Global biogeography of human infectious diseases

Kris A. Murray[a,b,1], Nicholas Preston[c], Toph Allen[d], Carlos Zambrana-Torrelio[d], Parviez R. Hosseini[d], and Peter Daszak[d]

[a]Grantham Institute–Climate Change and the Environment, Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, United Kingdom; [b]School of Public Health, Faculty of Medicine, Imperial College London, London W2 1PG, United Kingdom; [c]Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA 02215; and [d]EcoHealth Alliance, New York, NY 10001

Edited by Robert M. May, University of Oxford, Oxford, United Kingdom, and approved August 26, 2015 (received for review April 16, 2015)

The distributions of most infectious agents causing disease in humans are poorly resolved or unknown. However, poorly known and unknown agents contribute to the global burden of disease and will underlie many future disease risks. Existing patterns of infectious disease co-occurrence could thus play a critical role in resolving or anticipating current and future disease threats. We analyzed the global occurrence patterns of 187 human infectious diseases across 225 countries and seven epidemiological classes (human-specific, zoonotic, vector-borne, non–vector-borne, bacterial, viral, and parasitic) to show that human infectious diseases exhibit distinct spatial grouping patterns at a global scale. We demonstrate, using outbreaks of Ebola virus as a test case, that this spatial structuring provides an untapped source of prior information that could be used to tighten the focus of a range of health-related research and management activities at early stages or in data-poor settings, including disease surveillance, outbreak responses, or optimizing pathogen discovery. In examining the correlates of these spatial patterns, among a range of geographic, epidemiological, environmental, and social factors, mammalian biodiversity was the strongest predictor of infectious disease co-occurrence overall and for six of the seven disease classes examined, giving rise to a striking congruence between global pathogeographic and "Wallacean" zoogeographic patterns. This clear biogeographic signal suggests that infectious disease assemblages remain fundamentally constrained in their distributions by ecological barriers to dispersal or establishment, despite the homogenizing forces of globalization. Pathogeography thus provides an overarching context in which other factors promoting infectious disease emergence and spread are set.

infectious disease | biogeography | pathogeography | globalization | distribution

The distributions of the vast majority of infectious agents causing disease in humans have not been resolved (1). However, unknown and poorly known agents make a significant contribution to the global burden of disease (e.g., ref. 2) and are likely to underlie many future disease impacts (e.g., ref. 3). In the absence of disease- or pathogen-specific data, existing patterns of infectious disease occurrence may provide the first insights into the spatial distributions of many disease risks. These insights could be leveraged to research, survey, define, or manage burgeoning or poorly understood infectious disease risks more efficiently (4–8).

Previous studies have found some striking patterns in the geographic distributions of human infectious diseases, including richness gradients (e.g., more diseases in the tropics relative to higher latitudes, more diseases on larger islands relative to smaller ones) (9–11), nestedness patterns (e.g., disease assemblages at higher latitudes are subsets of larger assemblages toward the tropics) (9), and varying geographic range sizes according to latitude and for varying types of diseases (12–14). Despite these patterns, the biogeography of human infectious diseases remains poorly explored in comparison to other biological taxa (15–18) and biogeographic insights into human infectious diseases have been little explored for public or global health applications.

Biogeographic patterns have routinely underpinned efforts to discover, monitor, and manage global biodiversity for almost 150 y (19–21). Better understanding the broad biogeographical patterns of human infectious diseases, why diseases occur in some places but not others, or how the presence of a disease in one place might relate to the likelihood of its presence in another thus have considerable potential for decomposing, monitoring, and managing the risks currently faced by the global health community (4). Potential applications range from focusing outbreak investigations, pathogen discovery strategies, risk assessments, and disease surveillance to disease management and mitigation (5, 7, 8).

Here, we make use of the most comprehensive infectious disease occurrence database currently available at a global scale (22, 23) to analyze the occurrence patterns of 187 human infectious diseases in 225 geopolitical regions (hereafter countries) and for seven nonexclusive disease classes with varying epidemiological features: human-specific, zoonotic, vector-borne, non–vector-borne, bacterial, viral, and parasitic diseases (Table 1). We use a biodiversity metric, beta diversity, and a biogeographic framework (16, 24) to represent the change in infectious disease assemblages among countries spatially at a global level. From this metric, we derive a region-specific beta diversity measure, which we term the co-zone layer, for its ability to illustrate the connectivity in existing disease occurrence patterns among regions. We then test whether these patterns could be leveraged for baseline infectious disease risk assessments in the absence of disease-specific data, using outbreaks of Ebola virus as a topical model system for which data are available for validation (25). Finally, we explore a range of environmental (climate, latitude, land area, and mammalian biodiversity) and social (human flight traffic, health expenditure, population size, and biases in observation effort) factors for their abilities to explain these biogeographic patterns after controlling for spatial dependence (26), providing insights into the potential drivers of shared disease risks.

Significance

Understanding the distributions of infectious diseases is a central public and global health objective. We show that human infectious diseases exhibit striking biogeographic grouping patterns at a global scale, reminiscent of "Wallacean" zoogeographic patterns. This result is surprising, given the global distribution and unprecedented connectivity of humans as hosts and the homogenizing forces of globalization; despite these factors, infectious disease assemblages remain fundamentally constrained in their distributions by ecological barriers to dispersal or establishment. Biogeographic processes thus appear to provide an overarching context in which other factors promoting infectious disease emergence and spread are set. We use outbreaks of Ebola virus to illustrate how such patterns could be leveraged to provide a "head start" or added focus for risk management activities.

Author contributions: K.A.M. designed research; K.A.M., N.P., T.A., C.Z.-T., and P.D. performed research; K.A.M., T.A., and P.R.H. analyzed data; K.A.M., N.P., T.A., and C.Z.-T. wrote the paper; and P.D. contributed to interpretation of results and to revisions.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. Email: kris.murray@imperial.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1507442112/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1507442112

Table 1. **Composition of infectious disease data analyzed in this study**

Disease trait	Disease class	*n*
	All diseases	**187**
Agent	**Bacterium**	**61**
	Fungus	8
	Parasite	**52**
	Virus	**66**
Vector-borne status	**Vector-borne**	**63**
	Non–vector-borne	**124**
Host category	Environmental	5
	Human-specific	**57**
	Multihost	18
	Zoonotic	**107**

Disease classes analyzed are shown in boldface; nonboldfaced classes were excluded due to low sample sizes. Host category definitions are provided in *Materials and Methods*.

Results

Pathogeography. Human infectious diseases exhibit clear spatial grouping patterns at a global scale (Fig. 1). Similar patterns persisted when disease classes were analyzed separately (*SI Appendix*, Fig. S1). However, the strength and consistency of these patterns varied. This variation was linked to differences in the mean number of countries that diseases of each class occurred in, an index of their propensity to be more or less widespread. Human-specific diseases were the most cosmopolitan, whereas vector-borne and zoonotic diseases were the most restricted (Fig. 2*A*). Average Sørensen similarity ($1 - \beta_{sor}$), calculated by taking the mean of all pairwise $1 - \beta_{sor}$ values for each disease class, was related to this pattern, with disease classes with the lowest average similarity among countries being the most restricted (vector-borne and zoonotic diseases), and vice versa (Fig. 2 *A* and *B*). The turnover (β_{sim}) component of overall dissimilarity (β_{sor}) was greater for the more restricted disease classes, whereas nestedness (β_{nes}) was more pronounced for the more widespread disease classes (Fig. 2 *A* and *C*).

Co-zone Layers. Country-specific zoonotic disease co-zones for Thailand and the Democratic Republic of Congo (DRC) show how beta diversity, and hence the sharing of disease risks via co-occurrence, can be visualized with respect to any focal region (*SI Appendix*, Fig. S2). Additive co-zones for Ebola spillover-positive countries calculated across four time points (Fig. 3) became relatively stable from 1994 onward, indicating that newly infected countries after this date did not appreciably fall beyond the area considered to be within the Ebola co-zone from prior outbreaks. This stability extended to the 2014 West Africa outbreak, which cannot be considered inconsistent with historical co-occurrence patterns for zoonotic diseases in this region and the location of previous Ebola virus outbreaks, although other countries, notably Liberia, appear better candidates than Guinea as a source country (as per the 2000 model). As a result, Ebola co-zones calculated at each time point were qualitatively effective at shortlisting at-risk countries subsequently becoming Ebola-positive countries at the next time point (additional details are provided in *SI Appendix*, *S1*). In addition, the top 22 countries considered to be at risk from Ebola in this study at the final time point (2014; Fig. 3*D*) compared favorably (63.6% overlap) with the 22 at-risk countries identified by Pigott et al. (25) (*SI Appendix*, Table S1). Our top 12 countries all appeared on Pigott et al.'s list (25), and all countries on that list were in our top 36 countries (range of average $\beta_{sor} = 0.89–0.81$). As such, our co-occurrence index is qualitatively able to shortlist regions at elevated risk of Ebola outbreaks based purely on existing co-occurrence patterns of zoonotic diseases among Ebola spillover-producing countries (additional details are provided in *SI Appendix*, *S1*).

Pathogeographic Correlates. Combined multiple regression on distance matrices (MRDM) models, which tested for correlations between the extrinsic predictors and the similarity of infectious disease assemblages among countries while controlling for geographic distance (spatial autocorrelation), explained 54.0% of the variation in the data overall (Table 2), and from 32.6% (bacterial) to 52.8% (zoonotic) depending on disease class (*SI Appendix*, Table S2). Extrinsic factors, without exception, were together more explanatory than geographic distance (*SI Appendix*, Fig. S3).

Most of this explanatory power came from the similarity of mammal assemblages among countries, which was significantly and positively correlated with the similarity of disease assemblages among countries overall (Table 2) and for all disease classes when analyzed separately (*SI Appendix*, Fig. S4 and Table S2). After accounting for geographic distance, mammalian biodiversity explained 17.9% of the variation in the data overall (Fig. 4), and from 2.9% for bacterial diseases to 19.2% for parasitic diseases when analyzed separately (*SI Appendix*, Fig. S5). The only class of disease for which mammalian biodiversity was not the most explanatory variable was bacterial diseases (*SI Appendix*, Fig. S5). This result helps explain a striking similarity between pathogeographic and zoogeographic regions of the world (*SI Appendix*, Fig. S6), including regions often referred to as Nearctic, Neotropical, Ethiopian/African, elements of the Saharo-Arabian, elements of the Mediterranean, Palearctic/Eurasian, Oriental, Australian, and Oceanian (15–17), and suggests that biogeography provides a fundamental context in which all other factors that mediate disease emergence and spread are set.

Less consistent effects of other explanatory variables were also identified (*SI Appendix*, Table S2). Overall, all extrinsic variables, with the exception of the similarity in per capita health expenditure and latitude among countries, were significantly correlated with disease assemblage similarity among countries (Table 2). However, relative influence of the predictors varied considerably, with similarity in population size and land area among countries being the most important predictors after mammalian biodiversity (Fig. 4), whereas the explanatory value of the nonmammalian extrinsic predictors was highly variable when disease classes were analyzed separately (*SI Appendix*, Fig. S5 and Table S2).

Discussion

Pathogeography. Human infectious diseases exhibit clear biogeographic patterns at a global scale (27). The "pathogeographic" (28) patterns revealed here correlate with patterns of mammalian biodiversity and are broadly consistent with classic zoogeographic classifications, including regions reminiscent of Nearctic, Neotropical, Ethiopian/African, elements of the Saharo-Arabian, elements of the Mediterranean, Palearctic/Eurasian, Oriental, Australian, and Oceanian (15–17). Considering that mammal assemblages correlate with disease assemblages after controlling for all other factors, even for disease classes that have no contemporary connection to mammals, this spatial structure is likely governed by the same processes that govern patterns of biodiversity

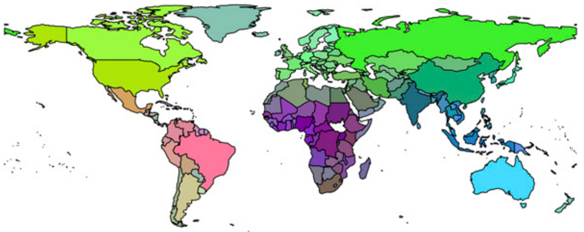

Fig. 1. Global human infectious disease pathogeographic patterns. Ordination analysis of β_{sor} of human infectious disease assemblages ($n = 187$ diseases). Similar colors indicate more similar disease assemblages. Separate disease classes and key to colors are presented in *SI Appendix*, Fig. S1.

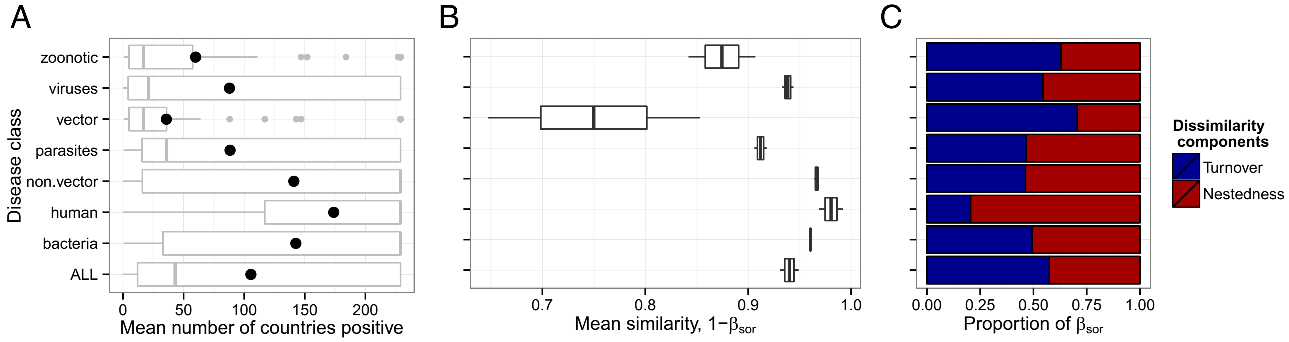

Fig. 2. (*A*) Mean number of countries positive for each disease class (black dots, box plots in gray). (*B*) Similarity ($1 - \beta_{sor}$) box plots for each disease class. (*C*) Turnover (β_{sim}) and nestedness (β_{nes}) components of total dissimilarity (β_{sor}). Turnover indicates the dissimilarity attributable to the replacement of diseases in one country relative to another. Nestedness indicates the dissimilarity attributable to diseases in one country being a subset of the diseases in another (24): Overall dissimilarity = Turnover + Nestedness($\beta_{sor} = \beta_{sim} + \beta_{nes}$).

more generally, which comprise a balance of historical events and local and regional processes, including diversification, geographic dispersal, and extinction (29). Biogeography thus provides a fundamental context in which all other factors contributing to the emergence and spread of infectious diseases are set, despite the unprecedented global availability and rising interconnectedness of humans as a host and the homogenizing forces of modern globalization (13, 18, 27).

Applied Pathogeography. Biogeographic patterns, such as those patterns originally described by Wallace (15), routinely underpin efforts to discover, monitor, and manage global biodiversity (19–21). In a similar way, pathogeographic structuring provides an as yet untapped source of prior information about the likelihood of disease co-occurrence that could inform a range of health-related activities. For example, we show that historical patterns of zoonotic disease occurrence can be used retrospectively to identify countries at elevated risk of Ebola virus outbreaks in the absence of any Ebola-specific information, as validated from outbreak data and by comparison with the most recent Ebola niche modeling study available (25). We stress that our objective was not to provide an applied risk assessment tool for Ebola virus outbreaks per se; rather, we aimed to evaluate how much information existing disease occurrence patterns contain that could be useful for applied problems in public and global health. The validated performance of our case study in identifying countries at elevated risk of Ebola outbreaks, based purely on existing patterns of zoonotic disease co-occurrence, strongly suggests that pathogeographic patterns could be exploited more broadly and at an early stage for a range of activities typically undertaken in data- or resource-poor settings. Examples include focusing surveillance for diseases, pathogens, or host and vector species; strengthening and targeting biosecurity capacities; responding to outbreaks; or optimizing pathogen discovery efforts (4–8).

Pathogeographic Specificity. Pathogeographic patterns varied somewhat by epidemiological class, which coincided with varying geographic range sizes, as broadly indicated by the number of countries in which diseases of each class occur. A relationship between geographic range size and the placement of biogeographic regions has been noted previously (16). On average, human-specific diseases were the most widespread, resulting in less apparent biogeographic structure, with greater overall similarity in disease assemblages among countries and more homogeneous disease assemblages at the global scale. Differences in disease assemblages in this group appear to be predominantly driven by nestedness, where the diseases in countries with lower disease richness are a subset of the diseases that occur in more disease-rich countries. This pattern could reflect true nestedness (i.e., reflecting gradients of pathogen richness) as reported in previous studies (9), but it could also be a residual factor of unequal sampling effort across countries, whereby countries with lower resources to fund disease research are more likely to have incomplete disease records. This suggestion is supported by our finding that disease assemblage similarity among countries may also be partially linked to relative observation effort (discussed below).

In contrast, vector-borne, zoonotic, and parasitic diseases were the least widespread, occurring in the lowest number of countries. This pattern coincides with clearer biogeographic structure, with lower overall similarity in disease assemblages among countries

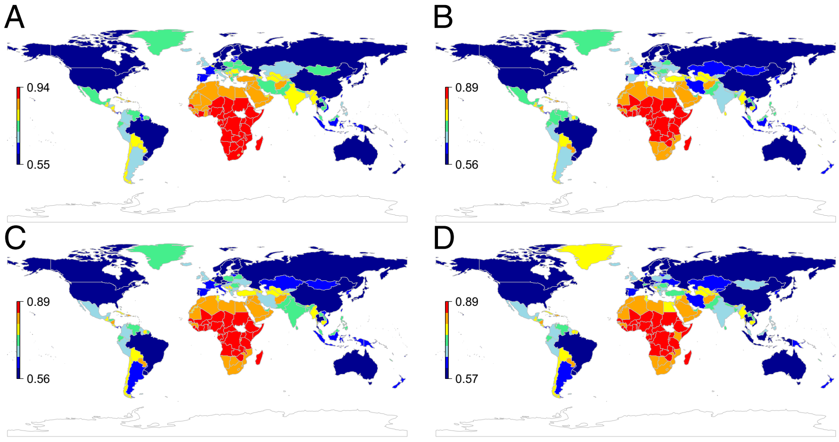

Fig. 3. Ebola co-zones through time (1976–2014). Additive co-zone models depict the average pairwise similarity ($1 - \beta_{sor}$; warmer colors indicate more similar) of zoonotic disease assemblages between countries recording primary Ebola (spillover) cases and all other countries calculated across four time points [1976 = DRC and Sudan (*A*), 1994 = *A* + Gabon and Cote d'Ivoire (*B*), 2000 = *B* + Republic of Congo and Uganda (*C*), and 2014 = *C* + Guinea (*D*)]. The scale is continuous but has been assigned categorical breaks to aid visualization.

Table 2. Correlates of the similarity of infectious disease assemblages among countries ("combined model," all diseases)

Variable	Coefficient	*P*
Mammal biodiversity	0.018	0.001***
Flight traffic	0.001	0.034*
Climate	−0.002	0.002**
Publications	−0.002	0.004**
Land area	−0.005	0.001***
Health expenditure	0.000	0.710
Population size	−0.007	0.001***
Latitude	0.001	0.208

Model: $R^2 = 0.54$. Tests of variable influence are presented in Fig. 3. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$. Variable descriptions are provided in *Materials and Methods*, and the full model output and separate disease class results are provided in *SI Appendix*, Fig. S5 and Table S1.

and more heterogeneous assemblages at the global scale. Differences in disease assemblages in these groups appear to be driven predominantly by turnover (the replacement of diseases from one country to another) and, in many cases, are likely primarily due to the geographic restrictedness of the reservoir, intermediate, or vector species on which they depend to complete transmission cycles. Although still considerably more homogenous than other taxa (e.g., mammals), the greater heterogeneity among these groups is likely driving the overall pattern for all diseases and illustrates that human infectious disease assemblages continue to be fundamentally constrained by biogeographic processes.

This is not to say that such processes present hard barriers, as the cosmopolitan distributions of some diseases (e.g., many human-specific diseases) or repeated cases of infectious disease emergence (e.g., influenza, severe acute respiratory syndrome, Middle East respiratory syndrome) clearly demonstrate. Rather, historic biogeographic barriers appear to vary somewhat in their porosity, mediated by a combination of geographic, ecological, social, and epidemiological factors. Nevertheless, the co-occurrence of human infectious diseases is likely to continue to be more prevalent within biogeographic basins than between them despite the risks of globalization for emergence and spread.

Pathogeographic Correlates. The effects of all extrinsic predictors were relatively weak in comparison to the pervasive effect of mammalian biodiversity (the only disease class for which this finding did not hold was bacterial diseases). The correlation between infectious diseases and mammalian biodiversity observed here is consistent with previous studies showing that infectious disease richness (the number of different types of diseases) is also correlated with mammalian plus avian biodiversity (30). Although host species richness or distributions may causally relate to pathogen species richness or distributions in some instances, these correlations do not imply causation. In contrast, biodiversity loss has been correlated with increases in human disease risks in some cases (31, 32) and the loss of ecosystem services more broadly, some with adverse impacts on health (30, 33). Although the generality of these trends remains unresolved (34), our results raise some additional questions regarding how shifts in and impacts on biodiversity might coincide with or contribute to future shifts in infectious disease risks. For example, our results suggest that faunal convergence among regions, which might arise as a result of nonrandom biodiversity loss or the spread of invasive species, could coincide with convergence in infectious disease assemblages and, in turn, disease risks.

Similarity in population size and similarity in land area were both correlated with similarity in infectious disease assemblages overall and for all disease classes when analyzed separately. Both factors could be analogous to "habitat availability," whereby the availability of more sites or hosts facilitates greater diversity, as suggested previously (30). Habitat availability, in turn, could drive similarly sized areas or populations toward more similar disease assemblages, after controlling for other factors.

Greater human flight traffic between countries was correlated with greater similarity of disease assemblages overall and for human-specific, vector-borne, and bacterial diseases. Although still a relatively weak predictor of overall similarity in disease assemblages among countries, and not evident across all disease classes, this finding supports previous studies showing that globalization is becoming an increasingly important factor for a range of disease risks (13, 35). With forecasts of increasing human migration globally (36), growth in human connectivity seems certain to play an increasingly important role in shaping human infectious disease assemblages in the future and driving disease assemblages toward greater homogeneity, particularly among well-connected countries. However, receptivity becomes the key issue following pathogen introduction, and our results suggest that the complete corrosion of biogeographic barriers (which also includes barriers to establishment) is unlikely. Nevertheless, diseases that are less constrained by spatially restricted factors (e.g., multihost pathogens and pathogens with invasive reservoir hosts or vectors) or diseases that are liberated from these limitations (e.g., via human-human transmission, access to a new pool of susceptible hosts, or because of rapid evolution or adaptation as for drug-resistant diseases) are more likely to be facilitated by the current growth in international connectivity.

Climatic similarity was correlated with disease similarity overall and for all disease classes except viruses when analyzed separately. Previous studies have documented clear latitudinal effects in the distribution and richness patterns of human infectious diseases, possibly related to climatic factors placing constraints on the transmission cycles of diseases (e.g., refs. 9, 30). Our results are consistent with these findings, showing that countries with more similar climates also share more similar diseases, although an effect of latitude was largely undetected (except for vector-borne and parasitic diseases) after accounting for climate and the effects of other variables. This result has implications for future disease shifts under climate change, whereby the convergence of climatic conditions among countries could trigger the convergence of disease assemblages via disease expansions or contractions.

Despite limited explanatory power overall, the similarity in per capita health expenditure was correlated with similarity in disease assemblages among countries for vector-borne, zoonotic, bacterial, and parasitic diseases. This result is consistent with previous studies showing a clear effect of health investment in limiting disease distributions and burdens (30), which supports ongoing investment in disease burden-reducing and biosecurity measures that could help limit establishment success (37). The lack of a stronger signal is nevertheless a surprise, given the divide in infectious disease burdens between higher and lower income countries, illustrating the distinction that must be made between the occurrence and burden of infectious diseases. To some extent, the lack of a clearer effect could also be due to

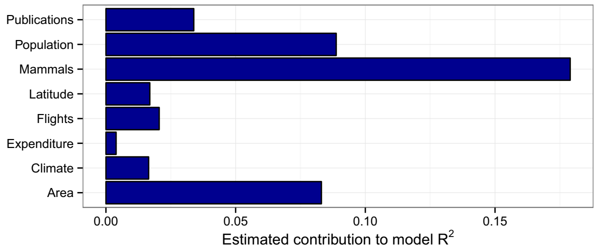

Fig. 4. Estimated relative influence of extrinsic predictors in explaining the similarity of human infectious disease assemblages among countries after accounting for the effect of geographic distance (spatial autocorrelation). Overall model: $R^2 = 0.540$. Descriptions of predictors are provided in *Materials and Methods*. Full model outputs are shown in Table 2 and *SI Appendix*, Table S1, and separate disease classes are shown in *SI Appendix*, Fig. S5.

having another more explanatory metric of investment in health expenditure (publication effort) in the models but, again, reinforces the overriding effect of global biogeographic processes in the spatial structuring of infectious disease distributions.

Limitations and Future Directions. Deficiencies in current availability and resolution of infectious disease data limit our ability to characterize, interpret, and exploit pathogeographic patterns more precisely. A recent systematic review similarly concluded that serious deficiencies in the current availability and quality of infectious disease data prevent comprehensive mapping for all but a few human infectious diseases (1). Although disease detection and coordinated reporting is a significant challenge, it is sobering to consider that we know more about the global distributions of the world's ~5,000 critically endangered plant and animal species, many of which are exceedingly rare, cryptic, and illusive (e.g., www.iucnredlist.org/details/21533/0), than we do about the 1,400 or so (1) infectious agents causing disease in humans.

For example, favoring coverage over resolution, we were restricted to a country-level dataset to conduct our global analysis. This dataset covered around 25% of all known human infectious diseases, but it is skewed toward diseases of greater current clinical relevance, which almost certainly includes a bias toward more widespread diseases. Of these infectious diseases we excluded nearly half due to the limited specificity of disease "taxonomy" or due to having no or limited information about causative agents. In addition, we detected a consistent signal of unequal observation effort that demonstrably biases our current picture of global disease distributions, even for the relatively well-known diseases that remained in the analysis. These issues pose problems for pathogeographic analyses, particularly in poorly studied regions and where larger countries dominate in global surface area or span historic biogeographic boundaries. Such limitations could inflate dissimilarity estimates for those countries, or otherwise obscure more biologically relevant biogeographic boundaries at subnational levels. Improving the specificity inherent in disease classification systems, passing those improvements on to disease occurrence databases, improving the spatial resolution of global databases to beyond the geopolitical region level, increasing and more evenly distributing observation effort, and better ground-truthing the proxies often used to represent observation biases (e.g., publication effort, health care spending, gross domestic product per capita) all have potential to improve our ability to parse out the signal from the noise and exploit pathogeographic patterns to improve public and global health outcomes.

Materials and Methods

Disease Data Source. We compiled a presence/absence matrix of human infectious diseases at the country level from data held in the Global Infectious Disease and Epidemiology Network (GIDEON) database (www.gideononline.com) (22, 23). To our knowledge, GIDEON is the most comprehensive infectious disease occurrence database currently available at a global scale and served as the basis for a recent systematic review on global disease mapping (1). GIDEON is updated monthly, and aggregates disease information from a very wide range of peer-reviewed and evidence-based sources. When last accessed for this study (August 25, 2014), GIDEON reported on 351 diseases of clinical interest derived from 630,728 data points compiled from 37,489 surveys and 424,543 references.

Disease Data. We expanded previous databases (5, 13) to link country-level disease occurrence data to specific disease epidemiological traits. For "agent type," a disease can be caused by a virus, bacterium, parasite, or fungus. For "vector-borne status," a disease is either vector-borne or not vector-borne. For "host category," a disease can be human-specific, zoonotic, multihost, or environmental (Table 1). Host categories were previously defined by Smith et al. (13), building on GIDEON's own definitions and host categories. Briefly, human-specific diseases are diseases currently restricted to human reservoir hosts (even where they originated in other species), zoonotic diseases are diseases for which development and reproduction are restricted to nonhuman hosts but dead-end human infections via spillover occur, multihost diseases are diseases that can use both animal and human hosts to complete the life cycle, and environmental diseases are diseases acquired directly from the environment.

To maximize geographic specificity for our analyses, we removed 164 (46.7%) diseases with unknown, nonspecific, or multiple causative agents (e.g., syndromes, disease complexes), leaving 187 diseases for analysis (Table 1). In addition to analyzing all diseases together, we used the epidemiological trait data to group further analyses into seven disease "classes" of interest, which are not mutually exclusive: human-specific, zoonotic, vector-borne, non–vector-borne, bacterial, viral, and parasitic diseases. We did not analyze three disease classes due to low sample sizes: multihost (n = 18), environmental (n = 5), and fungal (n = 8) diseases.

Measuring beta diversity. We adopted the concepts and metrics of beta diversity proposed by Baselga (24) and followed the framework of Kreft and Jetz (16) to visualize beta diversity and create co-zone layers for human infectious diseases. For each disease class (discussed above), we calculated pairwise "total dissimilarity" (beta diversity, β) of disease assemblages between all country pairs using the β_{sor} index. In our application, β_{sor}, a widely used index, encompasses the proportion of shared diseases between two countries, and is defined as:

$$\beta_{sor} = \frac{b+c}{2a+b+c},$$

where a is the number of diseases common to both countries, and b and c are the number of diseases that are unique to each of the two countries being compared, respectively (24).

Differences in disease assemblages between two countries (β_{sor}) could occur in two distinct ways, each with potentially different underlying mechanisms (24). The first is "turnover," which results from the replacement of diseases with others when comparing the assemblages of two countries. The second is "nestedness," which results when the diseases in one country form a subset of the diseases in another. We followed Baselga (24) in partitioning out these components of beta diversity. Baselga (24) shows that β_{sor} comprises these two additive components, reflecting the fractions of total dissimilarity derived from turnover and nestedness, respectively (*SI Appendix*, *S2*). We ran separate analyses on each of the three components, but we only report results from β_{sor} (overall dissimilarity) for brevity because we were mostly interested in the overall patterns of dissimilarity of diseases at the global level. We did, however, calculate the relative contribution of turnover and nestedness for each β_{sor} result to infer which component of dissimilarity is most important for driving patterns in each disease class analyzed. All dissimilarity metrics were calculated using the "beta.pair" function in the R package *betapart* (38). In some cases, we report assemblage similarity rather than dissimilarity, defined as $1 - \beta_{sor}$.

Visualizing beta diversity and the co-zone layer. For visualization of disease assemblage dissimilarities, we used ordination (nonmetric multidimensional scaling; NMDS) on the dissimilarity matrices, performed with the "metaMDS" function in the R package *vegan* (39) and specifying k = 3 dimensions. For plotting and mapping, we assigned each of the three NMDS axes one of the RGB (red, green, blue) color primaries scaled between minimum and maximum values, such that each country was assigned an additive color representing its relative position (according to its disease assemblage) in relation to other countries in 3D NMDS space [following Kreft and Jetz (16)]. In this framework, countries assigned more similar colors share more similar disease assemblages. We also used clustering (unweighted pair group method with arithmetic mean) to assess the qualitative fit between zoogeographic and pathogeographic patterns (*SI Appendix*, *S3*).

Applied pathogeography: Evaluating co-zones for risk assessment. Existing patterns of infectious disease co-occurrence could be used as a source of prior information to place baseline relative likelihoods of the occurrence of a pathogen or disease detected in one place also occurring in any other. This information could be exploited in public or global health applications in the absence of disease- or pathogen-specific data, for example, to identify at-risk or "target" regions rapidly for surveillance, pathogen discovery, or outbreak investigations.

We first plotted the similarity of disease assemblages for all countries relative to a single focal country. This approach is equivalent to visualizing beta diversity with respect to a reference point (16). We use the term co-zone layer to describe this source of information because it can be used to represent the connectivity of infectious disease risks between countries spatially based on patterns of existing infectious disease co-occurrence, as indicated by the similarity index ($1 - \beta_{sor}$) described above. We hypothesized that a novel pathogen or disease that is detected in the focal country is also more likely to occur within that country's co-zone than elsewhere. We first created co-zone layers for Thailand and the DRC as arbitrary example cases for visualization.

Second, we created an additive co-zone layer reflecting the regional similarity of disease assemblages with respect to a group of focal countries. This approach may be suited to situations where a novel pathogen or disease is detected across multiple countries, again with the implication that (in the absence of further information) a novel pathogen or disease that is detected in all of the focal countries is also more likely to occur within those countries' additive co-zone. To simulate a more realistic example and to provide an opportunity for model validation, we used the characteristics of Ebola virus outbreaks (primary cases only) in Africa as a test case. Results of the Ebola-specific co-zone model were used to rank countries at risk of recording Ebola cases in the future (further details and the validation approach are provided in *SI Appendix, S1*).

Correlates of infectious disease co-occurrence. For all diseases combined and for each of the seven separate disease classes, we used MRDM (26, 40) to analyze the correlates of global patterns of disease assemblage dissimilarity. MRDM is an extension of partial Mantel analysis often used for the analysis of spatial ecological data. Response and explanatory variables are matrices of "distances" (or dissimilarities), which can be used to represent spatial, ecological, and other factors of interest for their potential to explain variation in the response. Significance testing in MRDM is performed by permutation analyses, and the framework is well suited to quantifying and controlling for spatial autocorrelation at varying spatial scales with the use of distance lag matrices (26), a useful feature for biogeographic analyses. All MRDM models were implemented using the "MRM" function in the R package *ecodist* (41).

On the basis of being implicated in the observed distribution and richness patterns of certain groups of human infectious diseases in prior studies, we tested the relative influence of numerous environmental factors [biodiversity (e.g., refs. 30, 42), land area (30), climate (e.g., ref. 9), latitude (9), and geographic distance] and social factors [population size (30), human flight traffic (e.g., refs. 35, 43), health expenditure (30), and observation bias (e.g., refs. 44, 45)] for their abilities to explain disease assemblage similarities among countries. Methods of deriving explanatory variables for use in models are detailed in *SI Appendix, S4*.

To examine the relevance of these explanatory variables while controlling for the effects of spatial autocorrelation, we used MRDM incorporating distance lag matrices as described by Lichstein (26). Briefly, we first plotted disease similarity $(1 - \beta_{sor})$ by geographic distance (square root-transformed) between all country pairs to visualize any broad patterns. We then evaluated the presence of spatial autocorrelation (by which we mean the presence of a statistically significant spatial pattern, regardless of cause) with Mantel correlograms, implemented with the "mgram" function in the R package *ecodist* (41). Finally, we derived "distance lags" for use in MRDM, specifying distance classes derived from the Mantel correlogram bin categories (which are themselves derived according to Sturge's rule when using *ecodist*). Only significant lag categories were added to MRDM models. To estimate the relative influence of each extrinsic predictor variable relative to the others, we used the "calc.relimp" function in the R package *relaimpo* (46). We also estimated the variance components attributable to pure geographic distance, pure extrinsic factors, or shared factors by running separate extrinsic factor-only and geographic distance-only models for direct comparison of R^2 values between the three models (combined, extrinsic, and distance), following the method described by Lichstein (26).

ACKNOWLEDGMENTS. We thank the Editors and the anonymous referees for their comments and suggestions that have improved the manuscript. This project was partially supported by the US Agency for International Development Emerging Pandemic Threats-PREDICT program.

1. Hay SI, et al. (2013) Global mapping of infectious disease. *Philos Trans R Soc Lond B Biol Sci* 368(1614):20120250.
2. Hotez PJ, et al. (2014) The global burden of disease study 2010: Interpretation and implications for the neglected tropical diseases. *PLoS Negl Trop Dis* 8(7):e2865.
3. Morse SS, et al. (2012) Prediction and prevention of the next pandemic zoonosis. *Lancet* 380(9857):1956–1965.
4. Scheiner SM (2009) The intersection of the sciences of biogeography and infectious disease ecology. *EcoHealth* 6(4):483–488.
5. Murray KA, et al. (2012) Cooling off health security hot spots: Getting on top of it down under. *Environ Int* 48:56–64.
6. Bogich TL, Anthony SJ, Nichols JD (2013) Surveillance theory applied to virus detection: A case for targeted discovery. *Future Virol* 8(12):1201–1206.
7. Reperant LA (2010) Applying the theory of island biogeography to emerging pathogens: Toward predicting the sources of future emerging zoonotic and vector-borne diseases. *Vector Borne Zoonotic Dis* 10(2):105–110.
8. Peterson AT (2008) Biogeography of diseases: A framework for analysis. *Naturwissenschaften* 95(6):483–491.
9. Guernier V, Hochberg ME, Guégan J-F (2004) Ecology drives the worldwide distribution of human diseases. *PLoS Biol* 2(6):e141.
10. Cliff A, Haggett P, Smallman-Raynor M (2000) *Island Epidemics* (Oxford Univ Press, Oxford).
11. Cliff AD, Haggett P (1995) The epidemiological significance of islands. *Health Place* 1(4):199–209.
12. Smith KF, Guégan J-F (2010) Changing geographic distributions of human pathogens. *Annu Rev Ecol Evol Syst* 41(1):231–250.
13. Smith KF, Sax DF, Gaines SD, Guernier V, Guégan JF (2007) Globalization of human infectious disease. *Ecology* 88(8):1903–1910.
14. Guernier V, Guégan J-F (2009) May Rapoport's rule apply to human associated pathogens? *EcoHealth* 6(4):509–521.
15. Wallace AR (1876) *The Geographical Distributions of Animals, with a Study of the Relations of Living and Extinct Faunas as Elucidating the Past Changes of the Earth's Surface* (Macmillan, London).
16. Kreft H, Jetz W (2010) A framework for delineating biogeographical regions based on species distributions. *J Biogeogr* 37(11):2029–2053.
17. Holt BG, et al. (2013) An update of Wallace's zoogeographic regions of the world. *Science* 339(6115):74–78.
18. Martiny JBH, et al. (2006) Microbial biogeography: Putting microorganisms on the map. *Nat Rev Microbiol* 4(2):102–112.
19. Diamond JM (1976) Island biogeography and conservation: Strategy and limitations. *Science* 193(4257):1027–1029.
20. Channell R, Lomolino MV (2000) Dynamic biogeography and conservation of endangered species. *Nature* 403(6765):84–86.
21. Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature* 403(6772):853–858.
22. Berger SA (2005) GIDEON: A comprehensive web-based resource for geographic medicine. *Int J Health Geogr* 4(1):10.
23. Edberg SC (2005) Global Infectious Diseases and Epidemiology Network (GIDEON): A world wide Web-based program for diagnosis and informatics in infectious diseases. *Clin Infect Dis* 40(1):123–126.
24. Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Glob Ecol Biogeogr* 19(1):134–143.
25. Pigott DM, et al. (2014) Mapping the zoonotic niche of Ebola virus disease in Africa. *Elife* 3:e04395.
26. Lichstein J (2007) Multiple regression on distance matrices: A multivariate spatial analysis tool. *Plant Ecol* 188(2):117–131.
27. Just MG, et al. (2014) Global biogeographic regions in a human-dominated world: The case of human diseases. *Ecosphere* 5(11):art143.
28. Reichert I, Palti J (1967) Prediction of plant disease occurrence a patho-geographical approach. *Mycopathol Mycol Appl* 32(4):337–355.
29. Ricklefs RE (1987) Community diversity: Relative roles of local and regional processes. *Science* 235(4785):167–171.
30. Dunn RR, Davies TJ, Harris NC, Gavin MC (2010) Global drivers of human pathogen richness and prevalence. *Proc R Soc Lond B Biol Sci* 277(1694):2587–2595.
31. Keesing F, et al. (2010) Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* 468(7324):647–652.
32. Morand S, Jittapalapong S, Supputtamongkol Y, Abdullah MT, Huan TB (2014) Infectious diseases and their outbreaks in Asia-Pacific: Biodiversity and its regulation loss matter. *PLoS One* 9(2):e90032.
33. Costanza R, et al. (2014) Changes in the global value of ecosystem services. *Glob Environ Change* 26:152–158.
34. Salkeld DJ, Padgett KA, Jones JH (2013) A meta-analysis suggesting that the relationship between biodiversity and risk of zoonotic pathogen transmission is idiosyncratic. *Ecol Lett* 16(5):679–686.
35. Tatem AJ, Rogers DJ, Hay SI (2006) Global transport networks and infectious disease spread. *Adv Parasitol* 62:293–343.
36. Tatem AJ (2014) Mapping population and pathogen movements. *Int Health* 6(1):5–11.
37. Murray CJ, et al. (2012) Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380(9859):2197–2223.
38. Baselga A, Orme CDL (2012) betapart: An R package for the study of beta diversity. *Methods Ecol Evol* 3(5):808–812.
39. Oksanen J, et al. (2015) vegan: Community Ecology Package. R package version 2.3-0. Available at CRAN.R-project.org/package=vegan.
40. Legendre P, Lapointe F-J, Casgrain P (1994) Modeling brain evolution from behavior: A permutational regression approach. *Evolution* 48(5):1487–1499.
41. Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. *J Stat Softw* 22(7):1–19.
42. Jones KE, et al. (2008) Global trends in emerging infectious diseases. *Nature* 451(7181):990–993.
43. Hosseini P, Sokolow SH, Vandegrift KJ, Kilpatrick AM, Daszak P (2010) Predictive power of air travel and socio-economic data for early pandemic spread. *PLoS One* 5(9):e12763.
44. Yang K, et al. (2012) Global distribution of outbreaks of water-associated infectious diseases. *PLoS Negl Trop Dis* 6(2):e1483.
45. Hopkins ME, Nunn CL (2010) Gap analysis and the geographical distribution of parasites. *The Biogeography of Host-Parasite Interactions*, ed Krasnov SMaB (Oxford Univ Press, Oxford), pp 129–142.
46. Grömping U (2006) Relative importance for linear regression in R: The package relaimpo. *J Stat Softw* 17(1):1–27.

ECOLOGY