The coevolutionary mosaic of bat-betacoronaviruses spillover risk

Timothée Poisot [1,2,‡], Peregrin Took [3,4], Merriadoc Brandybuck [5,4,‡]

[1] Université de Montréal; [2] Québec Centre for Biodiversity Sciences; [3] Inn of the Prancing Pony; [4] Fellowship of the Ring; [5] Green Dragon Inn
[‡] These authors contributed equally to the work

Correspondance to:
Timothée Poisot — timothee.poisot@umontreal.ca

Purpose: This template provides a series of scripts to render a markdown document into an interactive website and a series of PDFs.
Motivation: It makes collaborating on text with GitHub easier, and means that we never need to think about the output.
Internals: GitHub actions and a series of python scritpts. The markdown is handled with pandoc.

Keywords:
bats
betacoronavirus
disease ecology
geographic mosaic theory of coevolution
phylogenetic diversity
viral sharing
SARS-CoV-2

Spillover risk is not unidimensional. From the standpoint of an animal community, i.e. a pool of possible hosts, there are a multiplicity of ecological factors that come into play (Plowright et al. 2017). The global richness of hosts is one such component commonly mentioned/analysed (Anthony et al. 2017), but there is an argument to be made that species who are not competent hosts of a specific virus genus may not factor into this (Plowright et al. 2015), or that species who are assumed to share viruses at different rates should be weighted accordingly (Albery et al. 2020). In mammals, key functional traits (for which phylogeny is a reasonable proxy) are determinants of the spillover potential (Olival et al. 2017); these include, notably, TK. Finally, especially when the pool of potential hosts spans the entire globe, there may be local host pools that are highly unique; not having been observed in other locations, these can act on the overall risk either by providing novel contact opportunities, reflecting unique host-environment combinations (Engering et al. 2013), or facilitating rapid changes in specialism (Agosta et al. 2010). In the specific case of generalist pathogens (as betacoronavirus clearly are), there is conceptual and empirical support to the idea that these community- level mechanisms are even more important in driving the overall risk (Power and Mitchell 2004).

Bats are important reservoir hosts for different classes of microorganisms (Chu 2008, Donaldson 2010, Li 2010), some of which can threaten human health. Especially concerning is the fact that bats are reservoirs for a variety of emerging viruses (Calisher 2006), making balancing the needs for bat conservation and disease prevention a potentially difficult act, especially in more densely populated areas (REF). Chiropterans emerged around 64 million years ago and are one of the most diverse mammalian orders, with an estimated richness of more than 12000 species, (Peixoto F et al, 2018) and 14325 known species Simmons & Cirranello. They exhibit a broad variety of habitat use, behaviour, and feeding strategies, resulting in their playing an essential role in the delivery of several ecosystem services (Kasso 2013), including economic benefits. Over two-thirds of bats are either obligate or facultative insectivorous mammals, therefore playing an important role in the regulation of insect pests that can affect crops (Williams-Guillen 2011), and vectors of diseases that put a risk on human health (Gonsalves 2013). Because bats are globally distributed and have a long evolutionary history, phylogeographic and biogeographic approaches are required to shed light on the extant distribution of coevolutionary processes between bats and the pathogens they carry. As a consequence, not all areas are facing a risk of human spillover, and those who do might not be facing risks of the same nature and magnitude.

Yet a comprehensive assessment of the risk of spillover of betacoronaviruses from bat hosts to humans is limited by the fact that we do not know the full diversity of viruses associated with every bat species. Predictive models can help fill in some of these gaps, by recommending hosts based on known host-virus associations BECKER REF. In this paper, we examine the biogeographic structure of bats-betacoronaviruses associations, based on a curated dataset of known and predicted hosts. We turn these associations into a spatially explicit additive mapping of zoonotic risk components, which reveals extreme heterogeneity of risk at the global scale; furthermore, we identify the Amazon and South-Eastern Asia as hotspots of phylogenetic distinctiveness of betacoronaviruses; surprisingly, current data suggest that viral sharing between hosts in high in the Amazon and low in South-Eastern Asia, which has the potential to result in different evolutionary dynamics between these two regions.

TK summary of the results

1

Methods

1.1. Known betacoronavirus hosts We downloaded the CoV reservoir database from https://www.viralemergence.org/betacov on Aug. 2021. This database was assembled by a combination of data mining, literature surveys, and application of an ensemble recommender system classifying hosts as either "Suspected" or "Unlikely" (REF BECKER). The hosts considered for this study were all hosts with a known record of a betacoronavirus, and all those with a "Suspected" status in the ensemble model. This resulted in a list of TK TP unique host species.

1.2. Bats occurrences We downloaded the rangemap of every extant bat species that was either classified as an empirically documented or a suspected host of beta-coronaviruses (Becker et al. 2020), according to recent IUCN data (IUCN 2021). The range maps were subsequently rasterized at a resolution of approximately TK TP. For every pixel in the resulting raster where at least one bat host of betacoronavirus was present, we extract the species pool, which was used to calculate the following risk assessment components: phylogenetic diversity, bat compositional uniqueness, and predicted viral sharing risk.

1.3. Bats phylogeography For every pixel, we measured Faith's Phylogenetic Diversity (Faith 1992) based on a recent synthetic tree with robust time calibration, covering about 6000 mammalian species (Upham et al. 2019). 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.

1.4. Bats compositional uniqueness For every species pool, we measured its Local Contribution to Beta-Diversity (Legendre and De Cáceres 2013); 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 (Legendre and Condit 2019) 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.

1.5. Viral sharing between hosts For all bat hosts of betacoronaviruses, we extracted their predicted viral sharing network (Albery et al. 2020). 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.

1.6. 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 (Seekell et al. 2018). 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. 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.

1.7. Viral phylogeography 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 v 1.4.0 (Katoh and Standley 2013, Supplemental X) and a maximum likelihood tree reconstructed in IQ-TREE v 1.6.12 (Nguyen et al. 2015) with ModelFinder (Kalyaanamoorthy et al. 2017) ultrafast bootstrap approximation (Hoang et al. 2018) and the following parameters (STEPH WILL ADD, Supplemental X).

1.8. Viral evolutionary diversification 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 (Isaac et al., 2007) 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.

1.9. 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 Kreft & Jetz (2010), 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 betacoronavirus tree by using the cophenetic function in 'ape'; 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.

1.10. Outbreaks data geo-referencing Finally, we provide a summary visualization of what available information describes the spillover of zoonotic betacoronaviruses of bat origin where data was available before and up through the COVID-19 pandemic. The SARS-CoV-2 outbreak was georeferenced to the initial case cluster in Wuhan, China; SARS-CoV was georeferenced based on the cave with the closest known viruses circulating in nature (Hu et al. 2017 PLoS Pathogens), and a nearby location where serological (antibody) evidence has indicated human exposure to SARS-like viruses (Wang et al. 2018 Virologica Sinica). For MERS-CoV, we presented the index cases available from a recently-published compendium of MERS-CoV cases (Ramshaw et al. 2019); these are largely if not all presumed to be camel-to-human transmission, and the precise origin point of MERS-CoV in bats is uncertain. Not

shown is a recent case of a recombinant canine coronavirus that showed the ability to infect humans, both because this study was published after the beginning of the COVID-19 pandemic and because bats' involvement in this cycle of transmission has been marginal to non-existent.

2 Results