The coevolutionary mosaic of bat betacoronavirus emergence risk

Norma Forero Rocio Munoz [1,2,‡] Renata L. Muylaert [3] Stephanie N. Seifert [4] Gregory F. Albery [5] Daniel J. Becker [6] Colin J. Carlson [7,8,9,‡] Timothée Poisot [1,2,‡]

[1] Université de Montréal [2] Québec Centre for Biodiversity Sciences [3] Molecular Epidemiology and Public Health Laboratory, School of Veterinary Science, Massey University, New Zealand [4] Paul G. Allen School for Global Health, Washington State University, Pullman, WA, United States [5] Department of Biology, Georgetown University, Washington, DC, USA [6] Department of Biology, University of Oklahoma, Norman, OK, USA [7] Department of Biology, Georgetown University, Washington, DC, [8] Center for Global Health Science and Security, Georgetown University Medical Center, Washington, DC, USA [9] Department of Microbiology and Immunology, Georgetown University Medical Center, Washington, DC, USA

[‡] These authors contributed equally to the work

Correspondance to:

Timothée Poisot — `timothee.poisot@umontreal.ca`

Last revision: *June 28, 2022*

Driven by the need to understand the ecological factors involved in the emergence of *Betacoronavirus* (the genus causing SARS, MERS, and COVID-19 in human) through bat hosts, we develop an approach to the assesment of spillover risk based on the Geographic Mosaic Theory of Coevolution. In doing so, we provide a global mapping of the spillover risk posed by betacoronaviruses, reflecting the fact that this risk is best understood through the multi-faceted prism of ecological and evolutionary mechanisms. Our framework reveals that host richness alone, although a component of viral hazard, is not a sufficiently integrative predictor of risk. We offer alternative insights based on viral sharing, host compositional uniqueness, and host phylogenetic diversity and phylogeographic regions. By comparing our aggregated measure of risk to a proxy for human density, namely the proportion of each pixel that is covered by urban or built land, we provide a synthetic risk map, allowing the identification of hotspots where the bat-betacoronavirus interaction network may facilitate the emergence of novel viruses and their spillover into human populations.

Disease emergence is complex, and is driven not only by animal-human contact, but also by the underlying evolutionary dynamics in viral reservoirs.[1] Although host richness is often used as a superficial proxy for spillover risk,[2,3] these approaches oversimplify the relevant interspecific heterogeneity in immunology, behavior, and other traits, and therefore overlook unique host pools that allow for the rapid evolution of highly divergent viruses.[4] In the case of generalist pathogens like betacoronaviruses, there is conceptual and empirical support to the idea that these community-level mechanisms are even more important,[5] particularly given that cross-species transmission may, as a rule, structure viral evolution more than co-divergence with hosts.[6] This creates a disconnect between coevolutionary theory and most existing ecological frameworks for mapping spillover risk.

The Geographic Mosaic Theory of Coevolution (GMTC) attempts to explicitly connect microevolutionary dynamics to the macroecology and biogeography of symbiotic interactions.[7] The GMTC posits that coevolutionary processes among pairs[8] or complexes[9] of species are structured in space by the rippling effects of abiotic conditions onto evolutionary mechanisms, giving rise to fragmented systems with different ecologies over large spatial extents.[10] The GMTC predicts a spatial fragmentation of coevolutionary dynamics under the joint action of three processes:[11] coevolutionary hot- and coldspots, which appear when the intensity of *interaction* (in terms of reciprocal fitness consequences) varies spatially; selection mosaics, wherein the intensity of *selection* varies across space, driven by both the biotic complexity of the community (locally diverse hosts and viruses are more biotically complex) and the local favorability of the environment;[12] and trait remixing, which occurs when coevolutionary dynamics change when community-level *functional traits* change through meta-community dynamics.

Here, we apply the GMTC to explore and explain the global biogeography of betacoronaviruses, the group that includes SARS-CoV, MERS-CoV, and SARS-CoV-2. In their bat reservoirs, coronaviruses evolve through a mix of host jumps, recombination among disparate lineages, and, to a lesser degree, co-divergence with their hosts—[2]a mix of mechanisms that creates a complex and nonlinear relationship between host diversity and viral emergence. Working from a recently published database of bat hosts of betacoronaviruses, we test whether spatial structure in bat-betacoronavirus coevolution is identifiable at a global scale. Aiming to explain these patterns, we develop a generalized framework for applying the GMTC to host-virus interactions, with a specific emphasis on the potential to create independent coevolutionary dynamics (and therefore spatial fragmentation in risk) through heterogeneity. We develop a trivariate risk assessment system that connects each GMTC mechanism to a quantifiable aspect of host-virus interactions: (i) viral sharing rates in host communities, representing the strength of potential interaction between viruses and any one host (i.e., places where viruses undergo constant host switching may be coevolutionary coldspots); (ii) the phylogenetic diversity of hosts, as a proxy for variation in the immunological mechanisms that antagonize viruses (i.e., the selection mosaic); and (iii) the local uniqueness of the bat community, representing the potential for viruses to be exposed to novel host traits (e.g., variation in receptor sequences). Together, we argue that these can be used to identify and map the evolutionary drivers that—in conjunction with transmission processes (e.g., viral prevalence in reservoirs and animal-human contact rates)— determine disease emergence risk.

Results and Discussion

Bat and betacoronavirus biogeography are broadly consistent

Most previous work has assumed that the presence or richness of key groups of bat hosts are predictive of coronavirus diversity.[2,3] Projecting bat and betacoronavirus phylogeny over space (fig. 1), we find support for the idea that bat community assembly is directly responsible for a global mosaic of viral evolution. The distinct groupings (represented by different colors, symbolizing positions in a subspace formed by the first two phylogenetic principal components) are essentially equivalent between the two groups, and can be coarsely delineated as (1) south and southeast Asia, (2) east Asia (including northern China), west Asia, and the Mediterranean coast; (3) Eurasia above a northing of 40; and (4) Africa and south America. In some cases, this diverges from expectations about coronavirus biogeography: for example, previous work has rarely flagged India as a region of interest, but for both bats and betacoronaviruses, the subcontinent falls into the same regions as the southeast Asian peninsula (and indeed, the region is home to known bat hosts of nobecoviruses, sarbecoviruses, and merbecoviruses).[3]

[Figure 1 about here.]

Overall, these results suggest that the boundaries of bat and betacoronavirus biogeographic regions are largely consistent. This may be surprising, given that cospeciation plays a minor role in coronavirus diversification,[2] a property that would theoretically allow for substantial broad divergence in their biogeography. However, host jumps at the family level or higher are relatively rare and significant events in coronavirus evolutionary history;[2,13] as a result, the mosaic of betacoronavirus phylogeography is assembled from a set of overlapping smaller coevolutionary systems, superimposed in space and filtered by the importance of different subgroups in local host communities. For example, the most speciose and cosmopolitan family of bats, the vesper bats (Vespertilionidae), are considered the primary reservoir of merbecoviruses;[3,13] but in the Americas, where merbecoviruses are the only lineage present, they have only been found in other bat taxa. At the coarsest scale, these heterogeneities are lost, and betacoronavirus biogeography tracks the deep rifts in bat evolutionary history—but within broad regions, the component coevolutionary systems may have very different dynamics.

Hotspots of bat and betacoronavirus biodiversity are distinct

Bats, the second most diverse groups of mammals, are found worldwide; gradients in their species richness generally track broader patterns of mammal diversity, with a striking Neotropical hotspot (especially in the Amazon basin) and a secondary hotspot centered in the southeast Asian peninsula. These hotspots of bat diversity are generally presumed to be hotspots of viral adaptive radiation, and therefore areas of concern for human health.[2,14] However, the hotspots of known bat betacoronavirus hosts show a distinct pattern, with primary hotspots (both in terms of size and higher values) of host richness situated in southeast Asia, parts of southern Europe, and to a lesser extent parts of Africa in the -25-0 range of latitudes (fig. 2; top). Although hundreds of species likely host undiscovered betacoronaviruses, machine learning predictions have suggested that these undiscovered reservoirs should follow the same diversity gradient.[15] In principle, these hotspots of locally-diverse, virus-rich bat communities should drive more adaptive diversification in their viruses.

[Figure 2 about here.]

However, we find that the global pattern of betacoronavirus phylogenetic distinctiveness is quite distinct from both bat host richness and phylogenetic distinctiveness (fig. 2; bottom). In contrast to the sparsity of Neotropical betacoronavirus hosts, South America has the most evolutionary distinct hosts *and* viruses, followed by secondary hotspots in southeast Asia and the Rift Valley region have mostly distinct viruses. Some degree of sampling bias may contribute to these patterns: for example, South-America is one of the places where the fewest bat betacoronavirus sequences have been generated,[2,14,16] resulting in a sparser phylogenetic tree, and artificially inflating distinctiveness; conversely, disproportionate research effort in eastern China[17] may have led to a more complete inventory of the local diversity of coronaviruses, again inflating these metrics relative to underlying patterns. Even accounting for these potential biases, though, there is obvious heterogeneity in betacoronavirus evolutionary distinctiveness that is distinct from overall bat diversity.

Overall, these patterns recapitulate the evolutionary history of both the order Chiroptera and the genus *Betacoronavirus*. Horseshoe bats (Rhinolophidae) are both the reservoirs of the SARS-like viruses (subgenus *Sarbecovirus*), the group of pandemic threats that have been of the greatest interest to researchers[13] (and so have been sampled most intensively).[17] The hotspots of host richness and viral diversity in southeast Asia—both of which are disproportionately high, considering the global landscape of bat species richness—are almost entirely driven by viral adaptive radiation through host switching within this clade[3,15]. In contrast, the Neotropical hotspot of viral distinctiveness is driven by isolation by host vicariance. Out of the four main groups of betacoronaviruses, only the subgenus *Merbecovirus* (MERS-like viruses) has been found in animals in the Americas—an introduction that is generally presumed to be ancient.[3] While comparatively understudied, New World merbecoviruses have been found in the ghost-faced bats (Mormoopidae), New World leaf-nosed bats (Phyllostomidae), and free-tailed bats (Molossidae).[18] The former two groups are endemic to the Neotropics, while the explosive adaptive radiations of the latter two (and particularly the phyllostomids) are responsible for the hotspot of bat diversity in the Amazon. Together, these clades of New World bats play host to a distinct regime of betacoronavirus coevolution.

Coevolution-informed emergence risk is different in space

As host richness, joint distinctiveness, or phylogeographic structure suggest that the bat-betacoronaviruses complex is globally fragmented enough to give rise to both different levels of risk (as evidenced by the spatial location of spillover events) and different types of co-evolutionary dynamics, we turn to the Geographic Mosaic Theory of Coevolution to provide a measure of risk accounting for multiple processes. In fig. 3, we overlapped three components of spillover risk: viral sharing, *i.e.* the chance that two bats will share viruses overall; Local Contribution to Beta Diversity, *i.e.* the fact that a bat community is compositionally unique compared to the average compositional similarity across the entire system; finally, host phylogenetic diversity, *i.e.* how dispersed the bats in a location are within the tree of life. This approach leads to the definition of broad biogeographic regions of risk, where the same color represents the same type of risk. By way of constrat to figures fig. 2 and fig. 1, these regions do not necessarily overlap with previous spatial partitions of the bat-betacoronaviruses complex.

[Figure 3 about here.]

From the perspective of spillover risk, the most important combination of factors is a high phylogenetic diversity of hosts with low viral sharing; this, essentially, means that very different betacoronaviruses could co-exist within the same place. This is particularly the case given that betacoronaviruses often evolve and even achieve host shifts through recombination, which requires the co-occurrence of sufficiently distinct viruses to be a major driver of emergence. In fig. 3, this corresponds to yellow to pale green areas, which are essentially limited to South-Eastern Asia, and to some part of Sub-Saharan Africa. Adopting a geographic mosaic theory perspective on risk, other regions of the world are of lesser concern (fig. 4). Our risk decomposition does not account for viral diversity or distinctiveness. The simple rationale behind it is that the acquisition of viral data is rarely disconnected from the acquisition of host data. There are more sources of information on hosts than on viruses, allowing to develop a host-centric perspective on risk (although this estimate would more accurate with viral traits related to *e.g.* ability to switch hosts or pathogenic potential). Areas with high bat diversity and high turnover *may* facilitate the evolutionary radiation of viruses, matching previous findings that the diversification of bat coronaviruses is driven largely by host shifts (inter-genus or higher levels of cross-species transmission) and, to a lesser degree, cospeciation and sharing, representing intra-genus cross-species transmission.[2] This diversification is not an actual risk factor for spillover itself, but acts downstream of a spillover event by increasing the random chance of the emergence of a virus with the raw genomic components required for the potential to infect humans.

[Figure 4 about here.]

From another perspective, areas of high host uniqueness and virus sharing (red-to-pink) could provide hotspots of *Betacoronavirus* risk through mixing of unique viruses (via codivergence) and in turn recombination. Under our framework, such a hotspot was identified in Madagascar, where most bat species are endemic following evolutionary divergence from sister species in both African and Asian continents.[19] Recent surveillance[20] has identified a novel *Betacoronavirus* (in the subgenus *Nobecovirus*) in

Madagascar-endemic pteropid bat species (*Pteropus rufus*, *Rousettus madagascariensis*), emphasizing strong proof of principle in model predictions.

Human landscapes filter the geography of emergence risk

The relationship between the underlying pathogen pool and emergence risk is mediated by both human-wildlife interfaces (the probability of spillover) and opportunities for onward transmission (the probability that spillovers become epidemics)[1]. As a proxy for both, we finally overlaid the risk component from the composite map (see above) with the proportion of built land, as a proxy for a mix of habitat disturbance, potential for bat synanthropy or contact with bridge hosts like livestock,[21,22] and human population density and connectivity[1,23,24] (fig. 5). Accounting for these factors, most of South America and Europe are at comparatively lower risk, as–although densely populated–settlements tend to be in areas with lower potential risk. Conversely, regions like Malaysia and the northern coast of Australia have a high evolutionary risk component, but should represent a relatively lower effective risk due to low human density. However, southeast Asia, the Indian subcontinent, and scattered hotspots in sub-Saharan Africa are at high risk due to the overlap between human populations and natural opportunities for cross-species transmission of betacoronaviruses.

[Figure 5 about here.]

Reassuringly, these predictions correspond to the geographic origins of the three bat-origin coronaviruses that have recently emerged in human populations. While available information puts the spillover of SARS-CoV-2 in a live animal market in Wuhan, China, the ultimate origin of the virus is almost certainly in a divergent lineage of sarbecoviruses from the Indochinese peninsula that was poorly characterized prior to the pandemic.[25–27] Similarly, the SARS-CoV outbreak began in Guangdong province in 2002, reaching humans through small carnivore bridge hosts, but was eventually traced back to a set of likely progenitor viruses found in cave-dwelling horseshoe bats in Yunnan province;[28] nearby, antibody evidence has indicated human exposure to SARS-like viruses.[29] MERS-CoV was originally detected in Saudi Arabia, accompanied by a nearly identical virus sequenced from an Egyptian tomb bat (*Taphozous perforatus*)[30], but is widespread in camels in East Africa and the Middle East, and may have reached its bridge host decades earlier than originally supposed;[31] as a result, the geography of the original bat-to-camel transmission is still widely regarded as uncertain. All of these are broadly consistent with the risk factors we identify. Notably, India and west Africa are additional hotspots that have yet to experience the emergence of a bat coronavirus into human populations, but may still be at risk—particularly given known gaps in bat surveillance,[17] and a dense population in both regions with global connectivity. In any of these regions, surveillance on viral reservoirs can be paired with targeted monitoring of high-risk human populations (i.e., those with regular wildlife contact)[32] for maximum impact.

Conclusion

Bats are important reservoir hosts for different classes of microorganisms, many of which a threat to human health.[33,34] Chiropterans emerged around 64 million years ago and are one of the most diverse mammalian orders, with an estimated richness of more than 1400 species.[35,36] They exhibit a broad variety of habitat use, behaviour, and feeding strategies, putting them at key positions in the delivery and provisioning of several ecosystem services, tied to important ecosystem-derived benefits to human.[37] For example, bats are an essential component of many seed-dispersal networks.[38] Over two-thirds of bats are know to be either obligate or facultative insectivores, therefore actively contributing for agricultural pest control,[39,40] and vectors of pathogens that put a risk on human health.[41,42] Because bats are globally distributed and have a long evolutionary history, phylogeographic and biogeographic approaches are required to shed light on the contemporary distribution of coevolutionary processes between bats and the pathogens they host. Not all areas in which bats, viruses, and human are co-occuring are facing a risk of spillover towards human populations, and the areas in which this risk exist may not be facing risks of the same nature and magnitude.

Here, we propose a simple freamework with broad explanatory power that helps contextualize discoveries like highly divergent nobecoviruses in Madagascar and the previously-neglected adaptive radiation of sarbecoviruses outside of southern China and throughout southeast Asia. In doing so, it advances ecological theory beyond the current state of the art for global maps of emergence risk. For example, previous studies that have used host richness as proxy have predicted a high diversity of unsampled bat viruses,[14] bat coronaviruses,[2] and even specifically betacoronaviruses[15] in both the Amazon and southeast Asia. While we find that both regions are characterized by highly divergent host and viral communities, our framework identifies key differences between the regions. We find that Latin America is a hotspot of both host and viral distinctiveness, suggesting that this branch of the bat-betacoronavirus complex may be undergoing independent evolutionary dynamics from the rest of the global pool, but with limited potential for viral diversification— a finding that is supported by previous work indicating a higher rate of codivergence in Latin America.[2] In contrast, in southeast Asia, host richness and viral distinctiveness are high but sharing is low; this suggests a different type of evolutionary dynamics that could generate high local diversity of viruses through host switching and viral recombination (see *e.g.*,[13] as well as the discovery of recombinant viruses that share genetic material from both the SARS-CoV and SARS-CoV-2 branches of the Sarbecovirus lineage).[43] Both of these regions are priority areas for sampling, especially given predictions that they contain many bat hosts of undiscovered betacoronaviruses.[15,17] However, both the evolutionary and ecological aspects of emergence risk are likely higher in southeast Asia—a fact that will only become more relevant, as bats track shifting climates and exchange viruses with other species, creating a hotspot of cross-species transmission unique to the region.[44]

The diversity and diversification potential of bats responds to anthropogenic factors others than shifting climates.[45] Land use changes could significantly decrease bat suitability, notably through effects on diet and availability of habitats.[46] As our results establish that the diversification of bats betacoronaviruses happens on top of processes affecting hosts, biogeographic variation in human population density and anthropogenic disturbances may feed into co-evolutionary dynamics. Increase in humans-hosts contacts also increase the risk of emergence of novel diseases,[47] so does the changes in landscape connectivity at local/regional scales.[48] This represents a challenge for both conservation strategies and disease ecology: some areas can a high emergence risk and more potential for the acquisition of zoonotic viruses through bat-human encounters.[49] In particular, the challenge ahead lies in the need to quantify actual exposure (and risk) accounting for several transmission scenarios, including both direct and indirect bat - human interactions, and feeding back into the provision of ecosystem services by bats.

Acknowledgements: We acknowledge that this study was conducted on land within the traditional unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and Omàmiwininiwak nations. This work was supported by funding to the Viral Emergence Research Initiative (VERENA) consortium including NSF BII 2021909 and a grant from Institut de Valorisation des Données (IVADO). This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca). NF is funded by the NSERC BIOS^2 CREATE program. TP and NF are funded by the Courtois Foundation. RLM was supported by Bryce Carmine and Anne Carmine (née Percival), through the Massey University Foundation. DJB was supported by the National Institute of General Medical Sciences of the National Institutes of Health (P20GM134973).

Methods

Known *Betacoronavirus* hosts

We downloaded the data on bats hosts of *Betacoronavirus* from `https://www.viralemergence.org/betacov` on Apr. 2022,[15] and filtered it to "known" hosts (established before the emergence of SARS-CoV-2) and "novel" hosts (confirmed through sampling and competence assays since the initial data collection). The original database was assembled by a combination of data mining and literature surveys, including automated alerts on the "bats" and "coronavirus" keywords to identify novel empirical evidence of bats-betacoronaviruses associations; this yielded a total of 126 known hosts, 47 of which were novel hosts.

Bat occurrences

We downloaded the rangemap of every current bat species that was classified as an empirically documented host of *Betacoronavirus* from the previous step, according to recent IUCN data.[50] The range maps were subsequently rasterized using the `rasterize` function from `GDAL`[51] at a resolution of approximately 100kmx100km. For every pixel in the resulting raster where at least one bat host of *Betacoronavirus* was present, we extract the species pool (list of all known bat hosts), which was used to calculate the following risk assessment components: bat phylogenetic diversity, bat compositional uniqueness, and predicted viral sharing risk.

Bat phylogenetic diversity

For every pixel, we measured Faith's Phylogenetic Diversity[52] based on a recent synthetic tree with robust time calibration, covering about 6000 mammalian species.[53] Faith's PD measures the sum of unique branches from an arbitrary root to a set of tips, and comparatively larger values indicate a more phylogenetic diverse species pool. We measured phylogenetic diversity starting from the root of the entire tree (and not from Chiroptera); this bears no consequences on the resulting values, since all branches leading up to Chiroptera are only counted one per species pool, and (as we explain when describing the assembly of the composite risk map), all individual risk components are ranged in [0,1]. This measure incorporates a richness component, which we chose not to correct for; the interpretation of the phylogenetic diversity is therefore a weighted species richness, that accounts for phylogenetic over/under-dispersal in some places.

Bat compositional uniqueness

For every species pool, we measured its Local Contribution to Beta-Diversity;[54] LCBD works from a species-data matrix (traditionally noted as **Y**), where species are rows and sites are columns, and a value of 1 indicates occurrence. We extracted the Y matrix assuming that every pixel represents a unique location, and following best practices[55] transformed it using Hellinger's distance to account for unequal bat richness at different pixels. The correction of raw community data is particularly important for two reasons: first, it prevents the artifact of richer sites having higher importance; second, it removes the effect of overall species richness, which is already incorporated in the phylogenetic diversity component. High values of LCBD indicate that the pixel has a community that is on average more dissimilar in species composition than what is expected knowing the entire matrix, i.e. a more unique community. Recent results by[56] shows that LCBD measures are robust with regards to spatial scale, and are therefore applicable at the global scale.

Viral sharing between hosts

For all bat hosts of *Betacoronavirus*, we extracted their predicted viral sharing network, generated from a previously published generalized additive mixed model of virus sharing by a tensor function of phylogenetic distance and geographic range overlap across mammals.[57] This network stores pairwise values of viral community similarity. To project viral sharing values into a single value for every pixel, we averaged the pairwise scores. High values of the average sharing propensity means that this specific extant bat assemblage is likely to be proficient at exchanging viruses.

Composite risk map

To visualize the aggregated risk at the global scale, we combine the three individual risk components (phylogenetic diversity, compositional uniqueness, and viral sharing) using an additive color model.[58] In this approach, every risk component gets assigned a component in the RGB color model (phylogenetic diversity is green, compositional uniqueness is red, and viral sharing is blue). In order to achieve a valid

RGB measure, all components are re-scaled to the [0,1] interval, so that a pixel with no sharing, no phylogenetic diversity, and no compositional uniqueness is black, and a pixel with maximal values for each is white. This additive model conveys both the intensity of the overall risk, but also the nature of the risk as colors diverge towards combinations of values for three risk components. Out of the possible combinations, the most risky in terms or rapid diversification and spillover potential is high phylogenetic diversity and low viral sharing,[59] in that this allows multiple independent host-virus coevolutionary dynamics to take place in the same location. In the colorimetric space, this correspond to yellow – because the HSV space is more amenable to calculations for feature extraction,[60] we measured the risk level by calculating the angular distance of the hue of each pixel to a reference value of 60 (yellow), and weighted this risk level by the value component. Specifically, given a pixel with colorimetric coordinates (h, s, v), its ranged weighted risk value is

$$v \times \left[1 - \frac{\left|\operatorname{atan}\left(\cos(\operatorname{rad}(h)), \sin(\operatorname{rad}(h))\right) - X\right|}{2\pi}\right],$$

where X is $\operatorname{atan}\left(\cos(\operatorname{rad}(60)), \sin(\operatorname{rad}(60))\right)$, a constant approximately equal to 0.5235.

Viral phylogeography and evolutionary diversification

To next represent phylogeography of betacoronaviruses in bats, we aggregated and analyzed betacoronavirus sequence data. We used the following query to pull all *Betacoronavirus* sequence data from the GenBank Nucleotide database except SARS-CoV-2; ("Betacoronavirus"[Organism] OR betacoronavirus[All Fields]) NOT ("Severe acute respiratory syndrome coronavirus 2"[Organism] OR sars-cov-2[All Fields]). We added a single representative sequence for SARS-CoV-2 and manually curated to remove sequences without the RNA-dependent RNA polymerase (RdRp) sequence or that contained words indicating recombinant or laboratory strains including "patent", "mutant", "GFP", and "recombinant". We filtered over-represented taxa including betacoronavirus 1, hCoV-OC43, Middle East respiratory syndrome coronavirus, Murine hepatitis virus, and hCoV-HKU1. Curated betacoronavirus RdRp sequences were then aligned using MAFFT[61] v1.4.0 (Algorithm FFT-NS-2, Scoring matrix 200PAM / k=2, gap open penalty 1.53m offset value 0.123) and a maximum likelihood tree reconstructed in IQ-TREE[62] v1.6.12 with ModelFinder[63] ultrafast bootstrap approximation[64] with a general time reversible model with empirical base frequencies and the 5-discrete-rate-category FreeRaye model of nucleotide substitution (GTR+F+R5).

We first tested the hypothesis that hotspots of viral diversification would track hotspots of bat diversification. To do so, we plotted the number of known bat hosts (specifically only those included in the phylogeny, so there was a 1:1 correspondence between data sources) against the "mean evolutionary distinctiveness" of the associated viruses. To calculate this, we derived the fair proportions evolutionary distinctiveness[65] for each of the viruses in the tree, then averaged these at the bat species level, projected these values onto their geographic distributions, and averaged across every bat found in a given pixel. As such, this can be thought of as a map of the mean evolutionary distinctiveness of the known viral community believed to be associated with a particular subset of bats present.

Co-distribution of hosts and viral hotspots

Subsequently, we tested the hypothesis that the biogeography of bat betacoronaviruses should track the biogeography of their hosts. To test this idea, we loosely adapted a method from,[66,67] who proposed a phylogenetic method for the delineation of animal biogeographic regions. In their original method, a distance matrix - where each row or column represents a geographic raster's grid cell, and the dissimilarity values are the "beta diversity similarity" of their community assemble - undergoes non-metric multidimensional scaling (NMDS); the first two axes of the NMDS are projected geographically using a four-color bivariate map. Here, we build on this idea with an entirely novel methodology. First, we measure the phylogenetic distance between the different viruses in the betacoronaviruses tree by using the cophenetic function in ape;[68] subsequently, we take a principal components analysis of that distance matrix (readily interchangeable for NMDS in this case) to project the viral tree into an n-dimensional space. We then take the first two principal components and, as with the evolutionary distinctiveness analysis, aggregated these to a mean host value and projected them using a four-color bivariate map.

References

1.

Plowright, R. K. *et al.* Pathways to zoonotic spillover. *Nature Reviews Microbiology* **15**, 502–510 (2017).

2.

Anthony, S. J. *et al.* Global patterns in coronavirus diversity. *Virus Evolution* **3**, (2017).

3.

Ruiz-Aravena, M. *et al.* Ecology, evolution and spillover of coronaviruses from bats. *Nature Reviews Microbiology* **20**, 299–314 (2022).

4.

Agosta, S. J., Janz, N. & Brooks, D. R. How specialists can be generalists: Resolving the 'parasite paradox' and implications for emerging infectious disease. *Zoologia (Curitiba)* **27**, 151–162 (2010).

5.

Power, A. G. & Mitchell, C. E. Pathogen Spillover in Disease Epidemics. *The American Naturalist* **164**, S79–S89 (2004).

6.

Geoghegan, J. L., Duchêne, S. & Holmes, E. C. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLOS Pathogens* **13**, e1006215 (2017).

7.

Thompson, J. N. *The Geographic Mosaic of Coevolution*. (University Of Chicago Press, 2005).

8.

Thompson, J. N. *The Coevolutionary Process*. (University of Chicago Press, 1994).

9.

Janzen, D. H. When is it Coevolution? *Evolution* **34**, 611–612 (1980).

10.

Price, P. W. *Macroevolutionary Theory on Macroecological Patterns*. (Cambridge University Press, 2002). doi:10.1017/CBO9780511615030.

11.

Gomulkiewicz, R. *et al.* Dos and don'ts of testing the geographic mosaic theory of coevolution. *Heredity* **98**, 249–258 (2007).

12.

Thrall, P. H., Hochberg, M. E., Burdon, J. J. & Bever, J. D. Coevolution of symbiotic mutualists and parasites in a community context. *Trends in Ecology & Evolution* **22**, 120–126 (2007).

13.

Latinne, A. *et al.* Origin and cross-species transmission of bat coronaviruses in China. *bioRxiv: The Preprint Server for Biology* 2020.05.31.116061 (2020) doi:10.1101/2020.05.31.116061.

14.

Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).

15.

Becker, D. J. *et al.* Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *The Lancet Microbe* (2022) doi:10.1016/S2666-5247(21)00245-7.

16.

Allen, T. *et al.* Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications* **8**, (2017).

17.

Cohen, L. E., Fagre, A. C., Chen, B., Carlson, C. J. & Becker, D. J. Sampling strategies and pre-pandemic surveillance gaps for bat coronaviruses. 2022.06.15.496296 (2022) doi:10.1101/2022.06.15.496296.

18.

Olival, K. J. *et al.* Possibility for reverse zoonotic transmission of SARS-CoV-2 to free-ranging wildlife: A case study of bats. *PLoS pathogens* **16**, e1008758 (2020).

19.

Shi, J. J. *et al.* A Deep Divergence Time between Sister Species of Eidolon (Pteropodidae) with Evidence for Widespread Panmixia. *Acta Chiropterologica* **16**, 279–292 (2014).

20.

Kettenburg, G. *et al.* Full Genome Nobecovirus Sequences From Malagasy Fruit Bats Define a Unique Evolutionary History for This Coronavirus Clade. *Frontiers in Public Health* **10**, (2022).

21.

Rulli, M. C., D'Odorico, P., Galli, N. & Hayman, D. T. Land-use change and the livestock revolution increase the risk of zoonotic coronavirus transmission from rhinolophid bats. *Nature Food* **2**, 409–416 (2021).

22.

Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology* **17**, 181–192 (2019).

23.

Muylaert, R. L. *et al.* Present and future distribution of bat hosts of sarbecoviruses: Implications for conservation and public health. *Proceedings of the Royal Society B: Biological Sciences* **289**, 20220397 (2022).

24.

Hassell, J. M., Begon, M., Ward, M. J. & Fèvre, E. M. Urbanization and disease emergence: Dynamics at the wildlife–livestock–human interface. *Trends in ecology & evolution* **32**, 55–67 (2017).

25.

Worobey, M. *et al.* The Huanan market was the epicenter of SARS-CoV-2 emergence. *Zenodo* (2022) doi:10.5281/zenodo.6299600.

26.

Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* **604**, 330–336 (2022).

27.

Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature microbiology* **5**, 1408–1417 (2020).

28.

Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS pathogens* **13**, e1006698 (2017).

29.

Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virologica Sinica* **33**, 104–107 (2018).

30.

Memish, Z. A. *et al.* Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerging infectious diseases* **19**, 1819 (2013).

31.

Müller, M. A. *et al.* MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983–1997. *Emerging infectious diseases* **20**, 2093 (2014).

32.

Xu, R.-H. *et al.* Epidemiologic Clues to SARS Origin in China. *Emerging Infectious Diseases* **10**, 1030–1037 (2004).

33.

Letko, M., Seifert, S. N., Olival, K. J., Plowright, R. K. & Munster, V. J. Bat-borne virus diversity, spillover and emergence. *Nature Reviews Microbiology* **18**, 461–471 (2020).

34.

Van Brussel, K. & Holmes, E. C. Zoonotic disease and virome diversity in bats. *Current Opinion in Virology* **52**, 192–202 (2022).

35.

Peixoto, F. P., Braga, P. H. P. & Mendes, P. A synthesis of ecological and evolutionary determinants of bat diversity across spatial scales. *BMC ecology* **18**, 18 (2018).

36.

Simmons, N. B. & Cirranello, A. L. Bat Species of the World: A taxonomic and geographic database. https://batnames.org/ (2020).

37.

Kasso, M. & Balakrishnan, M. Ecological and Economic Importance of Bats (Order Chiroptera). *ISRN Biodiversity* **2013**, e187415 (2013).

38.

Mello, M. A. R. *et al.* The Missing Part of Seed Dispersal Networks: Structure and Robustness of Bat-Fruit Interactions. *PLOS ONE* **6**, e17395 (2011).

39.

Bats in the Anthropocene: Conservation of Bats in a Changing World. (Springer International Publishing, 2016). doi:10.1007/978-3-319-25220-9.

40.

Williams-Guillén, K., Perfecto, I. & Vandermeer, J. Bats Limit Insects in a Neotropical Agroforestry System. *Science* **320**, 70–70 (2008).

41.

Gonsalves, L., Bicknell, B., Law, B., Webb, C. & Monamy, V. Mosquito Consumption by Insectivorous Bats: Does Size Matter? *PLOS ONE* **8**, e77183 (2013).

42.

Gonsalves, L., Lamb, S., Webb, C., Law, B. & Monamy, V. Do mosquitoes influence bat activity in coastal habitats? *Wildlife Research* **40**, 10–24 (2013).

43.

Wu, Z. *et al.* *A comprehensive survey of bat sarbecoviruses across China for the origin tracing of SARS-CoV and SARS-CoV-2.* (2021) doi:10.21203/rs.3.rs-885194/v1.

44.

Carlson, C. J. *et al.* Climate change increases cross-species viral transmission risk. *Nature* 1–1 (2022) doi:10.1038/s41586-022-04788-w.

45.

Alves, D. M. C. C., Diniz-Filho, J. A. F., da Silva e Souza, K., Gouveia, S. F. & Villalobos, F. Geographic variation in the relationship between large-scale environmental determinants and bat species richness. *Basic and Applied Ecology* **27**, 1–8 (2018).

46.

Treitler, J. T., Heim, O., Tschapka, M. & Jung, K. The effect of local land use and loss of forests on bats and nocturnal insects. *Ecology and Evolution* **6**, 4289–4297 (2016).

47.

Johnson, C. K. *et al.* Global shifts in mammalian population trends reveal key predictors of virus spillover risk. *Proceedings of the Royal Society B: Biological Sciences* **287**, 20192736 (2020).

48.

Gryseels, S. *et al.* When Viruses Don't Go Viral: The Importance of Host Phylogeographic Structure in the Spatial Spread of Arenaviruses. *PLOS Pathogens* **13**, e1006073 (2017).

49.

Amman, B. R. *et al.* *Investigating the role of bats in emerging zoonoses: Balancing ecology, conservation and public health interest*. (FAO, 2011).

50.

IUCN. The IUCN Red List of Threatened Species. `https://www.iucnredlist.org` (2021).

51.

Rouault, E. *et al.* *GDAL/OGR Geospatial Data Abstraction software Library*. (Zenodo, 2022). doi:10.5281/ZENODO.5884351.

52.

Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**, 1–10 (1992).

53.

Upham, N. S., Esselstyn, J. A. & Jetz, W. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology* **17**, e3000494 (2019).

54.

Legendre, P. & De Cáceres, M. Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecology Letters* **16**, 951–963 (2013).

55.

Legendre, P. & Condit, R. Spatial and temporal analysis of beta diversity in the Barro Colorado Island forest dynamics plot, Panama. *Forest Ecosystems* **6**, 7 (2019).

56.

Dansereau, G., Legendre, P. & Poisot, T. Evaluating ecological uniqueness over broad spatial extents using species distribution modelling. *Oikos* **n/a**, e09063 (2022).

57.

Albery, G. F., Eskew, E. A., Ross, N. & Olival, K. J. Predicting the global mammalian viral sharing network using phylogeography. *Nature Communications* **11**, 2260 (2020).

58.

Seekell, D. A., Lapierre, J.-F. & Cheruvelil, K. S. A geography of lake carbon cycling. *Limnology and Oceanography Letters* **3**, 49–56 (2018).

59.

Gomulkiewicz, R., Thompson, J. N., Holt, R. D., Nuismer, S. L. & Hochberg, M. E. Hot Spots, Cold Spots, and the Geographic Mosaic Theory of Coevolution. *The American Naturalist* **156**, 156–174 (2000).

60.

Keke, S., Peng, Z. & Guohui, L. Study on skin color image segmentation used by Fuzzy-c-means arithmetic. in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery* vol. 2 612–615 (2010).

61.

Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

62.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).

63.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587–589 (2017).

64.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**, 518–522 (2018).

65.

Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C. & Baillie, J. E. M. Mammals on the EDGE: Conservation Priorities Based on Threat and Phylogeny. *PLOS ONE* **2**, e296 (2007).

66.

Kreft, H. & Jetz, W. Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences* **104**, 5925–5930 (2007).

67.

Kreft, H. & Jetz, W. A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography* **37**, 2029–2053 (2010).

68.

Paradis, E. & Schliep, K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

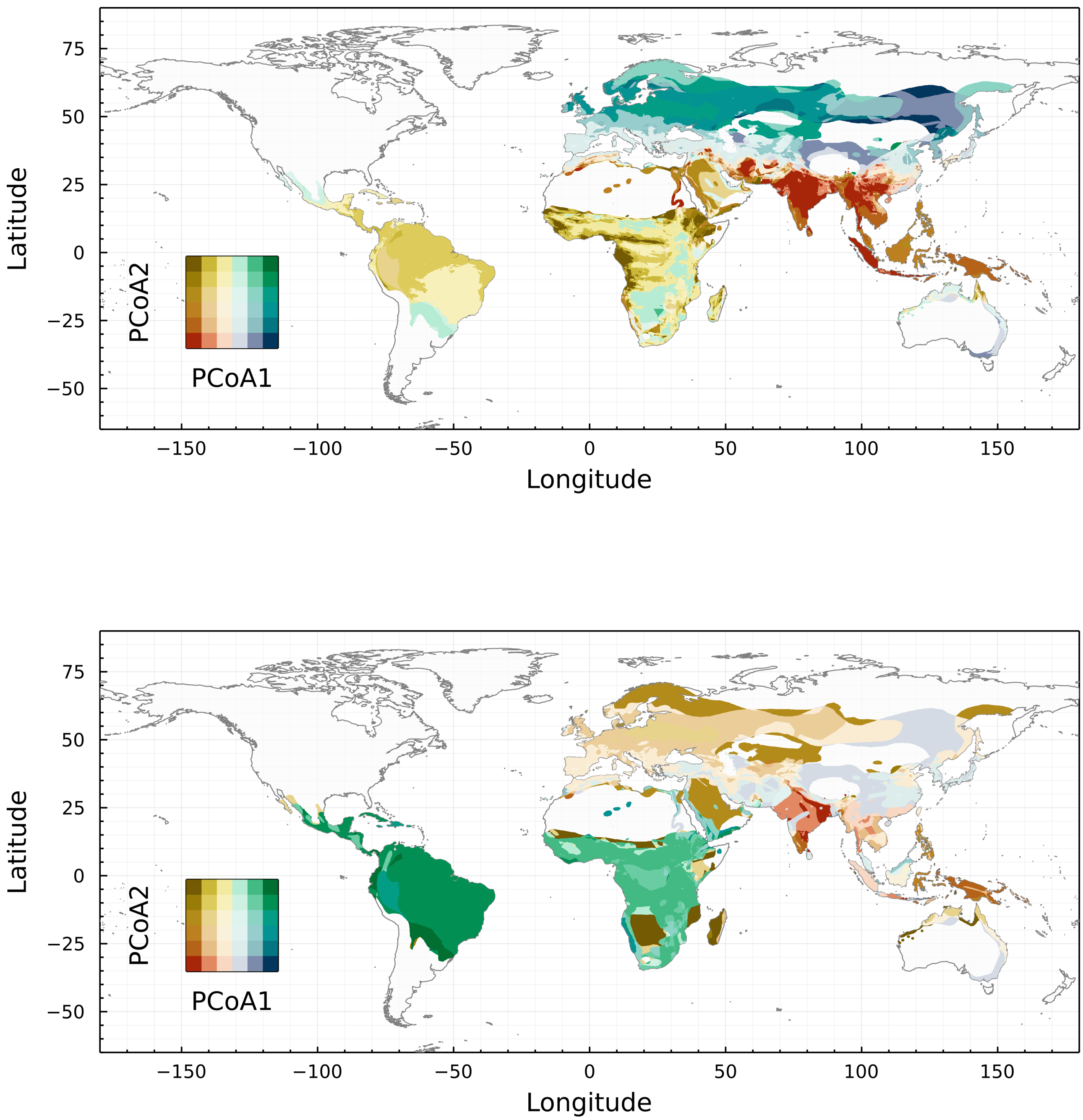

Figure 1: Phylogeographic regions of bats (top) and viruses (bottom) based on the joint analysis of their occurrence and phylogenetic relatedness. The different colors show tendencies to separate alongside the first two components of a PCoA. Note that the PCoA for the bats and viruses are independent, and so cannot be compared directly – that being said, the regions can be compared across maps.

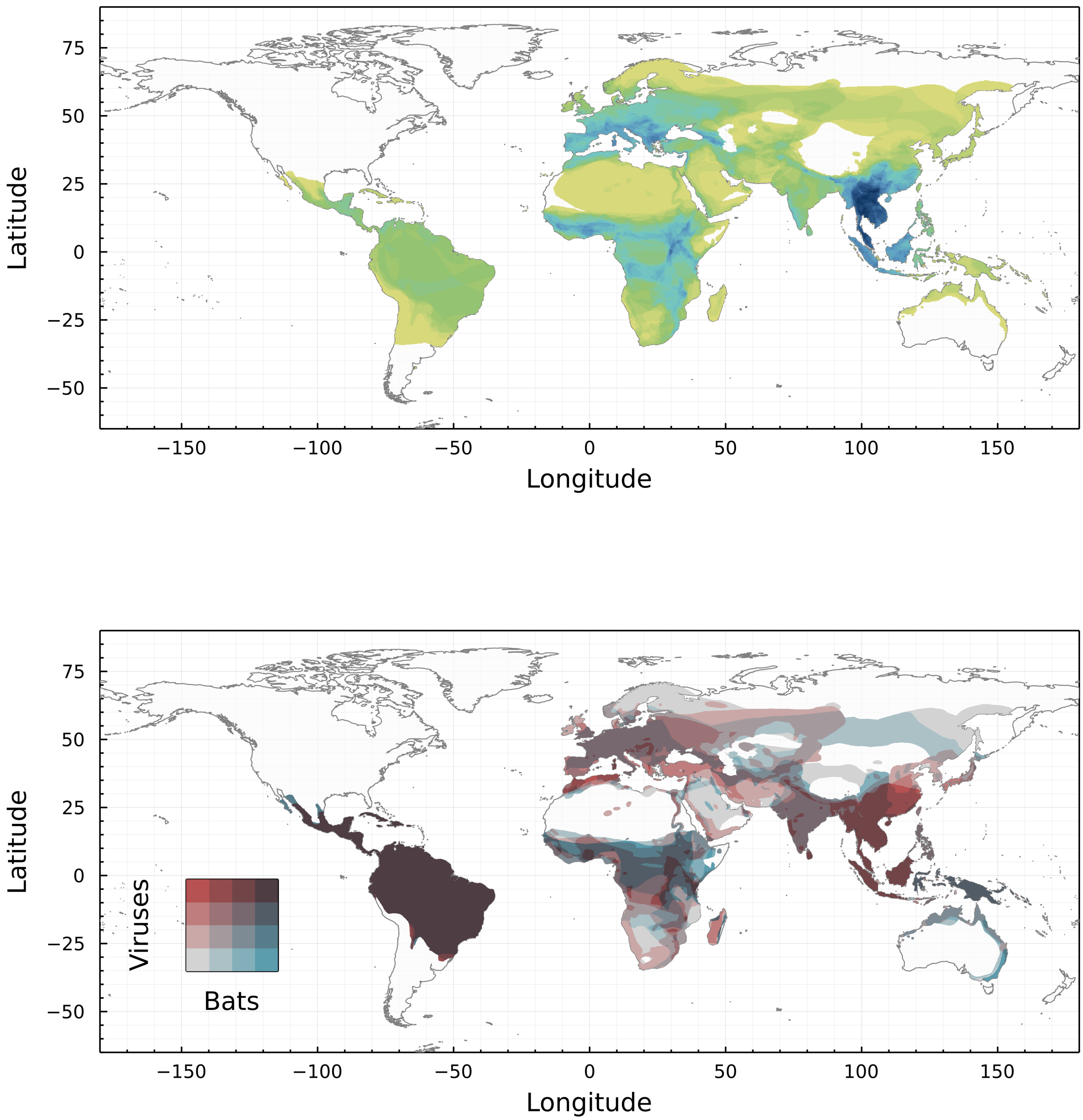

Figure 2: Top panel: relative diversity of known bat hosts of betacoronaviruses. This map shows that the region with the largest number of possible hosts is South-Eastern Asia. Bottom panel: congruence between the evolutionary distinctiveness of the hosts (grey to blue) and the viruses (grey to red).

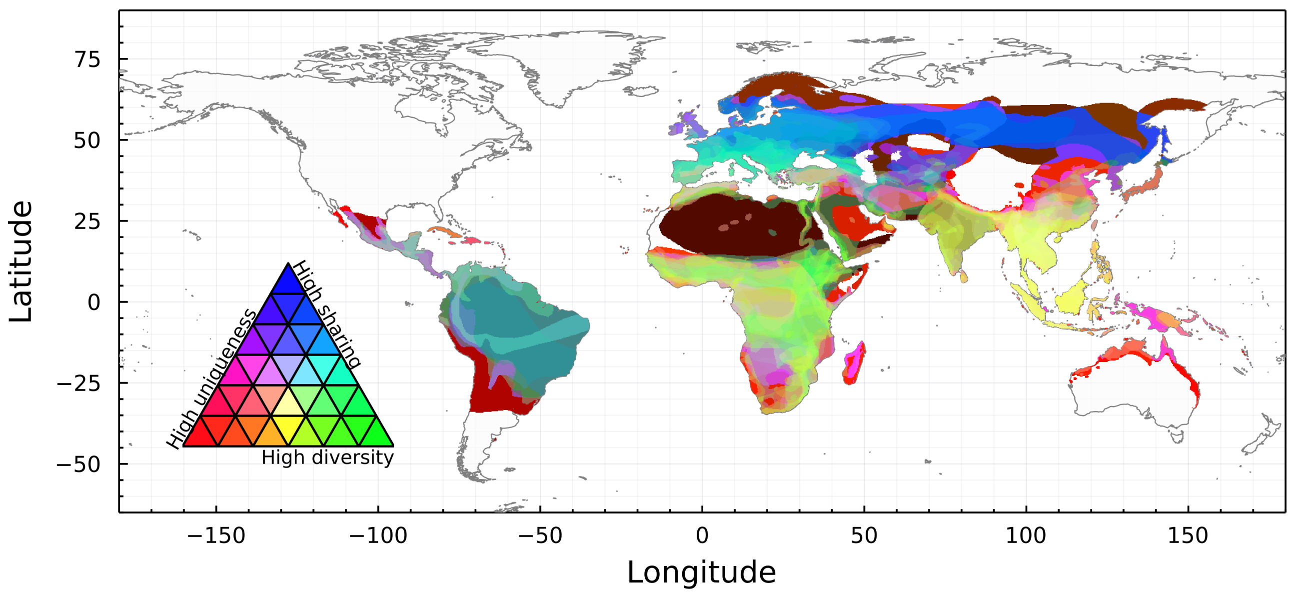

Figure 3: Trivariate additive mapping of the components of risk in the red/green/blue, where high virus sharing is encoded in the blue channel, host phylogenetic diversity in the green channel, and compositional uniqueness in the red channel. A pixel that would maximize all measures (highest possible risk) would be a pure white (specifically RGB(1.0, 1.0. 1.0)), and a pixel with the lowest possible values would be pure black (specifically RGB(0.0, 0.0, 0.0)). Therefore, lighter values (the sum of the three channels gets closer to 3) indicate higher risk, and the color indicates the proportional distribution of the factors making up the total risk.

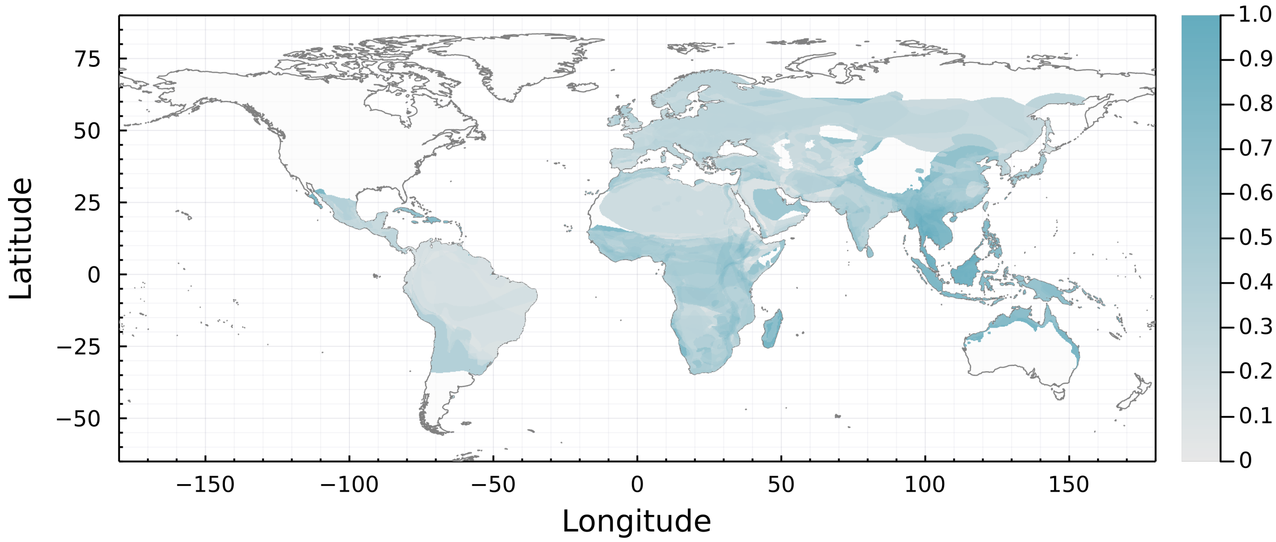

Figure 4: Extraction of a measure of *Betacoronavirus* spillover risk from bat hosts based on the colorimetric space from fig. 3. The risk is a composite measure of the color value and angular distance to the yellow hue, as defined in the methods, ranged in the unit space. Based on this analyses, regions at high risk of spillover are southeast Asia and Madagascar.

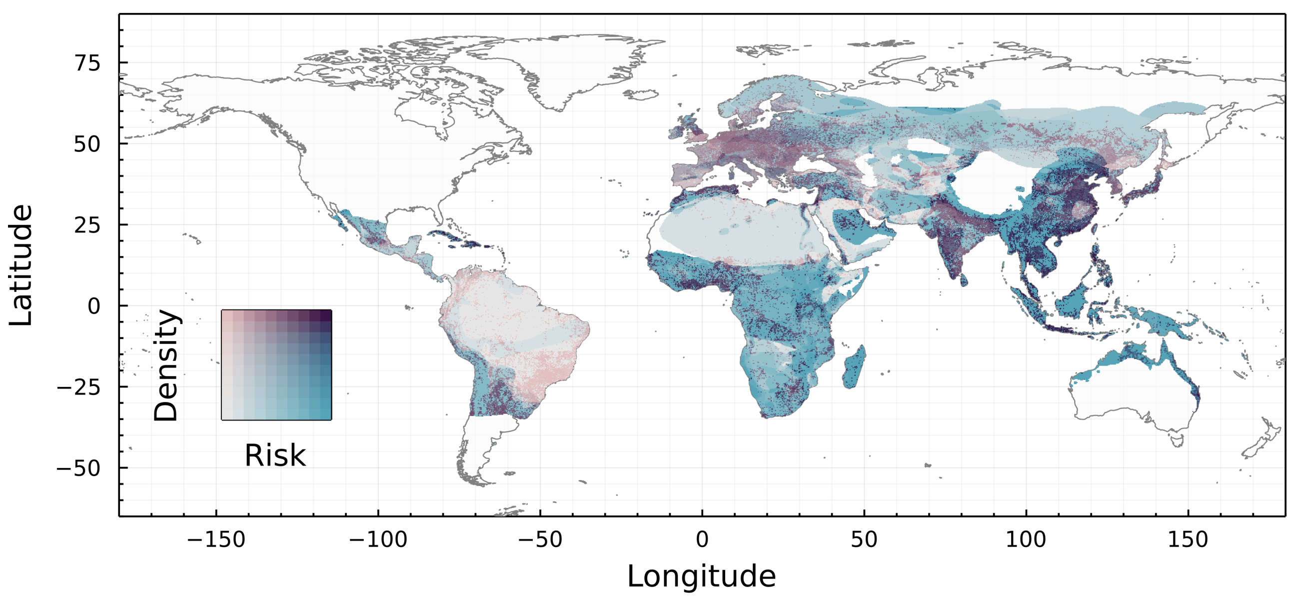

Figure 5: Overlap of the percent of each pixel occupied by urbanized structures, representing the degree of settlement, on the spillover risk map (where the risk comes only from wildlife, and ignores multi-hosts chains of transmissions including non-bats hosts). Darker pixels correspond to more risk, in that the GMTC-derived risk of fig. 4 is high *and* the pixel is densely occupied by human populations. This approach increases the relative risk of several regions in Africa, and highlights the risk in India, southeast China, and the Arabian peninsula where areas of high to moderate risk overlap with areas of denser population.