al. 2004a). Of the 51 mortality events considered here, 19 cases were temporal and spatially linked to recorded EVD outbreaks in humans, and/or were confirmed by PCR. The remaining 32 cases represented extensions of disease from confirmed events (Leroy et al. 2004a, Rouquet et al. 2005, Caillud et al. 2006, Lahm et al. 2007).

Only one of the 40 index cases of haemorrhagic fever syndrome considered in humans was not confirmed to be EVD through laboratory testing of clinical samples. This episode took place in Olloba (Republic of the Congo) in May 2002, and then sample collection was not possible (WHO 2002; Rouquet et al. 2005). However, a previous EVD index case in Olloba had been confirmed just five months before (Leroy et al. 2004a).
Environmental model of Ebola virus

Individual data points were represented as 5-km buffers around the original virus occurrence locations, and we overlapped these buffers to a 1°×1°-resolution grid covering the whole of Africa (n = 2,547 grid cells). With this buffer approach, we aimed to provide the occurrence-allocation in grid cells with some flexibility. In central Africa, 5 km is less than 0.05 times the side-length of a 1°×1° grid cell; given that the Ebola virus is hosted by animals, being less than 5-km far from a cell border is virtually equivalent to being located at the limits of two cells. The spatial resolution employed prevented autocorrelation that could exist as a consequence of spatial dependence among very close (< 1°) observations (Legendre & Legendre 1998). The final number of grid cells for detected presence of Ebola virus was 40, of which 33 corresponded to Zaire ebolavirus (Fig 1). From these, an environmental model of Ebola virus (i.e the distribution of environmentally favourable areas for the presence of Ebola virus) was made by employing the Favourability Function (Real et al. 2006, Acevedo & Real 2012):

\[
F = \frac{P}{1-P} \left( \frac{n_1}{n_0} + \frac{P}{1-P} \right)
\]

where \( F \) is environmental favourability (0-1), \( P \) is probability of occurrence, and \( n_1 \) and \( n_0 \) are presence and absence numbers, respectively (absences were considered to be those squares not included in the presences subset). \( P \) was calculated for the entire African continent through a forward-stepwise logistic regression, according to 24 predictor variables describing types of ecosystems, abiotic factors and anthropogenic pressures on wildlife (see variable descriptions and sources in Appendix S2 in Supporting Information). Land-cover variables were computed as cover percentages in every grid, and the rest of the variables were estimated by averaged grid-values. The logistic regression was based on the 40 Ebola virus
occurrence squares (= $n_1$), considering the four Ebola virus species together. The cases of serological or PCR positivity of EVD, of animal mortality related with EVD, and of index cases of EVD in humans were not differentiated in the occurrence data employed for model building—events of abnormal animal mortality not directly confirmed in laboratory only contributed to 5 presences, all of which were adjacent to the main core of EVD outbreaks in Gabon and Republic of Congo. Models based on the Favourability Function distinguish those localities with environmental conditions that favour the species’ existence from those with detrimental characteristics for its presence, irrespective of the species’ proportion of occurrences within the study area. This property is essential for our objectives, as it enables direct comparison between models when several species are involved in the analytical design, and allows for model combinations through fuzzy logic (Barbosa & Real 2012). Fuzzy logic provides metrics for fuzzy sets, i.e. classes of objects with a continuum of membership degrees; these sets are characterized by a membership function that assigns to each object a real number in the interval [0, 1] (Zadeh 1965). In biogeography, environmental favourability, chorotypes, biotic boundaries and species richness can be identified as fuzzy sets (Estrada et al. 2008, Olivero et al. 2011, 2013, Romero et al. in press), thus fuzzy logic can be easily applied to these concepts (see below).

Adding a spatial descriptor to this variable set allowed us to consider autocorrelation resulting from the spatial structure of the virus distribution (Sokal & Oden 1978), and so take into account the impact of dispersal barriers, geological history and biotic interactions as potential predictors (Fa et al. 2014). To this end, we followed the "trend surface approach" (Legendre & Legendre 1998). Thus, a series of spatial variables resulting from average X and average Y combinations for every square of the grid were examined through a backward
stepwise logistic regression. Then we used as spatial descriptor the lineal combination (logit) of spatial variables resulting from the logistic regression.

In order to control Type-I errors caused by the large number of variables employed in the process, we used Benjamini & Hochberg's (1995) False Discovery Rate (FDR). This control was performed before the stepwise variable selection, so that only significant variables under an FDR of \( q < 0.05 \) were accepted in a multivariate environmental model. To minimise multicollinearity, we also avoided Pearson correlation values higher than 0.8 between predictor variables.

Zoogeographic model of Ebola virus

We defined a zoogeographic model of Ebola virus as the distribution of favourable areas for the presence of Ebola virus according to mammalian chorotypes. A “chorotype” is a distribution pattern followed by one or several species, which can be operatively recognized within an area (Baroni-Urbani et al. 1978, Real et al. 2008); chorotypes, or types of distribution, represent the shared geographical, ecological and evolutionary context of several species (Real et al. 2008). The zoogeographic model of Ebola virus was produced in three steps: (1) we generated a list of mammal species to be considered; (2) we then defined the mammalian chorotypes with potential for explaining the Ebola virus range in Africa; and (3) we finally built a model sustained on a set of predictor variables based on the chorotypes.

Selection of mammal species to be considered

Two different criteria were used to select the list of mammal species based on the a-priori analysis of probable (sensu lato) links between mammals and Ebola virus: (1) taxonomic proximity; and (2) biogeographic coincidence. Our aim here was to describe
mammalian distribution patterns with enough potential to explain the distribution of Ebola virus; this does not mean that the species selected are proposed to be Ebola virus reservoirs or susceptible taxa.

Taxonomically related species might be susceptible to similar pathogens (Plyusnin & Morzunov 2001). Using the 28 mammal species recorded to be linked to the Ebola virus (see Appendix S1), we then generated a preliminary list of 216 species (Appendix S3), by considering congenerics: Mus, Praomys, Sylvisorex, Tadarida; and whole suprageneric groups if they were represented by more than 5% of their species (a lower proportion was considered to be anecdotic) among the probable host and susceptible species: Pteropodidae, Hystricidae, Hominidae, Cercopithecidae, Viverridae and Suidae families, and forest tribes of the family Bovidae. We also included pouched rats (Cricetomys) and grasscutters (Thryonomys) in our preliminary list, despite there is no current evidence of relationship with the Ebola virus. We considered that, because members of these rodent genera are so numerically important in the bushmeat trade in central and West Africa (Alexander et al. 2014), their inclusion in the analysis was warranted. Thus, our taxonomic criterion was defined so that it discarded groups very far from any suspect hosts (e.g. proboscideans, equids or felines), whilst it considered all those taxa including species whose membership of the Ebola virus host system cannot be categorically denied.

We then preselected, from our preliminary list of 216 species, those whose biogeography was potentially able to explain the distribution of Ebola virus. To this end, those species geographically coinciding with or nested within the Ebola virus range were chosen. Given that the area covered by the Ebola virus is poorly known, we estimated the geographical coincidence between the virus and mammalian distributions according to models defining the
spatial structure of their respective ranges. Thus, for every mammal species, and for Ebola virus, we built spatial models that described their distributions according to purely spatial variables (i.e. combinations of average X and average Y). These models were generated employing the Favourability Function (see above). For mammals, presences/absences were derived from polygon shapefiles for each species available from the IUCN website (IUCN 2012, compiled or modified in 2008) projected to a 1º×1º grid system, which is the maximum spatial resolution at which “extent-of-occurrence” range maps (such as those provided by IUCN) are suitable for analysis (Hurlbert & Jetz 2007). The spatial model of Ebola virus was based on the same presence/absence data set considered for the environmental model. The degree of geographic coincidence was quantified through the fuzzy overlap (or fuzzy similarity) index, whilst the degree of nesting was measured using the fuzzy inclusion index (Dubois & Prade 1980; Olivero et al. 2011); both measures ranged from 0 to 1. The mathematic formulation of the fuzzy overlap (i.e. intersection divided by union) is equivalent to Jaccard’s similarity index (Olivero et al. 2011), whose theoretical distribution is randomly distributed around 0.33 (Baroni-Urbani 1980). Thus, we considered a geographic coincidence to be significant when the fuzzy overlap was > 0.33 (i.e. when both Ebola virus and the mammal had more than one third of their ranges in common). There is no theoretical reference for estimating significance of fuzzy inclusion values hence, we considered that a mammal distribution was nested within that of Ebola virus when at least two thirds of the mammal range was included within the Ebola virus range (i.e. fuzzy inclusion > 0.66). A final list of 96 species was ultimately considered for entry in the zoogeographic model of Ebola virus (Appendix S3), 89 of which fulfilled the two selection criteria. Another 7 species were considered because there is published record of probable infection by Ebola virus.
**Defining and mapping mammalian chorotypes**

The distributions of the 96 above-selected species were classified hierarchically according to the Baroni-Urbani & Buser (1976) similarity index, using the UPGMA agglomerative algorithm (Sneath & Sokal 1973). All clusters in the resulting classification dendrogram were assessed for statistical significance with the method in Olivero et al. (2011), using the RMacoqui 1.0 software (http://rmacoqui.r-forge.r-project.org/). The number of resulting chorotypes was not predefined; all groups of distributions that were significantly clustered were considered chorotypes.

The resulting chorotypes were mapped following the accumulated favourability approach as in Fa et al. (2014). The accumulated favourability is considered a fuzzy-logic method for estimating species richness (Estrada et al. 2008). For all species forming part of the same chorotype, we built an environmental model using the method shown above for Ebola virus; then, in every grid cell of the study area, we added the favourability values defined by these environmental models. So, a cell showing high-accumulated favourability for the species of a chorotype is defined to have favourable conditions for the presence of a large number of these species. Mapping chorotypes this way allowed for further downscaling to a higher spatial resolution (see below).

**Building the zoogeographic model of Ebola virus**

To build the zoogeographic model of Ebola virus, we followed the same procedure as for the environmental model. In this case, the model was built using every chorotype as a predictor variable. The accumulated favourability for the species of a chorotype was employed as a variable.
Although every chorotype was based on a finite cluster of species (i.e. the chorotypical cluster), the distribution of every species has a certain degree of membership in all detected chorotypes (Olivero et al. 2011). We proffered a biogeographically-justified list of mammal species whose link with Ebola virus is worth investigating, based on a species membership degree > 0.5 in at least one of the chorotypes entered in the zoogeographic model of Ebola virus. The degree of membership of every species in each chorotype was calculated as the average of the Baroni-Urbani & Buser similarities between a species and all distributions in the chorotypical cluster. The theoretical distribution of this similarity index is randomly distributed around 0.5 (Baroni-Urbani & Buser 1976); we considered membership to be significant above this value.

Environmental/zoogeographic model assessment and combination

Both the environmental and zoogeographic models were assessed using calibration (Hosmer & Lemeshow 2000), i.e. testing whether favourability values reflected the existing observations of Ebola virus presence in wildlife. We also compared the performance of both models in relation to goodness of fit using -2×log-likelihood; discrimination capacity using the Area Under the receiver-operating-characteristic Curve (AUC) (Lobo et al. 2008); and classification capacity using sensitivity, specificity, correct classification rate (CCR), Cohen's Kappa (Fielding & Bell 1997), and under- and overprediction rates (Barbosa et al. 2013). Classification measures were based on the 0.5 favourability threshold because probability is equal to the overall prevalence at this level (Acevedo & Real 2012). In a calibrated model, Hosmer-Lemeshow index should be non-significant; for higher goodness of fit, -2×log-likelihood should be lower; for better discrimination, AUC should be higher; for better classification, sensitivity, specificity, CCR and Kappa should be higher whereas under-prediction and over-prediction should be lower.
The environmental and the zoogeographic models of Ebola virus were combined so that the final model gave favourability values based on the degree to which conditions are both environmentally and zoogeographically favourable for Ebola virus presence. To this end, we used the fuzzy intersection between the environmental model and the zoogeographic model of Ebola virus (Romero et al. in press), by measuring the minimum favourability value for either of the two models in each grid cell.

We analysed the relative contribution of environment and zoogeography to the combined model. To do this, we mapped where favourability values were derived from either the environmental or the zoogeographic model. We, thus, identified where, and to what extent, one of these models acted as a limiting factor whilst the other model showed a higher favourability for presence of Ebola virus. Finally, we employed the sensitivity index to assess the capacity of the combined model to classify recorded presences of Ebola virus and of the four African Ebola virus species separately.

We compared the distribution of known EVD outbreaks in humans with our resulting favourability maps for the Ebola virus. For this purpose, we overlapped the locations of index cases in humans with a favourability map divided into three regions depending on their favourability values, according to the thresholds proposed by Muñoz & Real (2006). If the predicted favourability was higher than 0.8, which means that the odds are more than 4:1 favourable to Ebola virus, the square was considered as highly favourable. Those areas with a favourability value lower than 0.2 (odds less than 1:4) were considered of low favourability for Ebola virus. The remaining squares were regarded as intermediate favourability areas.
The potential for contacts between human populations and wildlife could have influenced detection of Ebola virus infections beyond environmental and zoogeographic factors that could favour the presence of Ebola virus. Therefore, we tested whether Ebola virus occurrences poorly explained by the combined model of Ebola virus (i.e. not explained by environmental conditions nor by mammalian biogeography) could be accounted for by the presence of human populations. To do this, we performed a logistic regression of the 40 Ebola virus occurrence squares, using favourability values for Ebola virus as the independent variable. The residuals of this regression were then related to rural population density and with distance to roads using linear regression (for variable sources, see Appendix S2).

To increase the potential uses or our output for management, surveillance or analyses requiring an ecological context for Ebola virus, the combined model was downscaled to 0.1°×0.1° resolution squares, by employing the "direct downscaling approach" (Bombi & d’Amen 2012). This method was classified by Bierkens et al. (2000) as “downscaling based on mechanistic models through a deterministic [favourability] function”. To do so, we applied the favourability equations involved in the combined model to predictor variables at this resolution. A 10-fold shortening of the grain size (referring to pixel side length) does not severely affect predictions of species distributions (Bombi & d’Amen 2012).

RESULTS

The environmental model of Ebola virus

Three variables (terra-firme rain forests, natural vegetation/cropland mosaics and small annual temperature range) were significantly associated with the areas of high environmental favourability for Ebola virus presence ($P < 0.01$) (Figs 1 and 2). Pearson correlation values between these variables were lower than 0.51. Vegetation/cropland mosaics and constant
temperatures complemented the predictive power of forests, especially within deforested areas. Crucially, the large swamp forest areas along the lower course of the Congo River ("cuvette congolaise", where EVD outbreaks have so far not been recorded) did not appear as favourable as other terra firme central and West African forest areas (Fig 2).

359

The zoogeographic model of Ebola virus

Our analyses revealed the presence of 16 significant types of distributions, or chorotypes (Fig 3). Four of these chorotypes significantly supported the zoogeographic model of Ebola virus (P < 0.001): Rainforest chorotype (RF), West-African Forest chorotype (WAF), North-Western Congolian Forest chorotype (NWCF), and the single-species chorotype of Mus goundae (MG) (Fig 3). The zoogeographic model of Ebola virus, based on these chorotypes, showed high favourability values within the rain forests of Central and West Africa, with a decline in favourability towards the East, dropping even more dramatically South of the Congo River.

369

Having identified the four chorotypes that, according to the zoogeographic model, favour Ebola virus presence, we propose a list of 64 mammal species as a guide for future investigations in the search for Ebola virus hosts and potentially susceptible species (see Fig 3; Appendix S3).

374

Environmental/zoogeographic model assessment and combination

Both the environmental and zoogeographic models appear significantly well calibrated (Table 1). The zoogeographic model shows a better goodness of fit, higher discrimination and greater classification power than the environmental model. However, both provide significant complementary information about the virus distribution. Thus, our environmental model
suggests that waterlogged areas (swamps and swamp forests) could limit the presence of Ebola virus (Fig 2). Vegetation/cropland mosaics away from the rain forest block appeared unfavourable in the zoogeographic model, the only exception being a small area of the northern Central African Republic savannas; this area corresponds with the discovery of Ebola virus genetic sequences in rodents in gallery forests around Bohou River, 490 km North of Bangui (Morvan et al. 1999).

Zoogeography was the most extended limiting factor for the presence of Ebola virus, covering 75.6% of the African continent; instead, environment was the limiting factor in the remaining 34.4%. These percentages turned into 55.2% and 44.8%, respectively, if completely unfavourable areas for Ebola virus (favourability < 0.05) were excluded; and they became 47.7% and 52.3% within the intermediate or high favourability areas (favourability > 0.2) (see Appendix S4 in Supporting Information). The combined model inherited, from the zoogeographic model, the eastward decline of favourability values (Fig 2), where zoogeographic favourability was lower than environmental favourability (Appendix S4). Instead, the combined model excluded swamp forests as Ebola virus favourable areas as a legacy of the environmental model.

Due to the sensitivity of the combined model, more than 92% of the 1° x 1° squares with records of the Ebola virus (n = 40) were classified correctly in areas with favourability values higher than 0.5. The combined model explained 94% of the Zaire ebolavirus presences (n = 33), 100% of the Budibugyo ebolavirus presences (n = 2), 75% of the Sudan ebolavirus presences (n = 4), and 100% of the Tai Forest ebolavirus presences (n = 1). Up to 62.5% of Ebola virus presences appeared within the highly favourable areas (i.e. value > 0.8) (Fig 4a), largely coinciding with non-flooded rain forests. The 0.8-favourability threshold significantly
separated virus occurrences in humans from presences in other mammals (Fig 4a); the highly favourable region included a significantly higher proportion of presences in non-human mammals ($\chi^2_1 = 6.22, P < 0.05$), as well as in both humans and non-human mammals ($\chi^2_1 = 8.00, P < 0.01$). In contrast, presences recorded only in humans were significantly located within the intermediate favourability areas (i.e. value between 0.2 and 0.8) ($\chi^2_1 = 19.16, P < 0.001$). This regionalization reveals an emergent property that arises from the combination of the environmental and zoogeographic models. Residual analyses (to explore the differences between predicted and observed probabilities of presence) indicate that there is a significant positive correlation between Ebola virus points recorded outside the highly favourable region and both rural population density ($r = 0.589, P = 0.001$) and distance to roads ($r = -0.466, P = 0.014$). A downscaled geographical representation of the combined model, at a $0.1^\circ \times 0.1^\circ$ spatial resolution, is shown in Fig 4b.

A total of 17 countries contained high favourability areas (> 0.8) for Ebola virus (Fig 4b). In decreasing order, according to the extent of highly favourable areas in each, these were: Democratic Republic of the Congo, Gabon, Cameroon, Republic of Congo, Côte d’Ivoire, Liberia, Ghana, Central African Republic, Nigeria, Equatorial Guinea, Sierra Leone, Uganda, Benin, Angola, Tanzania, Guinea and Togo. Another four countries comprised areas of intermediate favourability (0.2 – 0.8): Rwanda, Burundi, South Sudan and Kenya.

**DISCUSSION**

Spatial modelling, applied to pathogen systems, can generate statistically robust predictions of the geographic distributions of the organism causing a disease, and its maintenance host(s) (Peterson 2006, Purse & Golding 2015). However, species distribution modelling applied to EVD transmission in Africa has so far underestimated the contribution
hosts and vulnerable species may make in sustaining the virus presence (Peterson et al. 2004, Pigott et al. 2014) (see above). Our models are the first to analyse the contribution of mammalian distribution patterns to the biogeography of Ebola virus in Africa. In our study, we considered the four African Ebola virus species together because of the scarcity of information available on the presence of species other than Zaire ebolavirus. The latter is the most lethal (Kuhn et al. 2011), was responsible for the EVD outbreaks in central Africa (Bausch & Schward 2014), and has also been associated with the current epidemic in West Africa (Dudas & Rambaut 2014). Our models have very high sensitivity (0.94) and specificity (0.90) to the presence of Zaire ebolavirus, but high favourability areas also adequately predicted the presence of Budibugyo ebolavirus, and of Taï Forest ebolavirus (sensitivity = 1). However, our models poorly predicted Sudan ebolavirus. With better knowledge of the distribution of all Ebola virus species more robust models could have been developed. In particular, models focused on the different Ebola species should be developed promptly, since some mammalian chorotypes that were not entered in the model could better explain the presence of these Ebola virus species e.g. chorotype NECF in the case of Budibugyo ebolavirus (compare Figs. 1 and 3).

Peterson et al. (2007) identified 55 groups of mammalian taxa of interest as potential reservoirs, based on a detailed inspection of distributional overlap patterns with known filovirus disease outbreaks. Our list of 68 species biogeographically-linked to Ebola virus differs from the list of likely reservoir species presented by Peterson et al. (2007). These authors generated an inventory of candidate reservoir species, assuming that the reservoir supports persistent, largely asymptomatic Ebola virus infections, which de facto eliminates most primates and ungulates. Our list, unlike Peterson’s, is 33% primates (families Hominidae and Cercopithecidae) and 20% ungulates (families Suidae and Bovidae).
Members of the fruit bat family (Pteropodidae), Molossid bats, viverrid carnivores, as well as the rodent *Praomys jacksoni* are included in both lists. In our list of species, however, we also included other fruit bats (*Casinycteris argynnis, Scotonycteris ophiodon* and *S. zenkeri*) as members of the chorotypes significantly linked to the Ebola virus distribution. Of the two rodent genera *Cricetomys* and *Thryonomys* included in our analyses given their prominence in the bushmeat trade, *C. emini* emerged as worth investigating. Thus, unlike Peterson's, our list is not a proposal of potential reservoir species; instead, we present a list of mammals biogeographically associated with Ebola virus, and thus with the geographic potential of being involved in the virus cycle as reservoirs, hosts or susceptible species. This is not a closed list, even though we have restricted the analysis to a limited number of species. This is because any species (i.e. also those ones not considered in the analysis) are members in all chorotypes to a certain degree (Olivero et al. 2011). Thus, it is possible to evaluate whether unassessed taxa can be considered part of the zoogeographic factor as defined in the Ebola virus model.

Of the 28 mammal species with published record of probable contact with Ebola virus, only 8 were not included in our 64-species list: *Rousettus aegyptiacus, Eidolon helvum, Epomophorus gambianus, Micropteropus pusillus, Tadarida condylura, Civettictis civetta, Papio anubis* and *Mandrillus leucophaeus* (Appendix S3). What these species have in common is that a large proportion of their populations are distant from where Ebola virus has been detected. We do not challenge the evidenced relationship of these species with Ebola virus, but note that they do not show strong overlap with it (even though there is overlap in some parts of the range). It could be argued that subspecific taxa of the same species may differ in their ability to serve as Ebola virus hosts. In the case of these 8 mammal species, however, all subspecies currently recognized (Kingdon et al. 2013) are partially distributed
outside the Ebola virus range, with the only exception of *T. c. osborni*. These species might be more interesting from a phylogeographic perspective, but their distributions cannot explain the geographic distribution of the African Ebola viruses in wildlife; the virus might potentially be hosted by individuals of these species, but factors other than the distribution of these hosts could explain the observed geographic limits of the Ebola virus in Africa (e.g. other hosts or environmental conditions).

We show, as Pigott et al. (2014) clearly indicated, that the high relative contribution of vegetation in the model might underscore clear limits to the presence of Ebola virus. As in the Pigott et al. (2014) and Murray et al. (2015) maps (and less patently in Peterson et al.’s 2004), we confirm West Africa as a highly favourable region for Ebola virus. Our map also suggests that the virus distribution may be even more widespread than previously suspected, given that additional favourable areas for the virus are found throughout the coastal areas of West and Central Africa, stretching from Cameroon to Guinea, and extending further East into the East African Lakes region (i.e. Uganda). Of the 17 countries with highly favourable areas (> 0.8) for Ebola virus occurrence in wildlife in our model, 16 of these are among the top 17 considered as high-risk from Ebola by both Pigott et al. (2014) or Murray et al. (2016)—10 of these common to both sources. The Democratic Republic of the Congo, Gabon, Republic of Congo, Côte d’Ivoire, Uganda and Guinea have a record of EVD index cases in humans, whereas Cameroon, Liberia, Central African Republic, Nigeria, Angola, Ghana, Sierra Leone, Benin, Tanzania and Togo have not. Identifying ‘at risk’ countries that are currently Ebola virus-free, according to various different evidence types, could be fundamental in achieving adequate levels of preparedness. For example, countries such as Equatorial Guinea, with 94% of its territory highly favourable for Ebola virus in our model, though ranked number 20 by Pigott et al. (2014), should not escape attention. Similarly,
Burundi, Rwanda and Kenya, containing (like South Sudan) areas of intermediate favourability — covering all of Burundi and Rwanda — but so far free from EVD cases, should be targeted for further investigation.

In our map, the favourability for the disease in areas south of the Congo River is remarkably lower than in other rainforest areas; a pattern not apparent in the Pigott map. A possible explanation may be a lack of adequate zoogeographic conditions for the virus to flourish linked to a lower diversity of potential host species in this area. There is evidence that the richness and population structure of mammals found in the central part of the Congo basin is considerably lower than in areas along the western, northern and eastern side of the Congo River (Fa et al. 2014). The limiting role of the zoogeographic factor south to the Congo River (Appendix S4) supports this possibility, but we cannot discard the chance of observation bias in a lowly accessible area as the central Congo.

Serological surveys among human populations support a rainforest origin of the potential reservoir of Ebola virus (Gonzalez et al. 1989, 2000). Living or spending significant time deep in the forest is positively correlated with seropositivity in Pygmies and in other human groups (Gonzalez et al. 2000, Becquart et al. 2010, Schoepp et al. 2014). This is consistent with our finding of a highest favourability for Ebola virus in the rainforest, and with the recorded concurrence of cases of spillover from animals to humans there even though this information was not used to build the model. However, epidemic outbreaks also originated outside of the rain forest block. This is the case for the first known outbreaks in humans, recorded in Zaire and Sudan in 1976 (WHO 1976, 1978a), of the last and most important episode in West Africa (Bausch & Schward 2014), and of some other cases in Uganda (Albariño et al. 2013), Congo (Pourrut et al. 2005), the Democratic Republic of the Congo.
(CDC 2009) and Sudan (Onyango et al. 2007). All of them have in common that the human patient zero were not documented to be related to cases of EVD in animals, and that they occurred in areas of intermediate favourability for Ebola virus according to our model. Thus, a main contribution of our model is the finding of a geographical context for exploring which mechanisms differentiate Ebola virus transmission to humans in highly favourable areas compared to areas of intermediate favourability for Ebola virus is of high epidemiological relevance. In the latter areas, gallery forests may play a significant role in harbouring and spreading the Ebola virus.

EVD epidemics seem to be related more with human behaviour, which increases the risk of contact with the reservoir, than with the emergence of a highly pathogenic viral strain (Gonzalez et al. 2000). The significant relation between human presence (i.e. population density and roads) and EVD cases in areas only moderately favourable to Ebola virus suggests that EVD transmission to humans in these areas could be influenced by anthropogenic factors. A more frequent contact between humans and forest fauna would amplify the chances of Ebola virus transmission to humans, even where environmental or zoogeographic conditions are not the most favourable for the virus. About half of the EVD index cases reported in the region of intermediate favourability happened in the limits of highly favourable areas (Fig 4), which could have facilitated contacts. Even so, some index cases have been located away from the forest domain. One possible explanation for the arrival of *Zaire Ebolavirus* in West Africa, during the 2013 Guinean outbreak, far from its usual haunts in central Africa, is that the virus was introduced by traveller bats (Bausch & Schward 2014). Sáez et al. (2015) also suggested that the index case in Guinea may have been infected by a colony of insectivorous free-tailed bats (*Mops condylurus*). Do Chiropterans have a more significant role in zoonotic spillover in the intermediate-
favourability areas, compared with the deep rainforest haunts of Ebola virus? In the first 
documented case of EVD in Sudan in 1976, the index case was located (by the World Health 
Organization) in a cotton factory far from the forest block, where the only wild significantly 
abundant species was an insectivorous bat species (WHO 1978b). Additional analysis and 
study of the human-wildlife interface are still required to delineate areas in which human 
populations are at-risk of zoonotic transmission of Ebola virus. In this task, our model could 
be used to delineate the geographic context of the analysis. As risk factors for zoonotic 
transmission of Ebola virus are better understood, it will be possible to incorporate this 
information into future risk mapping assessments and to develop mitigation or prevention 
measures.

ACKNOWLEDGEMENTS
This study was supported by USAID as part of the Bushmeat Research Initiative of the 
CGIAR research program on Forests, Trees and Agroforestry. We thank Dr. Kris Murray for 
his valuable help in improving the manuscript.

REFERENCES
Acevedo P, Real R (2012) Favourability: Concept, distinctive characteristics and potential 
Albariño CG, Shoemaker T, Khristova ML, Wamala JF, Muyembe JJ, Balinandi S, 
Tumusiime A, Campbell S, Cannon D, Gibbons A, Bergeron E, Bird B, Dodd K, 
Spiropoulou C, Erickson BR, Guerrero L, Knust B, Nichol ST, Rollin PE, Ströher U 
(2013) Genomic analysis of filoviruses associated with four viral hemorrhagic fever


virus disease in Africa. eLife 2014; 10.7554/eLife.04395. doi:
http://dx.doi.org/10.7554/eLife.04395.

Plyusnin A, Morzunov S (2001) Virus evolution and genetic diversity of hantaviruses and
their rodent hosts. Current Topics in Microbiology and Immunology 256: 47-75.

Pourrut X, Kumulungui B, Wittmann T, Moussavou G, Délicat A, Yaba P, Nkoghe D,
and Infection 7: 1005-1014.

Purse BV, Golding N (2015) Tracking the distribution and impacts of diseases with
biological records and distribution modelling. Biological Journal of the Linnean Society
115: 664-677.

Real R, Barbosa AM, Vargas JM (2006) Obtaining environmental favourability functions

Real R, Olivero J, Vargas JM (2008) Using chorotypes to deconstruct biogeographical and
biodiversity patterns: the case of breeding waterbirds in Europe. Global Ecology and
Biogeography 17: 735-746.

Romero D, Olivero J, Brito JC, Real R (in press) Comparison of approaches to combine
species distribution models based on different sets of predictors. Ecography In press; doi:
10.1111/ecog.01477.


Sáez AM, Weiss S, Nowak K, Lapeyre V, Zimmermann F, Düx A, Kühl H, Kaba M, Regnaut
S, Merkel K, Sachse A, Thiesen U, Villányi L, Boesch C, Dabrowski PW, Radonić A,
Nitsche A, Leendertz SAJ, Pettersson S, Becker S, Krähling B, Couacy-Hymann E,
Akoua-Koffi C, Weber N, Schaade L, Fahr J, Borchert M, Gogarten JF, Calvignac-


Fig 1. Animal species for which there is a record of naturally occurring Ebola virus infection either from serological or PCR positivity, or from increased mortality attributed to EVD. Top map: localities with a record of Ebola virus presence in wildlife. Bottom map: 1° x 1° squares with a record of Ebola virus presence (outlined in white); spatial model of Ebola virus (spatial favourability increases from white to black). Maps were generated using ArcGIS.

Fig 2. Modelling of the environmental/zoogeographic favourability for the presence of Ebola virus in wildlife. The environmental model is based on Terra-Firme Rain Forests (TFRF), Natural Vegetation/Cropland Mosaics (NVCM) and Annual Temperature Range (ATR, with increasing values from yellow to red). The zoogeographic model is composed by four types of mammalian distributions or chorotypes (see Fig 3). The two models are combined according to fuzzy-logic, requiring both environmentally and zoogeographically favourable conditions. Maps were generated using ArcGIS.

Fig 3. Classification of 96 mammal species according to distributional similarities (S = Baroni-Urbani & Buser similarity index). *: Branches defining significant chorotypes (P < 0.05). For every chorotype, accumulated favourability values (increasing from white to black) of all species are mapped. Acronyms name chorotypes. NWCF: North-Western Congolian Forest; SECF: South-Eastern Congolian Forest; NCF: Northern Congolian Forest; SCF: Southern Congolian Forest; CSBF: Cross-Sanaga-Bioko Forest; RF: Rain Forest; WAF: West-African Forest; SS: Sub-Saharan; RA: Rousettus aegyptiacus; TB: Tadarida bemmelen; MO: Mus oubangui; AS: Allochrocebus solatus; CD: Cercopithecus dryas; MG: Mus goundae; PM: Praomys mutoni; SK: Sylvisorex konganensis. Solid circles indicate the 64 mammals that are significant members (degree > 0.5) in chorotypes explaining the
zoogeographic model of Ebola virus (RF, WAF, NWCF and MG). Red circles indicate published records of probable contact with the Ebola virus. Maps were generated using ArcGIS.

Fig 4. **Combination of the environmental and the zoogeographic models of Ebola virus at the original 1° × 1° resolution (A) and downscaled to a 0.1° × 0.1° resolution.** Red tones distinguish regions with high (value > 0.8), intermediate (0.2-0.8), and low favourability (< 0.2) for the presence of Ebola virus. White areas are completely unfavourable for Ebola virus (< 0.05). Circles represent squares with cases of serological or PCR positivity of Ebola virus infection in animals, animal mortality attributed to EVD, and zoonotic transmission to humans. G: Guinea; SL: Sierra Leone; L: Liberia; CI: Cote d'Ivoire; GH: Ghana; T: Togo; Be: Benin; N: Nigeria; CA: Cameroon; CAR: Central African Republic; S: South Sudan; U: Uganda; R: Rwanda; B: Burundi; DRC: Democratic Republic of Congo; A: Angola; CO: Congo; GA: Gabon; EG: Equatorial Guinea. Maps were generated using ArcGIS.
Table 1. Model assessment and comparison based on calibration, goodness of fit, and discrimination and classification capacities. In a calibrated model, H-L should be non-significant; for higher goodness of fit, -2logL should be lower; for better discrimination, AUC should be higher; for better classification, sensitivity, specificity, CCR and Kappa should be higher whereas under-prediction and over-prediction should be lower.

<table>
<thead>
<tr>
<th>Indices</th>
<th>Environmental</th>
<th>Zoogeographic</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-L*</td>
<td>$\chi^2 = 13.02, P &gt; 0.05$</td>
<td>$\chi^2 = 7.03, P &gt; 0.05$</td>
</tr>
<tr>
<td>-2logL**</td>
<td>226.26</td>
<td>225.48</td>
</tr>
<tr>
<td>AUC***</td>
<td>0.943</td>
<td>0.965</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.925</td>
<td>0.975</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.885</td>
<td>0.889</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.180</td>
<td>0.196</td>
</tr>
<tr>
<td>CCR****</td>
<td>0.886</td>
<td>0.890</td>
</tr>
<tr>
<td>Under-prediction</td>
<td>0.00135</td>
<td>0.00045</td>
</tr>
<tr>
<td>Over-prediction</td>
<td>0.886</td>
<td>0.877</td>
</tr>
</tbody>
</table>

* Hosmer-Lemeshow calibration index
** -2 x ln(Likelihood)
*** Area Under the receiver-operating-characteristic (ROC) Curve
**** Correct Classification Rate
Supporting Information - online

- Appendix S1. Referenced locations of Ebola virus presence in wildlife.
- Appendix S2. Predictor-variable description and sources.
- Appendix S3. List and features of the 216 mammal species considered in this study.
Animal species for which there is a record of naturally occurring Ebola virus infection either from serological or PCR positivity, or from increased mortality attributed to EVD.
Modelling of the environmental/zoogeographic favourability for the presence of Ebola virus in wildlife.

272x125mm (150 x 150 DPI)
Modelling of the environmental/zoogeographic favourability for the presence of Ebola virus in wildlife.

157x110mm (300 x 300 DPI)
Combination of the environmental and the zoogeographic models of Ebola virus at the original 1° × 1° resolution (A) and downscaled to a 0.1° × 0.1° resolution. 205x199mm (300 x 300 DPI)